Artificial Intelligence Research and Development
M. Villaret et al. (Eds.)

doi:10.3233/FAIA210126

A Common Formal Framework for Factorial and Probabilistic Topic Modelling Techniques

Karina GIBERT [a,1], Yaroslav HERNANDEZ-POTIOMKIN [a]
[a] *Intelligent Data Science and Artificial Intelligence Research Group, Universitat Politècnica de Catalunya (Spain)*

Abstract. Topic modelling is nowadays one of the most popular techniques used to extract knowledge from texts. There are several families of methods related to this problem, among them 1) Factorial methods, 2) Probabilistic methods and 3) Natural Language Processing methods. In this paper a common conceptual framework is provided for Factorial and probabilistic methods by identifying common elements and describing them with common and homogeneous notation and 7 different methods are described accordingly. Under a common notation it is easy to make a comparative analysis and see how flexible or more or less realistic assumptions are made by the different methods. This is the first step to a wider analysis where all families can be related to this common conceptual framework and to go in depth in the understanding of stengths and weakenesses of each method and ellaboration of general guidelines to provide application criteria. The paper ends with a discussion comparing the presented methods and future research lines.

Keywords. Topic modelling, Multivariate methods, Probabilistic Methods

1. Introduction

Textual data analysis is a constantly growing field with many open research problems. Often, textual data analysis is used for 1) Understand the underlying topics of a set of documents and 2) Find the principal concepts that better summarize a given text.

There are many applications of topic modelling, such as enhanced document clustering [1], topic trend detection [14] high-dimensional classification [28] and dimensionality reduction and projection [18] among others.

However, currently there are some limitations of existing techniques, as for example considering semantic structures of a text; i.e. synonymy and polysemy, ortography, bias due to outliers and the fact that the topics are implicit and require subjective interpretation.

[1] Corresponding Author: Karina Gibert, Intelligent Data Science and Artificial Intelligence Research Group, Universitat Politècnica de Catalunya (Spain); E-mail: karina.gibert@upc.edu.

Topic modelling techniques mainly belong to three big families of methods, **1)** Factorial methods, **2)** Probabilistic methods and **3)** Natural Language Processing (NLP) methods. The former, aim to perform a decomposition over the multivariate design matrix in such a way that a given objective function is maximized and the given set of constraints are satisfied [6]. Several examples are Latent Semantic Analysis (LSA) [9] or Principal Components Analysis (PCA) [20], [30], Non-negative Matrix Factorization (NMF) [21], Canonical Correlation Analysis (CCA), Correspondence Analysis (CA) [20], Multiple Correspondence Analysis (MCA) [15], Non-linear Iterative Partial Least Squares algorithm (NIPALS) [30], Archetypal Analysis (AA) [6] among others.

Probabilistic methods, in turn, relies on statistical model definition, composed of probability model and a parameter's space definition. The parameters can be estimated either through the Maximum Likelihood function (*frequentist* approach), or *Bayesian* framework [12]. The probabilistic approaches are very valuable as they are generative, they provide clear interpretation, flexibility and extensibility. Research addressed to probabilistic topic modelling is widely applied to multi-document summarization [6], text classifiers [31], topics extraction [16] and topic-based document classification [23].

The third family, NLP, combines classical language analysis and statistical approaches [17] in order to provide very specific solutions that aim to tackle the problem of inferring the true meaning from text. A special attention is paid to linguistic annotations, *Treebanks* [22], [13], [11], [19], which intervene in part-of-speech (POS) tagging, syntactic parsing, morphological analysis and word sense disambiguation. NLP is used in topic modelling and it can be also considered as a pre-processing step for the other two families to provide more meaningful topic inference in terms of semantics. Due to space limitation this family will not be covered, but will be presented in future work.

Given that first two families of methods above come from very different origins (multivariate analysis, Bayesian analysis and AI, respectively), our goal is to present a unified notation and make it homogeneous over different techniques and authors. A general formalization contributes to a better understanding of the relationships between those families of methods and their common properties. This allows to have a common language for the existing state-of-the-art literature considerably simplifies the task of the researcher to identify the reasons of different results obtained with the different techniques over the same set of texts.

The structure of the paper is as follows. First, in Section §2, it is presented the proposed common structures and notation. Then, Factorial methods are presented in Section §3 along with their applications and extensions in the topic modelling domain. Then, Probabilistic PCA and Probabilistic Topic Modelling framework are presented in Sections §4 and §5, respectively. Finally, Section §6 contains a brief discussion among the different methods, conclusions and future work.

2. Methodology: a common notation framework for textual analysis methods

In this section the symbology associated to the different elements appearing in the presented methods is designed and it will be used in the methods presented in the following sections.

2.1. Numerization of the corpus

It appears to be a basic and very early operation in most of the methods from Factorial and Probabilistic families and it consists in representing a set of documents through numeric matrices that represent in each row the distribution of words in one document.

Given a set of documents $\mathscr{D}$ of size $n_{\mathscr{D}}$ and a set of terms $\mathscr{T}$ of size $n_{\mathscr{T}}$, a document is a sequence of words such that for each document $d_j \in \mathscr{D}$ with $j = 1 \ldots n_{\mathscr{D}}$, $d_j = (w_1, w_2, \ldots, w_{n_{d_j}})$ where n_{d_j} is the number of words in the document d_j and $w_{\mathscr{p}} \in \mathscr{T}$ with $\mathscr{p} = 1 \ldots n_{d_j}$.

The numerization of the corpus produces a matrix X of dimensionality $n_{\mathscr{T}} \times n_{\mathscr{D}}$, as shown below, where the rows correspond to terms $t \in \mathscr{T}$ (i.e. vocabulary used in the corpus $\mathscr{D}$), and columns correspond to documents $d \in \mathscr{D}$ of the given corpus. Therefore, the number of rows of X is the cardinal of the given vocabulary ($n_{\mathscr{T}}$) and the number of columns is the cardinal of $\mathscr{D}$, $n_{\mathscr{D}}$. Each cell (i, j) is the raw count of n_{ij} occurrences of the term t_i in the document d_j from the collection.

$$X = \begin{array}{c} \\ 1 \\ \vdots \\ n_{\mathscr{T}} \end{array} \begin{array}{c} \begin{array}{ccc} 1 & \ldots & n_{\mathscr{D}} \end{array} \\ \left[\begin{array}{c} \vdots \\ \cdots\ n_{ij}\ \cdots \\ \vdots \end{array} \right] \end{array} \tag{1}$$

Matrix X is often known as TDM, a term-document matrix, successfully applied in the field of information retrieval [2] and tf-idf computation [26].

Along this paper we will name as x^i the vector $x^i = (n_{i1}, \ldots, n_{in_{\mathscr{D}}})$, which represents the profile of a certain term in a corpus, that is, the distribution of the occurrences of term t_i in the documents of $\mathscr{D}$.

2.2. Binarization of documents

An alternative common representation of the corpus is the *Binarization* of documents. It consists in representing each single document $d_j \in \mathscr{D}$ as a binary matrix, $d^{(j)}$, describing the distribution of terms in the document.

Provided that n_{d_j} is the length of the document d_j, the matrix $d^{(j)}$, of dimensionality $n_{d_j} \times n_{\mathscr{T}}$, has $n_{\mathscr{T}}$ terms in columns and n_{d_j} rows representing the positions where terms can be placed along the document.$\mathscr{p} = 1 \ldots n_{d_j}$ indexes the positions of the terms that appear in the document d_j.

$$d^{(j)} = \begin{array}{cc} & \begin{array}{ccc} 1 & \dots & n_{\mathscr{T}} \end{array} \\ \begin{array}{c} 1 \\ \\ \vdots \\ \\ n_{d_j} \end{array} & \left[\begin{array}{ccc} & \vdots & \\ \cdots & d^{(j)}_{\not{p}i} & \cdots \\ & \vdots & \end{array}\right] \\ & \left[\begin{array}{ccc} n_{1j} & \dots & n_{\mathscr{T}j} \end{array}\right] \end{array} \tag{2}$$

The cell, $d^{(j)}_{\not{p}i}$, of the binary matrix, $d^{(j)}$, is defined as follows:

$$d^{(j)}_{\not{p}i} = \begin{cases} 1, & \text{if term } t_i \text{ appears at position} \not{p} \text{ of the document } d_j \\ 0, & \text{otherwise} \end{cases} \tag{3}$$

3. Factorial methods

3.1. Latent Semantic Analysis

The main goal of LSA [9] is to infer meaningful semantic structures of the terms in documents and to discard those attributable to noise. The internals of LSA reside in the two-way factor analysis using Singular Value Decomposition (SVD). Two is referred to the fact that both terms as well as documents are jointly represented in the same factorial space, thus allowing the analysis of relationships between them.

The matrix X, described in Section §2.1, or its weighted version (using tf-idf numerical statistic from [27]), can be decomposed, under SVD, into the product of three matrices as follows:

$$X = \mathscr{V}_{(n_{\mathscr{T}} \times \mathscr{K})} \Lambda^{\frac{1}{2}}_{(\mathscr{K} \times \mathscr{K})} (\mathscr{U}')_{(\mathscr{K} \times n_{\mathscr{D}})} \tag{4}$$

being $\mathscr{V}_{(n_{\mathscr{T}} \times \mathscr{K})}$ a matrix of eigenvectors of XX', $\mathscr{U}_{(n_{\mathscr{D}} \times \mathscr{K})}$ a matrix of eigenvectors of $X'X$, $\Lambda_{(\mathscr{K} \times \mathscr{K})}$ a diagonal matrix of eigenvalues and $\mathscr{K} = \min\{n_{\mathscr{T}}, n_{\mathscr{D}}\}$ the rank of X.

Let u_α be one of the $\mathscr{K}$ eigenvectors of the matrix $\mathscr{U}_{(n_{\mathscr{D}} \times \mathscr{K})}$, and it is a linear combination of the original set of "document-variables". The projection of X over u_α, $\Psi_\alpha = X u_\alpha$, is α-th principal component of the dataset, that can be thought as concept or topic. The associated eigenvalue $\lambda_{\alpha\alpha}$ measures the quantity of information retained by Ψ_α from the total information contained in X [3].

Joint representation (of terms and documents) onto factorial space is possible due to transition relations between $\mathscr{V}$ and $\mathscr{U}$ [10] and can be represented through *rescaling factor* or *biplot* representation [20].

In LSA several issues were identified: *synonymy*, *polysemy* and rare event detection. Last one tackled very elegantly in Correspondence Analysis [15], [20] by using a χ^2 metrics.

As an extension of LSA and to overcome the lack of context, in [25] authors present Distributional Semantic Model by extending the Vector Space Model representation in which they introduce the co-occurrence of the terms matrix, $C_{(n_{\mathscr{T}} \times n_f)}$ between all $n_{\mathscr{T}}$ terms and n_f pre-defined terms. The maximization expression becomes:

$$\max_{u_\alpha} u'_\alpha (X'C)'(X'C) u_\alpha \text{ s.t. } u'_\alpha u_\alpha = 1 \tag{5}$$

3.2. Archetypal Analysis

Archetypal Analysis (AA) [8] belongs to the same family of optimization problems such as LSA/PCA, *k*-means or NMF. For instance, in [6], authors present a framework to handle multi-document summarization problem. The formulation of the AA as an optimization problem is as follows:

$$\begin{aligned} &\underset{H,W}{min} \quad \|J - H_{(n_{\mathscr{T}} \times K)} W' J\|^2 \\ &s.t. \sum_k^K h_k^i = 1, h_k^i \geq 0, \forall i \in \{1 \dots n_{\mathscr{T}}\} \text{ and } \sum_i^{n_{\mathscr{T}}} w_k^i = 1, h_k^i \geq 0, \forall k \in \{1 \dots K\} \end{aligned} \tag{6}$$

where $Y_{(n_{\mathscr{S}} \times K)} = J'W$ is defined as the matrix of K archetypes in columns, which are built as convex combinations of observations. $W_{(n_{\mathscr{T}} \times K)}$ (estimated) determines the convex combination of J such that columns of Y are placed on the convex hull of data J. In turn, the observations can be approximated as convex combinations of archetypes. And matrix H describes convex combination of archetypes to approximate observations, such that $J \approx HY'$, or in other words, H is the weighting matrix that approximates the archetypal space into the transformed design matrix J. In contrast to NMF, AA decomposes the matrix J into sparser stochastic matrices. And the archetypes, the columns of Y, can be interpreted as topics or latent representation of the data.

4. Probabilistic Principal Component Analysis

In standard PCA, the approach is to maximize the projection of the original data space X (the individuals) onto the latent (unknown) factorial space Ψ. But in the probabilistic version, the idea is to first establish the link from latent space Ψ to original data space X and then, the reverse mapping is found by using the posterior distribution by Bayes theorem. A PPCA is a linear Gaussian *latent* variable model [29], [4].

A particular term profile x^i (defined in §2.1) is defined in [29] as stochastic linear combination of its corresponding projection in the latent space (see §3.1), namely ψ^i (is the i-th row of the matrix Ψ), plus a noise term

$$x^i|\psi^i = (\mu + W\psi^i) + \varepsilon^i, \qquad \varepsilon^i \sim \mathscr{N}(0, \sigma^2 I) \tag{7}$$

where μ is the global mean of, $X_{n_{\mathscr{T}} \times n_{\mathscr{D}}}$ and W is the parameter matrix that contains the *factor loadings*. After several simplifications, such as homoschedasticity hypothesis, the posterior distribution of the latent variables, ψ^i, is obtained with the Bayes Law and formulated as:

$$p(\psi^i|x^i) = \mathscr{N}\big(M^{-1}W'(x^i - \mu),\ \sigma^2 M^{-1}\big) \tag{8}$$

where $M_{K\times K} = \sigma^2 I + W'W$. Details have been omitted due to space constraints.

The marginal log-likelihood of the data, X, is formulated as:

$$\mathscr{L}(\mu, \sigma^2, W) = \sum_{i=1}^{n_{\mathscr{T}}} \log\left\{p(x^i)\right\} = -\frac{n_{\mathscr{T}}}{2}\left\{n_{\mathscr{D}} \log(2\pi) + \log|C| + \mathrm{tr}(C^{-1}S)\right\} \tag{9}$$

being $C = WW' + \sigma^2 I$ the model covariance matrix and S the empirical covariance matrix of $X_{n_{\mathscr{T}} \times n_{\mathscr{D}}}$ and $\pi = 3.1415\ldots$ By this technique, it can be obtained the approximation (by EM algorithm) to the same axes as in LSA or PCA.

5. Probabilistic topic modelling

The probabilistic mixture models [24], [4] are characterized by the fact that the data is generated by one of the mixture components. For example, a mixture of Gaussians can approximate any type of continuous distributions, including multimodal [7]. In this work, each mixture component is referred as topic.

Under the assumptions of independence between document size and words' sequence, the documents as an iid sample, and that the document can be expressed as mixture of the set of topics, the generic likelihood is as follows:

$$\begin{aligned}\mathscr{L}(\theta) = \prod_{j=1}^{n_{\mathscr{D}}} \sum_{k=1}^{K} \Bigg(& P(N = n_{d_j} | \mathrm{Z} = z_k \wedge \theta) \cdot \\ & \cdot P\Big(\bigwedge_{\not{p}=1:n_{d_j}} D_{\not{p}} = w_{\not{p}} | \mathrm{Z} = z_k \wedge \theta\Big) P(\mathrm{Z} = z_k|\theta)\Bigg)\end{aligned} \tag{10}$$

where Z is a discrete random variable with values in $\mathscr{Z} = \{z_1, \ldots, z_K\}$ and the associated probability space is defined as follows:

$$\langle \mathscr{Z}, \mathscr{P}(\mathscr{Z}), P_{\mathscr{Z}} \rangle \tag{11}$$

where $\mathscr{Z}$ is the sample space (set of topics), $\mathscr{P}(\mathscr{Z})$ is parts of $\mathscr{Z}$ and $P_{\mathscr{Z}}$ is the probability function associated to $\mathscr{P}(\mathscr{Z})$. $P_{\mathscr{Z}}$ is built on top of $p_{\mathscr{Z}} = P(Z = z_k)$ for $k = 1 \ldots K$, provided that $\mathscr{P}(\mathscr{Z})$ is a σ-algebra. $D_{\wp}$ is a random variable indicating which term is observed in any position $\wp$ of document d_j. And θ is the set of distributional parameters.

5.1. Generative model

In Generative model [23] the probability of a word remains constant for all documents in a corpus and independent of the words in other positions of same document, as well as independent of the position where it is observed, conditioned on the topic and a set of parameters. Therefore,

$$D_{\wp} | Z = z_k \wedge \theta \sim \mathrm{Cat}(\pi_{1k}, \ldots \pi_{n_{\mathscr{T}}k}) \tag{12}$$

where π_{ik} is the probability of occurrence of term $t_i \in \mathscr{T}$ given topic $Z = z_k$. Under this approach, the likelihood function in (10) is reformulated as follows:

$$\mathscr{L}(\theta) = \prod_{j=1}^{n_{\mathscr{D}}} \sum_{k=1}^{K} \left(P(N = n_{d_j} | Z = z_k \wedge \theta) \left(\prod_{i=1}^{n_{\mathscr{T}}} \pi_{ik}^{n_{ij}} \right) \cdot P(Z = z_k | \theta) \right) \tag{13}$$

Similarly, the likelihood function for the Multinomial model can be rewritten as follows:

$$\mathscr{L}(\theta) = \prod_{j=1}^{n_{\mathscr{D}}} \sum_{k=1}^{K} \left[P(N = n_{d_j} | Z = z_k \wedge \theta) \left(\frac{n_{d_j}!}{\prod_{i=1}^{n_{\mathscr{T}}} n_{ij}!} \prod_{i=1}^{n_{\mathscr{T}}} \pi_{ik}^{n_{ij}} \right) P(Z = z_k | \theta) \right] \tag{14}$$

And the likelihood function for the Multivariate Bernoulli model is as follows:

$$\mathscr{L}(\theta) = \prod_{j=1}^{n_{\mathscr{D}}} \sum_{k=1}^{K} \left[\left(\prod_{i=1}^{n_{\mathscr{T}}} \left(x_j^i \pi_{ik} + (1 - x_j^i)(1 - \pi_{ik}) \right) \right) P(Z = z_k | \theta) \right] \tag{15}$$

Having, for the given document d_j, the realization of the variable X_j^i is $x_j^i \in \{0, 1\}$, which states whether the term t_i is present in the specific document d_j or not.

5.2. Latent Dirichlet Allocation

The Latent Dirichlet Allocation (LDA) is a bayesian hierarchical model [5]. As opposite to models presented in Section §5, the documents are associated with multiple topics simultaneously [5]. Hence, the random vector D (defined in Section §5) with the sequence of words in the document, jointly appears with the random vector $\mathbb{Z}$ (sequence of topics in the document):

$$P(D = d \wedge \mathbb{Z} = (z_1, z_2, \ldots, z_{n_d})) = P(\bigwedge_{\mathscr{p}=1:n_d} (D_{\mathscr{p}} = w_{\mathscr{p}} \wedge Z_{\mathscr{p}} = z_{\mathscr{p}})) \tag{16}$$

The likelihood function of the parameters α and θ in the corpus $\mathscr{D}$ can be defined as follows:

$$\begin{aligned}\mathscr{L}(\alpha, \theta) &= \prod_{j=1}^{n_{\mathscr{D}}} P(D = d_j | \alpha \wedge \theta) \\ &= \prod_{j=1}^{n_{\mathscr{D}}} \int P(\zeta_j | \alpha) \left(\prod_{\mathscr{p}=1}^{n_{d_j}} \sum_{k=1}^{K} P(D_{\mathscr{p}} = w_{\mathscr{p}} | Z_{\mathscr{p}} = z_k \wedge \theta) P(Z_{\mathscr{p}} = z_k | \zeta_j) \right) d\zeta_j\end{aligned} \tag{17}$$

6. Conclusion

Although multivariate and probabilistic topic modelling families are very different approaches from the conceptual point of view, this work shows that using common notation, commonalities and differences among them can be analysed in detail.

In the multivariate setting no distributional assumptions are made, but it provides a very clear interpretation and geometrical representation of the associations between the topics, documents and terms, so that visual inspection of the results can provide a global overview of the interactions among these. These methods optimize a function related with the information of the original data set: residual sum of squares for AA, and maximize projected variance for PCA. In general, the optimal projection directions are found based on diagonalization techniques applied to combinations of the TDM or TSM.

The probabilistic methods, on the other side, do not provide geometric representation, but are more flexible while capturing associations between topics, documents and terms. Also, they assume a predetermined number of topics from the beginning, while the multivariate methods allow the determination of the relevant topics as an output, by analyzing the quantity of information preserved in each of the topics and keeping the significant ones. The flexibility of LDA, allowing every word in a document to be associated with a different topic, does not seem very realistic either.

The PPCA looks like a very interesting method as a combination of probabilistic and multivariate methods, nevertheless the Gaussian assumption does not correspond to the distribution of the terms in the document.

Therefore, although being a simple linear models, multivariate models look like the most conservative modelling schema as they do not make any distributional assumptions and the interpretation of the results is straightforward.

However, what this analysis is making evident is that all of the proposed methods provide elements to characterize the topics in terms of the documents in the topic, or the words more representative of the topic, but all of them rely on the comprehension of which topics really are, to the interpretational habilities of the analyst, thus pointing to a missing final step in the topic modelling research field which is to provide a concept (or a label) for each of the discovered topics.

Ongoing research is the generalization of the proposed common formal framework to include the main concepts of the NLP methods. And once the common framework has been established, the next step will be to apply the proposals of already existing and new methodologies on testing real cases. Also, the incorporation of Natural Language Processing into the pipeline of LSA or similar techniques, opens the door to introduce inductive reasoning and ontologies of words and terms to apply inductive learning principles to the extraction of explicit concepts associated to the discovered topics, so that a proposal for automatic interpretation of topics can be formalized.

References

[1] Parvin Ahmadi, Iman Gholampour, and Mahmoud Tabandeh. Cluster-based sparse topical coding for topic mining and document clustering. *ADAC*, 12(3):537–558, Sep 2018.

[2] Ricardo Baeza-Yates, Berthier Ribeiro-Neto, et al. *Modern information retrieval*, volume 463. ACM press New York, 1999.

[3] Jean-Paul Benzécri et al. Lanalyse des donnees, tome ii. *Lanalyse des correspondances. Dunod Press, Paris*, 1973.

[4] Christopher M. Bishop. *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag, Berlin, Heidelberg, 2006.

[5] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March 2003.

[6] Ercan Canhasi and Igor Kononenko. Multi-document summarization via archetypal analysis of the content-graph joint model. *Knowledge and Information Systems*, 41(3):821–842, Dec 2014.

[7] M. A. Carreira-Perpinan. Mode-finding for mixtures of gaussian distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1318–1323, Nov 2000.

[8] Adele Cutler and Leo Breiman. Archetypal analysis. *Technometrics*, 36(4):338–347, 1994.

[9] Scott Deerwester, Susan T. Dumais, George W. Furnas, Thomas K. Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal Of The American Society For Information Science*, 41(6):391–407, 1990.

[10] Luc Devroye, Ludovic Lebart, A Morineau, and J.-P Fenelon. Tratement des donnees statistiques: Methodes et programmes. *Journal of the American Statistical Association*, 75:1040, 12 1980.

[11] Christiane Fellbaum. *WordNet: An Electronic Lexical Database*. Bradford Books, 1998.

[12] Marti Font, Xavier Puig, and Josep Ginebra. Bayesian Analysis of the Heterogeneity of Literary Style. *Revista Colombiana de Estadstica*, 39(2):205–227, 2016.

[13] W. N. Francis and H. Kucera. Brown corpus manual. Technical report, Department of Linguistics, Brown University, Providence, Rhode Island, US, 1979.

[14] Wolfgang Gaul and Dominique Vincent. Evaluation of the evolution of relationships between topics over time. *Advances in Data Analysis and Classification*, 11(1):159–178, Mar 2017.

[15] Michael J. Greenacre. Correspondence analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2(5):613–619, 2010.

[16] Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, pages 50–57, New York, NY, USA, 1999. ACM.

[17] Nitin Indurkhya and Fred J Damerau. *Handbook of natural language processing*, volume 2. CRC Press, 2010.

[18] Serge Iovleff. Probabilistic auto-associative models and semi-linear pca. *Advances in Data Analysis and Classification*, 9(3):267–286, Sep 2015.

[19] S. Johansson. *Manual of Information to Accompany the Lancaster-Oslo/Bergen Corpus of British English, for Use with Digital Computers.* ICAME collection of English language corpora. Univ., Department of English, 1978.

[20] I.T. Jolliffe. *Principal Component Analysis.* Springer Series in Statistics. Springer, 2002.

[21] Daniel D. Lee and H. Sebastian Seung. Algorithms for non-negative matrix factorization. In *Proceedings of the 13th International Conference on Neural Information Processing Systems*, NIPS'00, pages 535–541, Cambridge, MA, USA, 2000. MIT Press.

[22] Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a large annotated corpus of english: The penn treebank. *Comput. Linguist.*, 19(2):313–330, June 1993.

[23] Kamal Nigam, Andrew Kachites Mccallum, Sebastian Thrun, and Tom Mitchell. Text classification from labeled and unlabeled documents using em. *Machine Learning*, 39(2):103–134, May 2000.

[24] D. Peña. *Análisis de datos multivariantes.* Mc Graw Hill, 2002.

[25] Martin Rajman and Romaric Besançon. Stochastic distributional models for textual information retrieval. *Proc. of 9th ASMDA*, pages 80–85, 1999.

[26] G. Salton and M.J. McGill. *Introduction to modern information retrieval.* McGraw-Hill computer science series. McGraw-Hill, 1983.

[27] Hinrich Schütze, Christopher D Manning, and Prabhakar Raghavan. *Introduction to information retrieval*, volume 39. Cambridge University Press, 2008.

[28] Dawit G. Tadesse and Mark Carpenter. A method for selecting the relevant dimensions for high-dimensional classification in singular vector spaces. *Advances in Data Analysis and Classification*, Jan 2018.

[29] M. E. Tipping and Christopher Bishop. Probabilistic principal component analysis. *Journal of the Royal Statistical Society, Series B*, 21/3:611622, January 1999.

[30] Svante Wold, Kim Esbensen, and Paul Geladi. Principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, 2(1):37 – 52, 1987.

[31] Yuhan Zhang and Haiping Xu. Sltm: A sentence level topic model for analysis of online reviews. In *SEKE*, pages 449–453, 2016.