University of Deusto

Explainability For Machine Learning

Thesis submitted by

Pau Figuera Vinué

as a requirement for the degree of
Doctor en Ingeniería.
With supervision by
Prof. PABLO GARCÍA BRINGAS

Author Director

Bilbao, May 2023

Dedication

Al meu pare Carles, per ensenyar-me que hi ha molt poques coses que un home sol no pot fer, i per la seva estima, de vegades molt sever i distant, però sempre present.
A la meva esposa la Maite, que sempre m'han fet costat amb amor i paciència durant el decurs de tota la recerca i l'escriptura final.
Als meus sogres, Maite i José María pel seu afecte, confiança i suport en tot moment.
També al nostre gos Tango, qui amb la seva alegria, en els llargs passejos que vàrem fer, vaig reflexionar amb profunditat.

Abstract

The leitmotif of this thesis is the search for interpretations with explainable content for Machine Learning. We interpret explainability using well-sound algebraic and statistics techniques. Our starting point is a probabilistic interpretation of the Singular Value Decomposition Theorem. This Theorem is fundamental in many branches of pure and applied mathematics, as well as statistics, and it is the starting point for a vast series of concepts and techniques that have given rise to countless algorithms in Machine Learning.

Initially, we consider the relationship between Probabilistic Latent Semantic Analysis (in its symmetrical formulation) and the Singular Value Decomposition Theorem. This analogy was introduced by Hofmann and studied by other authors, establishing it as a merely formal relationship without quantitative content. Our work is based on interpreting the dimension of the space of its factorization, since the latent or hidden variables are especially relevant. The first result we obtain is the conditions for equivalence: the orthonormalization of a data matrix transformed to the space of probabilities is the decomposition in singular values. This result requires that the dimensionality of the representation space of the non-negative matrix be greater than or equal to the dimensionality of the data matrix.

With these conditions, the search for the significance of the diagonal matrix, which is related to inertia, leads us to Fisher's kernel. The Fisher kernel covariance matrix is a diagonal matrix, and it is a measure of information, being its reciprocal covariance matrix. The non-negative input matrix algebra inherently supports these data structures. The result we derive is that the weight matrix kernel obtained by factorizing the non-negative matrices product is the Fisher information matrix.

To generalize this result, we use the Bregman divergence. This divergence

generalizes several types. In the limit case, it is the geodesic distance. Furthermore, each choice leads to the desired properties for the results. Also, this kernel provides another property: the margins are flat structures, so the misclassification is arbitrarily small.

One consequence we examine is the asymptotic behavior of the sequence of traces obtained with these matrices: its expectation is a statistic modeled by a density that obeys a gamma. The estimation reaches the Crámer-Rao bound, so it is efficient. Furthermore, it is the posterior of a Poisson distribution. We apply this result to the clustering problem, which allows us to build a validation criterion. The novel result is that inference for clustering validation is possible. Several examples demonstrate that the maximums of our densities are very similar to those obtained with the Silhouette and Gap statistic indices.

The statement is non-parametric and induces a metric, just as in the parametric case. Parameterization implies the existence of a differentiable manifold in the parameter space. In this case, parameters are the data matrices and/or vectors (their product allows reconstruction of the original data with no loss of information and are, therefore, non-sufficient minimum statistics).

Resumen

El leitmotiv de esta Tesis es la búsqueda de interpretaciones con contenido explicable para Machine Learning. La explicabilidad la interpretamos como la fundamentación en técnicas algebraicas y estadísticas sólidas. El punto de partida es la interpretación probabilísticas del Teorema de Descomposición en Valores Singulares. Fundamental en muchas ramas de las matemáticas puras y aplicadas, así como en la estadística, constituye el punto de partida de una vasta serie de conceptos y técnicas que han dado lugar a innumerables algoritmos en Machine Learning.

La relación entre el Probabilistic Latent Semantic Analysis (en su formulación simétrica) y el Teorema de Descomposición en Valores Singulares fue señalada por Hofmann y estudiada por otros autores. Se establece como una relación meramente formal y sin contenido cuantitativo. Nuestro trabajo se basa en la interpretación de la dimensión del espacio de su factorización, ya que las variables latentes u ocultas son especialmente relevantes. El primer resultado que obtenemos son las condiciones para la equivalencia: la ortonormalización de una matriz de datos transformada al espacio de probabilidades, es la descomposición en valores singulares. Este resultado requiere que la dimensionalidad del espacio de representación de la matriz no negativa sea mayor o igual a la dimensionalidad de la matriz de datos.

Con estas condiciones, la búsqueda del significado de la matriz diagonal, y que se relaciona con las inercias, nos conduce al kernel de Fisher. La matriz de covarianzas del kernel de Fisher es una medida de la información al ser el recíproco de la matriz de covarianzas. El álgebra de matrices de entradas no negativas soporta estas estructuras de datos de forma natural. El resultado que derivamos es que la matriz de peso asociada con el kernel que se obtiene con la factorización de matrices no negativas es la matriz de información de Fisher.

Para generalizar este resultado, usamos la divergencia de Bregman. Esta divergencia generaliza varios tipos. En el caso límite, es la distancia geodésica. Además, cada divergencia se puede elegir en función de las propiedades deseadas para los resultados. El kernel obtenido proporciona otra propiedad: los márgenes son estructuras planas, por lo que el error de clasificación es arbitrariamente pequeño.

Una consecuencia que examinamos es el comportamiento asintótico de la secuencia de las trazas que se obtienen con estas matrices: asintóticamente su esperanza es un estadístico modelado por una densidad que obedece a una gamma. La estimación alcanza la cota de Crámer-Rao por lo que es eficiente. Además, es la posterior de una distribución de Poisson. Aplicamos este resultado al problema de clustering, lo que permite construir un criterio de validación. El resultado novedoso es que permite inferencia en la validación de la clusterización. Algunos ejemplos ilustran que los máximos de nuestras densidades son muy similares a los que se obtienen con los índices Silhouette y Gap Statistic.

El enfoque en todo momento es no paramétrico e induce una métrica al igual que en el caso paramétrico. La parmetrización implica la existencia de una variedad diferenciable en el espacio de parámetros, que en este caso son las matrices y/o vectores de datos (su producto permite reconstruir los datos originales, sin pérdida de información, y son, por tanto, estadísticos no mínimos suficientes).

Acknowledgments

I wish to express my gratitude to the University of Deusto for accepting my intellectual lines; especially to Prof. Pablo García Bringas for the freedom he has given me to select, approach, and choose the methods in the research topics developed.

I am grateful to a large number of significant people.. Among them are Eng. Eduard Bosch for the advice that helped me to be better, the teachings of Eng. Juan Masip, and Dr. Eng. Antoni Pey † along with Eng. García-Duarte † for many helpful suggestions during my early career.

In the academic field, I would also like to thank Dr. Marta Alsina for teaching me the fundamentals of mathematical reasoning, Prof. Carles Cuadras for introducing me to Mathematical Statistics and Multivariate Analysis, and Professors Esteban Vegas and Ferran Reverter, who introduced me to Non-negative Matrix Factorization. Without the opportunity to receive lessons from them, I would never have been able to reason with originality and rigor.

Certain texts are also important. The connection that a reader establishes with an author is sometimes very special. Prof. Amari's text *Information Geometry and Its Applications* is especially important to me. In it, the profound theories relating geometrical concepts to parameter space are especially well synthesized, and it is one of the fundamentals of this thesis. Equally important is Prof. Cichocki's text *Non-negative Matrix and Tensor Factorization*, which is one of the most cited books in this thesis for standard definitions.

Another relevant text has been *Matrix Analysis and Applications* by Prof. Xian-Da Zhang. His monumental work includes almost all aspects related to algebra and matrix analysis. The clarity and simplicity of the topics

discussed in his book have accompanied me throughout all the reviewed topics. It has been a veritable bedside book and reference manual.

I am also indebted to writers of surveys and tutorials: they are paving the way to broader visions.

I apologize to all the people I have not mentioned here but who have, to some degree, contributed to the development of my ideas. My thanks go to them all. I also want to express my appreciation to all those who have preceded us in intellectual creation. Without them, the adventure of creation would not be possible.

Notation

Symbols and Mathematical Notation

Sets	
$\mathbb{Z}$	field of integer numbers
$\mathbb{Z}_+$	field of nonnegative integer numbers
$\mathbb{Z}_+^n$	set of nonnegative integer vectors of dimension n
$\mathbb{Z}_+^{m\times n}$	set of $m \times n$ integer entries matrices
$\mathbb{R}$	field of real numbers
$\mathbb{R}_+$	field of nonnegative real numbers
$\mathbb{R}_+^n$	set of nonnegative real vectors of dimension n
$\mathbb{R}_+^{m\times n}$	set of $m \times n$ real entries matrices
$\mathbb{C}$	complex field
$\mathbb{C}^n$	field of complex n-dimensional vectors
$\mathbb{C}^{m\times n}$	set of $m \times n$ complex matrices
$x \in A$	x belongs to the set A, i.e. x is an element of A
$x \notin A$	x is not an element of the A
$A \subseteq A$	A is a subset of B
$B \subset A$	A is a proper subset of B
$A \cup B$	union of sets A and B
$A \bigcap B$	intersection of sets A and B
iff	if and only if
$\Leftrightarrow$	if and only if

Norms	
$\lvert \cdot \rvert$	absolute value
$\lVert \cdot \rVert$	generic norm

$\|\cdot\|_p$	p-norm ($1 \leq p \leq +\infty$)
$\|\cdot\|_1$	Hibert norm
$\|\cdot\|_2$	Euclidean of Gaussian norm. Also Hibert-Scmidt norm

Vectors	
$\mathbf{x}$	vector s.t. $\mathbf{x} \in \mathbb{R}^n$
$\mathbf{x}'$	transpose of $\mathbf{x}$
$\widetilde{\mathbf{x}}$	vector s.t. $\|\widehat{\mathbf{x}}\|_{=}1$
$\mathbf{x}_j$	column vector of matrix $\mathbf{X}$
$\mathbf{x}_i$	row vector of matrix $\mathbf{X}$

Matrices	
$\mathbf{X}$	matrix $\mathbf{X} \in \mathbb{R}^{m \times n}$
$\mathbf{X}'$	transpose of $\mathbf{X}$
$\text{diag}(\mathbf{X})$	diagonal of matrix $\mathbf{X}$
$\text{rk}(\mathbf{X})$	rank of matrix $\mathbf{X}$
$\text{tr}(\mathbf{X})$	trace of square matrix $\mathbf{X}$
$[\mathbf{Y}]_{ij}$	probabilistic image matrix of $\mathbf{X} \in \mathbb{R}^{m \times n}$ accomplishing $\sum_{ij} [Y]_{ij} = 1$
$[\widetilde{\mathbf{Y}}]_{ij}$	column normalized probabilistic image matrix of $\mathbf{X} \in \mathbb{R}^{m \times n}$ accomplishing $\sum_{ij} [Y]_{ij} = n$
$\text{vec}(\mathbf{X})$	vector formed by stacking the columns of $\mathbf{X}$ into one vector
$\mathcal{P}_\Omega$	projection operator to the positive orthant
$\lambda_k(\mathbf{X})$	k-th eigenvalue of matrix $\mathbf{X}$
$\sigma_k(\mathbf{X})$	k-th singular value of matrix $\mathbf{X}$

Products	
$\langle , \rangle$	dot or inner product
$\frac{\mathbf{A}}{\mathbf{B}}$	Hadamard division of matrices $\mathbf{A}$ and $\mathbf{B}$
$\mathbf{A} \circledast \mathbf{B}$	Hadamard product of matrices $\mathbf{A}$ and $\mathbf{B}$

Functions and Derivatives	
$D(\cdot\|\cdot)$	generic divergence or distance
$D_{KL}(\cdot\|\cdot)$	KL divergence

$D_I(\cdot\|\cdot)$	I-divergence
$D_\phi(\cdot\|\cdot)$	Bregman divergence
$D_2(\cdot\|\cdot)$	Euclidean distance
∇	operator nabla
$\nabla_{\mathbf{x}}$	partial derivative with respect to vector $\mathbf{x}$
$\nabla_{\mathbf{X}}$	partial derivative with respect to matrix $\mathbf{X}$

Probability	
$P(x)$	probability of x
$P_{\theta\in\Theta}(x)$	parametric probability of x with sufficient minimal statistic θ of parameter class Θ
$P(x\|\, y)$	conditional probability of x, i.e. probability of x if y happens
$E(\cdot)$	expectation
$var(\cdot)$	variance

Acronyms and Abbreviations

AI	Artificial Intelligence
ALS	Alternating Least Squares
BD	Big Data
BSS	Blind Source Separation
EM	Expectation maximization
em	KL based algorithm
ERM	Empirical Risk Minimization
ICA	Independent Component Analysis
HALS	Hierarchical Alternating Least Squares
IR	Information Retrieval
iid	independent identically distributed
IG	Information Geometry
IR	Information Retrieval
kdf	kernel density function
KL	Kullback-Leibler
KKT	Karush-Kuhn-Tucker
LAPACK	Linear Algebra Package
LDA	Latent Dirichlet allocation

LSA	Latent semantic analysis
LSI	Latent semantic indexing
MAP	Maximum a Posteriori
MISE	Mean Integrated Standard Error
ML	Machine Learning
Mult	Multinomial
NMF	Non-negative matrix factorization
NN	Neural Networks
OBP	On-line belief propagation
OPL	Oblique projected Landweber
PCA	Principal component analysis
pdf	Probability Density Function
parafrac	parallel factor analysis
QM	Quantum Mechanics
QMat	Quantum Matrices
PLSA	Probabilistic latent semantic analysis
PSESOP	Projected Sequential Subspace Optimization
PLSI	Probabilistic latent semantic indexing
RKHS	Reproducing Kernel Hilbert Space
SDT	Spectral decomposition Theorem
SVD	Singular value decomposition
SVM	Support vector machine
tf-idf	term frequency-inverse document frequency
TL	Transfer Learning

List of Figures

List of Tables

List of Algorithms

Contents

Chapter 1

Introduction

1.1 Motivation

There exists an extensive bibliography on what explainability is in the context of Machine Learning (ML). We accept it as the set of techniques and methods used so that a human operator can understand the results provided by the machine. Explainability is also desirable, assuming a human operator can interpret the results.

ML methods are devoted to writing code to execute tasks as a human does, and it is a field that has seen a vertiginous increase in terms of concepts and applications in the last two decades. They constitute a branch of Computer Engineering that uses methods of a very diverse nature to learn a task. Formally, ML has been defined as the learning of a task $\mathcal{T}$, with a metric for its evaluation M based on an experience $\mathcal{E}$, and it is the intersection of Computer Engineering and Statistics [180, p. 1]. An important question in ML is the reliability of its conclusions [228]. The acceptance of these methods by many branches of the scientific community for purposes of compression, summarization, classification, or prediction has given rise to many competing techniques. According to our research, a problem shared for many of these methods is that although they work well, they cannot be related to others, and their construction may seem too *ad hoc* to provide good results for particular cases.

A formal approach consists of associating probability in situations in

which uncertainty exists. Probability is a measure of uncertainty, and it is credibility in Bayesian contexts. The algebraic structures induced by probabilistic approaches are supported by nonnegative-entries matrices, restricting entries in the $[0, 1]$ field, with summation condition of one. This algebra allow us to unify algebraic and probabilistic statements.

Nonnegative matrices allow the use of factorization techniques, also in the field of nonnegative reals. With suitable restrictions, this leads to multivariate densities. The dimension of the factorization space span is hidden or latent variables. Furthermore, factorization of nonnegative matrices always exists for any nonnegative-real-entries matrix. It is an approximative technique. There exist several families of algorithms for this purpose, briefly detailed in chapter 4. We used methods that minimize an objective (or loss) function. This function provides the desired properties for the solutions. The methods developed using these techniques are both interpretable and explainable. Interpretability comes from the existence of resources in statistics to present graphical results; explainability derives from the use of algebraic and probabilistic techniques.

The objective of this thesis is to develop some explainable methods. The Probabilistic latent semantic analysis (PLSA) interpretation as a probabilistic image of the Singular value decomposition (SVD) is the starting point. Interpretation of observations in terms of latent variables makes Principal component analysis (PCA) and Independent Component Analysis (ICA) useful tools. PCA is a direct application of SVD. In practice, both terms are used in a confusing way. In fact, the singular values are the variance or reciprocal of the information. In table 1.1 we show the main points of initial interest in our research.

Table 1.1: Milestones.

Year	Contribution	Remarks
2000	PLSA formulation	Conference proceedings [126, 127]. [128]. Comments on the connections among NMF, SVD, and information geometry.
2003	LDA	Criticism of PLSA: LDA formulation [30].
2003	Gaussian PLSA	Assumption of Gaussian mixtures [130].
2005	NMF	PLSA solves the NMF problem [104]. Introduction to stochastic matrices [73].
2008	Kernelization	Fisher kernel derivation from PLSA [129].
2008	Clustering	Equivalence between k-means and NMF [75].
2009	Randomized PLSA	Method to avoid overfitting [200].
2009	PCA	Comparison of NMF, PLSA, and PCA [153].
2012	Information Geometry	Relationship between Fisher information matrix and variance from the PLSA context [53].
2018	Neural Networks	Neural network interpretation of PLSA for transfer learning [158].

Main contributions on PLSA and NMF taken as start points.

Non-negative matrix factorization (NMF) with suitable normalizaton conditions is a similar technique to PLSA and leads to the probabilistic SVD image. In particular, our interest is in the diagonal matrix. Even though it has a secondary role, we have done the opposite. It allows evaluation of the information of the latent variables, conducting efficient estimators, and hence sufficiency of the factorization NMF, without hypotheses about the nature of distributions of parameters. The aim of the thesis will be to establish the conditions that ensure the equivalence of the results obtained in the NMF, as well as the significance and behavior of the diagonal matrix to explore those fields with explainable methods for probabilistic clustering.

1.2 Reflection on Methodological Issues

The appearance of ML complicates the epistemological question of how knowledge of the world is concerning to the world. This age-old question has been the subject of countless philosophical debates and is the center of any worldview. In our time, classic attempts to respond have hints of romanticism: the appearance of non-human entities, also capable of hoarding knowledge in a world where only humans existed, was not taken into account. It's all about the machines. This question was introduced into the epistemological debate by the mathematician Norbert Wiener in the 50s [237]. Thinking machines, or those provided with algorithms that learn, are another entity introduced into the current philosophical debate. This issue, still poorly understood by thinkers lacking scientific training, further increases the conceptual complexity of the philosophical debate.

Historically, the positions adopted regarding this problem are integrative (monism) or pluralistic. Integrative visions assume the existence of ontological categories derived from a single aspect of the thing (the thing has properties, but it is unique). Following Bunge [46], the main orientations of the monist positions are the materialist and idealist currents. Materialism has been said to be the conception of the practitioner of the basic sciences. The idealistic current presents in turn rationalist and empiricist conceptions. While rationalism postulates that reasoning allows deductive knowledge, empiricists assume that knowledge is built from observed facts. These last positions are still alive [46]. The non-integrative positions suppose the ontological division in various aspects of reality, the division between mind and matter being classical. For the case that concerns us, it translates into hardware and software, as entities of independent but related real existence. We note the identification of the mind with the software, while the physical support, which is the set of neurons and other organs that make up the brain, is identified with the hardware.

The methodologies derived from these philosophical approaches determine possible worldviews. Preferred methodologies derive from each worldview. The shared views adopted by the elites of each field are highly influential in the dominant paradigm. No intellectual activity is alien to this precedence. In the scientific context, these are fundamental criteria for the acceptability of knowledge, providing value criteria with respect to the methods used (acceptance or plausibility) that affects the results of a theoretical type (those in which the result for new knowledge is based on the reasoning of

necessity) or on experimentation.

All methodological questions are translated into what in philosophy has been called a mental map or map of representations. According to Bunge [46], a set of ideas Σ belonging to the entities k represents the set Ω of objects for K if the exhibition, memory, or imagination of the elements of Ω evokes a memory for K. We prefer to adhere, at least for the moment, to establishing that a mental map represents knowledge regarding a phenomenon in such a way that an equivalence relationship is established in such a way that it fulfills the property of symmetry when relating the conclusions drawn from the data with the ontological category considered, and vice versa. The reflection in this case would be evident. Reflexivity would be established by considering the consistency between the data, the results extracted from the data, and the properties of the considered ontological category.

An example of value judgments in the area of mathematics is what is considered a demonstration. It seems that inductive proofs, long rejected by the community, are now more accepted, and the same occurs with exemplifications. On the other hand, the experimentation seems to have suffered no changes due to the prestige of the so-called scientific method.

Experimentation progressively has a more relevant role in technique. It is tied to issues of functionality and reliability. This field has also experienced vertiginous growth with the industrial development of the 1930s and 1940s. Industrial experimentation has protocols (norms) related to the operating limits, the effect outside these limits, the mode of use, the conditions of use, and reliability. Also, measurement conditions (calibration) are a fully developed field.

The appearance of ML can be framed with the first expert systems in the 70s. This approach basically used logic as a tool. More recently, systems have appeared that use data to complete problems that cannot be solved exclusively with the use of logic. These methods are Bayesian [239]. This approach has been widely disseminated among statisticians, since Bayesian theory is also a probabilistic theory. In the current context in which massive data is processed, this approach is becoming somewhat obsolete. [1] This fact illustrates the progressive abandonment of these methods.

[1] A discussion of the progressive abandonment of Bayesian methods due to high computational cost is [239]. An example is PLSA, which is difficult to use despite its excellent properties and clear solutions (see chapter 3)

In the case of ML, the verification or adjustment of the postulates and the results is called experimentation. This type of work usually takes some reference datasets and uses the proposed methods comparing them with other existing ones. Experimentation on numerical data does not have systematic protocols to evaluate the effect of extreme values (outliers), zeros (stability), overlapping of variables, variation of the measured variables, their relevance, and the effect of changing them for others. For each of these issues, there are specific techniques for their treatment, and their use and selection seem to be linked to the author's needs and skills rather than being a set protocol. We believe that these issues will be addressed in the future, although we simply name it here as conjecture.

The current dominant paradigm in almost all sciences, and of which ML is no stranger, is of a positivist nature. It seems to sacrifice the issues of explainability and interpretability to the results and the ease of obtaining them for massive datasets. The predominance of results over explainability has been accentuated by the adoption by China of relatively free market laws [12] and its increasing rivalry with the US, as well as the propagandized use of the great effort dedicated to the development of COVID vaccines. In addition, in the current state of the art, it is not possible to formulate hypotheses that give rise to theories with thinking machines based on ML algorithms, although they can be dedicated to searching for data and verifying the accuracy of hypotheses.

In the basic sciences, the results obtained with mathematical tools do not admit discussion (discussion on these points would be a symptom of superficiality [46]). In the case of ML, the discussion also becomes complex. Well-founded theoretical developments may present problems in adjusting to the postulates. The reason may lie in the essence of the data: they do not fully satisfy the hypotheses for obtaining the stated results.

Without wanting to participate, at least actively and consciously, in the philosophical debate, the positions that we defend for our studies are modest. They concern explainability and interpretability. Explicability is derived from the foundation in algebraic techniques (sometimes we use some analytical tools) and the theory of probability; although in statistics, some results have led us to purely Bayesian positions. We understand programming techniques and the mathematical tools used in this sense provide *a posteriori* analytical knowledge. They are the result of mathematical operations from datasets, previously justified with the corresponding demonstrations. Interpretability is achieved by referring to the concepts developed within the framework

of statistics, for which there is a vast arsenal of graphic techniques that allow its interpretability. The consequence is that if, as we think, and regardless of the orientation taken unless they are radical denialists, it assumes the knowledge that is derived from any analysis of data, it provides an equivalence relationship. This equivalence relationship is between reality and the knowledge representation of reality, and it fulfills the symmetric and transitive properties. Transitivity is given by the constant change that is adopted in the different ways of thinking.

At this juncture, we point out that the statistical approaches of our work allow us to assign probabilities in the Bayesian sense of beliefs. Our results are not exclusive and allow discussion between the expert criteria in some classification and the quantitative, or those derived from the use of mathematical techniques. We also note that this is a consequence and not a goal.

We believe that without the need to clearly position ourselves in any epistemological trend, and without falling into the prevailing positivism, a reasonable position is to seek the explicability of the results, which in disciplines related to information and/or computing needs constructed probabilistic interpretations on algebraic and/or analytical bases.

1.3 State of the Art

The approach adopted to select works consists in interpreting NMF as a parameterization in probability space. The interest is in the interpretation of the diagonal matrix. Our start point is PLSA and the development of techniques to improve interpretability and some computational aspects.

The relationship between SVD and PLSA has been spanned in several works, perhaps looking for conditions of more solid equivalence than those stated by Hofmann when formulating PLSA. Using NMF to handle data, a demonstration of the equivalence between NMF and PLSA is provided by [104]. They argue that writing $\mathbf{W} = P(w|z)$, $\mathbf{H} = P(d|z)$, and introducing two diagonal matrices $\mathbf{S}$ and $\mathbf{D}$, the nonnegative matrix product can be written as the product of three matrices, being the central one diagonal. A consequence of this equality is that any local maximum solution of the PLSA is a solution of the NMF problem [104]. The work [75] deals with this same problem using theKullback-Leibler (KL) divergence, reaching the same

conclusions. NMF is closely related with classical methods like PCA. The applications of NMF are numerous, ranging from Artificial Intelligence (AI), text mining, text classification, analysis of linguistic knowledge acquisition, labeling, computational biology, clustering, classification of algorithms, and image classification, speech recognition, among others. In this way, NMF has provided new methods to analyses data in many experimental sciences, like Biology, Chemistry, Meteorology, Geophysics, Economics, Social Sciences, Optics and a wide range of knowledge areas that is impossible to list briefly. Due to the diaspora of publications, thematic orientations, and the interest they have to our research, we focus our interest on those that establish properties and/or analogies.

A work inspired by dimensionallity reduction [27], which provides a way to determine the appropriate number of retained components when NMF is compared with PCA. Using a local Gaussian representation of the posterior distribution provides the basis of a kernel as a positive factorization of a data matrix. Several authors provide an Expectation maximization (EM)-based algorithm interpretation of PCA. An attempt to provide the optimal decomposition dimension of both PLSA and Latent semantic analysis (LSA) is due [152], finding similarities between PLSA and LSA, and reducing the dimension of the decomposition. Then, the dimension problem reduction has a probabilistic signification.

Equivalences between the different NMF techniques have been obtained, in independent works, under KL divergence assumptions. The Latent Dirichlet allocation (LDA) and PLSA relationship is shown in [106] and [75]. The equivalence between NMF and PLSA is studied in [104], stating that (i) any local maximum likelihood solution of PLSA is a solution of NMF and (ii) any solution of NMF with KL divergence yields a local maximum likelihood solution of PLSA. In the same way, [195] introduces tensors to obtain a more general result in the same sense. Studying the correlation between question and answer via a PLSA model provides advantageous results. In [213], the data matrix requirements are relaxed to achieve a semi-positive entries decomposition. He shows the PLSA equivalence to the simplex in which the data points lie. In this sense, the paper is a generalization of PLSA.

Klingenberg [153] provides a geometrical interpretation comparing classical PCA with NMF. Since PCA solutions are not unique (they depend on the dimension reduction), is not guaranteed to extract the latent variables involved in a problem. A solution is to use prior knowledge on a factor

matrix. This approach connects with the KL divergence for multivariate probabilistic mass functions.

Studies on NMF matrices correlation have been carried out by He [121]. Chaudhuri [53] states that *PLSA is related to NMF with KL-divergence objective function... by examining the KL-divergence function and matrix L2 norm based error function.* Devarajan [71] *unifies various competing models and provides a unique theoretical framework for these methods... propose a unified algorithm for NMF and provide a rigorous proof of monotonicity for multiplicative updates and generalize the relationship between NMF and PLSI within this framework.*

All those proposals share computational problems, leading to another kind of work. This type of difficulty was revealed by Zitzler [252] in a talk at the Congress on Evolutionary Computation in Zurich (2001). The emergence of PLSA problems was noticed early: intrinsic computational difficulty which forces the introduction of alternatives to reduce computational cost. The proposal in [252], called simulated heating, is a trade-off between accuracy and run time. The use of the intrinsically slow EM algorithm in the iterative process, and the high number of iterations necessary to obtain quality results, lead to other authors searching for solutions based on different approaches.

A different proposal appears when the role of PLSA within NMF is compared to SVD in classical linear algebra. In this case, SVD is used to initialize the algorithm, reducing in this way the number of iterations [17]. Other methods to improve the speed of the algorithm are examined in chapter 3.

A widespread technique is kernelization. Kernel methods are crucial in ML. Although there exists many types of kernels, the Fisher kernel, introduced by Jaakkola [141], deserves special attention. It provides a metric for a probabilistic model, and a consistent estimator (the density converges in probability to the value of the parameter that generates the distribution) for the posterior (density of the data given the parameter) of the observed and unobserved instances of the statistical model selected to fit the data [226]. It has seen early success when applied to protein classification [140]. A major formulation of the Fisher kernel is obtained with PLSA. From approaches based on his work in which he introduces PLSA, and in the framework of Information Retrieval (IR), for the co-occurrences that take place for d_i documents of a corpus and w_j words of a thesaurus, the relative frequencies decompose as the product of mixtures $P(d_i, w_j) = P(d_i|\, z_k)P(z_k)P(w_j|\, z_k)$,

and that is the symmetric formulation. In this case $P(d_i|\, z_k)$ and $P(w_j|\, z_k)$ follow multinomial distributions of parameters $P(d_i|\, z_k) \sim \Psi$ and $P(w_j|\, z_k) \sim \Phi$, while $P(z_k)$ is a set of dummy variables with no statistical significance. A contribution to this approach is introduced in Chappelier [52], assuming that the distributions $P(d_i|\, z_k)$ and $P(w_j|\, z_k)$ are independent identically distributed (iid).

More recently, NMF techniques have allowed kernels to be obtained [245, 162]. Under proper normalization conditions, the obtained matrices are stochastic [75], allowing us to relate NMF to the Fisher kernel (in this case, the parameters are the matrices into which it decomposes). This statement, under Gaussian assumptions, leads to more understandable and stable classifications [203].

We also examined spectral clustering and its relationship with probabilistic clustering. Several studies focused on this viewpoint are attributed to Har [117], who introduced an indicator function for observations based on kernelization and used the null hypothesis as a classifier. Smyth [215] used likelihood cross-validation to infer information on the number of model components. Using a different approach, the similarity between clusters can be evaluated from the χ^2 statistic between probabilistic classifications [191]. By introducing an index and assuming that each cluster is generated by a parametric distribution, the minimum can be taken as the validation index [102]. More recent works include Olivares [189], which, in the scope of astronomical observations and under the hypothesis of normality and the existence of a correlation, presents an algorithm in which the posterior of the correlation follows a gamma *pdf*. The work of Usefi [116] presents a purely algebraic approach, in which elements are clustered by co-linearity. A review work, centered on the impact and importance of clustering validation in the context of the recent growth of bioinformatics is provided by Ullmann [227].

Furthermore, we can detect an interest with matrices. Quantum information represents the output of a system that provides classical information after processing distinguishable particles (observations). It relates to classical probabilities laws when assuming a finite set of observations [150]. A formulation of SVD from this point of view has been recently formulated by Kozhisseri [157]. Since this type of matrices seems attractive, they are Hermitian, complicating the problem in massive data treatments.

1.4 Lay-out

The content developed is methodological. It is based on the results obtained and summarized in Table 1.2. We note that chapters 2 and 3 are our conceptual starting point, from which we have developed the rest of the ideas. In the following chapters, we present the results obtained contextualized in their respondent area. An author index has been introduced. This is done to give a better view of the most relevant authors that circumscribe our work. Also, the same has been done for the index.

1.4.1 Singular Value Decomposition

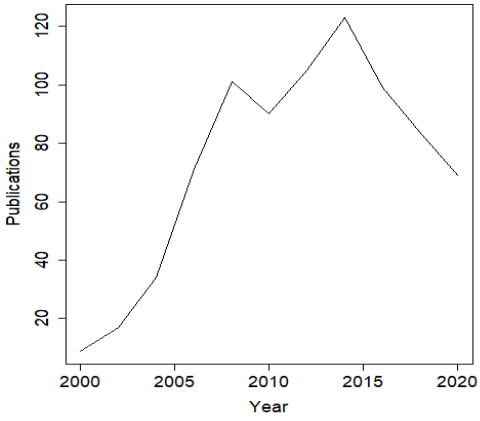
Chapter 2 revises SVD without using vectorization. It is intended to be as faithful as possible to historians [216, 171] and the tutorial [248]. First, we introduce the norm concept, especially Gaussian, spectral, and Hilbert norms.

After discussing SVD, we discuss the approximation theorems, especially Ky Fan's results. They constitute the basis of many of the works examined, and we used them for the results obtained in our works. Many existing works use approximation theorems and variational properties implicitly. We include some of the existing solutions for massive data. Among the selected applications is PCA, after explaining the Procrustes problem, which can be its logical antecedent, although historically this is not the case. ICA is introduced as a solution to certain problems of PCA, with the Amari result that to SVD.

1.4.2 Probabilistic Latent Semantic Analysis

Although PLSA was formulated as an information-retrieval technique, it has been used for diverse purposes. Here, we describe PLSA from the point of view of a maximum likelihood parameterization and by analogy with SVD. We emphasize the efforts made in the past decade to improve aspects related to the convergence of solutions. This search led to the existence of various algorithms. We omit some that are based on the use of purely computational techniques. Solutions have been built in some of the many versions of the

Figure 1.1: Number of works published on PLSA by year.



Source: *Web of Science*. Last accession: February 2023.

EM algorithm. Also included is the native application of this technique in the field of Transfer Learning (TL).

In Figure 1.1, we show the interest received by PLSA according to our bibliographical research.

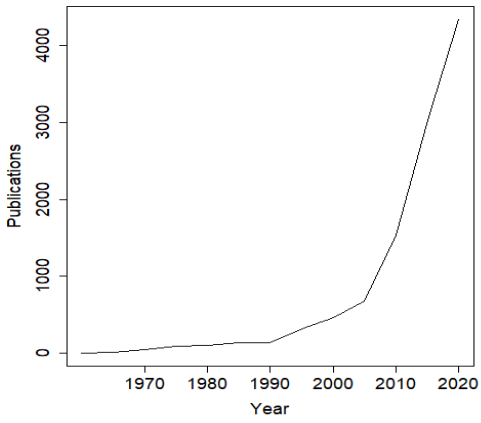
1.4.3 NMF

Stating a problem with uncertainty is an old problem in Statistics. NMF techniques are chosen for their ability for probabilistic interpretation of transformed data. In addition, they can be considered as a non-parametric approach and supposes the introduction of a density, with no hypotheses on the parameters of the underlying density. We reflect on the growing acceptance of these techniques in figure 1.2.

In this chapter, after discussing the conditions for the existence of a probabilistic image of any real matrix, we discuss the the main families of techniques and algorithms. We make special emphasis on solutions obtained with the KL based algorithm (em) algorithm, and that is, in many aspects, similar to the EM algorithm, providing maximum-likelihood solutions.

Notation in areas with strong mathematical content is a nontrivial ques-

Figure 1.2: Published works on NMF.



Source: *Web of Science*. Last accession: February 2023.

tion, and it has a secular history [47]. In many cases, the notation determines conceptual developments [25]. At this point, we note that classical matrix notation, attributed to Cayley [50], among others, remains useful today. However, in the NMF context, it is more convenient to specify the sub-indexes, since the dimension of the space span varies (index k). So, we write stochastic matrices within brackets, and the product $\mathbf{UV} = \sum_k u_{ik}v_{kj}$ as

$$[\mathbf{UV}]_{ij} = [\mathbf{U}]_{ik}[\mathbf{V}]_{kj}$$

making the dimension of span space explicit.

In addition, calling $\mathbf{X}$ the product $\mathbf{UV}$, matrix $\mathbf{U}$ is often the basis over which the column vectors of $\mathbf{V}$ expand. To be consistent with notation using rows for items or observations and columns for variables, the basis is placed as the right matrix herein (i.e., $\mathbf{V}$). The results are adapted to this convention, and with no loss of generality, we suppose that $m \geq n$.

In this chapter, our main aim is to establish the conditions for which the non-negative factorization converges to the probabilistic image of the data matrix. We obtain conditions for which the PLSA hidden variables constitute the diagonal matrix of the SVD.

1.4.4 Kernelization

The SVD theorem has come a long way to reach its current formulation. However, its direct computation presents computational problems. Several mathematicians present kernelization as a computational technique to alleviate such problems. Despite sharing this point of view, we perform a classical exposition. From the conditions that allow the existence of a well-defined kernel (Merccer's Theorem and Reproducing Kernel Hilbert Space (RKHS)), several extended kernels are explained. Then, identification of multivariate observations with vectors makes the dot product especially relevant, inferring a geometric structure and giving rise to a vast series of algorithms. These methods, known as *kernel methods*, are widely used and easy to understand, have solid mathematical properties, and constitute a fundamental of ML. We pay special attention to the Fisher kernel due to its statistical nature. Properties of the Fisher kernel are widely studied, providing a significant reduction in classification error, and good asymptotic properties.

Unlike in the original work of Jaakkola [141], we focus our attention on the Fisher information matrix for our non-parametric derivation providing maximum-likelihood estimations [92]. A reformulation of the Fisher kernel is provided by Hofmann [129] from approaches based on PLSA with the assumption of an underlying multinomial distribution. We generalized this result in a previous paper [89], and we studied the behavior of the Support vector machine (SVM) with our kernel in another [94]. We have shown that the behavior of the margins is a flat structure. Then, the misclassification error is asymptotically arbitrary. We interpret this property as a consequence of the efficiency of the Fisher kernel in computational terms.

1.4.5 Clustering

In chapter 6, we present a validation criterion based on the properties of the NMF traces. This result requires the identification of the space span of the factorization with clusters, converting the factorization problem into a classification problem, which can already be considered a classical problem. We showed that the limit distribution of the trace sequence is a gamma pdf in a previous paper [90]. Classically, the gamma density is obtained as a sum of exponentials [13, p. 179]. Also, it is possible to prove that the independence of sums and sums of squares implies that the distribution of

the coefficient of variation follows this distribution [138]. Our approach aims to evaluate the expectation of a trace sequence.

We generalize this result in [93] by stating that it is the expectation of the limiting distribution. In this last work, we also related the gamma as the posterior of a Poisson according to Bayesian theory. Furthermore, this estimation is efficient and reaches the Cramer-Rao bound.

The result referring to the span of the limit of the succession of traces allows us to relate it to clustering. This relationship appears when we examine the sense of space span as latent variables, which, according to works by other authors, are clusters. The limit of the sequence of traces is a differential equation that depends on two variables (parameters in our case). This equation admits a gamma pdf as a solution, and also a Poisson. This fact places us in a purely Bayesian context since both solutions are conjugate. We have examined the behavior of the proposed pdf solutions from a heuristic point of view, contrasting it with the results of the gap statistic and the Silhouette index.

An application to the validation problem appears when this development is contextualized in the framework of probabilistic clustering, verifying the choice of the number of clusters. An application of the previous results appears when this development is contextualized in the framework of clustering validations, verifying the choice of the number of clusters.

We compare the results of our validation criteria with the number of distributions that have generated the data and the results obtained with the Silhouette index and the gap statistic. We can affirm that the result is at least comparable. In addition to allowing the inference with the number of clusters, the graphical representation of the pdf allows a discussion of the results with expert criteria without the need for specific training in classification and validation techniques.

Table 1.2: Conceptual Development.

Subject	Publication
State of the Art	*Revisiting the Probabilistic Latent Semantic Analysis: The Method, Its Extensions and Its Algorithm* (Preprint)[94]
	Revisiting Probabilistic Latent Semantic Analysis: Extensions, Challenges and Insights (Review Article) [95]
NMF, SVD, and PLSA equivalence	*On the Probabilistic Latent Semantic Analysis Generalization as the Singular Value Decomposition Probabilistic Image* (Research Article) [91]
Kernelization	*A Non-parametric Fisher Kernel* (Conference Paper) [92]
	Generalized Fisher Kernel with Bregman Divergence (Conference Paper) [89]
	Non-negative Matrix Factorization Fisher Kernel (Preprint) [93]
	Exact Classification Fisher Kernel with Non-negative Matrix Factorization (Research Article) [94]
Clustering	*Probability Density Function for Clustering Validation* (Conference Paper) [90]
	A Theoretical Framework for Supporting Clustering Validation via Non-Negative-Matrix-Factorization Trace Sequences Over Probabilistic Spaces (Conference Paper) [65]
	On Clustering Validation Inference (Research Paper) [97]

Contribution of PLSA and SVD with NMF of article [91] is explained in sections 4.5 and 4.6. Contributions on kernelization are explained in sections 5.4, 5.5, 5.6, and 5.8. Contributions on clustering validation are explained in sections 6.3 and 6.4.

Chapter 2

SVD and its Implications

One of the most fruitful ideas in the theory of matrices is factorization into simpler matrices to handle data. This provides a profound interpretation and leads to latent variables. The richness of this idea has made SVD one of the most widespread techniques for data analysis. SVD plays a central role in algebra, constituting a field known as Eigenanalysis, and serves as a basis for matrix function theory. SVD arose from the efforts of several generations of mathematicians, from the nineteenth-century works of Beltrami and independently Jordan [171]. Contributions regarding inequalities between eigenvalues and matrix norms derived from the approximation theorem [208] were provided by Weyl [236], von Neumann [232], and Ky-Fan [87]. This research field still remains active. From the point of view of applications, it finds a place in almost all branches of Physics, structural engineering, and data analysis. Classical expositions of these contributions from a historical point of view have been reviewed by Stewart [216], with Martin's review focusing on applications [171].

SVD is currently featured in Linear Algebra Package (LAPACK), a set of libraries written in Fortran. LAPACK is implemented in many programming languages [7]. In addition, it redefines many properties of matrix algebra, leading to new approaches in the exposition of classical results.

Several applications, like PCA, the Procrustes problem, and ICA are explained. Since our interest is in a probability interpretation of SVD, also we comment on a result obtained with quantum matrices.

2.1 Matrix Norms

Norms induce metrics to evaluate the distance or (dis)similarity between elements of a vector or Banach space (independently of the space structure). For a matrix $\mathbf{A} \in \mathbb{R}^{m \times n}$ a norm is a function $\| \cdot \|$ satisfying

(i) $\|\mathbf{A}\| \geq 0$ and $\|\mathbf{A}\| = 0$ iff $a_{ij} = 0$ for all $a_{ij} \in \mathbf{A} = 0$

(ii) $\|\alpha\mathbf{A}\| = \alpha\|\mathbf{A}\|$ for all $\alpha \in \mathbb{R}$

(iii) $\|\mathbf{A} + \mathbf{A}\| \leq \|\mathbf{A}\| + \|\mathbf{B}\|$

Also, a norm is said consistent if

(iv) $\|\mathbf{AB}\| \leq \|\mathbf{A}\|\|\mathbf{B}\|$

The *induced norm* of the matrix $\mathbf{A}$ is defined by the vectors norm. For $\mathbf{x} \in \mathbb{R}^n$, the induced norm $\|\mathbf{Ax}\|$ is

$$\|\mathbf{A}\| = \sup\left\{ \frac{\|\mathbf{Ax}\|}{\|\mathbf{x}\|} \text{ s.t. } \mathbf{x} \in \mathbb{R}^n \text{ with } \mathbf{x} \neq 0 \right\} \tag{2.1}$$

$$= \sup\Big\{ \|\mathbf{Ax}\| \text{ s.t. } \mathbf{x} \in \mathbb{R}^n \text{ if } \|\mathbf{x}\| = 1 \Big\} \tag{2.2}$$

If a norm also satisfies $[\mathbf{a}_1 | \, 0 | \cdots | \, 0] = 1$ for $\mathbf{a}_i \in \mathbf{A}$, it is said to be a *gauge function*. Then, for matrices $\mathbf{A}$ and $\mathbf{B}$, taking values in the same set, $\|\mathbf{A}\|\|\mathbf{B}\| \leq \varphi(\mathbf{A})\varphi(\mathbf{B})$ is fulfilled for every gauge function [207].

L_p norms are a contribution by Schatten [207]. They are of the form

$$\|\mathbf{A}\|_p = \Big(\sum_{ij} | \, a_{ij}|^p \Big)^{1/p} \tag{2.3}$$

existing as several types.

For $p = 2$, it is the L_2 Euclidean, Gausian, Frobenius, or Hilbert-Schmidt norm

$$\|\mathbf{A}\|_2 = \left(|\, a_{ij}|^2\right)^{1/2} \tag{2.4}$$
$$= \langle \mathbf{A}, \mathbf{A} \rangle \tag{2.5}$$
$$= \left|\mathrm{tr}(\mathbf{A}'\mathbf{A})\right|^{1/2} \tag{2.6}$$

Another norm widely used to evaluate similarity in probability is the L_1 or Hilbert norm,

$$\|\mathbf{A}\|_1 = \sum_{ij} |\, a_{ij}| \tag{2.7}$$

Also, formula (2.1) is

$$\|\mathbf{A}\| = \max \sigma_i(\mathbf{A}) \tag{2.8}$$
$$= \max \frac{\|\mathbf{A}\|_2}{\|\mathbf{x}\|_2} \tag{2.9}$$

where σ are the eigenvalues of $\mathbf{A}$. This norm is useful for the study of sequences of matrices.

Norms are equivalents if there exist scalars α_1 and α_2 and norms $\|\mathbf{A}\|_{\beta_1}$ and $\|\mathbf{A}\|_{\beta_2}$, such that

$$\alpha_1 \|\mathbf{A}\|_{\beta_1} \leq \|\mathbf{A}\|_{\beta_2} \leq \alpha_2 \|\mathbf{A}\|_{\beta_1} \tag{2.10}$$

A norm is said *invariant*, and denoted as $|||\cdot|||$ if

$$|||\mathbf{A}||| = \|\sigma(\mathbf{A})\| \tag{2.11}$$

providing a one-to-one map between symmetric gauge functions and invariant norms.

2.2 SVD

The SVD of a matrix that takes its values in the real or complex field is fundamental. Currently, it is formulated as by Zhang [246, p. 274]

Theorem 1 *For matrix* $\mathbf{X} \in \mathbb{R}^{m\times n}$ *(or* $\mathbb{C}^{m\times n}$*) exists orthogonal (or unitary) matrices* $\mathbf{U} \in \mathbb{R}^{m\times m}$ *(or* $\mathbf{U} \in \mathbb{C}^{m\times m}$*) and* $\mathbf{V} \in \mathbb{R}^{n\times n}$ *(or* $\mathbf{V} \in \mathbb{C}^{m\times m}$*) such that*

$$\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}' \qquad \left(or\, \mathbf{X} = \mathbf{U}\Sigma\mathbf{V}^H\right) \qquad \Sigma = \begin{bmatrix} \Sigma_1 & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{bmatrix}$$

where $\Sigma_1 = \text{diag}(\sigma_1, \ldots, \sigma_r)$ *with diagonal entries*

$$\sigma_1 \geq, \ldots, \geq \sigma_r > 0 \quad r = rk(\mathbf{X})$$

and it can be written in vector form as

$$\mathbf{X} = \sum_i \sigma_i \mathbf{u}_i \mathbf{v}_i' \qquad (\mathbf{u}_i \in \mathbf{U} \text{ and } \mathbf{v}_i \in \mathbf{V}) \tag{2.12}$$

This decomposition is unique, existing for any matrix. One of the first proofs can be found in Eckart [82].

In statistics and data analysis, only real matrices are taken into account. Matrix $\mathbf{X}$ is usually referred as

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \cdots & x_{mn} \end{bmatrix} \tag{2.13}$$

Column vectors of $\mathbf{X}$ are written as $\mathbf{x}_j'$, and row vectors as $\mathbf{x}_i$.

2.3 QR Decomposition

QR decomposition is useful for solving over-determined equation systems. For $\mathbf{X} \in \mathbb{R}^{n \times n}$

$$\mathbf{X} = \mathbf{QR} \tag{2.14}$$

where $\mathbf{Q}'\mathbf{Q} = \mathbf{I}$ and $\mathbf{R}$ is an upper triangular matrix.

Construction of $\mathbf{Q}$ can be built-up with the Gram-Schmidt rule [151, p. 24] consisting of vectors $\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_n$ to obtain an orthogonal basis $\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$

$$\begin{aligned}
\mathbf{e}_1 &= \frac{\mathbf{v}_1}{\|\mathbf{v}_1\|_2} \quad (\mathbf{v}_1 = \mathbf{u}_1) \\
\mathbf{e}_2 &= \frac{\mathbf{v}_2}{\|\mathbf{v}_2\|_2} \quad (\mathbf{v}_2 = \mathbf{u}_2 - \frac{\mathbf{v}_1\mathbf{u}_2}{\|\mathbf{v}_1\|_2^2}\mathbf{v}_1 \\
&\vdots \\
\mathbf{e}_n &= \frac{\mathbf{v}_n}{\|\mathbf{v}_n\|_2} \quad (\mathbf{v}_n = \mathbf{u}_n - \sum_{i=1}^{m-1} \frac{\mathbf{v}_i\mathbf{u}_n}{\|\mathbf{v}_i\|_2^2}\mathbf{v}_i)
\end{aligned} \tag{2.15}$$

This construction is always possible for a non-singular matrix.

2.4 Derivation

If $\mathbf{X} \in \mathbb{R}^{n \times n}$, the scalars $\lambda \in \mathbb{R}$ satisfying

$$\det(\mathbf{X} - \lambda\mathbf{I}) = 0 \tag{2.16}$$

give rise to a polynomial equation of degree n, known as a *characteristic polynomial*. Values λ are the *eigenvalues*. If $(\mathbf{X} - \lambda\mathbf{I})$ is singular, the columns

of $\mathbf{A}$ are linear combinations of the diagonal matrix $\mathbf{I}$, which means that for any vector, $\mathbf{v} \in \mathbb{R}^n$ holds that

$$\mathbf{X}\mathbf{v} = \lambda\mathbf{v} \tag{2.17}$$

which can be written in matrix notation as

$$\mathbf{X}\mathbf{V} = \mathbf{V}\Gamma \tag{2.18}$$

and vectors $\mathbf{v} \in \mathbf{V}$ are called *eigenvectors*.

Because $\mathbf{V}$ and $\mathbf{V}'$ are orthogonal,

$$\mathbf{X} = \mathbf{V}\Gamma\mathbf{V}' \tag{2.19}$$

this result being the Spectral decomposition Theorem (SDT).

The generalization of SDT is SVD, providing a similar result for non-square matrices. Now, $\mathbf{X} \in \mathbb{R}^{m\times n}$. To achieve Theorem 1 it is necessary to preserve $\mathbf{V}\mathbf{V}' = \mathbf{V}'\mathbf{V} = \mathbf{I}$. Also, the eigenvalues of a real entries matrix are real.

Assuming the range $r = \mathrm{rk}(\mathbf{X})$, and writing $\Gamma = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_r)$ satisfying $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r$ and $\mathbf{U}_r = [\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_r]$, from SDT

$$\mathbf{U}_r'\mathbf{X}\mathbf{X}'\mathbf{U}_r = \Gamma_r \tag{2.20}$$

introducing $\mathbf{V}_r = [\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_r]$

$$\mathbf{V}_r = \mathbf{X}'\mathbf{U}_r\Gamma_r^{-1/2} \tag{2.21}$$

it is immediately seen that $\mathbf{V}_r'\mathbf{V}_r = \mathbf{I}$, and

$$\mathbf{X}' = [\mathbf{V}_r, \mathbf{0}]\Gamma_r^{-1/2} \oplus \mathbf{I}_{m-r}\mathbf{U}' \tag{2.22}$$

Hence, Theorem 1 is obtained writing $\sigma_i = \lambda_i^{1/2}$, ensuring $\sigma_1 \geq \sigma_2 \geq \ldots \sigma_r$. Transposing $\mathbf{X}'$ in (2.6) is the condensed SVD [248]. Those are the main results of Beltrami [19] and Jordan [146].

Another independent derivation is the work of Sylvester [219, p .650]. The goal is to diagonalize the bi-linear form

$$\mathbf{B} = \mathbf{x}'\mathbf{X}\mathbf{y} \tag{2.23}$$

considering the canonical form

$$M = \sum_i \Big(\frac{\partial \mathbf{B}}{\partial \mathbf{y}_i}\Big)^2 \tag{2.24}$$

$$= \sum_i \sigma_i^2 \xi_i \tag{2.25}$$

and assuming orthogonality obtains

$$\mathbf{M} - \sigma^2\mathbf{I} = \mathbf{0} \qquad (\mathbf{M} = \mathbf{X}\mathbf{X}') \tag{2.26}$$

$$\mathbf{N} - \sigma^2\mathbf{I} = \mathbf{0} \qquad (\mathbf{N} = \mathbf{X}'\mathbf{X}) \tag{2.27}$$

According to the referenced historians, Sylvester [219] did not know the works of Beltrami [19] and Jordan [146] [216]. Although he does not explicitly obtain the SVD, his work is central to the SVD interpretation, with a generalization that also connects to the SVD [154].

2.5 Approximation Theorem

We are indebted to Schmidt [208] for the approximation Theorem, which is fundamental for Eigenanalysis. From a practical point of view, it justifies the scalability of SVD and gives rise to many practical applications.

The works of Schmidt [208] start from the study of integral equations. Assuming that there exists a non-zero function $\varphi(s)$ satisfying

$$\varphi(s) = \int_a^b \mathcal{X}(s,t)\varphi(t)dt \tag{2.28}$$

shows

(i) The kernel $\mathcal{X}(s,t)$ has at least an eigenfunction.

(ii) The eigenvalues of the eigenfunctions $\varphi(s)$ are real.

(iii) Each eigenvalue of $\varphi(s)$ has a finite number of eigenfuctions.

(iv) The kernel $\varphi(s)$ can be expressed as a finite combination of orthonormal eigenfunctions $\varphi_j(s)$

(v) The eigenvalues satisfy

$$\int_a^b \int_a^b \left(\mathcal{X}(s,t)\right)^2 \geq \sum_i \frac{1}{\lambda_i^2}$$

(vi) A bi-linear form can be expressed as

$$\int_a^b \int_a^b \mathcal{X}(s,t)g(s)h(t)dsdt = \sum_i \frac{1}{\lambda_i} \int_a^b g(s)u_i(s)ds \int_a^b h(t)v_i(t)dt$$

where

$$g(s) = \sum_i \frac{u_i(s)}{\lambda_i} \int_a^b h(t)v_i(t)dt$$

A consequence is the low-rank approximation, whcih is of interest for practical applications. In addition, some techniques such as PCA are based on this approximation. Currently, it is stated as [81, 179]:

Theorem 2 *Given an $m \times n$ real matrix $\mathbf{X}$ of rank r ($r \leq \min(m,n)$) with $\mathbf{X} = \mathbf{U}\Sigma\mathbf{V}'$, the full SVD of $\mathbf{X}$, and $\mathbf{X}_k = \mathbf{U}_k\Sigma\mathbf{V}_k'$, where $\mathbf{U}_k$ and $\mathbf{V}_k$ are the first k columns of $\mathbf{U}$ and $\mathbf{V}$, and Σ_k the first k eigenvalues. Then, for an $m \times n$ real matrix $\mathbf{B}$ of rank $r \leq k$*

$$\|\mathbf{X} - \mathbf{X}_k\| \leq \|\mathbf{X} - \mathbf{B}\|$$

or equivalently,

$$\mathbf{X}_k = \arg \min_{rk(\mathbf{b}) \leq k} \|\mathbf{X} - \mathbf{B}\|$$

This theorem shows that the truncated SVD produces the best approximation of rank r and provides the spectral error bound norm [111]

$$\|\mathbf{X} - \mathbf{X}_r\| \leq \sqrt{\eta^2 + \sum_{i=k+1}^{p} \sigma_j^2(\mathbf{X})} \tag{2.29}$$

for $\eta > 0$.

2.6 Variational Properties

Additional properties of SVD are works related to stability and inequalities. On the one hand, the perturbation theorem due to [236]:

Theorem 3 *Let* $\mathbf{A} \in \mathbb{R}^{m \times n}$ *and* $\mathbf{B} \in \mathbb{R}^{m \times n}$ *such that* $rk(\mathbf{A}) < rk(\mathbf{B})$*, then*

$$\|\mathbf{A} - \mathbf{B}_k\|_2^2 \geq \sigma_{k+1}^2 + \cdots + \sigma_n^2 \tag{2.30}$$

If $\mathbf{A} = \mathbf{A}' + \mathbf{A}^{d\prime}$ *occurs that*

$$\sigma_{i+j-1} \leq \sigma_{i}\prime + \sigma_{j}\prime\prime \tag{2.31}$$

shows the effect that noise has on the eigenvalues.

For the symmetric matrices $\mathbf{M}$ and $\mathbf{N}$, von Neumann [232] introduces an orthogonal matrix $\mathbf{Q}$ such that [232]

$$\operatorname{tr}\big(\mathbf{Q}\Lambda_{\mathbf{M}}\mathbf{Q}'\Lambda_{\mathbf{N}}\big) \leq \sum_i \lambda_i(\mathbf{M}) \sum_i \lambda_i(\mathbf{N}) \tag{2.32}$$

being $\lambda = \operatorname{diag}(\lambda_i)$, and

$$\mathbf{N} = \begin{bmatrix} \mathbf{I}_k & 0 \\ 0 & 0 \end{bmatrix} \tag{2.33}$$

then

$$\sum_{i=1}^{k} \lambda_i(\mathbf{M})\lambda_i(\mathbf{N}) = \max_{\mathbf{QQ}'=\mathbf{I}} \operatorname{tr}\big(\mathbf{Q}'\mathbf{MQ}\big) \tag{2.34}$$

$$\sum_{i=n-k+1}^{n} \lambda_i(\mathbf{M})\lambda_i(\mathbf{N}) = \min_{\mathbf{Q}'\mathbf{Q}=\mathbf{I}} \operatorname{tr}\big(\mathbf{Q}'\mathbf{MQ}\big) \tag{2.35}$$

A generalization is the result of Fan [86]. Stating that for matrices $\mathbf{A} \in \mathbb{R}^{m\times n}$ and $\mathbf{A} \in \mathbb{R}^{m\times n}$, for their singular values

$$\sum_i \sigma_i(\mathbf{A})\sigma_i(\mathbf{B}) = \max_{\mathbf{X}'\mathbf{X}=\mathbf{I};\mathbf{Y}'\mathbf{Y}=\mathbf{I}} \left|\operatorname{tr}\big(\mathbf{X}'\mathbf{AYB}'\big)\right| \tag{2.36}$$

$$= \max_{\mathbf{X}'\mathbf{X}=\mathbf{I};\mathbf{Y}'\mathbf{Y}=\mathbf{I}} \operatorname{tr}\left(\mathbf{X}'\mathbf{AYB}'\right) \tag{2.37}$$

is reached for $\mathbf{X} = \mathbf{U}_{\mathbf{A}}\mathbf{U}'_{\mathbf{B}}$.

Also,

$$\sum_i \sigma_i = \max_{\mathbf{X}'\mathbf{X}=\mathbf{I};\mathbf{Y}'\mathbf{Y}=\mathbf{I}} \left|\operatorname{tr}\big(\mathbf{X}'\mathbf{AY}\big)\right| \tag{2.38}$$

$$= \max_{\mathbf{X}'\mathbf{X}=\mathbf{I};\mathbf{Y}'\mathbf{Y}=\mathbf{I}} \operatorname{tr}\big(\mathbf{X}'\mathbf{AY}\big) \tag{2.39}$$

achieved at $\mathbf{X} = \mathbf{U}_k$ and $\mathbf{Y} = \mathbf{V}_k$.

2.7 Generalized SVD

The generalization of SVD provides a basis for numerical techniques to solve some problems. It is formulated as [229]:

Theorem 4 *Given two matrices* $\mathbf{A} \in \mathbb{R}^{m \times n}$ *and* $\mathbf{B} \in \mathbb{R}^{m \times p}$ *with* $n \geq p$, *then exist two orthonormal matrices* $\mathbf{U}_A \in \mathbf{R}^{m \times m}$ *and* $\mathbf{U}_B \in \mathbf{R}^{n \times n}$, *and an invertible matrix* $\mathbf{X} \in \mathbb{R}^{p \times p}$ *such that*

$$\mathbf{U}'_A \mathbf{A} \mathbf{X} = \text{diag}(\alpha_1, \ldots, \alpha_q)$$
$$\mathbf{U}'_B \mathbf{B} \mathbf{X} = \text{diag}(\beta_1, \ldots, \beta_q)$$

being $\alpha_1 \geq \cdots \geq \alpha_q$ *and* $\beta_1 \leq \cdots \leq \beta_q$.

This theorem is useful for simultaneous diagonalization, and it has found applications in Bioinformatics.

2.8 Properties

From these results, there are useful properties from a computational perspective.

(i) $\text{rank}(\mathbf{A}) = \sigma_i \quad \text{s.t.}\, \sigma_i \neq 0$

(ii) $\|\mathbf{A}\|_2 = \sigma_1 \quad$ (spectral norm)

(iii) $\text{rank}(\mathbf{A}) = rk(\mathbf{A}'\mathbf{A}) = rk(\mathbf{U}_r)$

(iv) $\text{rank}(\mathbf{A}') = rk(\mathbf{A}\mathbf{A}') = \text{rank}(\mathbf{V}_r)$

(v) The eigenvalues of $\mathbf{A}'\mathbf{A}$ are σ_i^2, with $m - r$ zeros, corresponding to the right singular vectors $\mathbf{v}_i$.

(vi) The eigenvalues of $\mathbf{A}\mathbf{A}'$ are σ_i^2 with $m - r$ zeros.

(vii) Let $\mathbf{B} = \mathbf{U}_B \mathbf{\Sigma}_B \mathbf{V}_B$. Then $\mathbf{A} \oplus \mathbf{B} = (\mathbf{U} \oplus \mathbf{U}_B)(\Sigma \oplus \Sigma_B)(\mathbf{V}' \oplus \mathbf{V}'_B)$

(viii) If $\mathbf{A}$ is square and invertible, then $\mathbf{A}^{-1} = \mathbf{V}\Sigma^{-1}\mathbf{U}'$.

Also, the spaces angle is the cosines of the canonical angles as

$$\cos\,\theta_i = \sigma_i(\mathbf{U}'_{A,k}\mathbf{U}_{B,l}) \tag{2.40}$$

Currently, these properties redefine many properties of matrix algebra, and they are the starting point of numerical matrix algebra [6].

2.9 Large-scale SVD Methods

The exact computation of the SVD is a messy problem. For a $m \times n$ matrix, $\mathcal{O}(mn^2)$ $(m > n)$ computations are necessary. There exist several alternatives to alleviate this problem. According to Zhang [248], the main ones are randomized SVD, the CUR approach, and kernel methods.

2.9.1 Randomized SVD

The randomized methods consist of a random projection $\mathcal{P}$ of matrix $\mathbf{X}$ (without loss of generality, and assuming $m > n$, it results in a random selection of $m' < m$ row vectors), and then obtaining an orthonormal basis $\mathbf{Q} \in \mathbb{R}^{m' \times n}$ such that $\|\mathbf{X} - \mathbf{QR}\| \ll \|\mathbf{X} - \mathbf{X}_k\|$ [115, 34, 111].

For L_2 norm, the error bound is

$$\min_{\text{rank}(\mathbf{A}) \leq k} \|\mathbf{X} - \mathbf{QA}\| \leq (1 + \epsilon)\|\mathbf{X} - \mathbf{X}_r\| \tag{2.41}$$

For the spectral norm, it is necessary to choose a value ν satisfying

$$\left\|\mathbf{X} - \mathbf{UU}'\right\|_2^2 \leq \nu\|\mathbf{X} - \mathbf{X}_k\|_2^2 \tag{2.42}$$

leading to

$$\left\|\mathbf{X} - \mathbf{U}_k \Sigma_k \mathbf{V}_k'\right\|_2 \leq (n-k+1)\|\mathbf{X} - \mathbf{X}_k\|_2^2 \tag{2.43}$$

hence [115],

$$\left\|\mathbf{I} - \mathbf{U}_k \mathbf{U}_k' \mathbf{X}\right\| \leq (1+\epsilon)\sigma_{k+1}^2(\mathbf{X}) \tag{2.44}$$

2.9.2 Kernel Approximation

A sample of $\mathbf{X}$ approximates the eigenvalue problem

$$\begin{aligned} \mathbf{K} &= \mathbf{C}\mathbf{X}\mathbf{C}' && (2.45)\\ &= \mathbf{U}_C\big(\Sigma_c \mathbf{V}_c' \mathbf{X} \mathbf{V}_c \Sigma_c\big)\mathbf{U}_c' && (2.46)\\ &= \big(\mathbf{U}_c \mathbf{U}_z\big)\Lambda_z\big(\mathbf{U}_c \mathbf{U}_z\big)' && (2.47) \end{aligned}$$

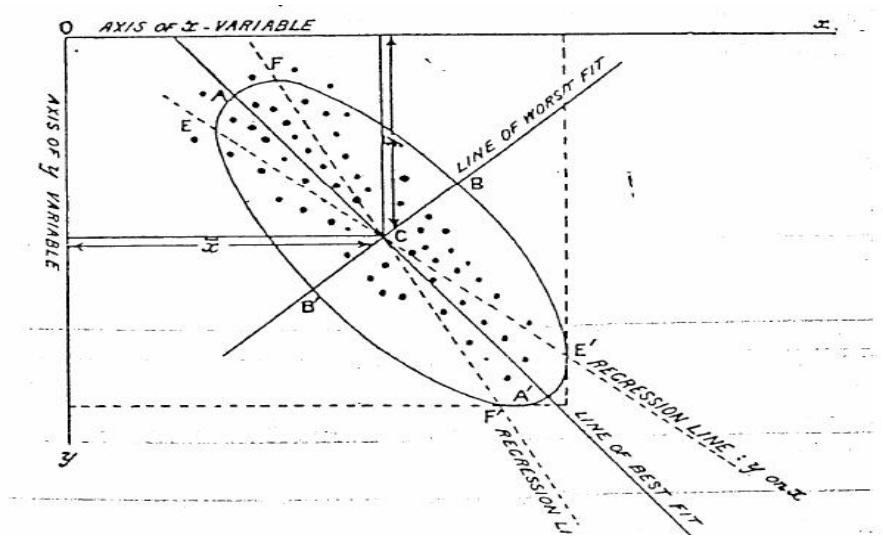
where $\mathbf{C} = \mathbf{U}_c \Sigma_c \mathbf{V}_C'$ is the SVD. The approach cost is $\mathcal{O}(nk^2)$.

According to [248], it is not possible to approximate the kernel for a randomized matrix. *The randomized SVD assumes that the matrix is fully observed; unfortunately, this is not true for kernel methods. Therefore, the primary objective of kernel approximation is to avoid forming the whole kernel matrix. The existing random projection methods all require the full observation of the matrix, so random projection is not a feasible option. We must use column selection in the kernel approximation problem* [248].

2.9.3 CUR Approximation

The CUR approach provides a solution when kernel methods do not. It involves extracting a matrix $\mathbf{C}$ into a lower dimensional space (for rows and columns) and using a compressed representation of the data, or a sample, obtaining the truncated SVD [108].

Figure 2.1: Principal Components.



Reproduced from the original paper of [194]. Principal components are lines that best fit 2-dim plane data projections.

2.10 The Procrustes Problem

The Procrustes problem consists of rotating a matrix $\mathbf{B} \in \mathbb{R}^{m \times k}$ into $\mathbf{X} \in \mathbb{R}^{m \times k}$. This problem is defined as [248]

$$\min_{\mathbf{Q} \in \mathbb{R}^{k \times k}} \|\mathbf{BQ}\|_2^2 \tag{2.48}$$

for $\mathbf{Q}'\mathbf{Q} = \mathbf{I}$, and the minimizer of the Procrustes problem is $\mathbf{UV}'$.

2.11 Principal Component Analysis

In applications, the terms SVD and PCA are sometimes interchanged. PCA is a particular case of SVD, in which the columns of $\mathbf{X}$ are centered as

$$\mathbf{X}_c = \mathbf{X} - \big[\mathbf{J}\bar{x}_1|\dots|\mathbf{J}\bar{x}_n\big]_{ij} \qquad \big(\text{with } \bar{x}_j = \frac{1}{m}\sum_i x_{ij} \text{ for all } j\big) \qquad (2.49)$$

where $\mathbf{J}$ is an $m \times 1$ dimension matrix of ones, the second term relation (2.49) is the expectation of $\mathbf{X}$, and $E(\mathbf{x}_i'\mathbf{x}_i) = 0$ (s.t. $\mathbf{x}_i \in \mathbf{X}_c$)

PCA is one of the most widespread tools for data analysis. Its basic concepts were formulated by Pearson [194]. The objective of PCA is to find an orthogonal axis system, called principal components in a Euclidean space, maximizing the projection (variance) of the zero-mean variables. In many cases, PCA is a dimension-reduction problem retaining as much as possible of the information or variance [1].

From (2.49), and for real entries-centered matrices, the covariance matrix $\mathbf{S}$ is related with Σ^2 of the SVD theorem as

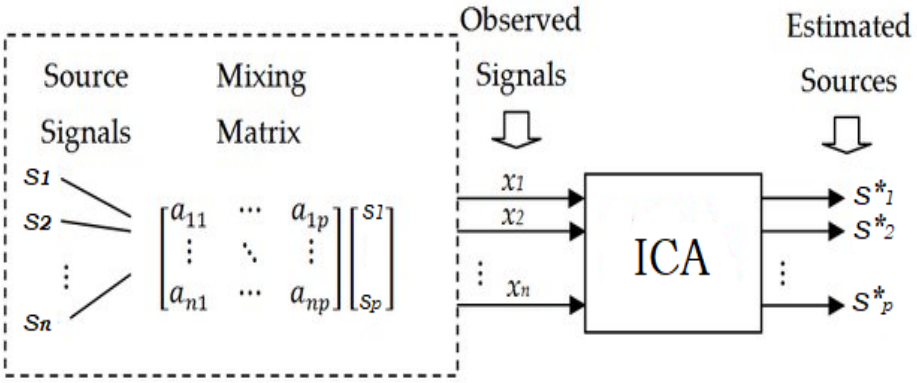
$$\begin{aligned} \Sigma^2 &= \mathbf{X}_c'\mathbf{X}_c && (2.50)\\ &= E\big(\mathbf{X} - E(\mathbf{X})\big)'E\big(\mathbf{X} - E(\mathbf{X})\big) && (2.51)\\ &= \mathbf{S} && (2.52) \end{aligned}$$

The trace of matrix $\mathbf{S}$ (which is diagonal) is often called *intertia* or *total intertia*. The first r elements are the *explained intertia* or percentage of *explained variance*.

PCA provides graphical representations for the orthogonal projections of observations on the planes formed by the consecutive pairs of columns of the matrix $\mathbf{V}$, which are orthogonal as a consequence of the SVD (i.e., $\Sigma^{1/2}\mathbf{V}$). If PCA is understood as a dimensional-reduction problem, there are several methods to determine the number of components to retain. A quantitative idea is based on the *condition number*, defining the GAP as in Cichocki [57, p. 72] taking the GAP as

[1] PCA dimension refers to the geometric multiplicity of the eigenvalues σ_r of the SVD theorem and corresponds to dim $E(\sigma)$, with $E(\sigma) = \{\mathbf{v} \in \mathbb{R}^m \text{ s.t } \mathbf{Y}\mathbf{v}_r = \sigma\mathbf{u}_y\}$ being $\mathbf{u}_r$ and $\mathbf{v}_r$ vectors of $\mathbf{U}$ and $\mathbf{V}$, respectively. The nonzero roots of σ such that $\det(\mathbf{Y} - \sigma\mathbf{I}) = 0$, or characteristic polynomial is the algebraic multiplicity . Both ideas play a fundamental role in the canonical forms [178, Chap. 10] and the interpretation of dimensionality in matrix analysis.

Figure 2.2: Independent Component Analysis.



Reproduced from [177]. Several signals sources are mixed in a matrix. Projections are the observed signals ICA consists to separate noise into observations, providing the informative or source signals.

$$GAP(k) = \frac{\text{var}\,\sigma_{r+1}^{K-1}}{\text{var}\,\sigma_{r}^{K-1}} \tag{2.53}$$

which is a more elegant formulation of the *elbow rule* (approximate according to the greater jump of the ordered eigenvalues). In Figure 2.1, we show the initial idea of reduction of variance direction and relation with projections, as in Pearson [194].

2.12 Independent Component Analysis

PCA is a variance-based representation and is a *low rank* approximation by taking the k largest eigenvectors associated with their corresponding eigenvalues. ICA provides a measure of independence other than variance, useful when the data (signals) depend on time. The non-existence of correlation means independence only if variables (column normalized matrix $\mathbf{X}$) are Gaussian, but not in other cases. ICA is introduced to help in such situations. The objective is to separate observations into the underlying signals. Figure 2.2 explains this idea.

In this case, the matrix $\mathbf{X}$ is transformed as

$$\mathbf{S} = \mathbf{B}\mathbf{X} \tag{2.54}$$

where $\mathbf{B}$ is a basis. An approximation is obtained by minimizing a divergence between $\mathbf{S}$ and the column-normalized matrix $\mathbf{Y}$.

Minor components are associated with noise. To extract them, a set of techniques known as Blind Source Separation (BSS) exists. This approach assumes the consideration of dynamic systems, in which the entries $\mathbf{X} = \mathbf{X}(t)$ are time-dependent (t is time) and centered. The classic approach, attributed to Oja [188], supposes an update of the matrix $\mathbf{W} = (w_1, \ldots, w_m)'$ with the update rules

$$\mathbf{W}(t) = \mathbf{W}(t-1) + \gamma(t)\mathbf{x}(t)\mathbf{x}(t)'\mathbf{W}(t-1) \tag{2.55}$$

where $\gamma(t)$ is a scalar representing the gain parameter, $\mathbf{x}(t)$ the systems inputs, and $\mathbf{W}$ are the constraints to maximize $E(\mathbf{w}_i'\mathbf{x})$, subject to orthogonality.

Orthonormalizing the expression (2.55) by introducing a suitable array $\mathbf{S}(t)$

$$\mathbf{W}_\perp(t) = \mathbf{W}(t)\mathbf{S}(t)^{-1} \tag{2.56}$$

$\mathbf{W}_\perp(t)$ is an orthonormal matrix.

Taking into account that the product $\mathbf{x}(t)\mathbf{x}(t)'$ is the covariance matrix, which is represented now as $\mathbf{V}$, the differential equations corresponding are obtained by simple differentiation of (2.55)

$$\dot{\mathbf{W}} = \mathbf{X}\mathbf{X}'\mathbf{W} - \mathbf{W}\mathbf{W}'\mathbf{X}\mathbf{X}'\mathbf{W} \tag{2.57}$$

$$= \mathbf{V}\mathbf{W} - \mathbf{W}\mathbf{W}'\mathbf{V}\mathbf{W} \tag{2.58}$$

The dot means, as usual, the time derivative. It is stable in the Lyapunov sense (a linear combination nearby solution differs a first-order infinitesimal).

Further works by Chen [55] established that the necessary and sufficient condition to extract the principal space (principal components) is that the initial condition $\mathbf{W}(0)$ must be full rank, and in this case, it occurs that

$$\mathbf{W}(t) \xrightarrow[t\to\infty]{} \mathbf{W} \tag{2.59}$$

and introducing $\mathbf{W} = \theta\mathbf{D}$ ($\mathbf{D}$ is diagonal) to orthogonalize, we have

$$\dot{\mathbf{W}} = \mathbf{VWD} - \mathbf{WW}'\mathbf{VW} \tag{2.60}$$

in this case, the SVD of $\mathbf{W}$ takes the form

$$\mathbf{W}(t) = \mathbf{U}(t)\mathbf{D}(t)\mathbf{V}(t) \tag{2.61}$$

with invariance properties for $\mathbf{U}$, $\mathbf{D}$, and $\mathbf{V}$ in (2.60) [5, p. 321].

A more recent approach focused on the detection of minor components, with illustrative examples, is that by Tan [221].

Related BSS methods have grown remarkably since the formulation of pioneering works by Oja [188] and Chen [55], finding application in physics, engineering, finance, and medicine, among many others. Without being unfair to the many excellent published works, we highlight the study by Cichocki, in which using non-negative matrices is recommended when the dimensionality of the latent variable space is unknown [58].

Chapter 3

The EM Approach: PLSA

PLSA is an unsupervised learning technique developed for information retrieval purposes. This method, also known as Probabilistic latent semantic indexing (PLSI), was introduced in a conference proceedings [126, 127], its classical reference being the paper *unsupervised learning by probabilistic latent semantic analysis* by Hofmann [129].

PLSA is based on the ideas of LSA [68] and is in fact a probabilistic version of this concept. LSA uses cross terms and documents of a corpus to obtain a count or a table of co-occurrences. Arranged as a matrix, the SVD space span is considered as a set of latent variables and has been interpreted as the aspect model [205]. PLSA associates counts with frequencies, and with decomposition as mixtures or aggregate Markov models [129], which are adjusted with the EM algorithm.

The similarity between documents can be considered as a distance. Two probabilistic decompositions are possible: asymmetric and symmetric formulations, as named by Hofmann [129]. The symmetric formulation is related to the SVD.

PLSA's versatility, clarity of results, and solid statistical properties have enabled a wide range of applications, in which the concepts of words and documents are assimilated into other discrete entities, thus enabling justification of the hypotheses on which PLSA relies. Examples can be found in bioinformatics [51, 172]; identification of genomic sequences with documents and some classes of genotype characteristics, such as as words; collaborative filtering, in which user ratings are used to construct suitable matrices

to perform PLSA algorithms [147]; and speech recognition, introducing a score concatenation matrix [135, 169]. PLSA is also a start point for image semantic analysis in *which visual patterns describe each object* [222].

Several variants are used in PLSA image analysis, such as co-regularized PLSA, for the recognition of observed images with different perspectives [143], among many others. For applications and developments in semantic image analysis, we refer to the Tian (2018) review [222] and its selected references.

3.1 Approaches and Solutions

The original formulation of PLSA according to Hofmann [129], provides a probabilistic solution to the problem of extracting a set of z_k ($k = 1, \cdots, k$) latent variables of a data frame $N(d_i, w_j)$, obtained from a corpus of d_i ($i = 1, \ldots, m$) documents when crossed with a thesaurus of w_j ($j = 1, \ldots, n$) words. The relative frequencies

$$n(d_i, w_j) = \frac{N(d_i, w_j)}{\sum_{ij} N(d_i, w_j)} \tag{3.1}$$

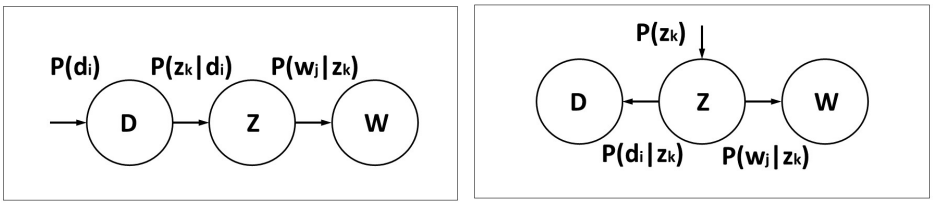
are estimated by the joint probability $P(d_i, w_j)$. A key idea in this method is decomposing this probabilistic approximation as the product of conditional distributions over a set of latent variables. After some manipulations, and using the Bayes rule,

$$P(d_i, w_j) = P(d_i) \sum_k P(w_j|z_k)P(z_k|d_i) \tag{3.2}$$

$$= \sum_k P(z_k)P(w_j|z_k)P(d_i|z_k) \tag{3.3}$$

where $P(d_i)$ and $P(z_k)$ are probabilities of the document d_i and the latent variable z_k, respectively. Formulas (3.2) and (3.3) are known as asymmetric and symmetric formulations [129], or formulations I and II [133], giving rise to the generative models shown in Figure 3.1.

Figure 3.1: PLSA generative models.



Reproduced from [129]. Generative models: at left asymmetric formulation; at right symmetric formulation.

The discrete nature of the documents identifies each one with the probabilities of $(d_1, \ldots, d_n)^t$ over the latent variables, and justifies the postulation that the mixtures $P(d_i|\, z_k)$ are k *iid* multinomials. Because the same occurs for the words, the objective is to determine the parameters θ and ϕ such that the conditional probabilities $P(w_j|\, z_k) \sim \text{Mult}(\theta_{jk})$ and $P(z_k|d_i) \sim \text{Mult}(\phi_{ki})$ for the asymmetric formulation (alternatively $P(w_j|\, z_k) \sim \text{Mult}(\theta_{jk})$ and $P(d_i|\, z_k) \sim \text{Mult}(\phi_{ik})$ for the symmetric case), with no hypothesis regarding the number or distribution of z_k, which is a set of *dummy* variables with no probabilistic sense.

The adjustment of mixtures, given by Formulas (3.2) and (3.3), is the other key idea for obtaining a reliable probabilistic interpretation. The method used for this purpose is the EM algorithm, which always converges [69]. The use of the EM algorithm is roughly equivalent to the problem of fitting $P(d_i, w_j)$ to $n(d_i, w_j)$, but ensuring a maximum likelihood estimation of the sufficient (not necessarily minimal) parameters θ and ϕ.

The EM algorithm supposes two steps: expectation and maximization. Expectation (E-step) is computed on the log-likelihood

$$\mathcal{L} = \sum_{ij} n(d_i, w_j) \log\, P(d_i, w_j) \tag{3.4}$$

and for parametrization (3.2) or (3.3) takes the forms

$$\mathcal{L} = \sum_{ij} n(d_i, w_j) \log \Big\{ P(d_i) \sum_k P(w_j|\, z_k) P(z_k|\, d_i) \Big\} \tag{3.5}$$

$$= \sum_{ij} n(d_i, w_j) \log \Big\{ \sum_k P(z_k) P(w_j|\, z_k) P(d_i|\, z_k) \Big\} \tag{3.6}$$

for the asymmetric and the symmetric cases, respectively. In both cases, the expectation of $\mathcal{L}$ is the posterior

$$E(\mathcal{L}) = P\left(z_k|\, d_i, w_j\right) \tag{3.7}$$

and after several manipulations

$$P(z_k|\, d_i, w_j) = \frac{P\left(z_k, d_i, w_j\right)}{P\left(d_i, w_j\right)} \tag{3.8}$$

The calculation of expectation $E(\mathcal{L})$ presents several complications related to the expression $P(z_k|\, d_i, w_j)$ of Formula (3.7). For computational purposes, the object supporting this data structure is an array containing the matrices with the estimates of $P(d_i, w_j)$, fixing for each one the values of z_k. Then each of the elements of the array is a matrix taking the form

$$[P(d_i, w_j)]_{ijk'} = \mathrm{vec}\big[P(\cdot|\, z_{k'})\big]\, \mathrm{vec}\big[P(\cdot|\, z_{k'})\big]^t \tag{3.9}$$

(for $k' = 1, \cdots, k$), indicating the primed index that is fixed. In this case, the *vec* notation has been used to better identify the scalar products of the vectors of probabilities $P(\cdot|\, z_{k'})$ obtained by varying z_k. The entire array is

$$[P(d_i, w_j)]_{ijk} = \Big[[P(d_i, w_j)]_{ij1} \Big| \ldots \Big| \, [P(d_i, w_j)]_{ijk} \Big] \tag{3.10}$$

Maximization (M-step) uses Lagrange multipliers λ_1, λ_2, and λ_3, and, for the asymmetric formulation, provides the functional $\mathcal{H}$

$$\begin{aligned}\mathcal{H} =& E\mathcal{L} + \lambda_1\big(1 - \sum_i P(d_i)\big) + \lambda_2 \sum_k P(w_j|\, z_k)\big) \\ &+ \lambda_3 \sum_i \big(1 - \sum_k P(z_k|\, d_i)\big)\end{aligned} \tag{3.11}$$

with derivatives

$$\begin{aligned}\frac{\partial\, H}{\partial\, P(d_i)} &= \sum_i \sum_j n(d_i, w_j)\, \frac{P\,(z_k|\, d_i, w_j)}{P(z_k)} - \lambda_1 \\ \frac{\partial\, H}{\partial\, P(w_j|\, z_k)} &= \sum_j n(d_i, w_j) \frac{P(z_k|\, d_i, w_j)}{P(x_j|\, z_k)} - \lambda_2 \\ \frac{\partial\, H}{\partial\, P(z_k|\, d_i)} &= \sum_i n(d_i, w_j) \frac{P(z_k|\, d_i, w_j)}{P(z_k|\, d_i)} - \lambda_3\end{aligned}$$

and equaling them to zero, leads to

$$\begin{aligned}\sum_i \sum_j n(d_i, w_j) P(z_k|\, d_i, w_j) - \lambda_1 P(w_j) &= 0 \\ \sum_j n(d_i, w_j) P(z_k|\, d_i, w_j) - \lambda_2 P(d_i|\, w_j) &= 0 \\ \sum_i n(d_i, w_j) P(z_k|\, d_i, w_j) - \lambda_3 P(w_j|\, z_k) &= 0\end{aligned}$$

the solutions of which are

$$P(d_i) = \frac{\sum_j \sum_k n\,(d_i, w_j)\, p(z_k|\, d_i, w_j)}{\sum_i \sum_j \sum_k n(d_i, w_j) P(z_k|\, d_i, w_j)} \tag{3.12}$$

$$P(w_i|\, z_k) = \frac{\sum_i n(d_i, w_j) P(z_k|\, d_i, w_j)}{\sum_j \sum_j n(d_i, w_j) P(z_k|\, d_i, w_j)} \tag{3.13}$$

$$P(z_k|\, d_i) = \frac{\sum_j in(d_i, w_j) P(z_k|\, d_i, w_j)}{\sum_k \sum_j n(d_i, w_j) P(z_k|\, d_i, w_j)} \tag{3.14}$$

For the case of symmetric formulation, the functional of the expectation is

$$\begin{aligned}\mathcal{H} =& E(\mathcal{L}) + \lambda_1\Big(1 - \sum_k P(z_k)\Big) + \lambda_2 \sum_k \Big(1 - \sum_i P(d_i|\, z_k)\Big) \\ &+ \lambda_3 \sum_k \Big(1 - \sum_j P(w_j|\, z_k)\Big) \end{aligned} \tag{3.15}$$

with derivatives

$$\begin{aligned}\frac{\partial\, H}{\partial\, P(z_k)} &= \sum_i \sum_j n(w_j, d_i) \frac{P(z_k|\, d_i, w_j)}{P(z_k)} - \lambda_1 \\ \frac{\partial\, H}{\partial\, P(d_i|\, z_k)} &= \sum_k n(w_j, d_i) \frac{P(z_k|\, d_i, w_j)}{P(d_i|\, z_k)} - \lambda_2 \\ \frac{\partial\, H}{\partial\, P(w_j|\, z_k)} &= \sum_g n(d_i, w_j) \frac{P(t|d_i, f)}{P(z_k|\, z_k)} - \lambda_3\end{aligned}$$

By solving these equations we have

$$P(z_k) = \frac{\sum_j \sum_i n(d_i, w_j) P(z_k|\, d_i, z_k)}{\sum_j \sum_i \sum_k n(d_i, w_j) P(z_k|\, d_i, z_k)} \tag{3.16}$$

$$P(d_i|\, z_k) = \frac{\sum_j n(d_i, w_j) P(t|\, d_i, z_k)}{\sum_i \sum_j n(d_i, w_j) P(z_k|\, d_i, z_k)} \tag{3.17}$$

$$P(w_j|\, z_k) = \frac{\sum_i n(d_i, w_j) P(z_k|\, d_i, z_k)}{\sum_j \sum_i n(d_i, w_j) P(z_k|\, d_i, z_k)} \tag{3.18}$$

which are for the symmetric formulation.

The adjustment of probabilities, in both formulations, involves selecting a value for k, initializing the distributions appearing in (3.2) or (3.3), and computing the E-step and M-step in an iterative process in which $P(d_i, w_j)$ is recalculated until a certain condition is achieved. Hofmann [129] has noted that the iterative process can end when there are no changes in the qualitative inputs, a condition called *early stop* [129].

Although the formal equivalence between models is clear, since from the Bayes rule

$$P(z_k|\, d_i) = \frac{P(d_i|\, z_k)P(z_k)}{P(d_i)}$$

its easy to see that

$$\begin{aligned} P(z_k|\, d_i)P(d_i) &= P(d_i|\, z_k)P(z_k) \\ P(w_j|\, z_k)P(z_k|\, d_i)P(d_i) &= P(w_j|\, z_k)P(d_i|\, z_k)P(z_k) \\ P(d_i)\sum_k P(w_j|\, z_k)P(z_k|\, d_i) &= \sum_k P(w_j|\, z_k)P(d_i|\, z_k)P(z_k) \end{aligned}$$

the interpretation of which is less clear. Hofmann [129] does not provide indications of when formulations are suitable.

Also, from our view point, the most important point is the equivalence of the symmetric formulation with the SVD. Hofmann [129] explicitly writes

$$[\mathbf{U}]_{ik} \sim P(d_i|\, z_k) \tag{3.19a}$$

$$\mathrm{diag}(\Sigma)_k \sim P(z_k) \tag{3.19b}$$

$$[\mathbf{V}]_{kj} \sim P(w_j|\, z_k) \tag{3.19c}$$

Indication that is a merely formal equivalence with no statistical significance. Is of our interest uniquely symmetric formulation.

3.1.1 Example

An example provided by Hofmann is reproduced in table 3.1 to illustrate the sense of word rank for interpretation of *the* 4 *aspects that most likely generate the word "segment," derived from a K=128 aspect model of the CLUSTER document collection. The displayed word stems are the most probable words in the class-conditional distributions* $P(w_j|\, z_k)$*, from top to bottom in descending order* [129].

In addition, we provide an artificial example to illustrate the effects of selection of k, consisting of a corpus of 5 ($d1$ to $d5$) documents containing

Table 3.1: Hofmann's example.

Aspect 1	Aspect 2	Aspect 3	Aspect 4
imag	video	region	speaker
SEGMENT	sequenc	contour	speech
color	motion	boundari	recogni
tissu	frame	descript	signal
Aspect1	scene	imag	train
brain	SEGMENT	SEGMENT	hmm
slice	shot	precis	sourc
cluster	imag	estim	speakerindep.
mri	cluster	pixel	SEGMENT
algorithm	visual	paramet	sound

letters $\{a, b, c, d, e, f\}$, which we assimilate into words in a thesaurus. The co-occurrences' data frame is N and the frequency matrix is

$$N(d_i, w_j) = \begin{array}{c} \\ d1 \\ d2 \\ d3 \\ d4 \\ d5 \end{array} \begin{array}{c} \begin{array}{cccccc} a & b & c & d & e & f \end{array} \\ \begin{pmatrix} 3 & 4 & 0 & 0 & 0 & 0 \\ 3 & 3 & 0 & 0 & 0 & 0 \\ 1 & 3 & 4 & 1 & 0 & 0 \\ 0 & 0 & 2 & 4 & 0 & 0 \\ 0 & 0 & 0 & 0 & 3 & 4 \end{pmatrix} \end{array}$$

and the correspondent frequency matrix is

$$n(d_i, w_j) = \begin{bmatrix} .086 & .114 & 0 & 0 & 0 & 0 \\ .086 & .086 & 0 & 0 & 0 & 0 \\ .029 & .086 & .114 & .029 & 0 & 0 \\ 0 & 0 & .057 & .114 & 0 & 0 \\ 0 & 0 & 0 & 0 & .086 & .114 \end{bmatrix}$$

If in this example, the objective is to classify documents by subject (or specialized words with the correspondent matters), simple visual inspection

indicates that there are 3. For the symmetric case formulas, running $p = 1000$ iterations in each case, the results for $k = 2$ are

$$P(d_i|\, z_k) = \begin{bmatrix} 0 & .411 \\ 0 & .588 \\ .333 & 0 \\ .278 & 0 \\ .167 & 0 \\ .222 & 0 \end{bmatrix} \quad n(d_i, w_j) = \begin{bmatrix} c & b \\ d & a \\ f & - \\ e & - \\ - & - \\ - & - \end{bmatrix}$$

for $k = 3$

$$P(d_i|\, z_k) = \begin{bmatrix} 0 & 0 & .462 \\ 0 & 0 & .538 \\ .545 & 0 & 0 \\ .454 & 0 & 0 \\ 0 & .429 & 0 \\ 0 & .571 & 0 \end{bmatrix} \quad n(d_i, w_j) = \begin{bmatrix} c & f & b \\ d & e & a \\ - & - & - \\ - & - & - \\ - & - & - \\ - & - & - \end{bmatrix}$$

and for $k = 5$

$$P(d_i|\, z_k) = \begin{bmatrix} .007 & 0 & 0 & .462 & 0 \\ .347 & 0 & 0 & .538 & 0 \\ .646 & .333 & 0 & 0 & 0 \\ 0 & .667 & 0 & 0 & 0. \\ 0 & 0 & .538 & 0 & .419 \\ 0 & 0 & .462 & 0 & .581 \end{bmatrix}$$

$$n(d_i, w_j) = \begin{bmatrix} c & d & e & b & f \\ b & c & f & a & e \\ a & - & - & - & - \\ - & - & - & - & - \\ - & - & - & - & - \\ - & - & - & - & - \end{bmatrix}$$

The characters' matrices are the ordination of the most likely words identifying each latent variable (informally, the subjects in our toy example). Lines represent probabilities close to zero and are not useful for classification. The effect of selecting K is clear in the comparison of columns 3 and 5, which are equivalent (for $k = 5$).

3.1.2 Training and Prediction

The PLSA algorithm can be executed for the entire dataset, providing results in the same manner as probabilistic clustering methods [2, Chap. 3]. However, to exploit the predictive power of PLSA, the model must be fitted on the available data (or training phase). Predictions for new observations are made by simply comparing them with the trained dataset. Both formulations have different features.

The asymmetric formulation, in the prediction phase, cannot assign probabilities for documents that are not in the training phase, because non-zero probabilities are needed. This problem has been solved in Brants [42] by splitting the dataset into a training group with the d_i observed documents and the new unobserved documents $q \in \mathcal{Q}$. By using probabilities $P(z_k|\, d_i)$ instead of $P(d_i|\, z_k)$ in (3.2) and expanding the logarithm, this equation can be rewritten as

$$\mathcal{L} = \sum_{ij} n(d_i, w_j) \log\, P(d_i) + \sum_{ij} n(d_i, w_j) \log\, P(w_j|\, d_i) \tag{3.20}$$

To avoid a zero probability of the unseen documents in the training phase, Brants [42] introduced $P(d_i) > 0$, stating that the log-likelihood can be maximized, taking into account only the second term of (3.20), and

$$P^{(new)}(Q) = \prod_{ij} P(w_j|\, q_i) \tag{3.21}$$

Brants [42] has pointed out that equation (3.20) *does not represent the true likelihood, but if the goal is likelihood maximization, the same parameter setting is found as that when the true likelihood had been maximized* [42]. The same article proposed other methods for estimating likelihood on the basis of marginalization and splitting. Brants [42] also proposed PLSA folding-in, a more refined derivation of this technique [40]. A further improvement, which is more computationally efficient and is protected by a patent, is [41], which involves estimating the log-likelihood by spiting the dataset in the training set, denoted $n'(d_i, w_j)$, and introducing the unknown documents one by one as the second term of

$$\mathcal{L} \propto \sum_{ij} n'(d_i, w_j) \log\, P(d) + \sum_{ij} \log\, P(w_j|\, d_i) \tag{3.22}$$

In the symmetric formulation, after training on the documents by using the formulas (3.16)-(3.18), new documents can be classified by simply alternating the expressions given by Masseroli [172]

$$P(z_k|\, d_i, w_j) = \frac{P(z_k|\, d_i)P(w_j|\, z_k)}{\sum_{k'} P\,(z_{k\prime}|\, d_i)P(w_j|\, z_{k\prime})} \tag{3.23}$$

$$P(z_k,\, d_i) = \frac{\sum_i n(d_i, w_j)P(z_k|\, d_i, w_j)}{\sum_{ij} n(d_i, w_j)P(z_k|\, d_i, w_j)} \tag{3.24}$$

In this case, binary data can be handled by entering a matrix $\mathbf{A}$ [172] such that

$$[\mathbf{A}]_{ij} = \begin{cases} 1 & \text{if } i \text{ is annotated to } j \\ 0 & \text{otherwise} \end{cases} \tag{3.25}$$

substituting $n(d_i, w_j)$ in equations (3.16)-(3.18).

PLSA can also be used as a semi-supervised learning tool in a process known as semi-supervised PLSA [251]. Using this mode requires entering labeled and non-labeled data in the EM iterative process, and being able to split the dataset into a portion in which the labels are assigned and a portion in which the labels are not assigned. A measure of similarity performs the rest of the task. Another related strategy involves introducing the link functions *must-link* and *cannot-link* in the training phase [187].

3.1.3 Continuous Data

A generalization of the PLSA for continuously evaluated responses has also been provided by Hofmann [129] in the context of collaborative filtering, as an alternative to the neighbor-regression method [130]. The method

construction assumes a set of person items y_j rated v for a set of persons u_i. Then,

$$P(v|\, u, y) = \sum_z P(z|\, u) P(v; \mu_{yz}, \sigma_{yz}) \tag{3.26}$$

where μ and σ are the expectation and variance, respectively, and assuming

$$P(v; u, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\Big\{ - \frac{(v-\mu)^2}{2\sigma^2} \Big\} \tag{3.27}$$

which is fitted with the EM algorithm.

Within the semantic image analysis field, the visual entities from a database are assimilated with the words from a thesaurus [31], but as discrete entities. This variant constitutes the Gaussian mixture model PLSA [134], and assumes a normal distribution of the descriptors f_j (the most relevant visual words) such that $f_h \sim N(f_h|\, \mu_k, \Sigma_k)$ $(h \leq j)$. Horster [134] has noted that this expression is difficult to train, and has proposed the alternative models 'shared Gaussian words PLSA' and 'fixed shared Gaussian words PLSA'. A more general treatment, in which normality is postulated for the mixtures $P(w_j|\, d_i)$, has been reported [165].

3.1.4 Tensorial Approach

Non-negative tensor factorization was introduced by Shashua [214] for n-way data structures. Peng [195] has established the relationship with PLSA in [195], noting that the case $n = 2$ corresponds to the NMF and illustrating that $n = 3$ allows for handling more complex data structures. Peng [195] has introduced a structure of the type

$$[\underline{\mathbf{F}}]_{ijl} \approx P(d_i, w_j, x_l) \tag{3.28}$$

called a tensor, and being now x_l $(l = 1, \ldots L)$ other probabilistic observations. The extension of these ideas to the PLSA is obtained by considering the factorizations for $k > r$

$$P(d_i, w_j, x_l) = \sum_p P(d_i|\, x_p)P(w_j|\, z_r)P(z_k|\, z_r)P(x_p, y_q, z_k) \tag{3.29}$$

$$= \sum_r P(d_i|\, x_r)P(w_j|\, x_r)P(z_k|\, x_r)P(x_r) \tag{3.30}$$

These decompositions are the tensorial cases of the asymmetric and symmetric formulations given by Formulas (3.2) and (3.3).

Two methods exist for adjusting Formulas (3.29) and/or (3.30): parallel factor analysis (parafrac) [118], assuming a linear approximation of the fibers (the one-dimensional structures that can be extracted from $P(d_i, w_j, z_k)$) and Tucker [14], a multiway PCA. Both methods provide different results, even when the objective function is the same [195], and indicates that the method is useful for determining the number of latent factors. An alternative formulation has been proposed by Yoo [244]. This decomposition has several implications, mainly related to neural network applications.

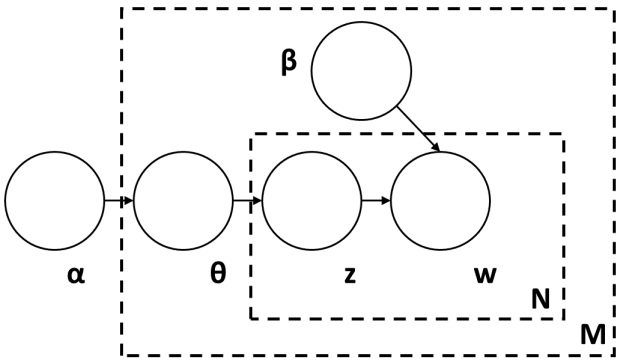
3.2 Other Formulations

One of the first criticisms was noted by Blei [30], who argued that *Hofmann's work is incomplete in that it provides no probabilistic model at the level of documents. This incompleteness leads to several problems: (i) the number of parameters grows linearly with the size of the corpus, thus resulting in severe problems with overfitting, and (ii) how to assign probabilities to a document outside the training set* is unclear; LDA has been proposed to solve this problem [30].

LDA introduces a generative Bayesian model that maps documents on topics such that the words of each document are captured by these topics. Each document is described by a topic distribution, and each topic is described by a word distribution. Introducing θ, a k dimensional Dirichlet with parameter α_k, and β as an array of initialization with values $P(w|\, z)$, and maintaining the notation of Formulas (3.2) and (3.3),

$$P(\theta, \mathbf{z}, \mathbf{w}|\, \alpha, \beta) = P(\theta|\, \alpha) \prod_k P(z_h|\, \theta)P(w_j|\, z_k, \beta) \tag{3.31}$$

Figure 3.2: LDA generative model.



Reproduced from [30]: Latent Dirichlet Allocation generative model.1. Choose a $N \sim Poiss(\alpha)$. 2. Choose $\theta \sim Dir(\beta)$. For each of the n words: (a) Choose a topic $z_k \sim Mult(\theta)$. (b) Choose a word w_i from $P(w_i|\, z_k, \beta)$.

The probabilities of a document and a corpus are obtained by marginalizing (integrating) over the complete collection. Further improvements to the model, also provided by Blei [30], include hierarchical LDA [28] and dynamic LDA [29].

LDA is a closely related technique that is different from PLSA. Its criticisms provided a starting point for several developments. Formal equivalence with PLSA has been shown by Girolami [106] and has led to several proposed solutions to those problems in the case of PLSA. The generative model is shown in Figure 3.2.

3.2.1 Algorithms Based on Expectation Maximization Improvement

PLSA must preserve maximum log-likelihood solutions, obtained with the EM algorithm. Also, the EM algorithm is one of the most studied in statistical environments, and many variants and simplifications exist [174]. A general description of the algorithm used in this section is provided by Roche [199]. Some versions or modifications of the EM have been exploited to achieve PLSA solutions.

3.2.2 Tempered EM

Tempered EM uses classical concepts of statistical mechanics for computational purposes [123]. Aside from its significance in physics, the primary idea is how to achieve a posterior (E-step) close to a uniform distribution. An objective function is introduced:

$$\begin{aligned}\mathcal{F}_\beta = &- \beta \sum_{ij} n(d_i, w_j) \sum_k \widetilde{P}(z_k|\, d_i, w_j) \log \left[P(d_i|\, z_k) P(w_j|\, z_k) P(z_k)\right] \\ &+ \sum_{ij} n(d_i, w_j) \sum_k \widetilde{P}(z_k|\, d_i, w_j) \log\, P(z_k|\, d_i, w_j) \end{aligned} \tag{3.32}$$

where $\widetilde{P}(z_k|\, d_i, w_j)$ is a variational parameter defined as

$$\widetilde{P}(z_k; d_i, w_j) = \frac{\left[P(z_k) P(d_i|\, z_k) P(w_j|\, z_k)\right]^\beta}{\sum_k \left[P(z_k) P(d_i|\, z_k) P(w_j|\, z_k)\right]^\beta} \tag{3.33}$$

and for $\beta < 1$, the convergence is faster [129].

3.2.3 Sparse PLSA

A proposal to improve the convergence has been based on sparse EM [184]. Assuming that only a subset of values is plausible for latent variables (in terms of probabilities), freezing non-significant avoids many calculations. PLSA is considered an algebraic optimization problem of the matrix $\mathbf{Y} \in \mathbb{R}^{m \times n}$ (which in this case is the data frame containing the relative frequencies $n(d_i, w_j)$) restricted to the constraint $\sum_r \lambda_r \mathbf{y}_r \mathbf{y}_r^t$ $(r < m)$, or unknown parameters, minimizing the divergence $D_q(\mathbf{Y} \| \sum_r \lambda_r \mathbf{y}_r \mathbf{y}_{r'})$ and satisfying [120]

$$\sum_r \lambda_r = \|\mathbf{y}_r\|_1 = \|\mathbf{y}'_r\|_1 = 1, \quad \text{for all } \mathbf{y}_r \in \mathbf{Y}\ (r' \neq\ r < n) \tag{3.34}$$

named Tsallis divergence [225], and computed for the r non-freezing column vectors of $\mathbf{Y}$ as [57, p.97]

$$D_q(\mathbf{y}_j \| \lambda_r \mathbf{y}_r \mathbf{y}_r^t) = \frac{1}{\kappa} \sum_i \left(\mathbf{y}_j \left(\mathbf{y}_j^\kappa - (\lambda_r \mathbf{y}_r \mathbf{y}_r^t)^\kappa \right) \right) - \sum_i \left(\mathbf{y}_j^\kappa \left(\mathbf{y}_j - \lambda_r \mathbf{y}_r \mathbf{y}_r^t \right) \right) \quad (\text{s.t. } \kappa \neq 0) \tag{3.35}$$

This divergence solves the optimization problem of adjusting $n(d_i, w_j)$ to $P(d_i, w_j)$ [148]. After adjustment, probabilistic factorizations of the considered parametrization must again be obtained.

3.2.4 Incremental PLSA

Instead of global maximization, simpler contributions can be maximized. This update procedure used in the E-step for the PLSA gives rise to the incremental PLSA algorithm [242], with which results can be obtained twice as quickly. Applications in image classification can be found in Wu and Li [241, 164].

A recursive algorithm, called recursive probabilistic latent semantic analysis, is based on the computation of the likelihood of a subset of words, as well as other words, recursively [17]. Its performance has been reported to be highly similar to that obtained with incremental PLSA.

3.2.5 Randomized PLSA

Randomized PLSA arose to address the problem of overfitting together with difficulties in optimizing the relation (3.31) of the LDA, thus providing a solution in the framework of PLSA [200]. Taking a random fraction of the trained datasets, the method proceeds by folding the training dataset $\mathcal{T} = \{T_1, \ldots, T_\Omega\}$ and the fraction T^l ($T^l < \Omega$) to run the PLSA algorithm with l samples. The average of the results is the provided output.

The basis for this statement is the work by Ho [125] on the subspace method. This method takes random subsets of the support vector machine to avoid computational complexity. In addition, the derived algorithm has been reported to be slower than the conventional PLSA implementation.

3.3 Search for Computational Efficiency

PLSA is considered an effective technique but has a notable drawback in its high consumption of computing resources, in terms of both execution and internal memory. This drawback has limited its practical applications [234] and is inherited from the EM. Herein, contributions to increasing the computational efficiency are examined according to the concepts on which they are based, their initialization conditions, and the use of EM algorithm variants. Efforts using purely computational techniques are also discussed.

The NMF with minimization of KL divergence inherits the same problems as the PLSA, so it is computationally slow. The EM convergence rate, is [235]

$$\| \theta^{(p+1)} - \theta^{\star} \| \leq c \| \theta^{(p)} - \theta^{\star} \| \tag{3.36}$$

where c is the largest eigenvalue of the data matrix. Also, PLSA is highly dependent on the initialization values [88, 247].

3.3.1 Algorithm Initialization

The dependence of the PLSA results on the initialization conditions has led to several variations. One possibility applicable only in the symmetric formulation, proposed by Farahat [88], initializes the algorithm with LSA solutions, which are the SVD solutions . Because some values can be negative, correction may be necessary (typically setting values to zero). Another strategy applicable in both formulations is execution for several random initialization distributions of the considered algorithm; after running, the higher log-likelihood value offers the best solution [247].

One algorithm is On-line belief propagation (OBP), which is based on a sequence of initializations on subsets of the data frame [247, 243]. OBP segments the data frame into several parts. After initialization of the first segmentation, solutions are obtained and used in the next initialization, and so on. This technique enables the use of PLSA on large datasets.

A fundamental of the OBP is stochastic initialization [33], which consists of defining a learning function as a risk function for which the difference in

conditional distributions describes a decreasing sequence between iterations [33]. The execution of this algorithm requires at least one iteration for the complete dataset and selection of the most significant contributions for the first partition.

3.3.2 Use of Computational Techniques

Difficulties in obtaining fast and reliable solutions for PLSA have also been approached through purely computational techniques. These advancements are a consequence of developments in computer architecture in recent decades: processing capabilities have been increased, thus resulting in a new branch of algorithms to reduce the computational time for PLSA. The introduction of multicore processors by Intel and Sun Microsystems, in 2005, for portable machines enabled a major step toward parallel computing [10], which is now the dominant paradigm.

Parallel computing involves simultaneous execution of tasks. It requires dividing a problem into independent pieces and executing each one in a separate processing unit. The use of parallel computing techniques for PLSA has been proposed in Hong [132]. A current and widespread technique to support parallel capabilities is Map Reduce [67]. This technique essentially consists of dividing tasks into two phases. The first phase is a map that partitions the input dataset and assigns labels to each one. The reduce phase supposes the execution of an operation on a set of previously labeled partitions. An algorithm exploiting the possibilities of Map Reduce for PLSA results has been proposed by Jin [144].

Furthermore, graphic processing units have increased the range of capabilities, and they are useful for a broad variety of applications, particularly the simulation of complex models [109]. These capabilities have been transferred to the PLSA algorithm [156] but have not yet yielded definitive results.

3.4 Latent Variables Sense: A TL Interpretation

Transfer learning can be defined as the machine learning problem of *trying to transfer knowledge from a source domain to a target domain* [220].

PLSA can be used from the point of view of neural networks for transfer learning purposes, by solving the problem for the case in which the source domain shares only a subset of its classes (column vectors of the data matrix) for an unlabeled target data domain [158]. The log-likelihood expression is thus [158]

$$\begin{aligned}\mathcal{L} = &\sum_{ij} n(d^S, w_j) \log \sum_{ij} P(d^S|\, z_k^S) P(w_j|\, z_k^S) P(z_k|\, z_k^S) \\ &+ \sum_{ij} n(d^T, w_j) \log \sum_{ij} P(d^T|\, z_k^T) P(w_j|\, z_k^T) P(z_k|\, z_k^T)\end{aligned} \tag{3.37}$$

where S indicates that a document is in the source, and T indicates the target domain. A detailed survey providing an introduction to neural networks is provided by Bozinovski [39]. A similar work on the problem of transfer learning is Zhao [249].

3.5 Problems

Despite the relatively well-sound PLSA, there are concerns of a diverse nature. On the one hand, there are the problems with polysemy and synonymy, problems related with the use of the EM algorithm, and the examined computational problems in section 3.3, for which in the current State of the Art there are no satisfactory solutions. This results is problematic in real-time applications, especially for moderate or large datasets, according to the current standards. Regarding this point, we point out the parallelism with SVD, which also presents computational problems for large data structures, being this analogy curios (if the symmetric version of the PLSA is compared with the SVD).

Problems related to synonymy and polysemy compromise the use of PLSA as an unsupervised technique. There are two solutions to this problem. On the one hand, it may require the adjustment of each term to a previously established thesaurus. This approach presents the problem that words used as metaphors would be classified in another category, so providing misclassification errors. In addition, the richness of each language in terms of indicating gradation would not be noticed. Another solution would be to filter the results obtained to later build up the thesaurus according to

the terms. This process would eliminate many of the appeared words, making non-formal texts difficult to classify. In both cases, there will exist bias, albeit with different natures. This problem can be also extended to other conceptualizations of PLSA, such as shape recognition. Similar object identification could be problematic since it could be difficult to distinguish object reproductions (toys) from original artifacts. The same happens for color.

Convergence problems are somewhat more delicate. Formally equivalent solutions can present results that in practice can be very different in a field of application. This would occur in situations where high precision is required. This is due to the nature of the EM algorithm. Convergence is better as the number of components increases, but it is not possible to determine their number *a priori*. It has been reported that the convergence limit does not necessarily occur at a global optimum [240] and does not necessarily converge to a point, but can converge on a compact set [38], thus providing sub-optimal results [114]. In addition, sparse data structures can cause failures in convergence [8].

These problems, which largely determine the relationship of the PLSA (in its symmetrical formulation) with the SVD, have been fundamental at the beginning of our work. Furthermore, we have studied the behavior of the diagonal matrix, which can be related to important well-known properties. We add that in the following chapters, we always refer to the symmetric formulation of the PLSA.

Chapter 4

Non-negative Matrix Factorization

NMF is widely attributed to Paatero [190], who called these techniques *positive matrix factorization*. However, many authors attribute it to Lee [161]. Historically, interest in non-negative entries matrices appeared with the Perron-Frőbenius theorem (for a matrix $\mathbf{X}$ s.t. $\mathbf{X}^k > 0\,(k > 1)$ there exists a unique eigenvalue greater than others with a positive product when multiplied at the left or right by the corresponding eigenvector), related to properties of free energy [26] (cumulant generating functions in statistics) and the n-bodies problem [103]. Further works rely on square matrices powers [167] and conditions for the existence of stochastic matrices [45]. Its factorization was introduced by Chen [54].

The formulation by Chen [54] supposes the approximation problem for the non-negative entries matrix $\mathbf{X} \in \mathbb{R}_{+}^{m \times n}$ of rank r $(0 < r < \min(mn)$ accomplishing

(i) $\mathbf{X}$ has a non-negative matrix factorization.

(ii) There exists a set of non-negative linearly independent vectors in the positive orthant.

(iii) There exists a non-singular matrix containing the vectors of the positive orthant.

(iv) Every vector of the positive orthant is expressable as a linear combination.

(v) There exists a simplex containing the linearly independent vectors in the positive orthant.

Currently, the standard formulation of the NMF is the problem of find matrices $\mathbf{V} \in \mathbb{R}^{m \times k}$ and $\mathbf{B} \in \mathbb{R}^{k \times n}$ approximating matrix $\mathbf{X}$ such that [57, p. 8]

$$[\mathbf{X}]_{ij} \approx [\mathbf{V}]_{ik}[\mathbf{B}]_{kj} \tag{4.1}$$

$$= [\mathbf{V}]_{ik}[\mathbf{B}]_{kj} + [\mathbf{E}]_{ij} \qquad \big(\text{s.t. } e_{ij} \in \mathbf{E} \geq 0\big) \tag{4.2}$$

minimizing an objective function for a certain value of k. $\mathbf{E}$ is the approximation error. The effect of k is the variation of the space span.

4.1 Probabilistic Image of a Real Entries Matrix

Probabilistic decomposition is a problem widely treated in the context of NMF. Efforts to infer probabilistic significance rely on the consideration that the entries of the matrix (2.13) are defined in the field of non-negative reals. This approach limits applicability to the cases of variables (in the notation followed, they are the columns of the matrix associated with a data frame) that are only frequencies or contingency tables. They constitute the major core of NMF applications.

Examples are the works of Shashua [214], who formulates a latent class model data collected in the matrix (2.13), and restricting entries on the non-negative integers. A similar treatment, with the same interpretation of probabilities, is PLSA[213], concerning the simplex formed by the data in the positive orthant. An extension introducing the idea of the probabilistic image of the data matrix is discussed in Salakhutdinov [181], using a normal distribution for score rates of movies. The I-divergence is used to generalize maximum likelihood estimates of the matrices in Devarajan (4.1) [71].

Also, several existing techniques allow the generalization of the NMF to a probabilistic image of a data matrix. With PLSA notation, and taking into account that transformation (3.1) allows writing the conditional pdf

$$P(d_i, w_j) = P(d_i|\, w_j)P(w_j) \tag{4.3}$$

Identifying the multivariate matrix $\mathbf{X} \in \mathbb{Z}_+^{m\times n}$ of (2.13) with the data frame $N(d_i, w_j)$, leads to

$$P(d_i, w_j) = \frac{1}{\sum_{ij} x_{ij}}\mathbf{X} \tag{4.4}$$

$$= [\mathbf{Y}]_{ij} \tag{4.5}$$

and conditional probabilities $P(d_i|\, w_j)$, in matrix notation are

$$P(d_i|\, w_j) = \mathbf{X}\mathbf{D}_{\mathbf{X}}^{-1} \qquad \Big(\text{s.t.}\mathbf{D}_{\mathbf{X}} = \text{diag}\big(\sum_i x_{i,j=1}, \ldots, \sum_i x_{i,j=n}\big)\Big) \tag{4.6}$$

$$= [\widetilde{\mathbf{Y}}]_{ij} \tag{4.7}$$

$$= \big[\widetilde{\mathbf{y}}_1|\cdots|\widetilde{\mathbf{y}}_n\big] \tag{4.8}$$

where the tilde symbol indicates that the vectors $\widetilde{\mathbf{y}}_j \in \widetilde{\mathbf{Y}}$ s.t. $\widetilde{\mathbf{Y}} \in \mathbb{R}_+^{m\times n}$ are column normalized. In this case, matrix $\widetilde{\mathbf{Y}}$ is a (non-square) stochastic column matrix.

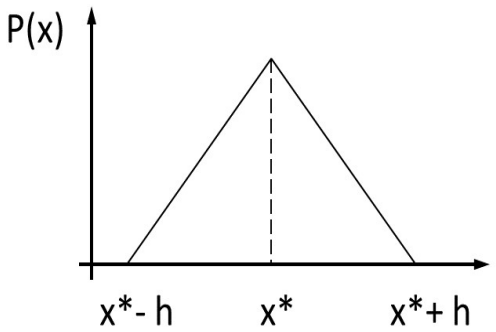
Conversely, for a wider data class, and with no non-negativity constraints ($x_{ij} \in \mathbb{R}$), matrix $\mathbf{X}$ can be written as the juxtaposition of column vectors

$$\mathbf{X} = \big[\mathbf{x}_{j=1}\big|\mathbf{x}_{j=2}\big|\ldots|\mathbf{x}_{j=m}\big] \qquad \big(\mathbf{x}_j \in \mathbb{R}^m\big) \tag{4.9}$$

Smoothing each column vector $\mathbf{x}_j \in \mathbf{X}$ as a mixture with the kernel density function (kdf)

$$\hat{f}_j(\mathbf{x}_j) = \frac{1}{n}\sum_i \Phi\Big(\frac{x - x_i}{h}\Big) \tag{4.10}$$

Figure 4.1: Triangular distribution.



Triangular distribution is based on triangular function $f(x) = 1 - |\,x|\,(x \in [-1, 1])$. It can be obtained from the sum of two independent standard uniform $U[0, 1]$ [13, Chap. 11].

provides a matrix containing the densities

$$[\widetilde{\mathbf{Y}}]_{ij} = \left[f_1(\mathbf{x}_1)\middle|\, f_2(\mathbf{x}_2)\middle|\, \ldots \,|\, f_n(\mathbf{x}_n)\right] \tag{4.11}$$

depending on the choice of the kdf and the *smooth parameter* h.

Consistency between results of (4.6) with (4.11) requires its obtention from (4.11), leading to

$$\mathbf{X}\mathbf{D}_{\mathbf{X}}^{-1} \sim \left[f_1(\mathbf{x}_1)\middle|\, f_2(\mathbf{x}_2)\middle|\, \ldots \,|\, f_n(\mathbf{x}_n)\right] \tag{4.12}$$

which can be achieved with a triangular kdf and $h = 1$.

The use of triangular kernels has been investigated by Kokonendji and Senga [155, 211], stating that it corresponds to a discrete pdf, while it is exposed as continuous in Balakrishnan [13, Chap. 13]. This question depends on the conditions of the definition of the variable domain and its support. It is illustrated in Figure 4.1.

The triangular kernel is [155]

$$\Phi(x^\star; h) = \begin{cases} \dfrac{h - |\, x - x_i|}{h^2} & x^\star - h \leq\ x \leq\ x^\star - h \\ 0 & \text{otherwise} \end{cases} \tag{4.13}$$

Taking the *grid* (the values at which the density is estimated) $\mathbf{x}_{j'} = (x_{1j'}, \dots\ x_{mj'})'$, indicating the value of subscript j' has been fixed and corresponds to a single column of the matrix (4.11), and writing the difference $|\, x - x^\star| = x^\star + |\, h|\, x^\star$ for $h = 1$, the interval $[x^\star - h, x^\star + h]$ contains a single point, and

$$\Phi(x^\star; h) = x_{ij'} \tag{4.14}$$

with density estimate at $x_{ij\prime} = x_{ij\prime}$ is

$$\hat{f}_j(x_{ij\prime}; h) = \frac{x_{ij'}}{(\mathrm{nr})_{ij'}} \tag{4.15}$$

being (nr) the number of observations with value $x_{ij'}$. The smoothed density is

$$\hat{f}_j(\mathbf{x}_j; h) = \frac{1}{n}\Big(\hat{f}_1(x_{1j\prime}; h) + \hat{f}_2(x_{2j\prime}; h) + \cdots + \hat{f}_n(x_{nj\prime}; h)\Big) \tag{4.16}$$

The construction of a probabilistic matrix can be assimilated to a random variable. Formally, it can be justified fixing a value for j and consider the set $\mathcal{B} = \{\mathbf{x}_i\}$, which takes values on all possible outcomes of set Ω, stating the problem in the triplet $(\Omega, \mathcal{B}, P)$, P being a measure or probability, and $\mathcal{B}$ a Borel σ-algebra. The map of the inverse image of $\mathcal{B}$ is $F^{-1}(\mathcal{B}) = \{\omega \,\text{s.t.}\, F(\omega) \in \mathcal{B}\}$ $(\omega \in \Omega)$. The Borel sets defined probabilities $P(\mathbf{x} \in \mathcal{B}) = P(F^{-1}(\mathcal{B}))$ with distribution $P(\mathbf{x} \leq\ x)$ and $P(\mathbf{x}_j) = 1)$, that estimates density f_j associated with the distribution P that generates the data [13, Chap. 1].

This construction can be concisely stated as

Theorem 5 *Let* $\widetilde{\mathbf{Y}} = [\widetilde{\mathbf{y}}_1, \widetilde{\mathbf{y}}_2, \dots, \widetilde{\mathbf{y}}_n]$ *the column stochastic matrix obtained from a contingency table or observed frequency matrix* $\mathbf{X}$ *and smoothing with*

a triangular kernel and smoothing parameter $h = 1$, evaluating the density in the grid $\mathbf{x}_{j\prime} = (x_{1j'}, \dots, x_{mj'})'$ gives the same result,

$$\mathbf{X}\mathbf{D}_{\mathbf{X}}^{-1} \sim \big[f_1(\mathbf{x})\big|\, f_2(\mathbf{x})\big|\, \dots \,|\, f_n(\mathbf{x})\big]$$

Also, matrices $\mathbf{X}\mathbf{D}_{\mathbf{X}}^{-1}$ and $\big[f_1(\mathbf{x})\big|\, f_2(\mathbf{x})\big|\, \dots \,|\, f_n(\mathbf{x})\big]$, are matrices that allow probabilistic interpretation.

We also note that matrix $\mathbf{Y}$ can be reconstructed from $\widetilde{\mathbf{Y}}$ for the integer case, but not for the continuous case. So, (4.3) is the mixture

$$P(x_{ij}) = \sum_j \alpha_j\, f_j \qquad \big(\text{with } \alpha_j \in \mathbf{R}_+ \text{ s.t. } \sum_j \alpha_j = 1\big) \tag{4.17}$$

$$= [\mathbf{Y}]_{ij} \qquad \big(\text{s.t. } \big\|[\mathbf{Y}]_{ij}\big\|_1 = 1\big) \tag{4.18}$$

Selection of coefficients (or weights) α_i is a classical problem in convex optimization [21]. In the machine learning context, this problem is related to the more relevant variables selection.

Choosing an uninformative *pdf* or discrete uniform density with $\alpha_) = 1/n$, in matrix notation

$$[\mathbf{Y}]_{ij} = [\widetilde{\mathbf{Y}}]_{ij}\, \mathbf{D}_{\mathbf{N}} \qquad \Big(\text{s.t. } \mathbf{D}_{\mathbf{N}} = \operatorname{diag}\big(1/n\big)\Big) \tag{4.19}$$

Taking the pseudoinverse

$$[\mathbf{Y}]_{ij}^{\dagger}[\mathbf{Y}]_{ij} = \big([\widetilde{\mathbf{Y}}]_{ij}'[\widetilde{\mathbf{Y}}]_{ij}\big)^{-1}[\widetilde{\mathbf{Y}}]_{ij}'[\widetilde{\mathbf{Y}}]_{ij}\mathbf{D}_{\mathbf{N}} \tag{4.20}$$

$$= \mathbf{I}\mathbf{D}_{\mathbf{N}} \tag{4.21}$$

We call matrices $\mathbf{Y}$ and $\widetilde{\mathbf{Y}}$ of (4.19) the *column probabilistic image* and *probabilistic image*, respectively. Despite them not being square, we shall refer to them as a stochastic data matrix and column stochastic matrix.

4.2 Non-negative Matrix Factorization

For the stochastic matrix $\mathbf{Y}$, formula (4.1) re-writes as

$$[\mathbf{Y}]_{ij} \approx [\mathbf{W}]_{ik}[\mathbf{H}]_{kj} \tag{4.22}$$

Also, [75] Ding demonstrates that normalization

$$[\mathbf{WH}]_{ij} = \frac{1}{k}[\widetilde{\mathbf{W}}]_{ik}[\widetilde{\mathbf{H}}]_{kj} \tag{4.23}$$

for

$$[\widetilde{\mathbf{W}}]_{ik'} \sim P(w_i|\, k') \tag{4.24}$$

$$[\widetilde{\mathbf{H}}]_{k'j} \sim P(h_j|\, k') \tag{4.25}$$

and for a fixed value of k denoted as k'

$$P(x_{ij}) = P(w_i|\, k')P(h_j|\, k') \tag{4.26}$$

We also refer to the columns of $\mathbf{W}$ and rows of $\mathbf{H}$ as *components*.

4.3 Objective Functions

To adjust the product (4.22), it is necessary to minimize some norm or divergence that evaluates the degree of approximation

$$\left\|[\mathbf{Y}]_{ij} - [\mathbf{WH}]_{ij}\right\| \leq \epsilon \tag{4.27}$$

where $\mathbf{W} \in \mathbb{R}^{m\times k}$ and $\mathbf{H} \in \mathbb{R}^{k\times n}$ are also probabilistic, into which it decomposes. Thus, $\epsilon > 0$ or, equivalently, the norm of matrix $\mathbf{E}$, is the approximation error.

Distances and divergences are applications of a space $\mathbb{R}^{m\times n}$ to the non-negative reals $\mathbb{R}_{+\cup 0}$. A distance d is a map that satisfies for vectors $\mathbf{a}$, $\mathbf{b}$, and $\mathbf{c}$ the conditions: (i) $d(\mathbf{a},\mathbf{b}) = d(\mathbf{b},\mathbf{a})$ (symmetry); (ii) $d(\mathbf{a},\mathbf{b}) = 0$ iff $\mathbf{a} = \mathbf{b}$ (identity); and (iii) $d(\mathbf{a},\mathbf{b}) \leq d(\mathbf{a},\mathbf{c}) + d(\mathbf{c},\mathbf{b})$ (triangular inequality). A divergence D does not satisfy one of these axioms, usually symmetry. These axioms are the same as for the norms given in 2.1. Furthermore, a divergence induces a metric if it satisfies the Cauchy-Scwartz inequality (axiom iv in 2.1).

The use of divergences was introduced in a study of population diversity by Rao [197, 198, 196] to measure similarity. So, it is necessary to transform observations into probabilities. Bregman divergence allows generalization of several ones. It is defined as [44]

$$D_f(\mathbf{y},\widehat{\mathbf{y}}) = f(\mathbf{y}) - f(\widehat{\mathbf{y}}) - \left\langle \mathbf{y}-\widehat{\mathbf{y}}, \nabla_{\widehat{\mathbf{y}}} f(\widehat{\mathbf{y}})\right\rangle \tag{4.28}$$

for $\widehat{\mathbf{y}}$, an estimate of $\mathbf{y}$. Also, the product $\mathbf{WH}$ is usually written as $\widehat{\mathbf{Y}}$ for simplicity reasons.

For matrices [57, p. 103]

$$\begin{aligned} D_f([\mathbf{Y}]_{ij}\|[\widehat{\mathbf{Y}}]_{ij}) = {} & f([\mathbf{Y}]_{ij}) - f([\widehat{\mathbf{Y}}]_{ij}) \\ & - \operatorname{tr}\Big(\nabla_{\widehat{\mathbf{Y}}} f([\widehat{\mathbf{Y}}]_{ij})'\big([\mathbf{Y}]_{ij} - [\widehat{\mathbf{Y}}]_{ij}\big)\Big) \end{aligned} \tag{4.29}$$

and exists for every sufficiently differentiable function f. Its significance is shown in Figure 4.2.

How to obtain different norms or divergences from it is a very studied topic. The most important from our point of view are.

Euclidean Distance

Choosing $f = \left\|\mathbf{Y}\right\|^2$ in (4.29)

Table 4.1: Distances and divergences.

Divergence	$f(\cdot)$	$D((X)\|\mathbf{Y})$
Euclidean	$\|\mathbf{X}\|^2$	$\|\mathbf{X}-\mathbf{Y}\|^2$
Mahalanobis	$\mathbf{X'AX}$	$(\mathbf{X}-\mathbf{Y})'\mathbf{A}(\mathbf{X}-\mathbf{Y})$
I-divergence	$\sum_{ij}[\mathbf{X}]_{ij}\log[\mathbf{X}]_{ij}$	$\sum_{ij}\Big([\mathbf{X}]_{ij}\log\frac{[\mathbf{X}]_{ij}}{[\mathbf{Y}]_{ij}} - [\mathbf{X}]_{ij}-[\mathbf{Y}]_{ij}\Big)$
KL-divergence	$\sum_{ij}[\mathbf{X}]_{ij}\log[\mathbf{X}]_{ij}$	$\sum_{ij}[\mathbf{X}]_{ij}\log\frac{[\mathbf{X}]_{ij}}{[\mathbf{Y}]_{ij}}$

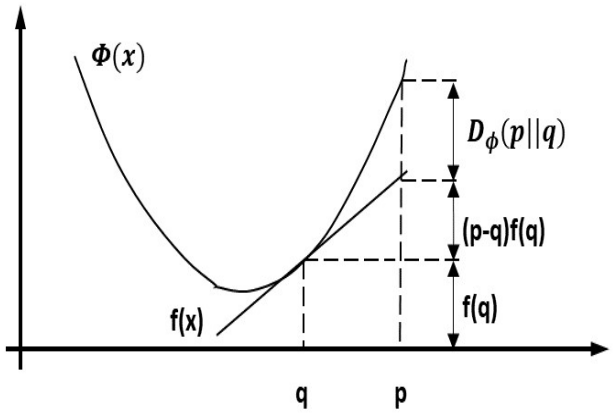
Most commonly used distances and divergences. The form of the function f in the Bregman divergence allows us to obtain several types of them, constituting a generalization.

$$
\begin{aligned}
D_\phi\big([\mathbf{Y}]_{ij}\|[\widehat{\mathbf{Y}}]_{ij}\big) &= \big\|[\mathbf{Y}]_{ij}\big\|^2 - \big\|[\widehat{\mathbf{Y}}]_{ij}\big\|^2 \\
&\quad - \operatorname{tr}\Big[\nabla_{\widetilde{\mathbf{Y}}}[\widehat{\mathbf{Y}}]'_{ij}\big([\mathbf{Y}]_{ij}-[\widehat{\mathbf{Y}}]_{ij}\big)\Big] \\
&= \big\|[\mathbf{Y}]_{ij}\big\|^2 - \big\|[\widehat{\mathbf{Y}}]_{ij}\big\|^2 \\
&\quad - \operatorname{tr}\Big[2[\widehat{\mathbf{Y}}]'_{ij}\big([\mathbf{Y}]_{ij}-[\widehat{\mathbf{Y}}]_{ij}\big)\Big] \\
&= \big\|[\mathbf{Y}]_{ij}\big\|^2 - \big\|[\widehat{\mathbf{Y}}]_{ij}\big\|^2 \\
&\quad - 2\operatorname{tr}\big([\widehat{\mathbf{Y}}]'_{ij}[\mathbf{Y}]_{ij}\big) + 2\operatorname{tr}\big([\widehat{\mathbf{Y}}]'_{ij}[\widehat{\mathbf{Y}}]_{ij}\big) \\
&= \big\|[\mathbf{Y}]_{ij}\big\|^2 - \big\|[\widehat{\mathbf{Y}}]_{ij}\big\|^2 \\
&\quad - 2\big\|[\widehat{\mathbf{Y}}]'_{ij}[\mathbf{Y}]_{ij}\big\| + 2\big\|[\widehat{\mathbf{Y}}]'_{ij}[\widehat{\mathbf{Y}}]_{ij}\big\| \\
&= \big\|[\mathbf{Y}]_{ij}\big\|^2 - \big\|[\widehat{\mathbf{Y}}]_{ij}\big\|^2 \\
&\quad - 2\big\|[\widehat{\mathbf{Y}}]'_{ij}[\mathbf{Y}]_{ij}\big\| + 2\big\|[\widehat{\mathbf{Y}}]_{ij}\big\|^2 \\
&= \big\|[\mathbf{Y}]_{ij}-[\widehat{\mathbf{Y}}]_{ij}\big\|^2
\end{aligned}
\tag{4.30}
$$

Mahalanobis Distance

For f the symmetric bi-linear form $\mathbf{Y'AY}$ the Bregman divergence takes the form

Figure 4.2: Bregman divergence.



Reproduced from [57, P.101]. Bregman divergence evaluates the similarity between two densities or distributions. It corresponds to the Taylor expansion at a point q close to p

$$\begin{aligned} D_\phi\big([\mathbf{Y}]_{ij} \| [\widehat{\mathbf{Y}}]_{ij}\big) &= [\mathbf{Y}]'_{ij}\mathbf{A}[\mathbf{Y}]_{ij} - [\widehat{\mathbf{Y}}]'_{ij}\mathbf{A}[\widehat{\mathbf{Y}}]_{ij} \\ &\quad - \operatorname{tr}\Big[\nabla_{\widehat{\mathbf{Y}}}[\widehat{\mathbf{Y}}]'_{ij}\mathbf{A}[\widehat{\mathbf{Y}}]_{ij}\big([\mathbf{Y}]_{ij} - [\widehat{\mathbf{Y}}]_{ij}\big)\Big] \\ &= [\mathbf{Y}]'_{ij}\mathbf{A}[\mathbf{Y}]_{ij} - [\widehat{\mathbf{Y}}]'_{ij}\mathbf{A}[\widehat{\mathbf{Y}}]_{ij} \\ &= \Big([\mathbf{Y}]_{ij} - [\widehat{\mathbf{Y}}]_{ij}\Big)'\mathbf{A}\Big([\mathbf{Y}]_{ij} - [\widehat{\mathbf{Y}}]_{ij}\Big) \end{aligned} \tag{4.31}$$

I and KL Divergence

For f the Shanon entropy $\mathbf{Y}\log\mathbf{Y}$ [11, chap. 1], substituting in (4.29)

$$
\begin{aligned}
D_I\big([\mathbf{Y}]_{ij}\|[\widehat{\mathbf{Y}}]_{ij}\big) &= \sum_{ij}[\mathbf{Y}]_{ij}\log[\mathbf{Y}]_{ij} - \sum_{ij}[\widehat{\mathbf{Y}}]_{ij}\log[\widehat{\mathbf{Y}}]_{ij} \\
&- \operatorname{tr}\Big(\nabla[\widehat{\mathbf{Y}}]_{ij}\log[\widehat{\mathbf{Y}}]_{ij}\big([\mathbf{Y}]_{ij} - [\widehat{\mathbf{Y}}]_{ij}\big)\Big) \\
&= \sum_{ij}[\mathbf{Y}]_{ij}\log[\mathbf{Y}]_{ij} - \sum_{ij}[\widehat{\mathbf{Y}}]_{ij}\log[\widehat{\mathbf{Y}}]_{ij} \\
&- \operatorname{tr}\Big(\big(1 + [\widehat{\mathbf{Y}}]_{ij}\big)\big([\mathbf{Y}]_{ij} - [\widehat{\mathbf{Y}}]_{ij}\big)\Big) \\
&= \sum_{ij}[\mathbf{Y}]_{ij}\log[\mathbf{Y}]_{ij} - \sum_{ij}[\widehat{\mathbf{Y}}]_{ij}\log[\widehat{\mathbf{Y}}]_{ij} \\
&- \operatorname{tr}(\big([\mathbf{Y}]_{ij} - [\widehat{\mathbf{Y}}]_{ij}\big) \\
&= \sum_{ij}[\mathbf{Y}]_{ij}\log\frac{[\mathbf{Y}]_{ij}}{[\widehat{\mathbf{Y}}]_{ij}} - \sum_{ij}[\mathbf{Y}]_{ij} - \sum[\widehat{\mathbf{Y}}]_{ij}
\end{aligned} \tag{4.32}
$$

Expression (4.32) is the I-divergence. If $\sum\widetilde{\mathbf{Y}} = \sum\mathbf{Y}$, it is the KL divergence [159]: [1].

$$
D_{KL}\big([\mathbf{Y}]_{ij}\|[\widehat{\mathbf{Y}}]_{ij}\big) = \sum_{ij}[\mathbf{Y}]_{ij}\log\frac{[\mathbf{Y}]_{ij}}{[\widehat{\mathbf{Y}}]_{ij}} \tag{4.33}
$$

Also, the KL divergence is related to Shannon's information, a result of expanding the logarithm of equation (4.33)

$$
D_{KL}\big([\mathbf{Y}]_{ij}\|[\widehat{\mathbf{Y}}]_{ij}\big) = \sum_{ij}[\mathbf{Y}]_{ij}\log[\mathbf{Y}]_{ij} - [\mathbf{Y}]_{ij}\log[\mathbf{WH}]_{ij} \tag{4.34}
$$

[1] Several authors have referred to the KL divergence as

$$
D_{KL}(\mathbf{Y}\|\mathbf{W}\,\mathbf{H}) = \sum_{ij}\Big([\mathbf{Y}]_{ij}\log\frac{[\mathbf{Y}]_{ij}}{[\mathbf{W}\,\mathbf{H}]_{ij}} - [\mathbf{Y}]_{ij} + [\mathbf{WH}]_{ij}\Big)
$$

which we prefer to call I-divergence or generalized KL-divergence, according to [57, P. 105], reserving the term KL divergence for the mean information, following the original nomenclature of S. Kullback and R.A. Leibler [159], and given by Formula (4.33), despite it differs a factor of log 2 from the original definition.

and identifying terms [53]

$$I([\mathbf{Y}]_{ij}|\,[\mathbf{WH}]_{ij}) = H([\mathbf{Y}]_{ij}) - H([\mathbf{W}]_{ik}|[\mathbf{H}]_{kj}) \tag{4.35}$$

where I is the mutual information. In this context, there are $r!$ representations (if the entries are labeled) that correspond to the indistinguishable entities (different entities that have the same values for all observational variables). A geometric interpretation of the information appears when the equivalence of the maximization of the likelihood is considered for the EM algorithm and/or the KL divergence. These results provide a stronger foundation for the probability space projection than for the orthogonal projection [48].

4.3.1 Minimization

The minimization technique used in our work is that based on the Karush-Kuhn-Tucker (KKT) conditions, and that for the matrix to be optimized supposes [57, p. 134]

$$[\mathbf{W}]_{ik} \geq 0 \tag{4.36}$$
$$\nabla_{\mathbf{W}} D\big([\mathbf{Y}]_{ij} \| [\mathbf{W}]_{ik}[\mathbf{H}]_{kj}\big) \geq 0 \tag{4.37}$$
$$[\mathbf{W}]_{ij} \circledast \nabla_{\mathbf{W}} D\big([\mathbf{Y}]_{ij} \| [\mathbf{W}]_{ik}[\mathbf{H}]_{kj}\big) = 0 \tag{4.38}$$

and

$$[\mathbf{H}]_{ik} \geq 0 \tag{4.39}$$
$$\nabla_{\mathbf{H}} D\big([\mathbf{Y}]_{ij} \| [\mathbf{W}]_{ik}[\mathbf{H}]_{kj}\big) \geq 0 \tag{4.40}$$
$$[\mathbf{H}]_{ij} \circledast \nabla_{\mathbf{H}} D\big([\mathbf{Y}]_{ij} \| [\mathbf{W}]_{ik}[\mathbf{H}]_{kj}\big) = 0 \tag{4.41}$$

4.4 Families of Algorithms

For an objective function, the obtention of matrices $\mathbf{W}$ and $\mathbf{H}$ gives rise to several algorithms. It is an iterative problem of existing classical methods: alternating least squares, gradient descent, Newton's methods, and multiplicative updates. These methods do not provide equivalent results. The convergence speed is critical in the choice of an alternative.

4.4.1 Alternating Least Squares

Alternating Least Squares (ALS) is the pioneering algorithm, used by Paatero [190] to factorize (4.22), minimizing the Euclidean distance

$$\frac{1}{2}\|[\mathbf{Y}]_{ij} - [\widehat{\mathbf{Y}}]_{ij}\|_2^2 = \frac{1}{2}\mathrm{tr}\big([\mathbf{Y}]_{ij} - [\widehat{\mathbf{Y}}]_{ij}\big)'\big([\mathbf{Y}]_{ij} - [\widehat{\mathbf{Y}}]_{ij}\big) \tag{4.42}$$

in the iterative process

$$[\mathbf{W}]_{ik}^{(p+1)} = \mathrm{argmin}_{\mathbf{W}}\big\|[\mathbf{Y}]_{ij} - [\mathbf{W}]_{ik}[\mathbf{H}]_{kj}^{(p)}\big\|_2^2 \tag{4.43}$$

$$[\mathbf{H}]_{kj}^{(p+1)} = \mathrm{argmin}_{\mathbf{H}}\big\|[\mathbf{Y}]'_{ij} - [\mathbf{W}]'_{ik}[\mathbf{H}^{(p)}]'_{kj}\big\|_2^2 \tag{4.44}$$

with solutions

$$[\mathbf{W}]_{ik} \leftarrow [\mathbf{Y}]_{ij}[\mathbf{H}]'_{kj}\big([\mathbf{H}]_{ij}[\mathbf{H}]'_{ij}\big)^{-1} \tag{4.45}$$

$$= [\mathbf{Y}]_{ij}[\mathbf{H}]^{\dagger}_{kj} \tag{4.46}$$

$$[\mathbf{H}]_{kj} \leftarrow \big([\mathbf{W}]'_{ik}[\mathbf{W}]_{ik}\big)^{-1}[\mathbf{W}]'_{ik}[\mathbf{Y}]_{ij} \tag{4.47}$$

$$= \big[\mathbf{W}\big]^{\dagger}_{ik}[\mathbf{Y}]_{ij} \tag{4.48}$$

This method has also been formulated as a regression problem: by taking the $\mathbf{y}_j$ column vectors of $\mathbf{Y}$ [99] minimizing the objective functions

$$\mathcal{F}_1 = \min\big\|\mathbf{y}_j - \mathrm{diag}\big([\mathbf{W}]_{ik}\mathbf{h}_k\big)\big\|_2^2 \tag{4.49}$$

$$\mathcal{F}_2 = \min\big\|\mathbf{y}_j - \mathrm{diag}\big([\mathbf{H}]_{kj}\mathbf{w}_j\big)\big\|_2^2 \tag{4.50}$$

These techniques are inefficient for large data sets. In addition, they have convergence problems when co-linearity exists.

Weighted ALS introduces a variance-covariance matrix Σ to stabilize the co-linearity effect in $\mathbf{X}$ of (2.13) [62, 60]

$$\frac{1}{2}\|[\mathbf{Y}]_{ij} - [\widehat{\mathbf{Y}}]_{ij}\|_{\Sigma}^{2} = \frac{1}{2}\mathrm{tr}\big([\mathbf{Y}]_{ij} - [\widehat{\mathbf{Y}}]_{ij}\big)'\Sigma\big([\mathbf{Y}]_{ij} - [\widehat{\mathbf{Y}}]_{ij}\big) \tag{4.51}$$

This parameterization is equivalent to the Mahalanobis distance. To increase computational efficiency, it is recommended to set $\mathbf{V}$ and $\mathbf{W}$ to $\mathbf{H}$ and $\mathbf{V}'$ of Theorem 1 [62].

Another solution, known as *line search*, involves interpolation [112]

$$[\mathbf{W}]_{ik} = [\mathbf{W}]_{ik}^{(p-2)} + \eta_{\mathbf{W}}\left([\mathbf{W}]_{ik}^{(p-1)} - [\mathbf{W}]_{ik}^{(p-2)}\right) \tag{4.52}$$

$$[\mathbf{H}]_{kj} = [\mathbf{H}]_{kj}^{(p-2)} + \eta_{\mathbf{H}}\left([\mathbf{H}]_{kj}^{(p-1)} - [\mathbf{H}]_{kj}^{(p-2)}\right) \tag{4.53}$$

in previous estimates. The differences in the second terms of (4.52) and (4.53) are the $p-th$ directions.

A solution for large-scale NMF problems is Hierarchical Alternating Least Squares (HALS) with a locally objective function [59].

4.4.2 Descendant Gradient

The gradient descent method involves iteration to achieve the minimum of a real function [107]

$$\mathbf{x}^{(p+1)} = x^{(p)} - \alpha^{(p)}\nabla_x f(\mathbf{x}^{(p)}) \quad (\mathbf{x} \in \mathbb{R}^n) \tag{4.54}$$

where α is the step between iterations. Typically, α is a sufficiently small positive real value.

Approximation (4.22) is an optimization problem in the positive orthant. To prevent the appearance of non-negative values, each iteration is projected in the positive orthant, with the help of the projection operator $\mathcal{P}_+$. The optimization problem is posed as [57, p. 268].

$$[\mathbf{W}]_{ik}^{(p+1)} = [\mathbf{W}]_{ik}^{(p)} - \boldsymbol{\alpha}_{\mathbf{W}}^{(p)}\mathcal{P}_+[\mathbf{W}]_{ik}^{(p)} \tag{4.55}$$

$$[\mathbf{H}]_{ik}^{(p+1)} = [\mathbf{H}]_{ik}^{(p)} - \boldsymbol{\alpha}_{\mathbf{H}}^{(p)}\mathcal{P}_+[\mathbf{H}]_{ik}^{(p)} \tag{4.56}$$

where $\boldsymbol{\alpha}$ is a step vector, also known as the learning factor.

A version of this method is the Oblique projcted Landweber (OPL). The OPL minimizes the expressions (4.55) and (4.56), calculating the gradient constrained to non-negativity [166]. Projected Sequential Subspace Optimization (PSESOP) uses an objective function [83]. The interior point algorithms introduce a scaling vector, and the learning factors must be adjusted to maintain non-negativity [16] conditions.

4.4.3 Quasi-Newton Method

Algorithms based on gradient methods are first-order approximations projected in the positive orthant. Taylor's development [61]

$$\begin{aligned} D\big([\mathbf{Y}]_{ij}\|([\mathbf{W}]_{ik}+\Delta[\mathbf{W}]_{ik})[\mathbf{H}]_{kj}\big) &= D([\mathbf{Y}]_{ij}\|[\widehat{\mathbf{Y}}]_{ij}) \\ &+ \mathrm{vec}(\nabla_{\mathbf{W}} D([\mathbf{Y}]_{ij}\|[\widehat{\mathbf{Y}}]_{ij})'\mathrm{vec}(\Delta[\mathbf{W}]_{ik}) \\ &+ \frac{1}{2}\mathrm{vec}(\Delta[\mathbf{W}]'_{ik})\mathcal{H}_{\mathbf{W}}\mathrm{vec}(\Delta[\mathbf{W}]_{ik}) \\ &+ \mathcal{O}\big((\Delta[\mathbf{W}]_{ik})'\big) \qquad (4.57) \\ D\big([\mathbf{Y}]_{ij}\|[\mathbf{W}]_{ik}([\mathbf{H}]_{kj}+\Delta[\mathbf{H}]_{kj})\big) &= D([\mathbf{Y}]_{ij}\|(\mathbf{W}\mathbf{W}) \\ &+ \mathrm{vec}(\nabla_{\mathbf{H}} D([\mathbf{Y}]_{ij}\|[\widehat{\mathbf{Y}}]_{ij})'\mathrm{vec}(\Delta[\mathbf{H}]_{kj}) \\ &+ \frac{1}{2}\mathrm{vec}(\Delta[\mathbf{H}]'_{kj})\mathcal{H}_{\mathbf{H}}\mathrm{vec}(\Delta[\mathbf{H}]_{kj}) \\ &+ \mathcal{O}\big((\Delta[\mathbf{H}]_{kj})'\big) \qquad (4.58) \end{aligned}$$

where $\mathcal{H}$ is the Hessian matrix, providing the model

$$\mathrm{vec}([\mathbf{W}]_{ik}) = \mathcal{P}_+\big[\mathrm{vec}([\mathbf{W}]_{ik} - \mathbf{W}\mathrm{vec}\nabla_{\mathbf{W}} D([\mathbf{Y}]_{ij})\|[\mathbf{W}\mathbf{H}]_{ij}))\big] \qquad (4.59)$$

$$\mathrm{vec}([\mathbf{H}]_{kj}) = \mathcal{P}_+\big[\mathrm{vec}([\mathbf{H}]_{kj} - \mathbf{H}\mathrm{vec}\nabla_{\mathbf{H}} D([\mathbf{Y}]_{ij})\|[\mathbf{W}\mathbf{H}]_{ij}))\big] \qquad (4.60)$$

relating the data curvature (or module of the Hessian matrix) with information, which we develop in chapter 6.

4.4.4 Multiplicative Updates

Multiplicative updates constitute a series of simple algorithms based on the convergence of an objective function. It is a positive semidefinite linear form that derives from the chosen divergence or distance.

The methods related to multiplicative updates allow the introduction of additional constraints, such as sparsity or normalization conditions, to the matrices. The structure of the solutions to the problem is

$$[\mathbf{W}]_{ik} = \arg\min_{\mathbf{W}\geq 0} D([\mathbf{Y}]_{ij} \| [\mathbf{WH}]_{ij}) \qquad \text{(for fixed } \mathbf{W}\text{)} \qquad (4.61)$$

$$[\mathbf{H}]_{kj} = \arg\min_{\mathbf{H}\geq 0} D([\mathbf{Y}]_{ij} \| [\mathbf{WH}]_{ij}) \qquad \text{(for fixed } \mathbf{H}\text{)} \qquad (4.62)$$

Ho [124] reports that the normalization conditions do not vary during the iterative process.

Solutions for matrices $\mathbf{W}$ and $\mathbf{H}$ with various objective functions are in Table 4.2.

4.4.5 Initialization and Stopping Criteria

Execution of the previous algorithms requires selecting a value of k, initializing matrices $\mathbf{W}$ and $\mathbf{H}$, and iterating until a certain condition

$$\|[\mathbf{Y}]_{ij} - [\mathbf{WH}]_{ij}\| \leq \epsilon \qquad (4.63)$$

is met for some real value $\epsilon \geq 0$.

NMF has a similar computational inefficiency to that of PLSA . There are two reasons for this. The first is that the iterative process is slow, especially in the case of multiplicative updates. The other issue is that the NMF problem is non-convex, so there can be several local optima. There are several solutions to circumscribe this problem: the spherical k-means algorithm [238], and initializing with SVD values, setting the negative values to zero [35]. However, for obtaining further properties in the 5 and 6 chapters, random initialization has several advantages.

Table 4.2: NMF Solutions.

Divergence	Solutions
Euclidean	$[\mathbf{W}]_{ik} \leftarrow [\mathbf{W}]_{ik} \odot \left(\frac{[\mathbf{W}]_{ik}^{\top}[\mathbf{Y}]_{ij}}{[\mathbf{W}]_{ik}^{\top}[\mathbf{W}]_{ik}[\mathbf{H}]_{kj}} \right)$ $[\mathbf{H}]_{kj} \leftarrow [\mathbf{H}]_{kj} \odot \left(\frac{[\mathbf{W}]_{ik}^{\top}[\mathbf{Y}]_{ij}}{[\mathbf{W}]_{ik}^{\top}[\mathbf{W}]_{ik}[\mathbf{H}]_{kj}} \right)$
Mahalanobis	$[\mathbf{W}]_{ik} \leftarrow [\mathbf{W}]_{ik} \odot \left(\frac{[\mathbf{Y}]_{ij}}{[\mathbf{W}\,\mathbf{A}\,\mathbf{H}]_{ij}} \left([\mathbf{A}]_{kk}[\mathbf{H}]_{kj} \right)^{\top} \right)$ $[\mathbf{H}]_{kj} \leftarrow [\mathbf{H}]_{kj} \odot \left(\left([\mathbf{A}]_{kk}[\mathbf{W}]_{ik}^{\top} \right)^{\top} \frac{[\mathbf{Y}]_{ij}}{[\mathbf{W}\,\mathbf{A}\,\mathbf{H}]_{ij}} \right)$
I-divergence	$[\mathbf{W}]_{ik} \leftarrow [\mathbf{W}]_{ik} \odot \left(\frac{[\mathbf{Y}]_{ij}}{[\mathbf{W}\mathbf{H}]_{ij}} [\mathbf{H}]_{kj} \right)$ $[\mathbf{H}]_{kj} \leftarrow [\mathbf{H}]_{kj} \odot \left([\mathbf{W}]_{ik} \frac{[\mathbf{Y}]_{ij}}{[\mathbf{W}\mathbf{H}]_{ij}} \right)$
KL-divergence	$[\mathbf{W}]_{ik} \leftarrow [\mathbf{W}]_{ik} \odot \left(\frac{[\mathbf{Y}]_{ij}}{[\mathbf{W}\,\mathbf{H}]_{ij}} [\mathbf{H}]_{kj}^{\top} \right)$ $[\mathbf{H}]_{kj} \leftarrow [\mathbf{H}]_{kj} \odot \left([\mathbf{W}]_{ik}^{\top} \frac{[\mathbf{Y}]_{ij}}{[\mathbf{W}\,\mathbf{H}]_{ij}} \right)$

Solutions of the NMF problem for referred distances/divergences. The symbol $\odot$ represents the Hadamard product.

4.4.6 Convergence

The convergence of families of algorithms depends on the optimization method. The algorithms given in the 4.4.1 section of the process assume an iterative least-squares adjustment. The methods of sections 4.4.2 and 4.4.3 are classical optimization techniques. A classic reference is Dahlquist [66].

Minimization of KL divergence by multiplicative updates is known as the *em* algorithm, and it is equivalent to the log-likelihood maximization with the EM algorithm. In both cases, the results are the same. However, the *em* results are asymptotes of the EM algorithm in the general case [3].

On the other hand, the I-divergence and KL-divergence belong to a larger

class of divergences and are the α-divergences, which are unique [4]. In the same paper [4], Amari points out that α-divergences belong to a larger class, calling them $f-$ divergences, and that they are the class of divergences that can be obtained with the Bregman divergence.

4.5 PLSA and NMF Relation

The explicit relationship between PLSA and NMF, stated by Gaussier [104], normalizes matrix $\mathbf{X}$ with formula (3.1), factorizing $\mathbf{Y}$ with I-divergence as the objective function and minimizing with KKT conditions. Then, introducing diagonal matrices $\mathbf{D}_1$ and $\mathbf{D}_2$

$$[\mathbf{WH}]_{ij} = [\big(\mathbf{W}\mathbf{D}_1^{-1}\mathbf{D}_1\big)]_{ik}\big[\big(\mathbf{D}_2\mathbf{D}_2^{-1}\mathbf{H}\big)\big]_{kj} \tag{4.64}$$

$$= [\big(\mathbf{W}\mathbf{D}_1^{-1}\big)]_{ik}\,\mathrm{diag}\,\big(\mathbf{D}_1\mathbf{D}_2\big)[\big(\mathbf{D}_2^{-1}\mathbf{H}\big)]_{kj} \tag{4.65}$$

Gaussier states that *any (local) maximum solution of PLSA is a solution of the NMF with KL-divergence* (I-divergence according to the nomenclature herein) [104].

Further work [76] with the same divergence has introduced normalization for matrices $\mathbf{W}$ and $\mathbf{H}$, such that the column stochastic matrix $\widetilde{\mathbf{W}} = [\widetilde{\mathbf{w}}_1, \cdots, \widetilde{\mathbf{w}}_K]$ and the row stochastic matrix $\widetilde{\mathbf{H}} = [\widetilde{\mathbf{h}}_1, \cdots, \widetilde{\mathbf{h}}_K]$ are obtained as

$$\widetilde{\mathbf{w}}_k = \frac{\mathbf{w}_k}{\sum_i w_{ik}} = 1 \tag{4.66}$$

$$\widetilde{\mathbf{h}}_k = \frac{\mathbf{h}_k}{\sum_j h_{kj}} = 1 \tag{4.67}$$

calling those conditions *probabilistic normalizations*, and writing

$$[\mathbf{Y}]_{ij} = [\widetilde{\mathbf{W}}]_{ik}\mathbf{D}_{\mathbf{W}}\,[\widetilde{\mathbf{H}}]_{kj}\mathbf{D}_{\mathbf{H}} \tag{4.68}$$

$$= [\widetilde{\mathbf{W}}]_{ik}\mathbf{S}[\widetilde{\mathbf{H}}]_{kj} \qquad (\text{s.t. } \mathbf{S} = \mathbf{D}_{\mathbf{W}}\,\mathbf{D}_{\mathbf{H}}) \tag{4.69}$$

where $\mathbf{D}_{\mathbf{W}}$ and $\mathbf{D}_{\mathbf{H}}$ are suitable diagonal matrices containing the column sums of the respective sub-index matrices. Both conclusions are similar..

4.6 SVD Probabilistic Image

Formal equivalence between the equations (3.19a)-(3.19c) and (4.65), requires a diagonal matrix, which plays the role of Σ in Theorem 1. Those relations correspond to our work [91].

Considering the minimal number of components which allows equality of results of (4.22), requires to take into account that the empirical density of $\mathbf{Y}$ is $\mathbf{Y}^\star$, and

$$\beta = \inf\ \left(y_{ij}\right) \qquad \left(y_{ij} \in \mathbf{Y}\right) \tag{4.70}$$

$$\beta^\star = \inf\ \left(y^\star_{ij}\right) = \frac{1}{m \times\ n} \qquad \left(y^\star_{ij} \in \mathbf{Y}^\star\right) \tag{4.71}$$

Since the convergence of $\beta^\star \to \beta$ is almost sure, the inequality

$$\frac{k}{\min\left(m \times\ n\right)}\ \beta \geq \beta \tag{4.72}$$

holds only if the number of model components is

$$k \geq \min(m, n) \tag{4.73}$$

ensuring the empirical distribution of $\widehat{\mathbf{Y}}$ is the same as $\mathbf{Y}$.

A direct consequence is that, using the Gramm-Scmidt procedure, the columns of $\mathbf{Y}$ are orthogonal, so the SVD and the symmetric PLSA are the same.

Because the centered matrix of $\mathbf{Y}$ is

$$[\bar{\mathbf{Y}}]_{ij} = [\mathbf{Y}]_{ij} - \mathbf{1}\bar{\mathbf{y}}_j \qquad \left(j = 1,\, \ldots,\, n\right) \tag{4.74}$$

where $\mathbf{1}$ is the ones matrix, and $\bar{\mathbf{y}}_j$ are the column means vector.

Choosing the diagonal matrix as

$$\operatorname{diag}\left(\mathbf{z}\right) = \frac{\operatorname{diag}\left(\left[\bar{\mathbf{Y}}\right]'_{ij}\left[\bar{\mathbf{Y}}\right]_{ij}\right)}{\operatorname{tr}\left(\left[\bar{\mathbf{Y}}\right]'_{ij}\left[\bar{\mathbf{Y}}\right]_{ij}\right)} \tag{4.75}$$

Introducing vector notation for the NMF matrices

$$\left[\mathbf{W}\,\mathbf{H}\right]_{ij} = \langle \mathbf{w}_i, \mathbf{h}_j \rangle \qquad \left(\text{s.t. } \mathbf{w}_k,\ \mathbf{h}_k \in \Re^K\right) \tag{4.76}$$

simple algebra manipulations

$$\langle \mathbf{w}_i, \mathbf{h}_j \rangle = \left\langle \frac{w_{ik'}}{\sqrt{z_{k'}}}\sqrt{z_{k'}}\ , \frac{h_{k'j}}{\sqrt{z_{k'}}}\sqrt{z_{k'}} \right\rangle \tag{4.77}$$

$$= \left[\frac{w_{ik'}}{\sqrt{z_{k'}}}\right] \operatorname{diag}\left(z_{k'}\right) \left[\frac{h_{k'j}}{\sqrt{z_{k'}}}\right] \tag{4.78}$$

and identifying

$$\left[\mathbf{G}\right]_{ik} = \left[\frac{w_{ik'}}{\sqrt{t_{k'}}}\right]_{ik} \tag{4.79}$$

$$\mathbf{Z}_D = \operatorname{diag}\left(\mathbf{z}\right) \tag{4.80}$$

$$\left[\mathbf{F}\right]^t_{jk} = \left[\frac{h_{k'j}}{\sqrt{t_{k'}}}\right]_{jk} \tag{4.81}$$

the equivalence

$$\left[\mathbf{G}\right]_{ik}\mathbf{Z}_D\left[\mathbf{F}\right]^{\top}_{jk} \sim \mathbf{U}\Sigma\mathbf{V}' \tag{4.82}$$

is achieved.

Also, the expectation of $\mathbf{z}$ is

$$\mathrm{E}\left[\mathbf{z}\right] = \frac{\operatorname{diag}\left(\operatorname{var}\left(\left[\mathbf{Y}\right]_{ij}\right)\right)}{\operatorname{tr}\left(\operatorname{var}\left(\left[\mathbf{Y}\right]_{ij}\right)\right)} \tag{4.83}$$

Since $\mathbf{z} = P(z_k)$

$$\mathrm{E}\left[P(\mathbf{z})\right] = \sum_k P(\mathbf{z}) z_k = \mathbf{z}'\mathbf{z} \tag{4.84}$$

Then, it is possible to prove that with the number of model components given by (4.75), the entries of diag($\mathbf{z}$) ordered as decreasing values, and the same permutation done in the columns of $\mathbf{G}$ and $\mathbf{F}$, the SVD of the orthonormalization (orthogonalization and column normalization) Gram-Schmidt of $\mathbf{Y}$, and the product of equations (3.19a)-(3.19a) provide the same result when the empirical distributions are reached, except for r repeated values of the data matrix $\mathbf{Y}_{[0,1]}$ obtained from $\mathbf{X}$. The norm $\|\mathbf{W}\,\mathbf{H}\|_1 = \|\mathbf{G}\mathbf{Z}_D\mathbf{F}^\top = 1$ is preserved [91].

A practical way to ensure that the empirical distribution equality $\widehat{\mathbf{Y}}^{\,\mathrm{emp}} = \mathbf{Y}^{\,\mathrm{emp}}$ is reached is to consider the Δ matrix operator. This operator compares the character matrices $\mathbf{N}$ and $\widehat{\mathbf{N}}$ obtained when a set of m labels substitutes the values of the columns, and they are arranged in decreasing order according to the numerical values of the entries of the obtained matrix (it can be increasing too, with no difference to the results). Those matrices can be seen as ordinal-named. Then, introducing as a definition, the Δ matrix operator is

$$\Delta_{ij}^{(p',p)} = \begin{cases} 0 & \text{if} \quad n_{ij}^{(p\prime)} = n_{ij}^{(p)} \quad \left(p\prime > p \quad \text{and} \quad n_{ij}^{(p)} \in \mathbf{N}^{(p)}\right) \\ 1 & \text{otherwise} \end{cases} \tag{4.85}$$

where p and p' are the iterations over which the comparison is done, and $n_{ij}^{(\cdot)} \in \mathbf{N}$ results from the column value substitution ordered according to their entries.

The degree of adjustment between the character matrices can be measured with the $\|\Delta_{ij}\|_1$ norm, which provides the number of non-coincidences between them. When it is zero, the total adjustment between both empirical matrices is obtained, and a consequence is that condition $\|\Delta_{\,ij}\|_1 = 0$ is equivalent to $\widehat{\mathbf{Y}}^{\,(\mathrm{emp})} = \mathbf{Y}^{\,(\mathrm{emp})}$.

In practical conditions, the zero bound is difficult to achieve, since the matrix $\mathbf{Y}$ can contain some identical entries. In this case, the label ordination admits permutations on the repeated values, and the lower bound is the

number of repeated values, named r. Also, the row labels can be substituted by the column labels, with identical results. In the case when the lower bound of $\Delta = r$ is achieved, the result is expressed concisely as

Theorem 6 *The probabilistic SVD image and the symmetric PLSA formulas are equal when $\Delta = r$ (r being the number of repeated values in $\mathbf{Y}$). The orthonormalized SVD is the same for them. In this case, the local basis which spans the general case PLSA equations is the orthonormalized basis which spans too the transformed data matrix $\mathbf{Y}$, obtained from $\mathbf{X}$.*

4.7 Example

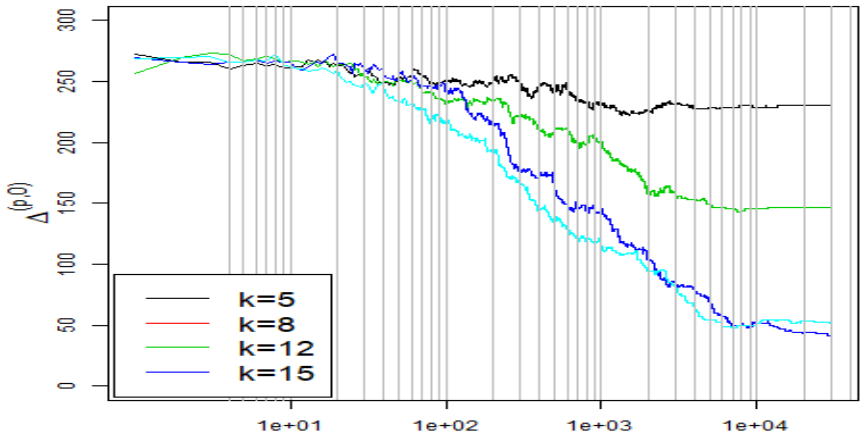
To see how the number of components affects the convergence, we examine how the Δ matrix works with the use of the em algorithm. A dataset based on Olympic decathlon results is used to drive the example. Included in several R packages, with some differences among them, the selected one is that included in the *FactoMineR* [160]. The data are the ranks of participants in the Athens 2012 Olympic Games men's decathlon competition. Additional reductions are done in the data by selecting 28 rows and 10 columns from the 41 and 13 original ones. Those are the athletes results only in the Olympic Games. Other meetings reference values are deprecated, and total points and classification are also omitted.

The start point is to run equations $\mathbf{W}$ and $\mathbf{H}$ of table 4.2 random initialized, varying the number of components. A estimation $\widehat{\mathbf{Y}}$ of $\mathbf{Y}$ defines $\mathbf{N}_c$. This qualitative matrix is obtained by substituting the numbers of the rows by the row-label (or athlete name). After a sufficient number of iterations, the Δ matrix is something similar to

$$\Delta_{ij}^{(1000,0)} = \begin{bmatrix} 0 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 1 \\ 1 & 0 & \cdots & 1 \\ \vdots & \vdots & \ddots & \vdots \\ 1 & 1 & \cdots & 0 \end{bmatrix}$$

where the zeros are the coincidences of the rank of the athlete in the column trial and a one appears when they are different. The L_1 norm gives the accuracy of the classification. In an ideal case, it should be zero, but repeated

Figure 4.3: Number of components vs. Delta Matrix.



Reproduced from [91] .The graphic shows how the KL divergence decreases as the number of iterations increases for different numbers of model components. Initial values are randomized. When Δ matrix has 47 non-coincidences, the limit is achieved; k is the number of components.

values appear. In this case, the 47 repeated values in the data matrix $\mathbf{Y}$ provides $r = 47$ permutations of the character matrix, being indistinguishable between them, and this is a limit for Δ.

One must be careful when ordering the results if significance is to be provided. There are two categories: more is better, which corresponds to distances achieved in jumping or throwing events; and less is better, which is the case for times achieved in running events. This correspondence does not always occur, and it should be done in a cautious way in all the cases if significance will be provided, but it is not important for computation purposes. The qualitative matrix $\mathbf{N}_c$ makes no algebraic sense. The ordination according to the obtained ranks can be omitted, and an ascendant or descendant one is sufficient for comparison purposes. This task is left to the analyst criteria and has no more importance than coherence with the data ordination.

The objective is to adjust the Δ matrix to $r = 47$ non coincidences (or fewer), ensuring the minimal number of model components to ensure

this condition k is fulfilled for $k \geq \min(m, n)$. For realistic data sets, this condition is difficult to meet since it is computationally expensive, as it can be seen in Figure 4.3. Figure 4.3 shows how the limit is reached when the components are $k \geq \min(m, n)$ but not in other cases. We use this case in further chapters, but not the low rank case.

Chapter 5

Kernelization

The kernel idea, attributed to Hilbert [105], was extended to solve certain equations of Mathematical Physics generalizing the vector basis to an orthonormal functions basis [216]. The conditions for a well-defined problem must satisfy Mercer's Theorem [175]. More rigorous conditions are the construction of the RKHS, proposed by Aronszajn [9]. Since the result are not equivalent, both allow the construction of well-defined kernels for classification and regression purposes. The use of kernels as classifiers was developed by Vapnik [230], who introduced the SVM. This approach works well in the case of separable spaces. For complex data structures, the *kernel trick* consists of mapping original observations, represented as vectors, to a higher-dimensional space. So, the kernel trick is a generalization of the dot product, requiring a map function ϕ to transform observations from the original space (*input space*) to another one, the *feature space*. There exist many types of kernels. A formal and detailed introduction is Ghojogh's review [105]. A kernel combining properties derived from a generative model that is asymptotically efficient is the Fisher kernel [141, 202]. First formulated by posing the problem of discriminative and logistic regression, it constitutes a statistical parametric model, introducing a differentiable manifold on the parameters space, with local metric given by the Fisher information.

A class of kernels admitting the non-parametric hypothesis are the NMF kernels [245, 162]. The probabilistic sense appears for suitable normalization conditions on the factorization. The idea of NMF is to adjust the data matrix as the product of two non-negative matrices, providing an approximation to the original data by minimizing an objective function. Then, the parameters

are those that factorize the matrix, which is a sufficient (not minimal) statistic. Bregman divergence allows generalization of several NMF cases by taking explicit forms for a sufficiently differentiable convex function f. Statistical properties depend on this function inferring properties of a Riemann metric. Consistency properties of Fisher's kernel are illustrated with the performed numerical experiments in section 5.9. They illustrate that efficiency translates into an asymptotically arbitrary misclassification error.

5.1 Kernelization

Let (2.13) be the data matrix, consisting of m items or observations evaluated with n variables or observational criteria. Kernelization is an application

$$\phi : \mathcal{H} \longrightarrow \mathbb{R} \tag{5.1}$$

$$\langle \mathbf{x}, \mathbf{x}_{i\prime} \rangle \longmapsto \langle \phi(\mathbf{x}), \phi(\mathbf{x}_{i\prime}) \rangle \tag{5.2}$$

$$= K(\mathbf{x}_i, \mathbf{x}_{i'}) \tag{5.3}$$

that generalizes the dot product from vectors $\mathbf{x}_i$ of the *input space* to vectors $\phi(\mathbf{x}_i)$ of the *feature space*.

The matrix containing the dot products of the row vectors of $\mathbf{X}$ is a Gramm matrix, and for reals $c_i\,, c_{i'}$ is a *positive semi-definite kernel*, accomplishing

$$\sum_{ii'} c_i, c_{i'} K(\mathbf{x}_i, \mathbf{x}_{i'}) \geq 0 \tag{5.4}$$

being zero for reals $c_i = c_{i'} = 0$. This construction is well-defined if it satisfies Mercer's Theorem or it is a RKHS, which we describe synthetically below.

Mercer's theorem [175] is perhaps the simplest way to build a kernel. The main idea is to define a map between the sets $[a, b] \times [a, b] \to \mathbb{R}$. It is satisfied for function

$$J = \int_a^b \int_a^b K(\mathbf{x}_i, \mathbf{x}_{i'}) f(\mathbf{x}) f(\mathbf{x}_{i'})\, d\mathbf{x} d\mathbf{x} \tag{5.5}$$

where f can be expressed as linear combinations

$$f(\mathbf{x}) = \sum_i \alpha_i \mathbf{x}_i \tag{5.6}$$

Since J takes non-negative values, expressing K as an orthonormal basis $\{\psi(\cdot)\}_{i=1}^{\infty}$ with corresponding eigenvalues $\{\lambda\}_{i=1}^{\infty}$, leads to

$$K(\mathbf{x}_i, \mathbf{x}_{i'}) = \sum_{ii'} \lambda_i \psi(\mathbf{x}_i)\psi(\mathbf{x}_{i'}) \tag{5.7}$$

Hence, Mercier's Theorem: *function ϕ of* (5.3) *lets us construct a kernel if it is symmetric and positive semi-definite.* The reciprocal statement is not necessarily accomplished [131]

More rigorous conditions are based on the RKHS [9]. The main idea is to consider the linear combinations functions $f(\cdot) = \sum_i \alpha_i K(\cdot, \mathbf{x}_i)$ and $g(\cdot) = \sum_{i'} \beta_{i'} K(\cdot, \mathbf{x}_i)$, with dot product

$$\langle f, g \rangle = \Big\langle \sum_i \alpha_i K(\cdot, \mathbf{x}_i), \sum_{i'} \beta_{i'} K(\cdot, \mathbf{x}_i) \Big\rangle \tag{5.8}$$

$$= \sum_{i,i'} \alpha_i \beta_{i'} K(\mathbf{x}, \mathbf{x}_{i'}) \tag{5.9}$$

Completeness of the space is achieved by substituting the kernel K orthonormal basis and the correspondent eigenvalues. So, the kernel defined in (5.3) is unique.

More recently, the representer theorem states conditions for dot products defined in a finite-dimensional Hilbert space [209]. Relevant properties are

(i) Symmetry: $K(\mathbf{x}_i, \mathbf{x}_{i'}) = K(\mathbf{x}_{i'}, \mathbf{x}_i)$, and $K(\mathbf{x}_i, \mathbf{x}_{i'}) = 0$ iff $\mathbf{x}_i = 0$.

(ii) Linearity: for $\alpha_1, \alpha_2, \ldots$ the linear combination of a semi-positive definite kernels $\alpha_1 K + \alpha_2 K + \ldots$ is semi-positive definite.

(iii) Product positiveness: $K_1 K_2 \geq 0$.

(iv) The direct sum of semi-positive definite kernels is semi-positive definite: $K_1 \oplus K_2 > 0$.

(v) The direct product of semi-positive positive kernels is semi-positive definite: $K_1 \otimes K_2$.

Demonstration of $(i) - (iv)$ is immediate. Demonstration of (v) and (vi) can be found in Berg [22].

5.2 Some Kernels

There exist many kernels for ML-related applications. In this section, we describe some of the most widely used. We dedicate the section 5.3 to the Fisher kernel and 5.4 to the NMF kernel. An introduction explaining the foundations of several kernels is found in Vert [231].

Linear Kernel

The linear kernel has the form

$$K(\mathbf{x}, \mathbf{x}_{i'}) = \langle \mathbf{x}, \mathbf{x}_{i'} \rangle \tag{5.10}$$

and transformation (5.2) does not change the feature space concerning the input space, allowing the dot product to evaluate the similarity between the observations.

Gaussian Kernel

The Gaussian kernel, also known as Radial basis function (RBS), is

$$K(\mathbf{x}, \mathbf{x}_{i'}) = \exp \frac{\|\mathbf{x} - \mathbf{x}_{i'}\|_2^2}{\sigma^2} \tag{5.11}$$

where σ^2 is the model variance.

A variety of this kernel is the Laplacian kernel, in which the L_1 norm is used. Algorithms derived from this type of kernel have been reported to have better properties than Gaussian kernels [105].

Polynomial Kernel

The polynomial kernel assumes its expansion in polynomial form with degree n. It is

$$K(\mathbf{x}, \mathbf{x}_{i'}) = a_n \langle \mathbf{x}, \mathbf{x}_{i'} \rangle + a_0 \tag{5.12}$$

where a_n is the shape parameter and a_0 the intercept.

Cosine Kernel

The angle between two vectors is one of the most used measures to evaluate the similarity between the data contained in a matrix. The cosine between two vectors naturally induces the kernel

$$K(\mathbf{x}, \mathbf{x}_{i'}) = \frac{\langle \mathbf{x}, \mathbf{x}_{i'} \rangle}{\|\mathbf{x}_i\|_2 \|\mathbf{x}_{i'}\|_2} \tag{5.13}$$

This kernel projects the points generated by the extrema of vectors onto a sphere in such a way that the angle is independent of the length of the vectors, as a consequence of the denominator.

Chi-square Kernel

The kernel χ^2 is based on the distance χ^2, being

$$K(\mathbf{x}_i, \mathbf{x}_{i'}) = \exp \frac{\left(\mathbf{x} - \mathbf{x}_{i'}\right)^2}{\mathbf{x} + \mathbf{x}_{i'}} \tag{5.14}$$

Anova Kernel

The Anova kernel, also known as $Linear Discriminant Analysis$(ADL), is a linear classifier, [1]. It is based on the analysis of variance. It supposes the introduction of the covariance matrix Σ, evaluating the similarity

$$K(\mathbf{x}_i, \mathbf{x}_{i'}) = \mathbf{x}_i \Sigma^{-1} \mathbf{x}_{i'} \tag{5.15}$$

PCA Kernel

The PCA kernel is an extension of the concept of principal components using kernelization techniques. The basic idea is to apply PCA to a kernel array.

[1] This type of kennel is also known as ADL and used as a linear classifier [210]

5.3 The Fisher Kernel

For data given by the matrix (4.9), and assuming a generative model, the kernel has statistical significance. Historically, it is a debt to Jaakkola [141] and is due to their study of covariance matrix properties.

Under the hypothesis of the existence of a *pdf* generating the data, the posterior $P(X|\theta)$, can be related to log-likelihood $l_X(\theta) = \log\, P(X|\theta)$, with gradient

$$U_X = \frac{\partial}{\partial\theta} \log\, P_{\theta\in\Theta}(X|\theta) \tag{5.16}$$

that provides directions of the maximum contribution of the parameters, and they are the Fisher scores. Also the Fisher information matrix $I_{\mathcal{F}}$ is

$$\begin{aligned} I_{\mathcal{F}} &= E(U_X) \\ &= U'_X\, U_X \end{aligned} \tag{5.17}$$

defining the Fisher kernel as [202]

$$\begin{aligned} K(\mathbf{x}_i, \mathbf{x}'_{i'}) = U_\theta(\mathbf{x}_i)\ , \mathcal{I}_F^{-1} U_\theta(\mathbf{x}_{i'}) \quad \text{s.t. (i) } U_\theta(\mathbf{x}_i) = -\frac{\partial}{\partial\,\theta} \log\, P(\mathbf{x}_i|\theta) \\ \text{(ii) } \mathcal{I}_F = E_{\mathbf{x}}\big[U_\theta U'_\theta\big] \end{aligned} \tag{5.18}$$

and mapping elements from the input space to the space of the gradients.

In Mathematical Statistics, the Fisher information matrix is a measure of the information contained in a sample and is related to the Cramer-Rao bound as [49]

$$var(\widehat{\theta}) \geq \frac{1}{I_{\mathcal{F}}} \tag{5.19}$$

$\widehat{\theta} \in \Theta$ being an estimation of $\theta \in \Theta$ and reaching equality if the statistic θ is sufficient. Formula (5.19) is the Cramer-Rao bound, and it is achieved for efficient estimators. Also, the Fisher kernel is an efficient estimator of the posterior. This result has been demonstrated by Tsuda [226].

5.4 NMF Kernel

This section and further are the results published in [92, 89, 90]. The NMF kernel was first investigated by Zhang [245]. It can be obtained from matrices

$\mathbf{W}$ and $\mathbf{H}$ of (4.22) assuming that $\mathbf{H}_\phi$ is a *learned basis* (one in which the representations are stable [163]). NMF is

$$[\mathbf{Y}]_{ij} = [\mathbf{W}]_{ik}[\mathbf{H}_\phi]_{ik} \tag{5.20}$$

and $\mathbf{Y}$ depends on the space span of $\mathbf{W}$ and $\mathbf{H}$. Hence,

$$\begin{aligned}[\mathbf{W}]_{ik}\mathbf{I} &= [\mathbf{Y}]_{ij}[\mathbf{H}_\phi]'_{kj}\big([\mathbf{H}_\phi]_{kj}[\mathbf{H}_\phi]'_{kj}\big)^{-1} &(5.21)\\ &= [\mathbf{Y}]_{ij}[\mathbf{H}_\phi]^\dagger_{jh} &(5.22)\end{aligned}$$

and

$$[\mathbf{W}]_{ik} = [\mathbf{Y}]_{ij}[\mathbf{H}_\phi]^\dagger_{jh}\mathbf{I} \tag{5.23}$$

and calling $\phi(\mathbf{Y}) = \mathbf{Y}\mathbf{H}_\phi{}^\dagger\mathbf{I}$

$$\begin{aligned}[\mathbf{W}]_{ik}[\mathbf{W}]'_{ik} &= \Big([\mathbf{Y}]_{ij}[\mathbf{H}_\varphi]^\dagger_{jk}\mathbf{I}\Big)\Big([\mathbf{Y}]_{ij}[\mathbf{H}_\Phi]^\dagger_{jk}\mathbf{I}\Big)'\\ &= \phi\big([\mathbf{Y}]_{ij}\big)\phi\big([\mathbf{Y}]'_{ij}\big) &(5.24)\end{aligned}$$

So, equation (5.24) leads to the NMF kernel

$$K\big(\mathbf{y}_i, \mathbf{y}_{i'}\big) = \big\langle \mathbf{w}_i, \mathbf{w}_{i'} \big\rangle \qquad \Big(\text{for } \mathbf{w}_i \in \mathbf{W}\Big) \tag{5.25}$$

5.5 NMF and Fisher Kernel Equivalence

The kernels obtained with NMF are Fisher kernels. This statement is immediate if we take into account that KL is equivalent to the the log-likelihood, and that it corresponds to the second term of (4.34). Then, it is immediate,

$$\mathcal{L}(\theta) = \log\, P(X|\theta) \tag{5.26}$$

Taking into account that $E(\mathbf{WH}) = E(\widehat{\mathbf{Y}}) = \mathbf{Y}$, leads to

$$I_{\mathcal{F}} = \frac{\partial}{\partial\widehat{\mathbf{Y}}} D_{KL}\big(\widehat{\mathbf{Y}}\|\mathbf{Y}\big) \tag{5.27}$$

These results allows to use the plug-in method. In this case, the plug-in consists of taking the results of the NMF of the table 4.2 and introducing them in the kernel (5.25). This kernel inherits properties of the objective function of the factorization to minimize the distance/divergence, and that also imposes the metric of the space, which in turn implies a norm.

5.6 NMF Kernel Classification Error

For a sufficiently derivable convex function f, for pdf p, and $q \approx p + \Delta p$, the Taylor expansion of $f(p)$ is

$$f(p) = f(q) + (p-q)f'(q) + \frac{1}{2}(p-q)^2 f''(q) + \cdots \tag{5.28}$$

and when linearized, it has an approximation error that can be estimated with the Lagrange remainder

$$f(p) = f(q) + (p-q)f'(q) + R_n(\xi) \tag{5.29}$$

for some ξ_i such that $\min(p_i, q_i) < \xi_i < \max(p_i, q_i)$. It is immediate

$$D_f(p \| q) = R_n(\xi_i) \tag{5.30}$$

and lets us interpret the Bregman divergence as the linearization error.

For the matrix case

$$R_n = \operatorname{tr}\Big[\nabla^2_{\widehat{\mathbf{Y}}} f([\widehat{\mathbf{Y}}]'_{ij})\big([\widehat{\mathbf{Y}}]_{ij} - [\mathbf{Y}]_{ij}\big)^2\Big] \tag{5.31}$$

$$= \operatorname{tr}\Big[\mathcal{H}\operatorname{var}(\widehat{\mathbf{Y}})\Big] \tag{5.32}$$

where $\mathcal{H}$ is the Hessian. So, by the Cauchy-Schwartz inequality

$$R_n \leq g_{ij}\operatorname{var}\big([\widehat{\mathbf{Y}}]_{ij}\big) \tag{5.33}$$

being

$$g_{ij} = \sum_j \mathcal{H} \qquad \big(\text{s.t. } \mathcal{H} \in \mathbb{R}^{p\times p}\big) \tag{5.34}$$

the well-known geodesic distance. So, the inequality

$$g_{ij} \geq \frac{1}{\Big|\operatorname{var}([\widehat{\mathbf{Y}}]_{ij})\Big|} \tag{5.35}$$

holds since $R_n < \text{var}(\widehat{\mathbf{Y}})$ and the well known geodesic distance is the Bregman divergence approximation error for every distance or entropy. The geodesic distance is related to the second derivative, which represents the curvature of the data. In this context, data represent information, so it is therefore their maximum variation or informational content.

5.7 NMF Kernel for SVM

SVM is a technique that uses kernels for classification. It is a supervised method if a part of the available data is used to classify the remaining observations, or a semi-supervised method if the available data is used to classify new observations. The classification is made on a qualitative or ordinal variable indication that belongs to a class identified with a label. The set of all observations is assimilated to a region of space on which a partition is established according to the value of the classifying variable, also called response.

The idea, due to Vapnik [32, 230], is to create a separation line between the two regions of space [2].

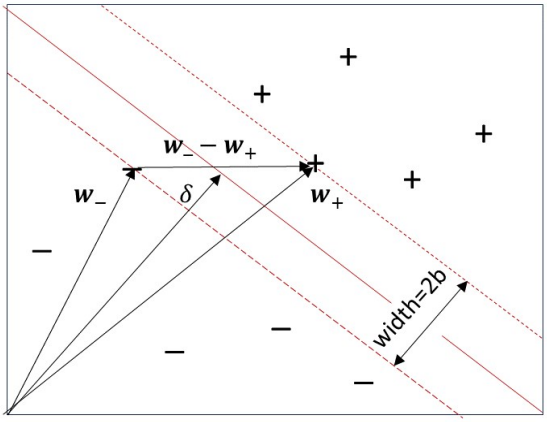
Restricting the problem to two regions, the elements belong to one of them and are graphically identified with the labels $\mathcal{Y} = \{-1, +1\}$. Construction of the SVM involves taking a vector δ to the center of the partitioning line and considering the closest vectors $\mathbf{w}_i$ of each region as support vectors. These vectors allow the creation of a separate band or street.

Because the dot product is a projection, the vectors located in the area above the separation line (observations identified with +), have a dot product greater than 1. The opposite occurs for the vectors in the zone of the points identified with − (their dot product is less than 1). This concept gives rise to a classification rule. Its graphic meaning is illustrated in Figure 5.1.

Under these conditions, we have for the support vectors (nearest vector to the street) $\mathcal{Y}_i(\mathbf{w}_i\delta + b) - 1 = 0$, being necessary to maximize the street.

[2] Some authors refer to the line perpendicular to the vector δ as the hyperplane that separates the two regions of the space. In our development, this question is not relevant. The orientation of a plane in a finite-dimensional space is a vector perpendicular to the surface. The intersection between the perpendicular plane hyperplane and the one containing the data is the separation

Figure 5.1: SVM margins construction.



Reproduced from [94]. Representation of vector δ and band or street. Vectors identified with a minus sign have a dot product less than one. Support vectors are those that construct the separation band or street between margins. The SVM minimizes the margins difference.

Taking the difference $\mathbf{w}_+ - \mathbf{w}_-$ as the street width, it can be expressed as

$$\text{width} = \mathbf{w}_+ - \mathbf{w}_- \tag{5.36}$$

$$= \mathbf{w}_+ - \mathbf{w}_- \frac{\delta}{\|\delta\|} \tag{5.37}$$

where $\delta/\|\delta\|$ has unit norm. Observing that $\mathbf{x}_+ = 1 - b$ and $\mathbf{x}_- = 1 + b$, so (5.37) is

$$\text{width} = \frac{2}{\|\delta\|} \tag{5.38}$$

The street maximization is equivalent to the minimization of $\mathbf{w}$, which is achieved with the Lagrange multipliers. For Lagrangian $\mathcal{L}$

$$\mathcal{L} = \frac{1}{2}\|\delta\|^2 + \sum_i \alpha_i\big(\mathcal{Y}_i \delta \mathbf{w}_i + b\big) - 1 \tag{5.39}$$

where the factor 1/2 is to simplify the derivatives, which are

$$\frac{\partial L}{\partial \mathbf{w}} = \mathbf{w} - \sum_i \alpha_i \mathcal{Y}_i \mathbf{w}_i$$
$$\frac{\partial L}{\partial b} = -\sum_i \alpha_i \mathcal{Y}_i$$

and equating to zero gives the solutions

$$\mathbf{w} = \sum_i \alpha_i \mathcal{Y}_i \mathbf{w}_i \tag{5.40}$$
$$\sum_i \alpha_i \mathcal{Y}_i = 0 \tag{5.41}$$

Substituting (5.39) y (5.40) in (5.39) leads to

$$\mathcal{L} = \sum_i \alpha_i + \sum_{ii'} \alpha_i \alpha_{i\prime} \mathcal{Y}_i \mathcal{Y}_{i\prime} \mathbf{y}_i \mathbf{y}'_{\mathbf{i}} \tag{5.42}$$

The expression (5.41) depends solely on the scalar product $\mathbf{x}_i \mathbf{x}_{\mathbf{i}}^{\mathbf{P}}\mathbf{rime}$, which justifies the introduction of 4.81 transformations to achieve space separability. Among the many possible transformations, perhaps the most widespread are those explained in the 5.2 section. Using these transformations to achieve space separability is known as the kernel trick.

The generalization to the case where $\mathcal{Y}$ has more than two levels is due to Cortes [64]. The exposed margins construction corresponds to hard margins. The relaxation of this construction to the case in which the existence of points on the street is allowed is known as soft margins. Figure 5.2 shows the meaning of the margins

5.8 NMF Kernel Margins Behavior

The behavior of SVM with the proposed kernel in (5.25) presents some peculiarities related to margins. The convergence (4.73) has important consequences. The margins depend on the value of k and are a flat structure.

The proof is simple when considering the matrix containing the vectors $\mathbf{w}_i$ of (5.25). For any two vectors, identified by the subscript i and i', and obtained with different numbers of components k and k', respectively, for which $k \geq \min(m, n)$. As a consequence of the convergence of (4.73), the expectation of both vectors are the same, and

$$\frac{\sum_{k=1}^{k} w_{ik}}{k} = \frac{\sum_{k=1}^{k'} w_{ik\prime}}{k'} \tag{5.43}$$

since $k' > k$, the sum of the second member is greater than the sum of the first. It follows

$$w_{ik} - E(w_{ik}) > w_{ik'} - E(w_{ik\prime}) \tag{5.44}$$

and

$$w_{ik'} - w_{ik'} \xrightarrow[k\to\infty]{} 0 \tag{5.45}$$

Showing that the difference between the margins is zero.

5.9 Examples

The examples have the sole purpose of illustrating the features of the proposed kernel; in particular, the arbitrary misclassification error rate. The datasets were selected from the UCI repository [79], having sufficiently few criteria to be easily handled by our kernel, and the existence of two or more evaluation criteria of each observation. Those data sets are: Seeds (210×7 non-negative real valued), Ecoli 336×8 non-negative real valued, Glass (214×10 non-negative real valued), and Bupa (345×23 non-negative real/integer valued).

To use this kernel, there are several parameters to choose. They are related to the transformation to probabilistic space, its NMF, involving the selection of values of k, and the support vectors after training phase. These steps are indicated in algorithm 1

Since the selected datasets have non-negative inputs, to estimate the density, we do simple manipulations taking

$$\widehat{\mathbf{y}}_j = \frac{x_{ij}}{\sum_i x_{ij}} \qquad \text{(for each } j\text{)}$$

Algorithm 1: Learned Basis.

Input: Data Matrix $\mathbf{X}$
Approximation condition ϵ
Number of model components k
(**NMF** :)
1: Transform $\mathbf{X}$ to $\mathbf{Y}$
initialize $\mathbf{W}$ and $\mathbf{H}$
2: Factorize $\mathbf{Y}$ obtaining $\mathbf{W}$ and $\mathbf{H}$
Output: $\mathbf{W}$, $\mathbf{H}$

This criterion is consistent with those explained in section 4.1.

Parameters related to the non-negative factorization, consists to select a value of k, initializing matrices $\mathbf{W}$ and $\mathbf{H}$ for a particular distance or divergence and iterate between them until a condition $\|\mathbf{Y}-\mathbf{WH}\|_1 < O(1/m)$ is fulfilled. For all cases, we randomly initialize matrices, and we use values $k = \min(m, n), \cdots$,, which allows an arbitrary approximation of $\mathbf{Y}$ to $\widehat{\mathbf{Y}}$ (for a high enough number of iterations), as is well-known [91, 153].

The selection of support vectors can be a laborious task, after training the data. We use the extended method consisting in introducing the output of the vanilla kernel for the input the proposed kernel with the help of the *e1071* R package [176]. This package consists in a iterative procedure with parameters μ (means of the data, a rotation matrix $\mathbf{Q}$, randomly generated before the first iteration with a linear cost function ($C = 1$).

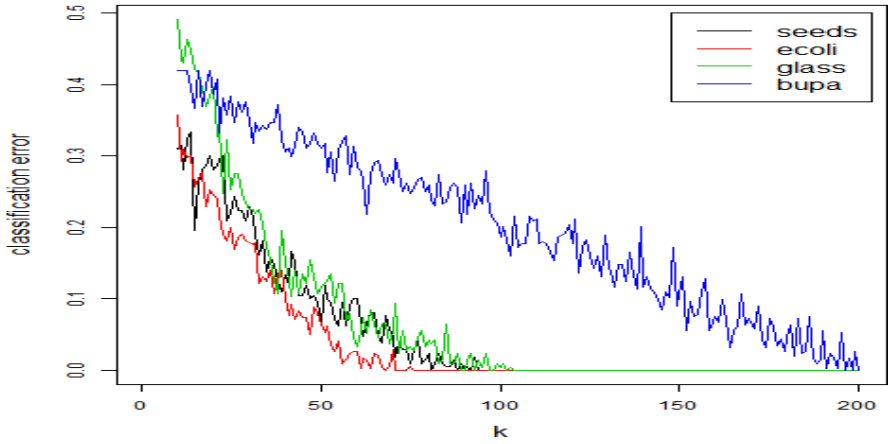
For new observations, its use is shown in algorithm 2

Algorithm 2: Classification Error.

Input: $\mathbf{W}$
q_1, q_2, ..., q_q (labels)
1: Transform $\mathbf{X}$ to $\mathbf{Y}$
initialize $\mathbf{W}$ and $\mathbf{H}$
2: Factorize $\mathbf{Y}$ obtaining $\mathbf{W}$ and $\mathbf{H}$
for $q = 1, \ldots, q$ **do**
1:$\mathcal{Y}_i = (q_i, \cup q_{i+1})$
2: compute $E(q_i|\mathbf{W})$
Output: $(\mathbf{w}_i, l_q)$

To better illustrate the potential of the Fisher kernel, we run the algorithm.

Figure 5.2: NMF-Fisher Kernel Misclassification Error.



Misclassification error rate vs. number of components. Number of components is identified with the value of k in the expressions of $\mathbf{W}$ and $\mathbf{H}$. The execution parameter for the algorithm is $C = 1$. Error is obtained from the confusion matrix. We eliminate the diagonal, or correct classifications, and then we sum the remaining entries.

This procedure is useful to compare results. We run the kernel for different number of model components ($k = 10, \ldots, 200$), obtaining 190 confusion matrices for each dataset. From confusion matrices, the misclassification rate (or the result of subtracting to one the number of valid classifications divided by the total number of observations), is plotted versus k in Figure 5.2. The sizes of the training-prediction partition are the extended 8020.

Results shown in Figure 5.2 are the misclassification error rate vs. the number of model components. The error classification rate decreases indefinitely with the number of components. In fact, it is a result depending on the convergence speed, which increases with the number of model components, achieving the same limit if $k \geq \min(m, n)$ [91].

Chapter 6

Clustering Validation

Clustering can be defined as *the formal study of algorithms and methods for grouping objects according to measured or perceived intrinsic characteristics or similarities* for purposes such as classification of underlying data structures, natural classification, data compression, and summarization [142]. It is attributed to MacQueen [168], who introduced the classical k-means method. It is one of the most widespread techniques in ML and has seen significant growth in recent years. Its main idea is to group observations into groups so that they are as similar as possible.

Various criteria coexist, and they give rise to the families of methods *hard-clustering*, in which each entity belongs exclusively to a group, and *soft-clustering*, in which they have different degrees of membership to each group. Independent of the classification method, an important step is determining the quality of the partition, known as cluster validation. In most of cases, this means determining the number of clusters. This step is relevant for classification and critical for unsupervised cases. Thus, the inferred number of clusters depends only on the data structure.

6.1 Clustering Methods Overview

The data frame or matrix $\mathbf{X}$ of Formula (2.13) represents the data. Each row corresponds to the measure of (some) characteristics of interest, called a *low-level property* or *feature*, for the corresponding i-th item (also, observation

or instance) providing the numerical results $(x_{i1}, \ldots, x_{in})$ after evaluation. The geometrical interpretation of these results are points in the $\mathbb{R}^n$ space. For convenience, we associate a vector with each point, and the set of all observations $\mathbf{x}_i$ is a matrix.

Formally, clustering is the task of assigning each observation $\mathbf{x}_i \in \mathbf{X}$ to a subset $l_k \in \mathcal{Y} = \{l_1, \ldots, l_K\}$. Each l_k is known as a *label* or *cluster*. Typically $K < n$. The assignment task assumes a similarity measure or a map $h(\cdot)$ such that [228]

$$h : \mathbf{x_i} \longrightarrow \mathcal{Y} \qquad (\mathbf{x}_i \in \mathbf{X}) \tag{6.1}$$

$$h(\mathbf{x}_i) \mapsto l_k \tag{6.2}$$

and for each item we have the label assignment

$$\big\{(\mathbf{x}_1, l_k), \ldots, (\mathbf{x}_m, l_{k'})\big\} \tag{6.3}$$

The shape of $h(\cdot)$ characterizes a clustering method, of which two families of approaches exist.

6.1.1 Hard-clustering

Hard-clustering methods assume grouping the data into K disjoint groups. Obtention involves assigning k geometric estimates recalculated in an iterative process in such a way that they minimize a distance d

$$d = \sum_i \| \mathbf{c}_k - \mathbf{x}_i \|_2^2 \quad \text{(for each } k\text{)} \tag{6.4}$$

where $\mathbf{c}_k \in \mathbb{R}^n$ is the position vector the geometric center of the groups k. This is *k-means* [168]. If medians are taken for $\mathbf{c}_k$, the method is *k-medians*. Also, medioids, or existing points as class representative, for c_k is *k-medioids* [182]. In this variant, the task is usually performed with an L_1 metric, as in Park and Ackermann (6.3) [192, 1].

Hierarchical algorithms are based on the use of a distance matrix to study the connectivity of items [145]. These techniques suffer a drawback in

that they are difficult to implement for large datasets, and so they are not scalable, remaining of interest to provide graphical and intuitive results, and there are many variants.

Agglomerative clustering is based on a tree construction, and the distance between all the items is computed, followed by averaging the closest ones and repeating the procedure until a single cluster is achieved. This method is known as AGNES [217]. If the reverse order to the top-down procedure is chosen, the method is called DIANA [170].

6.1.2 Soft-clustering

Unlike hard-clustering methods, soft-clustering is based on the assumption that each entity has some degree of membership in more than one group. There are several subfamilies of methods for this purpose. Perhaps the most representative methods are those that belong to the probabilistic methods.

Probabilistic clustering is attributed to Har [117] and independently to Cohn [63], who introduced an indicator function for observations based on kernelization and used the null hypothesis as a classifier. There are two concepts of probabilistic clustering: the first concept supposes that each entity belongs to each cluster with a different probability [78], while the second concept classifies the observations that most likely belong to each distribution [70]. Both approaches are based on determining the suitable mixtures or *components*; in number and parameter, $f(\mathbf{x}|\Theta) = \sum \alpha_j g_j(\mathbf{x}|\theta_j)$ (where $\mathbf{x}$ is the vector corresponding to an observed entity, g is a density with parameter θ, and α gives the mixing weights assigned to each density; the sum is performed on the overall set of mixtures), converting an unsupervised classification problem into a parameter estimation problem that assigns probabilities of membership to each item in each cluster.

Within this family of methods, other approaches are important. On the one hand, the creation of a grid for one (or several) data points is a division of the space that surrounds each data point, allowing the evaluation of a cell density, and then ordering them according to their density. The identification of centers of the cells is the clusters [110]. These methods, known as *grid clustering*, are intuitively justified by *the main reason why we recognize clusters is that within each cluster we have a typical density of points that is considerably higher than outside of the cluster* [85], proposing

the DBSCAN [85] algorithm. These algorithms work well for well-defined shapes.

Also, the work of Ng [185] introduces spectral clustering and the algorithm for well-defined shapes CLARANS [186]. These algorithms clusterize according the eigenvalues of matrices derived from the data.

6.1.3 NMF Clustering

NMF can be used for clustering purposes, regardless of the type and/or algorithm used to obtain factorization (4.22). This relation admits additional interpretation when considering the products $\mathbf{w}_i\mathbf{h}_j$ as the linear combinations of the underlying or latent variables [2, p. 8]. The factorization of (4.22) involves transforming the coordinates of each vector of observations concerning a base $\mathbf{H}$, obtaining a new representation $\mathbf{w}_i \in \mathbb{R}^K$. In this case, a high value of any of the components of $\mathbf{w}_i$ is identified as a high dependency or relationship. The pioneering work in this sense was provided by Lee [161].

Considering the data frame $\mathbf{X}$ as a matrix, the interpretation of NMF as a co-clustering technique is immediate. Co-clustering, initially proposed by Hartigan [119], is a consequence of this type of algebraic structure. Factorizing (4.1) provides simultaneous classification of items and observational variables. An example in the context of this chapter is the work of Cho [56]. The use of NMF appears implicitly when using genomic data.

From the matrix structure, it follows that each entity can be assigned to more than one group, or in other words, more than one l_k label can be assigned to the items. Furthermore, (4.5) transformations provide a probabilistic interpretation of (2.13). In this way, the columns of the matrix $\mathbf{W}$ are equivalent to the densities $P(x_i|\, z_1), \ldots\ P(x_i|\, z_k)$ expressed concerning the base $\mathbf{H}$, and identifying the dimension of the space span with the latent variables, which in turn we identify with the clusters. This probabilistic algebraic equivalence means that clustering methods based on NMF have some theoretical advantages when related to other methods.

NMF has important theoretical consequences. From the ML point of view, equivalence with other techniques is relevant. The equivalence of NMF with PLSA is immediate if the relationships are taken into account [129, 74, 75]. Also, it can be shown that classification obtained with NMF factorization is equivalent to *k-means* by using a Bayes classifier [74], and

equivalence to spectral clustering supposes to introduce a weights matrix [73].

6.2 Clustering Validation

Validation is the evaluation of the quality of the assignment. Validation requires that several properties (sensibility, cluster number impact, and invariance) be determined in order to establish the capability of a cluster method for various data structures. There exist many works that justify these criteria, such as Aggarwal [2, chap. 23] and Franti [101], among others.

For the case of soft-clustering (each item has some degree of membership to various classes), there exist qualitative methods based on graphical criteria. These methods are usually based on the assumption $k \leq n$. The graphic representation of eigenvalues obtained with the aid of SVD provides good estimations but suffers from drawbacks in terms of providing quantitative results [24].

Formally, it supposes the minimization of a *loss function*

$$L = \left\| f_\theta(x;\theta) - \hat{f}_{\theta^\star \in \Theta}(x;\theta^\star) \right\|_2 \tag{6.5}$$

where $\theta^\star$ is the estimate of θ, and $\hat{f}$ estimates f, both multivariate. They are the posteriors representing the credibility (or degree of belonging) of each observation to a class.

Parameters minimizing (6.5) are obtained as

$$\theta^\star = \arg\min_{\theta \in \Theta} L \tag{6.6}$$

and they are the best choice.

Credibility is the quality of classification, which evaluates the posterior probability of the number of mixtures [215]. Some contributions are a debt to Smyth [215] who uses likelihood cross-validation to infer information on

the number of model components. Using a different approach, the similarity between clusters can be evaluated from the χ^2 statistic between probabilistic classifications [191]. By introducing an index and assuming that each cluster is generated by a parametric distribution, the minimum can be taken as the validation index [102]. More recent works include Olivares [189], which, in the scope of astronomical observations and under the hypothesis of normality and the existence of a correlation, presents an algorithm in which the posterior of the correlations follows a gamma probability density function (*pdf*). The work of Usefi [116] presents a purely algebraic approach in which elements are clustered by co-linearity. The use of probability as proximity measure is attributed to Israel [20].

Other studies focused on validation and sharing this viewpoint of probabilistic clustering are attributed to Har [117], who introduced an indicator function for observations based on kernelization and used the null hypothesis as a classifier. Smyth [215] used likelihood cross-validation to infer information on the number of model components. Belkin [18] studied empirical risk minimization when over-fitting. A discussion on the shape of $\hat{f}$ and f is provided by Huber [137]. Wade [233] proposes a Bayesian parametric estimation of the uncertainty of the cluster assignment. By introducing an index and assuming that each cluster is generated by a parametric distribution, the minimum can be taken as the validation index [102]. More recent works include Olivares [189], which, in the scope of astronomical observations and under the hypothesis of normality and the existence of a correlation, presents an algorithm in which the posterior of the correlations follows a gamma probability density function (*pdf*). The work of Usefi [116] presents a purely algebraic approach, in which elements are clustered by co-linearity. A review work, centered on the impact and importance of clustering validation in the context of the recent grow of bioinformatics is provided by Ullmann [227].

Extended and good methods for validations are the silhouette index and the gap statistic. The silhouette value is a measure of how similar an object is to its own cluster (cohesion) compared to other clusters (separation). It can be used to study the separation distance between the resulting clusters. The silhouette plot displays a measure of how close each point in one cluster is to points in the neighboring clusters and thus provides a way to assess parameters like number of clusters visually [201]. The gap statistic compares the within-cluster dispersion to the expectation of a reference distribution [224].

6.3 Sequence of Traces

From formula (4.22), the trace

$$\operatorname{tr}\big(\widehat{\mathbf{Y}}'\widehat{\mathbf{Y}}\big) = \sum_{k} \operatorname{diag}\big([\mathbf{W}]_{ik}[\mathbf{H}]_{kj}\big)'\big([\mathbf{W}]_{ik}[\mathbf{H}]_{kj}\big) \tag{6.7}$$

$$= \Big\|\operatorname{diag}\,\big([\mathbf{W}]_{ik}[\mathbf{H}]_{kj}\big)'\big([\mathbf{W}]_{ik}[\mathbf{H}]_{kj}\big)\Big\|_1 \tag{6.8}$$

leads to a sequence with varying k in the NMF factorization space span.

$$\Big\{z_{[k]}\Big\}_k = \Big\{\operatorname{tr}\,\big([\mathbf{W}]_{i1}[\mathbf{H}]_{1j}\big)'([\mathbf{W}]_{i1}[\mathbf{H}]_{1j}\big),$$
$$\operatorname{tr}\,\big([\mathbf{W}]_{i2}[\mathbf{H}]_{2j}\big)'\big([\mathbf{W}]_{i2}[\mathbf{H}]_{2j}\big), \dots\Big\} \tag{6.9}$$

$$= \Big\{z_1, z_2, \dots\Big\} \tag{6.10}$$

Here, the sub-index brackets indicate that terms are given in increasing order.

We denote z as

$$z = \operatorname{tr}\,\big([\mathbf{Y}]'_{ij}[\mathbf{Y}]_{ij}\big) \tag{6.11}$$

From formulas (6.10) and (6.11), the quotient

$$\tilde{z} = \Big\{\frac{z_{[k]}}{z}\Big\}_k \tag{6.12}$$

leads to a monotonically decreasing sequence, which can easily be explained by considering the following cases.

If $k \geq \min(m, n)$ occurs, then $\widehat{\mathbf{Y}} \to \mathbf{Y}$ during the iteration process, as has been demonstrated. By imposing the same approximation condition for ϵ in formula (4.27) for all products of $\mathbf{WH}$ obtained with different values of k and introducing

$$\left\|[\mathbf{W}]_{ik}[\mathbf{H}]_{kj}\right\|_1 = \frac{1}{k}\left\|[\widetilde{\mathbf{W}}]_{ik}[\widetilde{\mathbf{H}}]_{kj}\right\|_1 \tag{6.13}$$

we find that the inequality

$$\frac{1}{k^2}\text{tr}\left([\widetilde{\mathbf{W}}]_{ik}[\widetilde{\mathbf{H}}]_{kj}\right)'\left([\widetilde{\mathbf{W}}]_{ik}[\widetilde{\mathbf{H}}]_{kj}\right) > \frac{1}{(k')^2}\text{tr}\left([\widetilde{\mathbf{W}}]_{ik'}[\widetilde{\mathbf{H}}]_{k'j}\right)'\left([\widetilde{\mathbf{W}}]_{ik'}[\widetilde{\mathbf{H}}]_{k'j}\right) \tag{6.14}$$

holds only if $k' > k$.

If $k < \min(m, n)$, the convergence of $\widehat{\mathbf{Y}}$ to $\mathbf{Y}$ does not generally occur. Thus, to observe this behavior, it is necessary to impose a wider condition on inequality (4.27) (i.e., $\epsilon = O(1/m)$), which is the jump in the empirical univariate distribution of the columns of $\widetilde{\mathbf{Y}}$. Because the convergence of the empirical distribution to each component (column) of $\widetilde{\mathbf{Y}}$ is almost sure, the difference is bounded. Taking the maximum difference, a decreasing behavior appears when we divide by increasing values of k.

6.3.1 Trace Sequence Limit Behavior

The construction of the trace sequence is the first step in obtaining the results that are exposed in the following sections and that constitute the publications [90, 93].

For sequence (6.12) obtained from a full rank matrix, the function

$$\varphi(z_{[k]}) = \left(\frac{z_{[k]}}{z}\right)^{-z} \tag{6.15}$$

represents the inverse of the (non-logarithmic) likelihood. This function can be written as

$$\varphi(z_{[k]}) = \left(1 + \frac{z_k - z}{z}\right)^{-z} \tag{6.16}$$

By introducing the following transformations

$$\lambda = z_1 - z \tag{6.17}$$

$$\frac{1}{\nu} = \frac{z_k - z}{z_1 - z} \qquad (k \neq 1) \tag{6.18}$$

with Jacobian

$$|\, J \,| = \det \begin{bmatrix} \dfrac{\partial\, z_1}{\partial \lambda} & \dfrac{\partial\, z_1}{\partial \nu} \\ \dfrac{\partial\, z_k}{\partial \lambda} & \dfrac{\partial\, z_k}{\partial \nu} \end{bmatrix} \tag{6.19}$$

$$= \det \begin{bmatrix} 1 & 0 \\ 0 & \dfrac{\lambda}{\nu^2} \end{bmatrix} \tag{6.20}$$

$$= \frac{\lambda}{\nu^2} \tag{6.21}$$

we can express (6.12) as function of the new variables as follows:

$$\varphi(\lambda, \nu) = \frac{\lambda}{\nu^2}\Big(1 + \frac{\lambda}{z\nu}\Big)^{-z} \qquad \Big(1 \leq \nu < +\infty\Big) \tag{6.22}$$

Formula (6.17) is merely a displacement, and relation (6.18) transforms the domain of z_k to a set with lower bound 1 but no upper bound. This transformation does not change the dimension of the space because ν depends on λ.

Hence,

$$\varphi(\lambda, \nu) = \frac{z}{z-1}\frac{\partial}{\partial \nu}\Big(1 + \frac{\lambda}{z\nu}\Big)^{1-z} \tag{6.23}$$

Because

$$\Big(1 + \frac{\lambda}{z\nu}\Big) \xrightarrow[\nu \to \infty]{} \exp\Big(\frac{\lambda}{z\nu}\Big) \tag{6.24}$$

substituting (6.24) in (6.23) and taking into account that z is constant, we obtain

$$\varphi(\lambda,\nu) = c\,\frac{\partial}{\partial\nu}\,\exp\Big(-\frac{\lambda}{c\nu}\Big) \qquad \Big(\text{s.t. } c = \frac{z}{z-1}\Big) \tag{6.25}$$

With the sole purpose of facilitating further calculations and avoiding an incomplete inverse gamma, we perform a change of variables $y = \nu - 1$ that ensures the variation domain $(0, +\infty)$, with no effect on the scale (Jacobian is one). Then, we take

$$x = \frac{1}{cy} \tag{6.26}$$

Function (6.25) can be rewritten as follows:

$$\varphi(\lambda, x) = \frac{1}{c}\,\frac{\partial}{\partial\,x}\frac{\partial^2}{\partial\lambda^2}\,e^{-\lambda\,x} \tag{6.27}$$

More generally,

$$\frac{\partial^{r-1}}{\partial\,x^{r-1}}\frac{\partial^{p-2}}{\partial\lambda^{p-2}}\varphi(\lambda,x) = \frac{1}{c}\frac{\partial^r}{\partial\,x^r}\frac{\partial^p}{\partial\lambda^p}\,e^{-\lambda\,x} \quad (p \geq 2 \text{ and } r \geq 1) \tag{6.28}$$

6.3.2 Expectation of Trace Sequence Limit

Because exponential functions are sufficiently regular to interchange the derivative signs, from formula (6.28), we introduce

$$\zeta(\lambda, x) = \frac{\partial^r}{\partial\,x^r}\varphi(\lambda, x) \tag{6.29}$$

After taking a derivative, (6.28) is now

$$\frac{\partial^{p-2}}{\partial\lambda^{p-2}}\zeta(\lambda,x)=\frac{\lambda}{c}(-1)^p x^p e^{-\lambda\,x} \tag{6.30}$$

The Laplace transform of (6.30) is

$$\zeta_{(p-2)}(s)=\frac{\lambda}{c}(-1)^p\int_0^{+\infty}e^{-sx}x^p e^{-\lambda\,x}\,dx \qquad (s>0) \tag{6.31}$$

$$=\frac{\lambda}{c}(-1)^p\Big(\frac{1}{s+\lambda}\Big)^{p+1} \tag{6.32}$$

where we indicate the order of the derivative in parentheses to avoid confusion with the exponents.

From (6.32), we find

$$\zeta(s)=\frac{\lambda}{c}\Big(\frac{1}{s+\lambda}\Big) \tag{6.33}$$

The relationship between (6.32) and (6.33) provides the following recursive formula:

$$\zeta(s)=\frac{(-1)^{p+1}}{(p+1)!}\zeta_{(p-2)}^{(p+1)}(s) \tag{6.34}$$

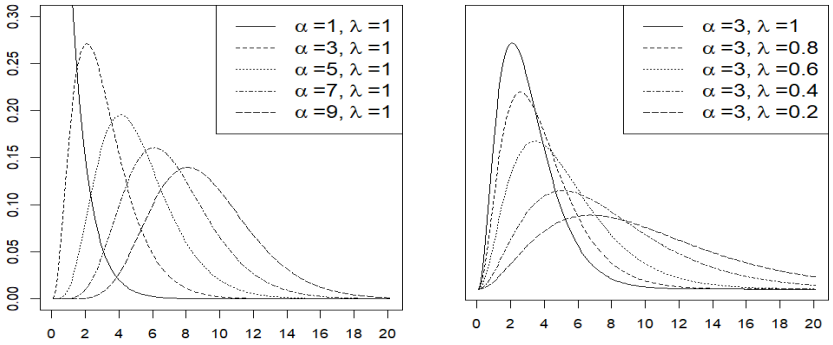
It follows that

$$\int_0^{+\infty}e^{-sx}\zeta(\lambda,\,x)\,dx=\frac{(-1)^{p+1}}{(p+1)!}\int_0^{+\infty}e^{-sx}x^p e^{-\lambda\,x}\,dx \tag{6.35}$$

Hence,

$$\zeta(x;p,\lambda)=\frac{(-1)^{p+1}}{(p+1)!}x^p e^{-\lambda\,x} \tag{6.36}$$

Figure 6.1: Gamma pdf.



.
Left panel shows the effect of parameter α with fixed lambda. Right panel fixes α and varies λ. Both figures shares the same axis scale.

By reversing the change given by (6.29) for $r = p + 1$, we obtain

$$\varphi(x; p, \lambda) = \frac{\partial^p}{\partial\, x^p}\zeta(x; p, \lambda) \tag{6.37}$$

The negative signs cancel, and we find

$$\varphi(x; \alpha, \lambda) = \frac{1}{\Gamma(\alpha)}\lambda^\alpha x^{\alpha-1}e^{-\lambda\, x} \qquad (\alpha = p + 1) \tag{6.38}$$

Formula (6.38) gives the solution to equation (6.28) and is the main result of this chapter. Because the Laplace transform can be viewed as an expectation, this relation can be interpreted as the expectation of the limit function, obtained from the sequences of a non-negative matrix trace, which follows a gamma *pdf*.

6.3.3 Gamma Parameter Selection

The adjustment of an unbiased (gamma) density for sequence (6.15) implicitly assumes

$$\varphi(x_m;\alpha,\lambda) = \max\ c\left(\frac{z_{[k]}}{z}\right)^{-z} \quad \left(\text{s.t. } x_m = \arg\max_x \varphi(x;\alpha,\lambda)\right) \tag{6.39}$$

for c given by formula (6.25) and imposes values for parameters α and λ. Also, the maximum can be obtained in a closed form as

$$x_m = \frac{\alpha - 1}{\lambda} \tag{6.40}$$

By introducing the classical transformation in (6.39),

$$t = \lambda\, x \qquad \left(\text{with } \left|\frac{\partial\, x}{\partial\, t}\right| = \frac{1}{\lambda}\right) \tag{6.41}$$

leads to

$$\varphi(t;\alpha) = \frac{1}{\Gamma(\alpha)} t^{\alpha-1} e^{-t} \tag{6.42}$$

which is a standard gamma density with the following maximum

$$\arg\max_t \varphi(t;\alpha) = \alpha - 1 \tag{6.43}$$

and expectation α.

Figure 6.1 shows the effect of parameters on the gamma pdf. If (6.42) and (6.38) must reproduce the same shape, it is necessary to adjust the expectation of (6.38) to the maximum of (6.42). Then

$$\alpha = \frac{\alpha - 1}{\lambda} \tag{6.44}$$

and

$$\lambda = 1 - \frac{1}{\alpha} \tag{6.45}$$

6.4 Properties

Another solution for equation (6.28) is obtained by writing

$$\frac{\partial^r}{\partial x^r}\frac{\partial^2}{\partial \lambda^2}\varphi(\lambda, x) = \frac{(-1)^r}{c}\lambda^r \frac{\partial^2}{\partial \lambda^2}\varphi(\lambda, x) \tag{6.46}$$

and denoting $\psi(\lambda, x) = \partial^2\varphi(\lambda, x)/\partial \lambda^2$, the relation between derivatives with respect to x is

$$\frac{\partial^r}{\partial x^r}\psi(\lambda, x) = \frac{(-1)^r}{c}\lambda^r \psi(\zeta, x) \tag{6.47}$$

By following the same reasoning described in section 6.3 and considering that $\partial^2\varphi(\lambda, x)/\partial \lambda^2 = x^2 \exp(-\lambda x)$, the Laplace transforms of $\psi^{(p)}$ and ψ are

$$\psi_{(r)}(s) = (-1)^r \frac{\lambda^r}{c}\int_0^{+\infty} e^{-sx}\psi^{(r)}(\lambda, x)dz \tag{6.48}$$

$$= (-1)^r \frac{\lambda^r}{c}\Big(\frac{1}{s+\lambda}\Big)^{r+1} \tag{6.49}$$

$$\psi(s) = \frac{1}{c}\int_0^{+\infty} e^{-sx}\psi(\lambda, x)dx \tag{6.50}$$

$$= \frac{1}{c}\Big(\frac{1}{s+\lambda}\Big) \tag{6.51}$$

respectively. Comparing (6.49) and (6.51), we find

$$\psi(s) = \frac{(-1)^r}{r!}\lambda^r \psi^{(r)}(s) \tag{6.52}$$

after simplifying, the minus signs cancel, which leads to

$$\psi(x; \lambda) = \frac{1}{r!}\lambda^r e^{-\lambda x} \tag{6.53}$$

6.4.1 Relationship with Bayesian Conjugate Analysis

By comparing equations (6.38) and (6.53) for a value of $r = p+1$, we achieve an equality by introducing a factor $\prod_p x$, and

$$\frac{1}{\Gamma(\alpha)}\lambda^{\alpha}x^{\alpha-1}e^{-\lambda x} = \Big(\prod_p x\Big)\,\frac{1}{(p+1)!}\lambda^{p+1}e^{-\lambda x} \tag{6.54}$$

The Bayesian theorem can be written as follows:

$$P(\theta|\,x) = L(\theta)\,P(x|\theta) \tag{6.55}$$

where $P(\theta|\,x)$ is the posterior, $L(\theta)$ is the likelihood, and $P(x|\theta)$ is the prior. We identify the factors as

$$P(\theta|\,x) = \frac{1}{\Gamma(\alpha)}x^{p-1}e^{-\lambda x} \tag{6.56}$$

$$P(x|\theta) = \frac{1}{(p-1)!}\lambda^{p-1}e^{-\lambda x} \tag{6.57}$$

$$L(\theta) = \prod_p x \tag{6.58}$$

$$= x^{\alpha-1} \tag{6.59}$$

where $\theta = \{\alpha, \lambda\}$ is the minimal sufficient statistic.

6.4.2 Efficiency

From (4.22), the Hessian is

$$\mathcal{H} = \frac{\partial^2}{\partial \mathbf{y}'_j \mathbf{y}_{j'}}\mathrm{tr}\Big([\widehat{\mathbf{Y}}]'_{ij}[\widehat{\mathbf{Y}}]_{ij}\Big) \qquad (\mathbf{y}_j \in \mathbf{Y}) \tag{6.60}$$

$$= \frac{\partial^2}{\partial \mathbf{y}'_j \mathbf{y}_{j'}}\,\delta_{jj'}[\widehat{\mathbf{Y}}]'_{ij}[\widehat{\mathbf{Y}}]_{ij} \qquad \text{with}\,\delta_{jj'} = \begin{cases} 1 & \text{if}\, j = j' \\ 0 & \text{otherwise} \end{cases} \tag{6.61}$$

$$= \mathcal{I}_F \tag{6.62}$$

where $\mathcal{I}_F$, the Fisher information matrix, has the following expectation:

$$\mathcal{I}_F = E\,\mathcal{H} \tag{6.63}$$
$$= \mathcal{I} \tag{6.64}$$

In contrast, the covariance matrix is

$$\Sigma = \left(\widehat{\mathbf{Y}} - E\,\widehat{\mathbf{Y}}\right)'\left(\widehat{\mathbf{Y}} - E\,\widehat{\mathbf{Y}}\right) \tag{6.65}$$

For $E\,\widehat{\mathbf{Y}} = \mathbf{Y}$ (which occurs when $k \geq \min(m, n)$), because all of the involved entries of Σ are in the compact set $[0, 1]$

$$\|\mathcal{I}_F\|_1 \geq \|\Sigma\|_1 \tag{6.66}$$

and we immediately find

$$\frac{1}{\mathcal{I}_F} \leq \Sigma \tag{6.67}$$

which achieves the Cramer-Rao bound.

6.4.3 Application to the Clustering Problem

In the NMF context, the cardinality of $\mathcal{Y}$ is the dimension of the space span of factorization of (4.22) (value of k). They are latent variables, which in turn are clusters [57, p. 439]. Also, the transformation of $\mathbf{X}$ to $\widetilde{\mathbf{Y}}$ discussed in section 4.1 leads to the probabilistic image for the data matrix, and it makes suitable the use of the loss function (6.5). Next, we justify that (6.15) is a loss function equivalent to (6.5).

The approach developed in section 6.3, functional $z = f(\mathbf{W}\,\mathbf{H})$ (where $f(\cdot)$ is the trace of the involved matrix product) depends on k, and since

$z \geq z_k$, as has been demonstrated, it takes non-negative values. Then, by taking logarithms,

$$\log \varphi(z_k) = z \log\, z - z \log\, z_k \tag{6.68}$$

So, Formula (6.15) can be thought as

$$L = \exp\left(D_{KL}(z \| \, z_k)\right) \tag{6.69}$$

From this point on, the procedure we follow involves finding the expectation of the limit of sequence (6.15), and with no statistical assumptions, the result is a *pdf* depending on k. This density represents the credibility of the cardinality of K, or the number of clusters. Taking into account that only positive integers make sense in the clustering validation problem, Formula (6.38) must be re-written as

$$\varphi(x; \alpha, \lambda) = \mathbf{J}' \, \frac{1}{\Gamma(\alpha)} \lambda^{\alpha} x^{\alpha-1} e^{-\lambda x} \qquad (\alpha = p + 1) \tag{6.70}$$

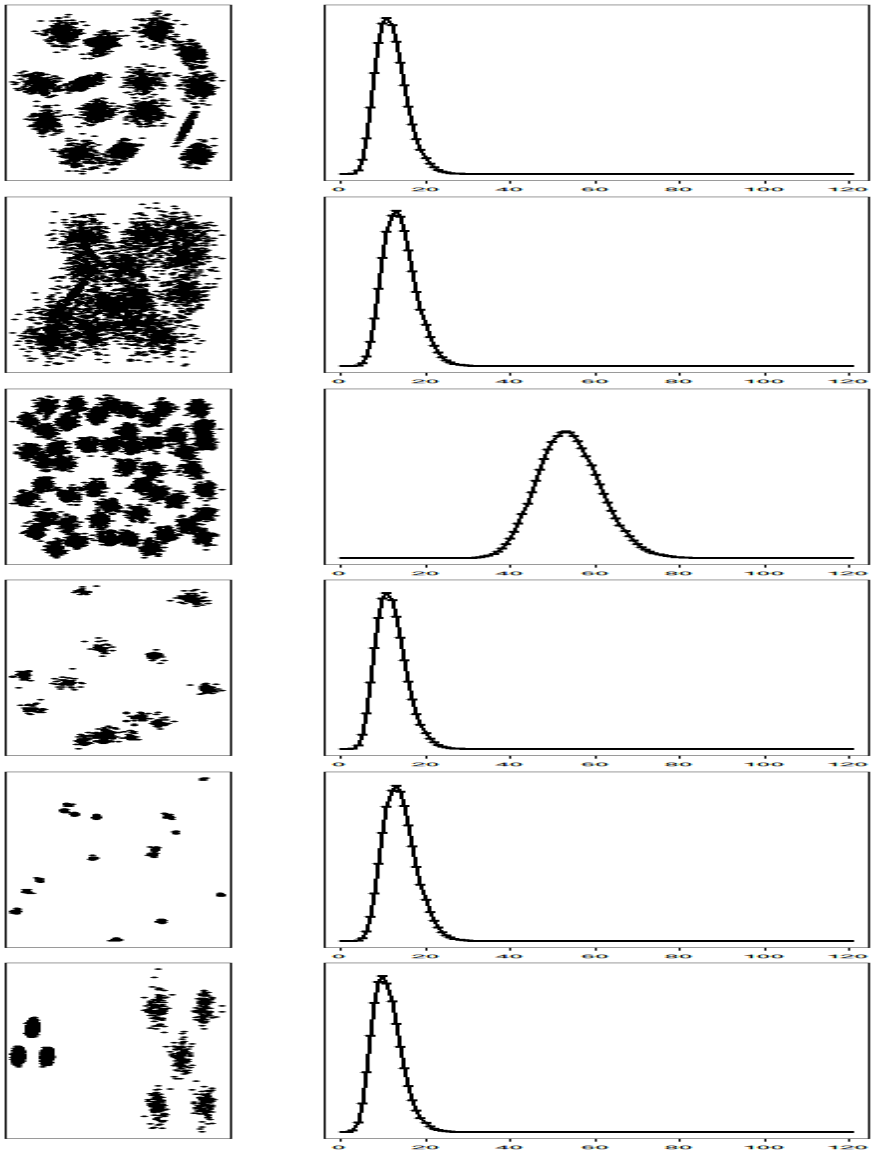
being $\mathbf{J} = (1, 1, \dots)'$ a suitable one-column ones matrix.

This result is the application of the construction in section 6.3 to the clustering validation problem, and that is: cardinality (number of clusters) follows a gamma *pdf*.

6.5 Synthetic Examples

We selected several synthetic datasets created by Franti [101] to study multiple validation criteria (available at $http://cs.joensuu.fi/sipu/datasets/$). These datasets were generated by a certain number of distributions, in which the overlaps and numbers of generating distributions vary. The datasets include four types of artificial data configurations in which the overlapping effect increases (datasets $a1$, $a3$), the number of clusters varies (dataset $s3$), and the dimension varies (datasets $dim032$, $dim512$), and an *unbalance* dataset to examine the effect of different numbers of observations per cluster.

Figure 6.2: Studied Data Sets Densities.



Reproduced from [93]. The left panel of each data set is the representation of the points in the XY plane. In the case of datasets, *dim* is the biplot (first two dimensions representation). Solid lines are the density of the number of clusters given by Formula (6.38). From top to bottom datasets a1, a3, s3, dim32, dim512, and inb. (unbalanced). Parameters are in Table 6.1.

Table 6.1: Parameters for Gamma Density.

	a2	a3	s3	d32	d512	unb.
α	10	13	53	11	13	10
λ	.909	.929	.981	.917	.929	.909

To proceed, several parameters involved in the reliability of the estimation are required. It is necessary to reproduce the shape, especially the maximum, and a sufficient number of values for the queue, which is related to the value of the ancillary statistic (k). The computational cost to obtain a reliable approximation of (4.22) is high; thus, for (6.15). An alternative approach is to relax the approximation of (4.27) and re-compute the terms with q random re-initialization of matrices in formulas of table 4.2 (otherwise, the re-estimation makes no sense). Then, by selecting a statistic $z^{\star}$ such that $E\, z^{\star} = z$, we obtain an estimate sequence. Overall results are shown in Figure 6.2 and parameters for gamma densities in Table 6.1.

Algorithm 3: Sequence of Traces and Parameters.

Input: Data Matrix $\mathbf{X}$; Approximation condition ϵ; Number of model components k; Number of re-estimations q

Data Parameters:

1: Dimension of $\mathbf{X}$: m and n;
2: Transform $\mathbf{X}$ to $\mathbf{Y}$;
3: Compute trace of $\mathbf{Y}$;
4: Obtain a full range matrix $\mathbf{Y}_r$ from $\mathbf{Y}$;
4: Obtain n_r $\big($number of columns of matrix $\mathbf{Y}_r\big)$;
5: Compute determinant of correlation matrix $dC = \mathrm{cor}\big(\mathbf{Y}_r\big)$;
6: Determine $\max(oV)$ $\big($maximum overlapping of columns of $\mathbf{Y}_r\big)$;
7: **Define Variables** z (vector containing estimations $z_1, z_2, \ldots, z_K$);
$\mathbf{Z}$ (matrix of dimension $q \times K$ containing q estimations of vector z);

for *each* $q = 1, \cdots, q$ $\big($re-estimations$\big)$ **do**
  for *each* $k = 1, \cdots, K$ $\big($*construction of sequence*$\big)$ **do**
    1: Random initialize $\mathbf{W}$ of dimension $m \times k$ and $\mathbf{H}$ of dimension $k \times n$;
    2: Factorize $\mathbf{Y}$ obtaining $\mathbf{W}$ and $\mathbf{H}$
    3: Compute the trace $z_k = \mathrm{tr}\mathbf{WH}$;
  in each row of $\mathbf{Z}$ put z_k

Output: $\mathbf{Z}$; n_r; dC; oV

Also, a major hypothesis is linear independence of components, being necessary to extract from the data a matrix accomplishing this.

We selected the same parameters for each case: $k = 120$, $q = 25$, no condition on the degree of approximation, and $p = 3$ iterations in the process of switching equations of table 4.2. To determine the rank of the matrix, we take the number of relevant eigenvalues of the matrix $\mathbf{Y}$ using the *condition number* (the quotient of the largest involved eigenvalue), fixing it to 10. We divide the process into three phases to clarify the effect of the parameters. In these phases, we (i) obtain the sequence of traces, (ii) evaluate the limit sequence, and (iii) adjust the gamma density.

The first step is to obtain matrices $\mathbf{W}$ and $\mathbf{H}$ with random initialization, varying k from one to the ancillary statistic in increasing order. For each of these matrices, the sequence (6.10) is computed. This process is repeated $q = 25$ times, with the results given in a matrix. In this phase, parameters obtained from matrix $\mathbf{Y}$ are also obtained, including the matrix dimension, the matrix rank, the value of z given by (6.11), the correlation matrix, and the degree of overlapping. To determine the overlapping between *iid* densities, we use the R *overlapping* package [193]. This package smooths empirical distributions, for which we have selected a Gaussian kernel and the default parameter h ($h = 1.06$). For cases in which the overlap between variables is greater than 0.85, we correct the effect by multiplying $c_r J$ by the number of overlap variables. These remarks are summarized in algorithm 3.

The next step involves selecting an estimator for the matrix containing the trace sequences. The estimator is calculated for each value of k. When the estimator is computed, we handle the results by using the local likelihood [223], with the help of the *sm R* package [36].

Finally, it is necessary to determine the parameters of the pdf that provide the gamma density. This step is simple and requires estimating the maximum as explained in section 6.3.3 and the value of the parameter λ, shown in the algorithm 5.

To apply the previous results to the validation clustering problem, one must interpret regions or intervals taken on the density as credibility in the cluster assignment. Obtaining this assignment involves estimating the sequence (6.25) on the support $(x_1, x_2, \dots)$ such that each one of these points is related to the corresponding component k by the factor scale c given in (6.25). At this point, it is necessary to consider that (6.7) has an upper

Algorithm 4: Expectation of Traces Sequence Limit.

Input: $\mathbf{Z}$; linear independent components n_r; overlapping oV
for *each column of* $\mathbf{Z}$ **do**
  $\mathbf{s} = \big(1/q\big) \sum_k \mathbf{Z}$ (vector of re-estimation means)
support$\big(z/(1-z)\big)\sqrt{mn_r}\,(0, 1, \ldots, K)$
if Correlation exists and ovelap exists **then**
  support$=$ n_r support
else
  support$=$ $n_r$support
1: Define $\mathbf{s}(0) = 0$
2: Assign pairs $S = (\text{support}, \mathbf{s})$
3: $S = \text{normalize}(S)$ $\quad \big(\text{as } S = S/\sum S\big)$
Output: S

Algorithm 5: Gamma pdf Parameters

Input: S
  1: $\alpha = \max(S) + 1$
  2: $\lambda = 1 - 1/\alpha$
Output: α,λ $\big($parameters of gamma *pdf*$\big)$

bound of 1 and to ensure that the solution is appropriate in the space of positive integers. Here, it is necessary to re-scale as $c_r = \frac{z}{1-z}\sqrt{mn_r}$, where n_r is the rank of matrix $\mathbf{Y}$.

$$\kappa_{opt} = \max_{x \in \mathcal{Z}_+ \cup \{0\}} \varphi(x; \alpha, \lambda) \tag{6.71}$$

summarizing the comparative in Table 6.2.

Credible regions are defined in many books on Bayesian statistics as follows:

$$\int_C \varphi(x; \alpha, \lambda) = 1 - \beta \tag{6.72}$$

where C is a compact region of non-negative real numbers and β is a probability.

Table 6.2: Validation Methods Comparative.

Data Set	m	n	D	silhouette	gap	gamma
a2	5000	2	15	18	18	10
a3	5000	2	15	14	17	13
s3	7500	2	50	57	57	53
dim032	1024	32	16	17	69	11
dim512	1024	512	16	18	28	13
unbalance	6500	2	8	5	15	10

D is the number of clusters that generates the dataset (provided by Franti in [101]). m and n refer to the size of the data matrix (rows and columns). Values for the silhouette index and gap statistic are given in the respective columns. Results for the maxima of relation (6.38) (gamma density) are also provided.

Also, hypothesis tests can be done constructing the classical acceptance regions as

$$\mathcal{R}(x) = \frac{\int_{\mathcal{C}} f(x;\alpha,\lambda)d\theta}{\int_{\theta\leq\theta_0} f(x;\alpha,\lambda)d\theta} \tag{6.73}$$

Chapter 7

Conclusions

The results explained in chapters 4, 5, and 6 could be synthesized in an axiomatic development. The start point would be the probabilistic image of the data matrix $\mathbf{X}$, which is $\mathbf{Y}$, constructed according section 4.1. The NMF can be obtained in a more general context with the Bregman divergence as the objective function [1]. This framework allows solutions with the desired properties, depending on f of (4.28) or (4.29). Furthermore, for the case of $k \geq \min(m, n)$, the convergence of Formula $\mathbf{WH} \to \mathbf{Y}$ is *almost sure* (it is a consequence of the convergence explained in section 6.3). In this case, $P\big(|\widetilde{\mathbf{Y}}| - \mathbf{Y}| < \epsilon\big) = 1$ for any $\epsilon > 0$ as n grows. Then, the theorem 6 is immediate. Furthermore, it has been shown (without mentioning it explicitly) that (4.83) is a Poisson.

For the case of maximum likelihood, solutions are sought (and obtained by imposing for f of (4.29) the form (4.33)), the relation (5.25) is a Fisher kernel. The consistency can be reformulated by showing that the classification error is arbitrarily small. Furthermore, the construction of an SVM has a flat structure for margins. Another consequence of the posterior, contextualized in a ML problem, requires writing the isomorphic relation 6.15 of the KL

[1] Some authors point out that that α-divergences can be obtained from Bregman divergence [4, 218], while others relates the class of β divergences [122], which takes the form [57, p. 113]

$$D_\beta(p\|\, q) = \sum_i \Big(p_i \frac{p_i^\beta - q_i^\beta}{\beta} - \frac{p_i^{\beta+1} - q_i^{\beta+1}}{\beta + 1}\Big)$$

for $\alpha \neq 0$ or -1.

divergence, allowing the inferential estimation of the number of clusters. Another consequence according to Bayesian results (the posterior of a Poisson is a gamma), has been contextualized within the ML problem, and opening the door to inferential validation.

Also, NMF with suitable normalization conditions does not make any hypothesis on the nature of the parameters, and is therefore non-parametric, with sufficient parameter $\mathbf{W}$ to a base $\mathbf{H}$. At this point, it is necessary to take into account that a density or a distribution function is a sufficient not minimal statistic. Broader interpretations (from an information theory point of view) indicate that information is not lost.

Axiomatic conceptualization would reduce the length of the exposition, and also making more clear in what our developed methods are explainable. However, we have chosen to contextualize the problems according to the intellectual development of our ideas.

7.1 Open Questions

Our contributions are not without problems. On the one hand, the matrix interpretation of the relation (4.19) presents a problem in that the diagonal matrix $\mathbf{D}_N$ of (4.21) is not unique, neither a full solution to this problem, which is related to the choice of the number of components in a mixture. Furthermore, it is necessary to examine the case $k \leq \min(m, n)$, which would place our developments on the low-rank case. Currently, we are working on this issue.

The remaining open questions are

(i) An unsatisfactory relation of the covariance matrix of the original data space and the probabilistic image space. The same occurs for the correlation matrix.

(ii) We have not minimized the variance of the proposed gamma distribution. In this sense, it seems that this distribution is a beta, but at the moment, we have not been able to obtain a formal derivation.

(iii) The interpretation of our kernel for some class of equations in the time domain.

In considering these problems, we have also realized that our probabilistic transformations place us in the Riemann manifolds, in which the evaluation of the (dis)similarity of sufficient statistics puts us in the field of information geometry, where the similarities in the space of the parameters for a Riemann metric are evaluated.

7.2 Achievements

Although we have opened more questions than we have closed, we have achieved some results:

(i) Established the convergence conditions for the probabilistic image of the multivariate data matrix and its NMF. Furthermore, we have shown that this convergence is almost sure.

(ii) An error bound for NMF in the qualitative matrix approximation sense.

(iii) The distributional sense of the Fisher information matrix can be obtained with kernelization techniques, generalizing the result for several divergences.

(iv) A kernel with arbitrary misclassification error.

(v) The asymptotic behavior of SVM margins when binding an NMF kernel.

(vi) A statistic from traces of the NMF. The expectation of the limit distribution of this statistic follows a gamma distribution. This statistic lets us estimate the credibility of partitions in clustering problems.

(vii) The convergence conditions of the probabilistic image of the data matrix allow us to situate ourselves within the framework of a well-defined problem. Previous works by several authors have indicated that this convergence is to a local optimum or an open interval. Obtaining this result has allowed us to examine more closely the behavior of Theorem 6 for the NMF case.

We recall that the result of (6.38) can also be obtained by normalizing as

Table 7.1: Algorithms Description.

Algorithm	Description
Learned Basis	Computes a stable NMF basis (pg. 91).
Classification error	Module of confusion matrix misclassification (pg. 91).
Sequence of Traces and Parameters	NMF sequence of traces (pg. 111).
Expectation of Traces Sequence Limit	Posterior of NMF sequence of traces (pg. 113).
Gamma Probability Density Parameters	Computes a stable NMF basis (pg. 113).

Algorithms developed to perform numerical experiments.

$$f(x;\alpha,\lambda) = \frac{f(x;\alpha,\lambda)}{\int_0^{+\infty}\int_0^1 f(x;\alpha,\lambda(d\lambda\,dx} = \frac{1}{\Gamma(\alpha)}\lambda^\alpha x^{\alpha-1}e^{-\lambda x}$$

obtaining for $\alpha = r = p+1$

$$f(x;\alpha,\lambda) = \frac{1}{\Gamma(\alpha)}\lambda^\alpha x^{\alpha-1}e^{-\lambda x}$$

In this case, the result does not differ, but is less general. Since we stated that the gamma pdf is the expectation of the limit density, normalization as done previously is merely the limit density. This nuance has no practical implications, only theoretical.

Also, the algorithm we developed to factorize is faster than any currently existing, as we show in Table 7.1.

7.3 Algorithms

To develop our work, the algorithms listed in Table 7.2 have been developed.

Table 7.2: Computational Speed.

	Time Cycle (ms)	Memory usage (MB)
PLSA	14,379	132.0
NMF-PLSA	1,350	106.6

Reproduced from [91]. Computations performed using an Intel 2.30 GHz processor with 8.00 GB RAM (time for 10^4 iterations). NMF-PLSA corresponds to the proposed extraction diagonal method of Chapter 4.

These algorithms were not the main objective of the Thesis, but were developed for numerical experiments. In the development of the work that constitutes this Thesis, we have paid special attention to the consistency of theoretical results in practical situations. In the case of densities adjustment for clustering validation, a more in-depth study regarding the overlap is important, the adjusted dimensionality seems to be low. This could be for two reasons. The first would be that the number of distributions that generates the data is smaller in such cases. The other would be the inability of the proposed method when overlap exists. We lean towards the first option.

The peculiarities of the NMF mean the proposed validation method can also be used as a classification method. This would be the classical use of this technique for clustering. In this case, for the arrangements observations and characteristics, the columns of $\mathbf{W}$ of (6.13) would correspond to the classification in the k clusters and they are placed in the columns. This case corresponds to soft-clustering. From this result, a hard-clustering could be obtained by simply introducing a Bayes classifier, as proposed by Ding [75].

We note that equivalence is not complete, since the concept of a geometric center for the k-means method is diluted when probabilistic concepts are used. The introduction of the Bayes classifier does not provide any geometric center.

7.4 Current Paradigm Contextualization

In addition to the open questions, the current paradigm in ML requires the processing of large amounts of data. The iterative nature of the methods

used and the slowness to converge seems to limit the applicability in Big-Data environments. To this, we add that the non-parametric nature of our solutions further complicates this issue. Currently, we think that to maintain the desired properties of our solutions, we should confine the problem to parametric approaches, and simplifying estimations of densities. This path should alleviate this problem. This approach forces us to consider the parametric case as an approximation, and this is in opposition to the orthodox positions.

At the moment of closing this Thesis, we are aware that the parameterization with densities provides a transformation in an invariant space, placing our results in the field of Information Geometry with a non-parametric development.

Bibliography

[1] Ackermann, Marcel R, Blömer, Johannes, and Sohler, Christian. Clustering for metric and nonmetric distance measures. *ACM Transactions on Algorithms (TALG)*, 6(4):1–26, 2010.

[2] Aggarwal, Charu C. *Algorithms and Applications*. CRC Press Taylor and Francis Group, 2014.

[3] Amari, Shun-Ichi. Information geometry of the em and em algorithms for neural networks. *Neural networks*, 8(9):1379–1408, 1995.

[4] Amari, Shun-Ichi. alpha-divergence is unique, belonging to both f-divergence and bregman divergence classes. *IEEE Transactions on Information Theory*, 55(11):4925–4931, 2009.

[5] Amari, Shun-ichi. *Information geometry and its applications*, volume 194. Springer, 2016.

[6] Golub, Gene H and Reinsch, Christian. Singular value decomposition and least squares solutions. In *Linear algebra*, pages 134–151. Springer, 1971.

[7] Anderson, Edward, Bai, Zhaojun, Bischof, Christian, Blackford, L Susan, Demmel, James, Dongarra, Jack, Du Croz, Jeremy, Greenbaum, Anne, Hammarling, Sven, McKenney, Alan, et al. *LAPACK Users' guide*. SIAM, 1999.

[8] Archambeau, Cédric, Lee, John Aldo, Verleysen, Michel, et al. On convergence problems of the em algorithm for finite gaussian mixtures. In *ESANN*, volume 3, pages 99–106, 2003.

[9] Aronszajn, Nachman. Theory of reproducing kernels. *Transactions of the American mathematical society*, 68(3):337–404, 1950.

[10] Asanovic, Krste, Bodik, Ras, Catanzaro, Bryan Christopher, Gebis, Joseph James, Husbands, Parry, Keutzer, Kurt, Patterson, David A, Plishker, William Lester, Shalf, John, Williams, Samuel Webb, et al. The landscape of parallel computing research: A view from berkeley, 2006.

[11] Ash, Robert B. *Information theory*. Dover Publications, 1990.

[12] Bai, Chong-En, Hsieh, Chang-Tai, and Song, Zheng Michael. Crony capitalism with chinese characteristics. *University of Chicago, working paper*, pages 39–58, 2014.

[13] Balakrishnan, Narayanaswamy and Nevzorov, Valery B. *A primer on statistical distributions*. John Wiley & Sons, 2004.

[14] Balažević, Ivana, Allen, Carl, and Hospedales, Timothy M. Tucker: Tensor factorization for knowledge graph completion. *pre-print*, 2019.

[15] Banerjee, Arindam, Merugu, Srujana, Dhillon, Inderjit S, Ghosh, Joydeep, and Lafferty, John. Clustering with bregman divergences. *Journal of machine learning research*, 6(10), 2005.

[16] Bardsley, Johnathan M and Vogel, Curtis R. A nonnegatively constrained convex programming method for image reconstruction. *SIAM Journal on Scientific Computing*, 25(4):1326–1343, 2004.

[17] Bassiou, Nikoletta and Kotropoulos, Constantine. Rplsa: A novel updating scheme for probabilistic latent semantic analysis. *Computer Speech & Language*, 25(4):741–760, 2011.

[18] Belkin, Mikhail, Hsu, Daniel, Ma, Siyuan, and Mandal, Soumik. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

[19] Beltrami, Eugenio. Sulle funzioni bilineari. *Giornale di Matematiche ad Uso degli Studenti Delle Universita*, 11(2):98–106, 1873.

[20] Ben-Israel and Cem Iyigun. Probabilistic d-clustering. *Journal of Classification*, 25, 2011.

[21] Ben-Tal, Aharon and Nemirovski, Arkadi. Robust convex optimization. *Mathematics of operations research*, 23(4):769–805, 1998.

[22] Berg, Christian, Christensen, Jens Peter Reus, and Ressel, Paul. *Harmonic analysis on semigroups: theory of positive definite and related functions*, volume 100. Springer, 1984.

[23] José M Bernardo and Adrian FM Smith. *Bayesian theory*, volume 405. John Wiley & Sons, 2009.

[24] Bholowalia, Purnima and Kumar, Arvind. Ebk-means: A clustering technique based on elbow method and k-means in wsn. *International Journal of Computer Applications*, 105(9), 2014.

[25] Barry David Biletch, Hongji Yu, and Kathleen Rose Kay. An analysis of mathematical notations: for better or for worse, 2015.

[26] Birkhoff, Garrett and Varga, Richard S. Reactor criticality and nonnegative matrices. *Journal of the Society for Industrial and Applied Mathematics*, 6(4):354–377, 1958.

[27] Christopher M Bishop. Bayesian pca. *Advances in neural information processing systems*, pages 382–388, 1999.

[28] Blei, David M, Griffiths, Thomas L, Jordan, Michael I, Tenenbaum, Joshua B, et al. Hierarchical topic models and the nested chinese restaurant process. In *NIPS*, volume 16, 2003.

[29] Blei, David M and Lafferty, John D. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120, 2006.

[30] Blei, David M., Ng, Andrew Y., and Jordan, Michael I. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[31] Bosch, Anna, Zisserman, Andrew, and Muñoz, Xavier. Scene classification via plsa. In *European conference on computer vision*, pages 517–530. Springer, 2006.

[32] Boser, Bernhard E, Guyon, Isabelle M, and Vapnik, Vladimir N. A training algorithm for optimal margin classifiers. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 144–152, 1992.

[33] Bottou, Léon et al. Online learning and stochastic approximations. *On-line learning in neural networks*, 17(9):142, 1998.

[34] Boutsidis, Christos, Drineas, Petros, and Magdon-Ismail, Malik. Near-optimal column-based matrix reconstruction. *SIAM Journal on Computing*, 43(2):687–717, 2014.

[35] Boutsidis, Christos and Gallopoulos, Efstratios. Svd based initialization: A head start for nonnegative matrix factorization. *Pattern recognition*, 41(4):1350–1362, 2008.

[36] Bowman, A. W. and Azzalini, A. *R package* `sm`*: nonparametric smoothing methods (version 2.2-5.6)*. University of Glasgow, UK and Università di Padova, Italia, 2018.

[37] Bowman, Adrian W and Azzalini, Adelchi. *Applied smoothing techniques for data analysis: the kernel approach with S-Plus illustrations*, volume 18. OUP Oxford, 1997.

[38] Boyles, Russell A. On the convergence of the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 45(1):47–50, 1983.

[39] Bozinovski, Stevo. Reminder of the first paper on transfer learning in neural networks, 1976. *Informatica*, 44(3), 2020.

[40] Thorsten Brants. Test data likelihood for plsa models. *Information Retrieval*, 8(2):181–196, 2005.

[41] Brants, T.H., Tsochantaridis, I., Hofmann, T., and Chen, F.R. Computer controlled method for performing incremental probabilistic latent semantic analysis of documents, involves performing incremental addition of new term to trained probabilistic latent semantic analysis model, 2006. US Patent Number US2006112128-A1.

[42] Brants, Thorsten, Chen, Francine, and Tsochantaridis, Ioannis. Topic-based document segmentation with probabilistic latent semantic analysis. In *Proceedings of the eleventh international conference on Information and knowledge management*, pages 211–218, 2002.

[43] Bredensteiner, Erin J. and Bennett, Kristin P. Multicategory classification by support vector machines. In *Computational Optimization*, pages 53–79. Springer, 1999.

[44] Bregman, Lev M. The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming. *USSR computational mathematics and mathematical physics*, 7(3):200–217, 1967.

[45] Brualdi, Richard A, Parter, Seymour V, and Schneider, Hans. The diagonal equivalence of a nonnegative matrix to a stochastic matrix. *Journal of Mathematical Analysis and Applications*, 16(1):31–50, 1966.

[46] Mario Bunge. *Matter and mind: A philosophical inquiry*, volume 287. Springer Science & Business Media, 2010.

[47] Florian Cajori. *A history of mathematical notations*, volume 1. Courier Corporation, 1993.

[48] Caroline Uhler. Geometry of maximum likelihood estimation in gaussian graphical models. *The Annals of Statistics*, 40(1), feb 2012.

[49] Casella, George and Berger, Roger L. *Statistical inference*. Cengage Learning, 2007.

[50] Arthur Cayley. Remarques sur la notation des fonctions algébriques., 1855.

[51] Chang, Jia-Ming, Su, Emily Chia-Yu, Lo, Allan, Chiu, Hua-Sheng, Sung, Ting-Yi, and Hsu, Wen-Lian. Psldoc: Protein subcellular localization prediction based on gapped-dipeptides and probabilistic latent semantic analysis. *Proteins: Structure, Function, and Bioinformatics*, 72(2):693–710, 2008.

[52] Chappelier, Jean-Cédric and Eckard, Emmanuel. Plsi: The true fisher kernel and beyond. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 195–210. Springer, 2009.

[53] Chaudhuri, A. R. and Murty, M. N. On the relation between k-means and plsa. *2012 21st International Conference on Pattern Recognition*, 2012.

[54] Chen, Ji-Cheng. The nonnegative rank factorizations of nonnegative matrices. *Linear algebra and its applications*, 62:207–217, 1984.

[55] Chen, Tianping, Amari, Shun Ichi, and Lin, Qin. A unified algorithm for principal and minor components extraction. *Neural networks*, 11(3):385–390, 1998.

[56] Cho, Hyuk, Dhillon, Inderjit S, Guan, Yuqiang, and Sra, Suvrit. Minimum sum-squared residue co-clustering of gene expression data. In *Proceedings of the 2004 SIAM international conference on data mining*, pages 114–125. SIAM, 2004.

[57] Cichocki, A., Zdunek, R., Phan, A.H., and Amary, S.I. *Nonnegative Matrix and Tensor Factorizations*. John Willey and Sons Ltd, 2009.

[58] Cichocki, Andrzej and Georgiev, Pando. Blind source separation algorithms with matrix constraints. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, 86(3):522–531, 2003.

[59] Cichocki, Andrzej, Phan, Anh Huy, and Caiafa, Cesar. Flexible hals algorithms for sparse non-negative matrix/tensor factorization. In *2008 IEEE Workshop on machine learning for signal processing*, pages 73–78. IEEE, 2008.

[60] Cichocki, Andrzej and Zdunek, Rafal. Regularized alternating least squares algorithms for non-negative matrix/tensor factorization. *Lecture Notes in Computer Science*, 4493:793, 2007.

[61] Cichocki, Andrzej, Zdunek, Rafal, and Amari, Shun-ichi. Csiszár's divergences for non-negative matrix factorization: Family of new algorithms. *ICA'06*, page 32–39, 2006.

[62] Cichocki, Andrzej, Zdunek, Rafal, and Amari, Shun-ichi. Hierarchical als algorithms for nonnegative matrix and 3d tensor factorization. In *Independent Component Analysis and Signal Separation: 7th International Conference, ICA 2007, London, UK, September 9-12, 2007. Proceedings 7*, pages 169–176. Springer, 2007.

[63] Cohn, David A, Ghahramani, Zoubin, and Jordan, Michael I. Active learning with statistical models. *Journal of artificial intelligence research*, 4:129–145, 1996.

[64] Cortes, C. and Vapnik, V. Support-vector networks. *Machine Learning*, 20:273–297, 1995.

[65] Alfredo Cuzzocrea, Pau Figuera, Mojtaba Hajian, and Pablo García Bringas. A theoretical framework for supporting clustering validation via non-negative-matrix-factorization trace sequences over probabilistic spaces. In *2023 IEEE Intl Conf on Dependable, Autonomic and Secure Computing, Intl Conf on Pervasive Intelligence and Computing, Intl Conf on Cloud and Big Data Computing, Intl Conf on Cyber Science and Technology Congress (DASC/PiCom/CBDCom/CyberSciTech)*, pages 1080–1087. IEEE, 2023.

[66] Dahlquist, Germund and Björck, Åke. *Numerical methods*. Courier Corporation, 2003.

[67] Dean, Jeffrey and Ghemawat, Sanjay. Mapreduce: simplified data processing on large clusters. *Communications of the ACM*, 51(1):107–113, 2008.

[68] Deerwester, Scott, Dumais, Susan T, Furnas, George W, Landauer, Thomas K, and Harshman, Richard. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.

[69] Dempster, A.P., Laird, N.M., and Rubin, D.B. Maximum likelihood from incomplete data via the em agorithm. *Journal of the Royal Statistical Society, Series B, Methodological*, 1977.

[70] Deng, Hongbo and Han, Jiawei. Probabilistic models for clustering. In *Data Clustering*, pages 61–86. Chapman and Hall/CRC, 2018.

[71] Devarajan, K., Wang, G.L., and Ebrahimi, N. A unified statistical approach to non-negative matrix factorization and probabilistic latent semantic indexing. *Machine Learning*, 2015.

[72] Dhillon, Inderjit S and Tropp, Joel A. Matrix nearness problems with bregman divergences. *SIAM Journal on Matrix Analysis and Applications*, 29(4):1120–1146, 2008.

[73] Ding, Chris, He, Xiaofeng, and Simon, Horst D. On the equivalence of nonnegative matrix factorization and spectral clustering. In *Proceedings of the 2005 SIAM international conference on data mining*, pages 606–610. SIAM, 2005.

[74] Ding, Chris, Li, Tao, and Peng, Wei. Nonnegative matrix factorization and probabilistic latent semantic indexing: Equivalence chi-square statistic, and a hybrid method. In *AAAI*, volume 42, pages 137–43, 2006.

[75] Ding, Chris, Li, Tao, and Peng, Wei. On the equivalence between non-negative matrix factorization and probabilistic latent semantic indexing. *Computational Statistics & Data Analysis*, 52(8):3913–3927, 2008.

[76] Ding, Chris HQ, Li, Tao, and Jordan, Michael I. Convex and semi-nonnegative matrix factorizations. *IEEE transactions on pattern analysis and machine intelligence*, 32(1):45–55, 2008.

[77] Donghwan Kim. *kfda: Kernel Fisher Discriminant Analysis*, 2017. R package version 1.0.0.

[78] Dougherty, Edward R. and Brun, Marcel. A probabilistic theory of clustering. *Pattern Recognition*, 37(5):917–925, 2004.

[79] Dua, Dheeru and Graff, Casey. UCI machine learning repository, 2017.

[80] Durgesh, K Srivastava and Lekha, B. Data classification using support vector machine. *Journal of theoretical and applied information technology*, 12(1):1–7, 2010.

[81] Eckart, Carl and Young, Gale. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.

[82] Eckart, Carl and Young, Gale. A principal axis transformation for non-hermitian matrices. *Bulletin of the American Mathematical Society*, 45(2):118–121, 1939.

[83] Elad, Michael, Matalon, Boaz, and Zibulevsky, Michael. Coordinate and subspace optimization methods for linear least squares with non-quadratic regularization. *Applied and Computational Harmonic Analysis*, 23(3):346–367, 2007.

[84] Elkan, Charles. Deriving tf-idf as a fisher kernel. In *International Symposium on String Processing and Information Retrieval*, pages 295–300. Springer, 2005.

[85] Ester, Martin, Kriegel, Hans-Peter, Sander, Jörg, Xu, Xiaowei, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, 1996.

[86] Fan, Ky. Maximum properties and inequalities for the eigenvalues of completely continuous operators. *Proceedings of the National Academy of Sciences*, 37(11):760–766, 1951.

[87] Fan, Ky and Hoffman, Alan J. Some metric inequalities in the space of matrices. *Proceedings of the American Mathematical Society*, 6(1):111–116, 1955.

[88] Farahat, Ayman and Chen, Francine. Improving probabilistic latent semantic analysis with principal component analysis. In *11th Conference of the European Chapter of the Association for Computational Linguistics*, 2006.

[89] Pau Figuera, Alfredo Cuzzocrea, and Pablo García Bringas. Generalized fisher kernel with bregman divergence. In *Hybrid Artificial Intelligent Systems*, pages 186–194. Springer International Publishing, 2022.

[90] Pau Figuera, Alfredo Cuzzocrea, and Pablo García Bringas. Probability density function for clustering validation. In *Hybrid Artificial Intelligent Systems*, pages 133–144. Springer Nature Switzerland, 2023.

[91] Pau Figuera and Pablo García Bringas. On the probabilistic latent semantic analysis generalization as the singular value decomposition probabilistic image. *Journal of Statistical Theory and Applications*, 19(2):286–296, 2020.

[92] Pau Figuera and Pablo García Bringas. A non-parametric fisher kernel. In *Hybrid Artificial Intelligent Systems*, pages 448–459. Springer International Publishing, 2021.

[93] Pau Figuera and Pablo García Bringas. Non-parametric nonnegative matrix factorization fisher kernel. *Available at SSRN 4585853*, 2023.

[94] Pau Figuera and Pablo García Bringas. Revisiting the probabilistic latent semantic analysis: The method, its extensions and its algorithms. *Preprints*, 2023.

[95] Pau Figuera and Pablo García Bringas. Revisiting probabilistic latent semantic analysis: Extensions, challenges and insights. *Technologies*, 12(1):5, 2024.

[96] Pau Figuera and Pablo García Bringas. Exact classification fisher kernel with non-negative matrix factorization. *International Journal of Approximate Reasoning*, Submited 2023.

[97] Pau Figuera and Pablo García Bringas. On clustering validation inference. *Mathematics*, Submited 2024.

[98] Fine, Shai, Navratil, Jiri, and Gopinath, Ramesh A. A hybrid gmm/svm approach to speaker identification. In *2001 IEEE International Conference on Acoustics, Speech, and Signal Processing. Proceedings (Cat. No. 01CH37221)*, volume 1, pages 417–420. IEEE, 2001.

[99] Franc, Vojtěch, Hlaváč, Václav, and Navara, Mirko. Sequential coordinate-wise algorithm for the non-negative least squares problem. In *Computer Analysis of Images and Patterns: 11th International Conference, CAIP 2005, Versailles, France, September 5-8, 2005. Proceedings 11*, pages 407–414. Springer, 2005.

[100] Franke, Beate, Plante, Jean-François, Roscher, Ribana, Lee, En-shiun Annie, Smyth, Cathal, Hatefi, Armin, Chen, Fuqi, Gil, Einat, Schwing, Alexander, Selvitella, Alessandro, et al. Statistical inference, learning and models in big data. *International Statistical Review*, 84(3):371–389, 2016.

[101] Fränti, Pasi and Sieranoja, Sami. K-means properties on six clustering benchmark datasets. *Applied Intelligence*, 48(12):4743–4759, 2018.

[102] Fred, Ana LN and Jain, Anil K. Cluster validation using a probabilistic attributed graph. In *2008 19th International Conference on Pattern Recognition*, pages 1–4. IEEE, 2008.

[103] Garrod, Claude and Percus, Jerome K. Reduction of the n-particle variational problem. *Journal of Mathematical Physics*, 5(12):1756–1776, 1964.

[104] Gaussier, E. and Goutte, C. Relation between plsa and nmf and implications. *In Proceedings 28th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR'05)*, 2005.

[105] Ghojogh, Benyamin, Ghodsi, Ali, Karray, Fakhri, and Crowley, Mark. Reproducing kernel hilbert space, mercer's theorem, eigenfunctions, nystr\" om method, and use of kernels in machine learning: Tutorial and survey. *arXiv preprint arXiv:2106.08443*, 2021.

[106] Girolami, M. and Kabøn, A. On an equivalence between plsi and lda. *SIGIR 03 Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval*, 2003.

[107] Gong, Pinghua and Zhang, Changshui. Efficient nonnegative matrix factorization via projected newton method. *Pattern Recognition*, 45(9):3557–3565, 2012.

[108] Goreinov, Sergei A, Tyrtyshnikov, Eugene E, and Zamarashkin, Nickolai L. A theory of pseudoskeleton approximations. *Linear algebra and its applications*, 261(1-3):1–21, 1997.

[109] gpgpu.org. General-purpose computation graphics hardware. https://web.archive.org/web/20051231024709/http://www.gpgpu.org/, 2006.

[110] Grabusts, Peter and Borisov, Arkady. Using grid-clustering methods in data classification. In *Proceedings. International Conference on Parallel Computing in Electrical Engineering*, pages 425–426. IEEE, 2002.

[111] Gu, Ming. Subspace iteration randomization and singular value problems. *SIAM Journal on Scientific Computing*, 37(3):A1139–A1173, 2015.

[112] Guan, Naiyang, Tao, Dacheng, Luo, Zhigang, and Yuan, Bo. Nenmf: An optimal gradient method for nonnegative matrix factorization. *IEEE Transactions on Signal Processing*, 60(6):2882–2898, 2012.

[113] Gudovskiy, Denis, Hodgkinson, Alec, Yamaguchi, Takuya, and Tsukizawa, Sotaro. Deep active learning for biased datasets via fisher kernel self-supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9041–9049, 2020.

[114] Gupta, Mithun Das. Additive non-negative matrix factorization for missing data. *pre-print*, 2010.

[115] Halko, Nathan, Martinsson, Per-Gunnar, and Tropp, Joel A. Finding structure with randomness: Probabilistic algorithms for constructing approximate matrix decompositions. *SIAM review*, 53(2):217–288, 2011.

[116] Hamid Usefi. Clustering, multicollinearity, and singular vectors. *Computational Statistics & Data Analysis*, 173:107523, 2022.

[117] Har-Even, Michael and Brailovsky, Victor L. Probabilistic validation approach for clustering. *Pattern Recognition Letters*, 16(11):1189–1196, 1995.

[118] Harshman, Richard A et al. Foundations of the parafac procedure: Models and conditions for an explanatory multimodal factor analysis. *University of California at Los Angeles Los Angeles, CA*, 1970.

[119] Hartigan, John A. Direct clustering of a data matrix. *Journal of the american statistical association*, 67(337):123–129, 1972.

[120] Hazan, Tamir, Hardoon, Roee, and Shashua, Amnon. Plsa for sparse arrays with tsallis pseudo-additive divergence: noise robustness and algorithm. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8. IEEE, 2007.

[121] Yuan He, Cheng Wang, and Changjun Jiang. Correlated matrix factorization for recommendation with implicit feedback. *IEEE Transactions on Knowledge and Data Engineering*, 31(3):451–464, 2018.

[122] Hennequin, Romain, David, Bertrand, and Badeau, Roland. Beta-divergence as a subclass of bregman divergence. *IEEE Signal Processing Letters*, 18(2):83–86, 2011.

[123] Hinton, Geoffrey E and Zemel, Richard S. Autoencoders, minimum description length, and helmholtz free energy. *Advances in neural information processing systems*, 6:3–10, 1994.

[124] Ho, Ngoc-Diep and Van Dooren, Paul. Non-negative matrix factorization with fixed row and column sums. *Linear Algebra and its Applications*, 429(5-6):1020–1025, 2008.

[125] Ho, Tin Kam. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):832–844, 1998.

[126] Hofmann, T. Probabilistic latent semantic indexing. *SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, 1999.

[127] Hofmann, Thomas. Probabilistic latent semantic analysis. *Uncertainty in Artificial Intelligence, Prodeedings*, 1999.

[128] Hofmann, Thomas. Learning the similarity of documents: An information-geometric approach to document retrieval and categorization. In *Advances in neural information processing systems*, pages 914–920, 2000.

[129] Hofmann, Thomas. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 2001.

[130] Hofmann, Thomas. Collaborative filtering via gaussian probabilistic latent semantic analysis. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 259–266, 2003.

[131] Hofmann, Thomas, Schölkopf, Bernhard, and Smola, Alexander J. Kernel methods in machine learning. *The annals of statistics*, pages 1171–1220, 2008.

[132] Hong, Chuntao, Chen, Wenguang, Zheng, Weimin, Shan, Jiulong, Chen, Yurong, and Zhang, Yimin. Parallelization and characterization of probabilistic latent semantic analysis. In *2008 37th International Conference on Parallel Processing*, pages 628–635. IEEE, 2008.

[133] Hong, Liangjie. A tutorial on probabilistic latent semantic analysis. *pre-print*, 2012.

[134] Hörster, Eva, Lienhart, Rainer, and Slaney, Malcolm. Continuous visual vocabulary modelsfor plsa-based scene recognition. In *Proceedings of the 2008 international conference on Content-based image and video retrieval*, pages 319–328, 2008.

[135] Hsieh, Chia-Hsin, Huang, Chien-Lin, and Wu, Chung-Hsien. Spoken document summarization using topic-related corpus and semantic dependency grammar. In *2004 International Symposium on Chinese Spoken Language Processing*, pages 333–336. IEEE, 2004.

[136] Hsu, Chih-Wei and Lin, Chih-Jen. A comparison of methods for multiclass support vector machines. *IEEE transactions on Neural Networks*, 13(2):415–425, 2002.

[137] Huber-Carol, Catherine, Balakrishnan, Narayanaswamy, Nikulin, M, and Mesbah, M. *Goodness-of-fit tests and model validity*. Springer Science & Business Media, 2012.

[138] Hwang, Tea-Yuan and Hu, Chin-Yuan. On a characterization of the gamma distribution: The independence of the sample mean and the sample coefficient of variation. *Annals of the Institute of Statistical Mathematics*, 51(4):749–753, 1999.

[139] Inokuchi, Ryo and Miyamoto, Sadaaki. Nonparametric fisher kernel using fuzzy clustering. In *International Conference on Knowledge-Based and Intelligent Information and Engineering Systems*, pages 78–85, 2006.

[140] Jaakkola, Tommi, Diekhans, Mark, and Haussler, David. A discriminative framework for detecting remote protein homologies. *Journal of computational biology*, 7(1-2):95–114, 2000.

[141] Jaakkola, Tommi S., Haussler, David, et al. Exploiting generative models in discriminative classifiers. *Advances in neural information processing systems*, pages 487–493, 1999.

[142] Jain Anil, K. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8, SI):651–666, JUN 1 2010. 19th International Conference on Pattern Recognition (ICPR 2008), Tampa, FL, DEC 08-11, 2008.

[143] Jiang, Yu, Liu, Jing, Li, Zechao, Li, Peng, and Lu, Hanqing. Co-regularized plsa for multi-view clustering. In *Asian Conference on Computer Vision*, pages 202–213. Springer, 2012.

[144] Jin, Yan, Gao, Yang, Shi, Yinghuan, Shang, Lin, Wang, Ruili, and Yang, Yubin. P 2 lsa and p 2 lsa+: Two paralleled probabilistic latent semantic analysis algorithms based on the mapreduce model. In *International Conference on Intelligent Data Engineering and Automated Learning*, pages 385–393. Springer, 2011.

[145] Joe H. Ward Jr. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, 58(301):236–244, 1963.

[146] Jordan, Camille. Mémoire sur les formes bilinéaires. *Journal de mathématiques pures et appliquées*, 19:35–54, 1874.

[147] Kagie, Martijn, Van Der Loos, Matthijs, and Van Wezel, Michiel. Including item characteristics in the probabilistic latent semantic analysis model for collaborative filtering. *Ai Communications*, 22(4):249–265, 2009.

[148] Kanzawa, Yuchi. On tsallis entropy-based and bezdek-type fuzzy latent semantics analysis. In *2018 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 3685–3689. IEEE, 2018.

[149] Kassambara, Alboukadel and Mundt, Fabian. factoextra: Extract and visualize the results of multivariate data analyses, 2019. R package version 1.0.6.

[150] Keyl, Michael. Fundamentals of quantum information theory. *Physics reports*, 369(5):431–548, 2002.

[151] Khuri, André I. *Advanced calculus with applications in statistics*. John Wiley & Sons, 2nd edition, 2003.

[152] Dongsoon Kim and In-Beum Lee. Process monitoring based on probabilistic pca. *Chemometrics and intelligent laboratory systems*, 67(2):109–123, 2003.

[153] Klingenberg, Bradley, Curry, James, and Dougherty, Anne. Non-negative matrix factorization: Ill-posedness and a geometric algorithm. *Pattern Recognition*, 42(5):918–928, 2009.

[154] Kogbetliantz, EG. Solution of linear equations by diagonalization of coefficients matrix. *Quarterly of Applied Mathematics*, 13(2):123–132, 1955.

[155] Kokonendji, CC, Senga Kiesse, Tristan, and Zocchi, Silvio S. Discrete triangular distributions and non-parametric estimation for probability mass function. *Journal of Nonparametric Statistics*, 19(6-8):241–254, 2007.

[156] Kouassi, Eli Koffi, Amagasa, Toshiyuki, and Kitagawa, Hiroyuki. Efficient probabilistic latent semantic indexing using graphics processing unit. *Procedia Computer Science*, 4:382–391, 2011.

[157] Kozhisseri, Shyju and Surov, Ilya A. Quantum-probabilistic svd: complex-valued factorization of matrix data. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 22(3):567–573, 2022.

[158] Krithara, Anastasia and Paliouras, Georgios. Tl-plsa: Transfer learning between domains with different classes. In *2013 IEEE 13th International Conference on Data Mining*, pages 419–427. IEEE, 2013.

[159] Kullback, Solomon and Leibler, Richard A. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[160] Sébastien Lê, Julie Josse, and François Husson. Factominer: an r package for multivariate analysis. *Journal of statistical software*, 25:1–18, 2008.

[161] Lee, Daniel D and Seung, H Sebastian. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401(6755):788–791, 1999.

[162] Lee, Hyekyoung, Cichocki, Andrzej, and Choi, Seungjin. Kernel nonnegative matrix factorization for spectral eeg feature extraction. *Neurocomputing*, 72(13-15):3182–3190, 2009.

[163] Lewicki, Michael S and Sejnowski, Terrence J. Learning overcomplete representations. *Neural computation*, 12(2):337–365, 2000.

[164] Li, Ning, Luo, Wenjuan, Yang, Kun, Zhuang, Fuzhen, He, Qing, and Shi, Zhongzhi. Self-organizing weighted incremental probabilistic latent semantic analysis. *International Journal of Machine Learning and Cybernetics*, 9(12):1987–1998, 2018.

[165] Li, Zhixin, Shi, Zhiping, Liu, Xi, and Shi, Zhongzhi. Modeling continuous visual features for semantic image annotation and retrieval. *Pattern Recognition Letters*, 32(3):516–523, 2011.

[166] Lin, Chih-Jen. Projected gradient methods for nonnegative matrix factorization. *Neural computation*, 19(10):2756–2779, 2007.

[167] London, David. Two inequalities in nonnegative symmetric matrices. *Pacific Journal of Mathematics*, 16(3):515–536, 1966.

[168] MacQueen, James et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, pages 281–297, 1967.

[169] Madsen, Rasmus Elsborg, Larsen, Jan, and Hansen, Lars Kai. Part-of-speech enhanced context recognition. In *Proceedings of the 2004 14th IEEE Signal Processing Society Workshop Machine Learning for Signal Processing, 2004.*, pages 635–643. IEEE, 2004.

[170] Mann, Amandeep Kaur and Kaur, Navneet. Review paper on clustering techniques. *Global Journal of Computer Science and Technology*, 2013.

[171] Martin, Carla D and Porter, Mason A. The extraordinary svd. *The American Mathematical Monthly*, 119(10):838–851, 2012.

[172] Masseroli, Marco, Chicco, Davide, and Pinoli, Pietro. Probabilistic latent semantic analysis for prediction of gene ontology annotations. In *The 2012 international joint conference on neural networks (IJCNN)*, pages 1–8. IEEE, 2012.

[173] McCarty, Charles. Constructivism in mathematics. *Philosophy of Mathematics*, pages 311–343, 2009.

[174] Meng, Xiao-Li and Van Dyk, David. The em algorithm—an old folk-song sung to a fast new tune. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 59(3):511–567, 1997.

[175] Mercer, James. Xvi. functions of positive and negative type, and their connection the theory of integral equations. *Philosophical transactions of the royal society of London. Series A, containing papers of a mathematical or physical character*, 209(441-458):415–446, 1909.

[176] Meyer, David, Dimitriadou, Evgenia, Hornik, Kurt, Weingessel, Andreas, and Friedrich Leisch. *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071), TU Wien*, 2021. R package version 1.7-6.

[177] Mika, Dariusz, Budzik, Grzegorz, and Jozwik, Jerzy. Single channel source separation with ica-based time-frequency decomposition. *Sensors*, 20(7):2019, 2020.

[178] Mirsky, L. Introduction to linear algebra, 1965.

[179] Mirsky, Leon. Symmetric gauge functions and unitarily invariant norms. *The quarterly journal of mathematics*, 11(1):50–59, 1960.

[180] Mitchell, Tom Michael. *The discipline of machine learning*, volume 9. Carnegie Mellon University, School of Computer Science, 2006.

[181] Mnih, Andriy and Salakhutdinov, Russ R. Probabilistic matrix factorization. *Advances in neural information processing systems*, 20:1257–1264, 2007.

[182] Murtagh, Fionn and Contreras, Pedro. Methods of hierarchical clustering. *arXiv preprint arXiv:1105.0121*, 2011.

[183] Naik, Ganesh R. *Non-negative matrix factorization techniques*. Springer, 2016.

[184] Neal, Radford M and Hinton, Geoffrey E. A view of the em algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models*, pages 355–368. Springer, 1998.

[185] Ng, Andrew, Jordan, Michael, and Weiss, Yair. On spectral clustering: Analysis and an algorithm. *Advances in neural information processing systems*, 14, 2001.

[186] Ng, Raymond T. and Han, Jiawei. Clarans: A method for clustering objects for spatial data mining. *IEEE transactions on knowledge and data engineering*, 14(5):1003–1016, 2002.

[187] Niu, Lingfeng and Shi, Yong. Semi-supervised plsa for document clustering. In *2010 IEEE International Conference on Data Mining Workshops*, pages 1196–1203. IEEE, 2010.

[188] Oja, Erkki. Principal components, minor components, and linear neural networks. *Neural networks*, 5(6):927–935, 1992.

[189] Olivares, J, Sarro, LM, Bouy, H, Miret-Roig, N, Casamiquela, L, Galli, PAB, Berihuete, A, and Tarricq, Y. Kalkayotl: A cluster distance inference code. *Astronomy and Astrophysics*, 644:A7, 2020.

[190] Paatero, Pentti and Tapper, Unto. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2):111–126, 1994.

[191] Pallis, George, Angelis, Lefteris, Vakali, Athena, and Pokorny, Jaroslav. A probabilistic validation algorithm for web users' clusters. In *2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No. 04CH37583)*, volume 5, pages 4129–4134. IEEE, 2004.

[192] Park, Hae-Sang and Jun, Chi-Hyuck. A simple and fast algorithm for k-medoids clustering. *Expert systems with applications*, 36(2):3336–3341, 2009.

[193] Pastore, Massimiliano. Overlapping: A R package for estimating overlapping in empirical distributions. *J. Open Source Softw.*, 3(32):1023, dec 2018.

[194] Pearson, Karl. Liii. 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.

[195] Peng, Wei and Li, Tao. On the equivalence between nonnegative tensor factorization and tensorial probabilistic latent semantic analysis. *Applied Intelligence*, 35(2):285–295, 2011.

[196] Rao, C Radakrishna. Differential metrics in probability spaces. *Differential geometry in statistical inference*, 10:217–240, 1987.

[197] Rao, C Radhakrishna. Diversity and dissimilarity coefficients: a unified approach. *Theoretical population biology*, 21(1):24–43, 1982.

[198] Rao, C Radhakrishna. Diversity: Its measurement, decomposition, apportionment and analysis. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 1–22, 1982.

[199] Roche, Alexis. Em algorithm and variants: An informal tutorial. *pre-print*, 2011.

[200] Rodner, Erik and Denzler, Joachim. Randomized probabilistic latent semantic analysis for scene recognition. In *Iberoamerican Congress on Pattern Recognition*, pages 945–953. Springer, 2009.

[201] Rousseeuw, Peter J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65, 1987.

[202] Jaakkola, Tommi S and Haussler, David. Probabilistic kernel regression models. In *Seventh International Workshop on Artificial Intelligence and Statistics*. PMLR, 1999.

[203] Salazar, Diego, Rios, Juan, Aceros, Sara, Flórez-Vargas, Oscar, and Valencia, Carlos. Kernel joint non-negative matrix factorization for genomic data. *IEEE Access*, 9:101863–101875, 2021.

[204] Salcedo-Sanz, Sancho, Rojo-Álvarez, José Luis, Martínez-Ramón, Manel, and Camps-Valls, Gustavo. Support vector machines in engineering: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 4(3):234–267, 2014.

[205] Saul, Lawrence and Pereira, Fernando. Aggregate and mixed-order markov models for statistical language processing. *arXiv preprint cmp-lg/9706007*, 1997.

[206] Scetbon, Meyer and Harchaoui, Zaid. A spectral analysis of dot-product kernels. In *International conference on artificial intelligence and statistics*, pages 3394–3402. PMLR, 2021.

[207] Schatten, Robert. A theory of cross-spaces.(am-26), volume 26. In *A Theory of Cross-Spaces.(AM-26), Volume 26*. Princeton University Press, 2016.

[208] Schmidt, Erhard. Zur theorie der linearen und nichtlinearen integralgleichungen. In *Integralgleichungen und Gleichungen mit unendlich vielen Unbekannten*, pages 190–233. Springer, 1989.

[209] Schölkopf, Bernhard, Herbrich, Ralf, and Smola, Alex J. A generalized representer theorem. In *Computational Learning Theory: 14th Annual Conference on Computational Learning Theory, COLT 2001 and 5th European Conference on Computational Learning Theory, EuroCOLT 2001 Amsterdam, The Netherlands, July 16–19, 2001 Proceedings 14*, pages 416–426. Springer, 2001.

[210] Scholkopft, Bernhard and Mullert, Klaus-Robert. Fisher discriminant analysis with kernels. *Neural networks for signal processing IX*, 1(1):1, 1999.

[211] Senga Kiesse, Tristan and Cuny, HE. Discrete triangular associated kernel and bandwidth choices in semiparametric estimation for count data. *Journal of Statistical Computation and Simulation*, 84(8):1813–1829, 2014.

[212] Sharma, Alok and Paliwal, Kuldip K. Rotational linear discriminant analysis technique for dimensionality reduction. *IEEE Transactions on Knowledge and Data Engineering*, 20(10):1336–1347, 2008.

[213] Shashanka, Madhusudana. Simplex decompositions for real-valued datasets. In *2009 IEEE International Workshop on Machine Learning for Signal Processing*, pages 1–6. IEEE, 2009.

[214] Shashua, Amnon and Hazan, Tamir. Non-negative tensor factorization with applications to statistics and computer vision. In *Proceedings of the 22nd international conference on Machine learning*, pages 792–799, 2005.

[215] Smyth, Padhraic. Model selection for probabilistic clustering using cross-validated likelihood. *Statistics and computing*, 10(1):63–72, 2000.

[216] Stewart, Gilbert W. On the early history of the singular value decomposition. *SIAM review*, 35(4):551–566, 1993.

[217] Struyf, Anja, Hubert, Mia, and Rousseeuw, Peter. Clustering in an object-oriented environment. *Journal of Statistical Software*, 1(4):1–30, 1997.

[218] Wolfgang Stummer and Anna-Lena Kißlinger. Some new flexibilizations of bregman divergences and their asymptotics. In *Geometric Science of Information: Third International Conference, GSI 2017, Paris, France, November 7-9, 2017, Proceedings 3*, pages 514–522. Springer, 2017.

[219] Sylvester, James Joseph. *The Collected Mathematical Papers of James Joseph Sylvester...*, volume 3. University Press, 1909.

[220] Tan, Chuanqi, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. A survey on deep transfer learning. *arXiv preprint*, 2018.

[221] Tan, Kok Kiong, Lv, Jian Cheng, Yi, Zhang, and Huang, Sunan. Adaptive multiple minor directions extraction in parallel using a pca neural network. *Theoretical computer science*, 411(48):4200–4215, 2010.

[222] Tian, Dongping. Research on plsa model based semantic image analysis: A systematic review. *Journal of Information Hiding and Multimedia Signal Processing*, 9:1099–1113, 09 2018.

[223] Tibshirani, Robert and Hastie, Trevor. Local likelihood estimation. *Journal of the American Statistical Association*, 82(398):559–567, 1987.

[224] Tibshirani, Robert, Walther, Guenther, and Hastie, Trevor. Estimating the number of clusters in a data set via the gap statistic. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 63(2):411–423, 2001.

[225] Tsallis, Constantino. Possible generalization of boltzmann-gibbs statistics. *Journal of statistical physics*, 52(1):479–487, 1988.

[226] Tsuda, Koji, Akaho, Shotaro, Kawanabe, Motoaki, and Müller, Klaus-Robert. Asymptotic properties of the fisher kernel. *Neural computation*, 16(1):115–137, 2004.

[227] Ullmann, Theresa, Hennig, Christian, and Boulesteix, Anne-Laure. Validation of cluster analysis results on validation data: A systematic framework. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, page e1444, 2022.

[228] Valiant, LG. A theory of the learnable. *Comunications of the ACM*, 27(11):1134–1142, 1984.

[229] Van Loan, Charles F. Generalizing the singular value decomposition. *SIAM Journal on numerical Analysis*, 13(1):76–83, 1976.

[230] Vapnik, Vladimir N. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999.

[231] Vert, Jean-Philippe, Tsuda, Koji, and Schölkopf, Bernhard. A primer on kernel methods. *Kernel methods in computational biology*, 47:35–70, 2004.

[232] Von Neumann, J. Some matrix-inequalities and metrization of matric-space. tomsk univ. rev. 1, 286–300 (1937). reprinted in collected works, 1962.

[233] Wade, Sara and Ghahramani, Zoubin. Bayesian cluster analysis: Point estimation and credible balls (with discussion). *Bayesian Analysis*, 13(2):559–626, 2018.

[234] Wan, Raymond, Anh, Vo Ngoc, and Mamitsuka, Hiroshi. Efficient probabilistic latent semantic analysis through parallelization. In *Asia Information Retrieval Symposium*, pages 432–443. Springer, 2009.

[235] Watanabe, Michiko and Yamaguchi, Kazunori. *The EM algorithm and related statistical models*. CRC Press, 2003.

[236] Weyl, Hermann. Inequalities between the two kinds of eigenvalues of a linear transformation. *Proceedings of the national academy of sciences*, 35(7):408–411, 1949.

[237] Wiener, Norbert. *Cybernetics or Control and Communication in the Animal and the Machine*. MIT press, 1949.

[238] Wild, Stefan, Curry, James, and Dougherty, Anne. Improving non-negative matrix factorizations through structured initialization. *Pattern recognition*, 37(11):2217–2232, 2004.

[239] Williamson, Jon. The philosophy of science and its relation to machine learning. In *Scientific data mining and knowledge discovery: Principles and foundations*, pages 77–89. Springer, 2009.

[240] Wu, CF Jeff. On the convergence properties of the em algorithm. *The Annals of statistics*, pages 95–103, 1983.

[241] Wu, Hu, Wang, Yongji, and Cheng, Xiang. Incremental probabilistic latent semantic analysis for automatic question recommendation. In *Proceedings of the 2008 ACM conference on Recommender systems*, pages 99–106, 2008.

[242] Xu, J., Ye, G.T., Wang, Y., Herman, G., Zhang, B., and Yang, J. Incremental em for probabilistic latent semantic analysis on human action recognition. *6th IEEE International Conference on Advanced Video and Signal Based Surveillance*, 2009.

[243] Ye, Yun, Gong, Shengrong, Liu, Chunping, Zeng, Jia, Jia, Ning, and Zhang, Yi. Online belief propagation algorithm for probabilistic latent semantic analysis. *Frontiers of Computer Science*, 7(4):526–535, 2013.

[244] Yoo, Jiho and Choi, Seungjin. Probabilistic matrix tri-factorization. In *2009 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1553–1556. IEEE, 2009.

[245] Zhang, Daoqiang, Zhou, Zhi-Hua, and Chen, Songcan. Non-negative matrix factorization on kernels. In *Pacific Rim International Conference on Artificial Intelligence*, pages 404–412. Springer, 2006.

[246] Zhang, Xian-Da. *Matrix analysis and applications*. Cambridge University Press, 2017.

[247] Zhang, Yu-Fang, Zhu, Jun, and Xiong, Zhong-Yang. Improved text clustering algorithm of probabilistic latent with semantic analysis [j]. *Journal of Computer Applications*, 3, 2011.

[248] Zhang, Zhihua. The singular value decomposition, applications and beyond. *arXiv preprint arXiv:1510.08532*, 2015.

[249] Zhao, Rui and Mao, Kezhi. Supervised adaptive-transfer plsa for cross-domain text classification. In *2014 IEEE International Conference on Data Mining Workshop*, pages 259–266. IEEE, 2014.

[250] Zhou, Yuhao, Shi, Jiaxin, and Zhu, Jun. Nonparametric score estimators. In *International Conference on Machine Learning*, pages 11513–11522. PMLR, 2020.

[251] Zhuang, Liansheng, She, Lanbo, Jiang, Yuning, Tang, Ketan, and Yu, Nenghai. Image classification via semi-supervised plsa. In *2009 Fifth International Conference on Image and Graphics*, pages 205–208. IEEE, 2009.

[252] Zitzler, Eckart, Teich, Jrgen, and Bhattacharyya, Shuvra S. Optimizing the efficiency of parameterized local search within global search: A preliminary study. In *Proceedings of the 2000 Congress on Evolutionary Computation. CEC00 (Cat. No. 00TH8512)*, volume 1, pages 365–372. IEEE, 2000.

Index of authors