Non-negative matrix factorization for discovering latent features in approval ratings data

William Clocksin
Department of Computer Science,
University of Hertfordshire
william.clocksin@cantab.net

January 2022

Abstract

We describe here the use of non-negative matrix factorization (NMF), an algorithm for the decomposition of large data sets into a small number of latent features. Coupled with a model selection mechanism, NMF is an efficient method for identification of latent features and provides a powerful method for the discovery of latent patterns in data. We use convex NMF together with a model selection policy based on consensus clustering to analyze data about approval ratings of world leaders. This can provide insight into developing general methods for discovering information about relationships between individuals.

1 Introduction

This study is part of a project that is concerned with how intelligent systems might reason about the values and relationships that are most significant to them. Central to this concern is the concept of an *affinity*, the degree of an individual's opinion or belief about a proposition. Because individuals survive in communities, we are interested in how affinities develop through relationships within the community.

As a small initial step within this project, here we apply non-negative matrix factorisation [4] to data sets resulting from opinion surveys. The aim of this analysis is to find latent or 'hidden' patterns in the data. Such latent patterns may be informative about the affinities of individuals or populations, and any inter-relations between affinities and individuals. For example, in some countries, loyalty to a football club correlates with the political views of fans, and these correlations can be estimated by statistical methods [6, 3].

2 The Data Set

Because matrix factorisation in general has been used for a wide variety of applications in which the data take different forms, we first define here a uniform terminology. The data are represented by an *opinion matrix* $\boldsymbol{V}$ of size of $N \times M$, where each $\boldsymbol{V}_{ij}$ represents the *opinion* expressed by *entity* i about *proposition* j. Depending upon how the data are collected, entities could be individuals, communities, or samples from larger populations.

In the small data set used here for illustration [5], the N *samples* (rows of matrix $\boldsymbol{V}$) are approval ratings of a set of five world leaders by the sampled populations of 32 countries. The matrix entry $\boldsymbol{V}_{ij}$ is the approval rating of leader j by a sample of the population of country i. The approval rating is a percentage approval, so for example, $\boldsymbol{V}_{ij} = 0.34$ means that the 34% of the population of country i approves of world leader j. The values of matrix $\boldsymbol{V}$ are adapted from [5], and are shown together with illustrative entity and proposition labels in Table 1.

	Trump	Merkel	Macron	Putin	Xi
Canada	0.28	0.73	0.68	0.29	0.33
Poland	0.51	0.46	0.36	0.15	0.18
Slovakia	0.34	0.37	0.44	0.49	0.20
Hungary	0.33	0.28	0.18	0.28	0.14
Italy	0.32	0.44	0.26	0.38	0.24
U.K.	0.32	0.69	0.55	0.26	0.34
Czechia	0.28	0.25	0.30	0.33	0.17
Bulgaria	0.26	0.50	0.32	0.62	0.33
Greece	0.25	0.22	0.31	0.52	0.17
Netherlands	0.25	0.82	0.70	0.24	0.38
Spain	0.21	0.69	0.60	0.21	0.28
France	0.20	0.74	0.48	0.28	0.23
Sweden	0.18	0.86	0.69	0.17	0.23
Germany	0.13	0.74	0.73	0.36	0.28
Ukraine	0.44	0.66	0.41	0.11	0.28
Russia	0.20	0.34	0.23	0.73	0.59
Philippines	0.77	0.58	0.62	0.61	0.58
India	0.56	0.28	0.28	0.42	0.21
S. Korea	0.46	0.69	0.56	0.25	0.25
Japan	0.36	0.60	0.41	0.26	0.14
Australia	0.35	0.69	0.65	0.27	0.39
Indonesia	0.30	0.35	0.35	0.36	0.32
Israel	0.71	0.52	0.33	0.36	0.35
Lebanon	0.23	0.46	0.50	0.40	0.41
Tunisia	0.12	0.54	0.40	0.41	0.44
Turkey	0.11	0.24	0.14	0.35	0.29
Kenya	0.65	0.46	0.46	0.39	0.58
Nigeria	0.58	0.45	0.45	0.41	0.61
S. Africa	0.42	0.37	0.35	0.36	0.52
Brazil	0.28	0.34	0.27	0.17	0.24
Argentina	0.22	0.29	0.27	0.30	0.35
Mexico	0.08	0.32	0.27	0.28	0.34

Table 1: Small data example adapted from [5]. Entities are countries, and propositions are world leaders. Each entry of the table represents the approval rating of a world leader expressed by a sample of a country's population. Percentage approval has been divided by 100 to give a number between 0.0 and 1.0 without loss of generality. For example, the table shows that 65% of the sample of Australia's population approved of Macron. Following [5], we exclude the data for Lithuania for methodological reasons. These are the values of $\boldsymbol{V}$.

A number of applications have used this format of matrix $\boldsymbol{V}$ for analysis. In recommender systems, the $\boldsymbol{V}_{ij}$ represent ratings of M products N reviewers. In document analysis, the $\boldsymbol{V}_{ij}$ represent the frequency weights of M terms in a set of N documents. In gene analysis, the $\boldsymbol{V}_{ij}$ represent expression levels of M observations of N genes. Using our uniform terminology, the $\boldsymbol{V}_{ij}$ represent the opinions on M propositions as expressed by N entities.

The aim is to find a small number R of latent features, which when combined in linear combination, will approximate $\boldsymbol{V}$ and thereby represent an 'explanation' of the data. Several standard techniques are used to find features in data, including analysis of variance, principal components analysis, vector quantization, Bayesian estimation, and singular value decomposition. However, we use the more recent technique of non-negative matrix factorization (NMF) [4] because non-negativity is inherent in the data being considered, and NMF automatically clusters the data making interpretation easier. Depending on the cost function to be minimised, NMF is formally equivalent to K-means clustering and probabilistic latent semantic analysis.

3 Non-Negative Matrix Factorization

Given a non-negative matrix $\boldsymbol{V}$ and a desired number of features R, an NMF algorithm iteratively calculates non-negative matrix factors $\boldsymbol{W}$ and $\boldsymbol{H}$ such that $\boldsymbol{V} \approx \boldsymbol{W}\boldsymbol{H}$. When $\boldsymbol{V}$ is a $N \times M$ matrix, and the number of features chosen for the approximation is R, where R is smaller than M and N, then $\boldsymbol{W}$ is a $N \times R$ features matrix and $\boldsymbol{H}$ is a $R \times M$ coefficients matrix. R is the desired rank of the factorization problem.

The algorithm starts by initializing $\boldsymbol{W}$ and $\boldsymbol{H}$ to small random values, which are iteratively updated to minimize a cost function. Our algorithm implements both alternative formulations of NMF introduced by [4], and additionally enforces a convexity constraint on $\boldsymbol{W}$, which improves the quality of the clustering result [2]. The two alternative cost functions in [4] are the squared error (or Frobenius norm) and an extension of the Kullback–Leibler divergence to positive matrices. The squared error version minimises the function $F(\boldsymbol{W}, \boldsymbol{H}) = \|\boldsymbol{V} - \boldsymbol{W}\boldsymbol{H}\|_F^2$ where $\|A - B\|_F^2 = \sum_{ij}(A_{ij} - B_{ij})^2$ using the following iterative update rules:

$$\boldsymbol{H}_{rm} \leftarrow \boldsymbol{H}_{rm} \frac{(\boldsymbol{W}^T\boldsymbol{V})_{rm}}{(\boldsymbol{W}^T\boldsymbol{W}\boldsymbol{H})_{rm}} \quad \text{and} \quad \boldsymbol{W}_{nr} \leftarrow \boldsymbol{W}_{nr} \frac{(\boldsymbol{V}\boldsymbol{H}^T)_{nr}}{(\boldsymbol{W}\boldsymbol{H}\boldsymbol{H}^T)_{nr}} \tag{1}$$

The divergence version minimises the function $D(\boldsymbol{V}\|\boldsymbol{W}\boldsymbol{H})$ where $D(\boldsymbol{A}\|\boldsymbol{B}) = \sum_{ij}(\boldsymbol{A}_{ij} \log \frac{\boldsymbol{A}_{ij}}{\boldsymbol{B}_{ij}} - \boldsymbol{A}_{ij} + \boldsymbol{B}_{ij})$ using the following update rules:

$$\boldsymbol{H}_{rm} \leftarrow \boldsymbol{H}_{rm} \frac{\sum_i \boldsymbol{W}_{ir}\boldsymbol{V}_{im}/(\boldsymbol{W}\boldsymbol{H})_{im}}{\sum_j \boldsymbol{W}_{jr}} \quad \text{and} \quad \boldsymbol{W}_{nr} \leftarrow \boldsymbol{W}_{nr} \frac{\sum_m \boldsymbol{H}_{rm}\boldsymbol{V}_{nm}/(\boldsymbol{W}\boldsymbol{H})_{nm}}{\sum_j \boldsymbol{H}_{rj}} \tag{2}$$

For large problems, number of desired features R can be significantly smaller than both M and N. In the illustrative example of Table 1, M and N are already small, so we set the desired number of features $R = 3$.

We impose a convexity constraint on the problem by normalising each column of $\boldsymbol{W}$ to sum to 1 after each iteration of the chosen update rule. The benefits of the convexity constraint in providing a sharper distinction between principal and non-principal component subspaces are discussed in [2].

A given factorization $\boldsymbol{V} \approx \boldsymbol{W}\boldsymbol{H}$ can be used for data analysis in various ways. The prediction error $|\ \boldsymbol{V} - \boldsymbol{W}\boldsymbol{H}\ |$ can be used to evaluate whether some combinations of entity and proposition are 'easier' to predict than others. Matrix $\boldsymbol{H}$ can be used to group the M samples into k clusters. Each sample is placed into a cluster corresponding to the maximum feature in the sample; that is, sample j is placed in cluster i if $\boldsymbol{H}_{ij}$ is the largest entry in column j. Matrix $\boldsymbol{H}$ can be used in a similar way, using maximum values in each row. For clustering to be reliable, a method for *model selection* is recommended.

4 Model Selection

For any value of R, the NMF algorithm groups the samples into clusters, and it is useful to determine whether a given R decomposes the samples into clusters that have a meaningful interpretation. Because the NMF algorithm is not guaranteed to find the lowest error solution but only a local minimum, the algorithm may or may not converge to the same solution on each run, depending on the random initial conditions.

We have adapted a method for consensus clustering [1] to ensure that groupings into clusters are consistent and stable over many runs of the NMF algorithm with different initial conditions. Our version of this method creates a consensus matrix $\boldsymbol{C}$ of size $M \times M$, with entry $\boldsymbol{C}_{ij}$ representing the probability that propositions i and j belong to the same cluster. For each run of NMF, the coefficients matrix $\boldsymbol{H}$ provides the cluster membership. If $\boldsymbol{H}_{kj} > \boldsymbol{H}_{ij}$ for all $i \neq j$, this suggests that proposition j belongs to cluster k. While a given run may assign a given proposition to a different cluster because of different initial conditions, the key observation is that a set of propositions will consistently belong to the same clusters over many runs. The cluster consistency method therefore abstracts away from cluster membership of a cluster k from a given run, and identifies consistent membership of the same propositions within a cluster on each run.

Each run of NMF updates $\boldsymbol{C}$ by incrementing all $\boldsymbol{C}_{ij}$ for which proposition i and j belong to the same cluster according to matrix $\boldsymbol{H}$. At the end of all the runs, we normalise $\boldsymbol{C}$ by dividing its values by the total number of runs. The $\boldsymbol{C}_{ij}$ therefore range from 0 to 1 and reflect the probability that samples i and j are in the same cluster. If a clustering is stable, we expect that $\boldsymbol{C}$ will tend not to vary among runs, and that the entries of $\boldsymbol{C}$ will be close to 0 or 1. In our experiments we run NMF 100 times to construct matrix $\boldsymbol{C}$.

5 Results

NMF can reveal a variety of latent information from the data matrix. Here we look at two types of information: Prediction and Clustering. In the experiments described here, we used the approval rating data shown in Table 1, set $R = 3$, and iterated the update rules until the root mean square error between $\boldsymbol{V}$ and $\boldsymbol{WH}$ dropped to below 10^{-7}. For the squared error rule, convergence was achieved in around 800 iterations. For the divergence rule, convergence was achieved in around 500 iterations. For the model selection, 100 runs were performed to construct the consensus matrix. There was no difference in cluster results between using the squared error update rule or the divergence update rule.

5.1 Prediction

Using the matrix product $\boldsymbol{WH}$, each $(\boldsymbol{WH})_{ij}$ can serve as a prediction of the data value $\boldsymbol{V}_{ij}$. This property is used by recommender systems to predict the rating a customer may give to an item they have not yet reviewed, and it is also used to impute values for missing data. Here, because the data are complete (for each country there is an approval rating for each world leader), we interpret the absolute error $\mid \boldsymbol{V} - \boldsymbol{WH} \mid$ as an indication of how 'easy' it is to make a prediction. An error very near zero suggests that the prediction comes very close to the actual data. The larger the error is, this indicates that the available data are not sufficient to make a reliable prediction.

Table 2 shows the six most significant digits of the absolute error $\mid \boldsymbol{V} - \boldsymbol{WH} \mid$. Certain values stand out. Looking at the column sums, the total error for Trump is lower than for other leaders, suggesting that it is easier to predict Trump's approval rating; By contrast, Xi is the hardest to predict from the data given. Looking at the row sums, it is easier to predict Sweden's total approval ratings (the lowest total error) than Greece's (the largest total error).

	Trump	Merkel	Macron	Putin	Xi	Sum
Canada	0.000421	0.040766	0.042053	0.001000	0.000094	0.084334
Poland	0.002232	0.017789	0.014154	0.006063	0.011906	0.052144
Slovakia	0.005160	0.047285	0.080402	0.102067	0.140497	0.375411
Hungary	0.006603	0.040240	0.030325	0.050302	0.066887	0.194357
Italy	0.004703	0.067974	0.060590	0.045019	0.057232	0.235518
U.K.	0.001295	0.006821	0.018179	0.017970	0.030577	0.074841
Czechia	0.002753	0.034909	0.052693	0.052622	0.073263	0.216239
Bulgaria	0.004120	0.067967	0.055284	0.087431	0.104498	0.319300
Greece	0.003804	0.045946	0.073664	0.118846	0.150552	0.392812
Netherlands	0.002093	0.025643	0.003566	0.029612	0.053769	0.114683
Spain	0.000583	0.020262	0.015478	0.006965	0.012331	0.055619
France	0.002479	0.081234	0.064169	0.030646	0.050220	0.228747
Sweden	0.000858	0.012733	0.001390	0.008841	0.020963	0.044784
Germany	0.001977	0.052199	0.077586	0.038336	0.065768	0.235867
Ukraine	0.005245	0.056918	0.091449	0.029068	0.068887	0.251567
Russia	0.001703	0.033264	0.052042	0.018737	0.035391	0.141137
Philippines	0.003743	0.059344	0.060774	0.019528	0.021618	0.165007
India	0.007070	0.004651	0.022714	0.070164	0.095368	0.199967
S. Korea	0.002989	0.015650	0.002749	0.018420	0.034304	0.074111
Japan	0.008227	0.068092	0.034522	0.056075	0.097873	0.264789
Australia	0.003308	0.051048	0.033523	0.032614	0.053393	0.173887
Indonesia	0.000806	0.024939	0.026252	0.004305	0.003644	0.059946
Israel	0.003001	0.063445	0.069711	0.000372	0.002810	0.139339
Lebanon	0.001787	0.061679	0.057712	0.026851	0.032410	0.180441
Tunisia	0.001385	0.023430	0.036232	0.039778	0.053738	0.154562
Turkey	0.000269	0.036180	0.041186	0.009446	0.013982	0.101063
Kenya	0.011460	0.042409	0.019169	0.104239	0.138763	0.316039
Nigeria	0.010771	0.042079	0.017955	0.117410	0.152098	0.340313
S. Africa	0.007457	0.023838	0.003764	0.102074	0.129409	0.266542
Brazil	0.002587	0.003521	0.015552	0.026921	0.041484	0.090064
Argentina	0.002479	0.014717	0.005339	0.047983	0.059675	0.130193
Mexico	0.002667	0.005159	0.006701	0.050702	0.065051	0.130280
Sum	0.116036	1.192130	1.186878	1.370406	1.938453	

Table 2: The absolute error $|\boldsymbol{V} - \boldsymbol{W}\boldsymbol{H}|$, with entity and proposition labels, and sums of rows and columns.

5.2 Clustering

The model selection method gave 100% consistent clustering over 100 runs. The three consistent clusters of world leaders consisted of {Trump}, {Macron, Merkel}, and {Putin, Xi}. This clustering does not seem to reflect clusters of absolute approval ratings across countries, but a more subtle combination of relative ratings between countries.

The model selection method was also used to find clusters among countries using matrix $\boldsymbol{W}$, from the observation that if $\boldsymbol{W}_{jk} > \boldsymbol{W}_{ji}$ for all $i \neq j$, this suggests that entity j belongs to cluster k. Not all entries of the consistency matrix showed 100% consistent clustering over 100 runs, but most countries were in clusters with 100% consistency. For example, one cluster was {U.K., Canada, Netherlands, Spain, France, Sweden, Germany, Australia}, and another cluster was {Bulgaria, Greece, Russia, Indonesia, Turkey, Argentina, Mexico}. There is clearly some connection of countries within these clusters, for example the first cluster are Western European or Anglophone countries. While one might expect these countries to have similar world leader approval ratings, it is not the purpose of this paper to draw any specific conclusions about political leaders and countries.

6 Conclusion

We have shown how the use of non-negative matrix factorization can be used together with a model selection mechanism for identification of latent features in data. As an example, we have used convex NMF together with a model selection policy based on consensus clustering to analyse public data about approval ratings of world leaders. The insights provided by this preliminary study are currently being used for discovering latent information about relationships between individuals.

Acknowledgements

This work is supported by a grant on Understanding Spiritual Intelligence from the Templeton World Charity Foundation to the International Society for Science and Religion, TWCF0542.

References

[1] Jean-Philippe Brunet et al. “Metagenes and molecular pattern discovery using matrix factorization”. In: *PNAS* 101.12 (2004), pp. 4164–4169.

[2] C. Ding, T. Li, and M.L. Jordan. “Convex and semi-nonnegative matrix factorizations”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32 (2010), pp. 45–55.

[3] João Pedro Silva Francisco. “Football and politics : preference for football teams in Portugal correlates with political party identification”. MA thesis. Universidade Católica Portuguesa, 2017.

[4] D.D. Lee and H.S. Seung. “Algorithms for non-negative matrix factorisation”. In: *Proceedings of the 13th International Conference on Neural Information Processing Systems*. Ed. by T. Leen, T. Dietterich, and V. Tresp. Vol. 13. MIT Press, 2000, pp. 535–541.

[5] Pew Research. *Spring 2019 Global Attitudes Survey, Q38a-e,* 2020. URL: `https://www.pewresearch.org/global/2020/01/08/trump-ratings-remain-low-around-globe-while-views-of-u-s-stay-mostly-favorable/`.

[6] Thanaporn Sriyakul, Anurak Fangmanee, and Kittisak Jermsittiparsert. “Whether loyalty to a football club can translate into a political support for the club owner: An empirical evidence from Thai League”. In: *Journal of Politics and Law* 11.3 (2018), pp. 47–52.