Non Negative Matrix Factorization using separability assumption for Topic Modeling

Durga Harish Dayapule
Oregon State University
dayapuld@oregonstate.edu

1. Abstract

Non Negative matrix factorization is a well studied technique for Topic Modeling problems. Solving this problem is considered as NP-Hard (i.e) no polynomial time solution. Recent work on solving this problem based on separability condition makes this problem to solve in polynomial time. In this project, a) I formally introduce the mathematical formulation of NMF problem and how it is used to solve topic modeling problem. b) Explain the separability assumption c) Replicate the results for using XRAY algorithm with synthetic dataset and four real world dataset which were not considered in the main paper. This report further include run time analysis of the algorithm in parallel mode. The XRAY algorithm shows a good anchor recovery rate and robustness to the noise. I have used cvxpy package in python to implement this algorithm.

2. Introduction

The challenge to develop tools which can comprehend data from web pages, newspaper articles, images, user rating has been a well studied problem in machine learning field. Topic modeling is an approach that has proved successful in all the aforementioned problems. In order to learn structure one has to posit the existence of structure, and in topic models one assumes a generative model for a collection of documents. Specifically, each document is represented as a vector of word-frequencies (the bag of words representation). Papers in theoretical CS [3] and machine learning [2] suggested that documents arise as a convex combination of (i.e. distribution on) a small number of topic vectors, where each topic vector is a distribution on words (i.e. a vector of word-frequencies). Each convex combination of topics thus is itself a distribution on words, and the document is assumed to be generated by drawing N independent samples from it. Subsequent work makes specific choices for the distribution used to generate topic combinations the well-known Latent Dirichlet Allocation (LDA) model of [4] hypothesizes a Dirichlet distribution (see Section 4). For example, fitting to a corpus of newspaper articles may reveal 50 topic vectors corresponding to, say, politics, sports, weather, entertainment etc., and a particular article could be explained as a (1/2, 1/3, 1/6)-combination of the topics politics, sports, and entertainment.

The work of [3] which states the problem of topic modeling as, there is an unknown topic matrix W with non-negative entries with dimension m x r and stochastically generated unknown matrix H with dimension (r x n). Each column of X = WH can be viewed as a probability distribution on columns of matrix W with weights corresponding to rows in matrix H. The problem of finding non-negative matrices W, H with a small inner-dimension r is called non-negative matrix factorization (NMF) and this problem is Np-hard [6].

Lets formalize the definition of NMF and an elegant property which help to solve this problem in polynomial time. Figure 1 shows geometry of the NMF problem. Each point in the space corresponds to a column vector in data Matrix X. Each column vector in X $\in \mathbb{R}^{\mathrm{mxn}}$ can be viewed as a point $\in \mathbb{R}^{\mathrm{m}}$. A conical combination of vectors $w_1, w_2, w_3 ... w_r$ is $h_1 w_1 + h_2 w_2 + .. h_r w_r$ where $h_i \geq 0 \forall i = 1 ... r$. One can construct basis vectors $w_1, w_2 .. w_r$ as a matrix (W) such that the conical combination of this matrix contains all the columns of X. In other words, $cone(\mathbf{X}) \subset cone(\mathbf{W}) \subset \mathbb{R}^{\mathrm{m}}_{+}$. These kind of polyhedral nesting problems studied in computational geometry are known to be NP-hard. Faced with such results, almost the en- tire algorithmic focus in the NMF literature, has centered on treating the problem as an instance of general non-convex programming, leading to heuristic procedures that lack optimality guarantees beyond convergence to a stationary point of the objective function for approximate NMF. Recently, in a series of papers [6], promising alternative approaches have been developed based on certain *separability* assumption on the data which enables the NMF problem to be solved in polynomial time and assure completeness. Separability property assumes that the columns of matrix W are obtained by selecting columns from data Matrix X.

Geometrically, the assumption states the following: all columns of X reside in a cone generated by a small subset of r columns of X. In algebraic terms, $X = WH = X_A H$ so that the r columns of W are hidden among the columns of X (indexed by an unknown subset of indices A). Equivalently, a corresponding subset of r columns of H happen to constitute the rxridentity matrix. We refer to these columns as anchors [6]. Informally, in the context of topic modeling problems where X is a document-word matrix and W, H are document- topic and topic-term associations respectively, the separability assumption equivalently posits the existence of special anchor words in the vocabulary, whose occurence uniquely identifies the presence of a topic, and whose usage across the corpus is collectively predictive of the usage of all the other words. The separability assumption was investigated earlier by Donoho Stodden [5] in the context of deriving uniqueness conditions for NMF.

3. Problem Formulation

For this project, I replicated the XRAY algorithm from original paper [1]. But, I implemented cyclic Coordinate descent and a parallel computation approach to solve the Matrix regression problem, this is one of the contribution to this project. The Algorithm is as follows: For a data Matrix X = WH can be factorized in to two non-negative matrices W and H, where the columns of W contain the some columns of X, H is a special weight matrix which has two parts (i) having a lower dimension permutation matrix (ii) and a weight matrix. The Intuition behind XRAY algorithm is to find r column vectors in X which contains all the column vectors of X (i.e) in other words the conical combination of r selected vectors has to span all the X column vectors. Figure 1 provides a geometric intuition underlying the XRAY algorithm. The algorithm executes r iterations. In each iteration a new anchor column is identified. This corresponds to expanding current cone one extreme ray at a time, until the entire dataset is eventually contained in the cone defined by the full set of anchors. Figure 2 illustrates one step of the algorithm where there is an existing cone defined by three extreme rays (marked 1 to 3).

To identify the next extreme ray, the algorithm picks a point outside the current cone (a green point) and projects it to the current cone to compute a residual vector (this is called projection step). This residual vector separates the current cone from at least one non-selected extreme ray that can be found by maxi- mizing a specific selection criteria (this is called detection step). Intuitively, the algorithm picks a face of the current cone (spanned by rays 1 and 3 in Figure 2) that sees exterior points and rotates this face towards the exterior until it hits the last point. In the example shown in Figure 2, ray 4 is identified as a new extreme ray.

A Background of Cones, Extreme Rays: Recall that a cone C is a non-empty convex set that is closed with respect to taking conic combinations (i.e., linear combinations with non-negative coefficients) of its elements. A ray in C generated by a vector $x \in C$ is the set of all vectors $\{tx : t \geq 0\}$. A ray R is an extreme ray if its generators cannot be expressed by taking conic combinations of elements in C that do not themselves belong to R. A cone is called finitely generated if its elements are conic combinations of a finite set of vectors, and pointed if it does not contain both a vector x as well as its negation x.

Furthermore, the generators of these extreme rays are a subset of the finite set of vectors used to originally express the cone. In the NMF context, note that any cone contained in Rm+ is pointed. This implies that cone(X) can also be described by a minimally compact set of generators, i.e., $cone(X) = cone(X_A)$ where A uniquely indexes the extreme rays (anchors). Thus, a non-negative matrix X admits a separable NMF with inner-dimension r if the number of extreme rays of cone(X), i.e. size of A, coincides with r. A face of a cone is the intersection between the cone and a supporting hyperplane.

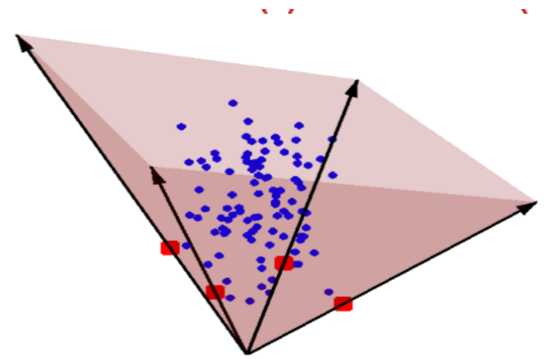

Figure 1. Geometry of NMF Problem, The red dots indicate the basis vectors for conical hull, The colored region is a cone.

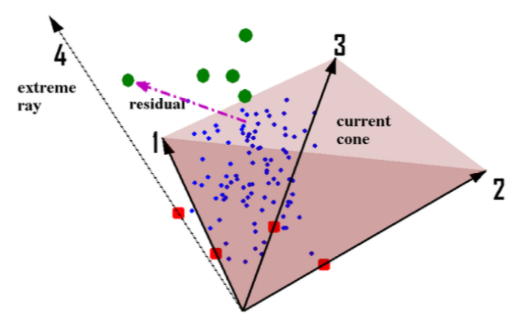

Figure 2. Figure Depicts one Step update process for XRAY algorithm. The Green dot indicates the points outside the current cone which is in colored region.

Algorithm

Algorithm 1 details the steps for Xray Algorithm. Each iteration consists of two steps: (i) a detection step: This

ALGORITHM 1: XRAY : Algorithm for Separable NMF

Input: $X \in \mathbb{R}_{+}^{\mathrm{mxn}}$, Topic Dimension r
Output: Topic Matrix : $W \in \mathbb{R}_{+}^{\mathrm{mxr}}$, Weight Matrix : $H \in \mathbb{R}_{+}^{\mathrm{rxn}}$, where r indices are selected from matrix X such that : X = WH
Initialize : Residual Matrix : $R \leftarrow X$, Index Set: $A \leftarrow Set(\{\})$
while $|A| < r$ **do**
  1. Detection Step : Finding an Extreme Ray which is not in current Cone ;
    $j^* = \arg\max_j \frac{R_i^T X_j}{p^T X_j}$ for any i : $||R_i||_2 > 0$;
  Exterior point Selection : ;
    max : i = $\arg\max_j ||R_k||_2$;
    dist: i = $\arg\max_j ||(R_k^T X)_+||_2$;
  2. Update the Index Set : $A \leftarrow A \cup \{j^*\}$;
  3. Projection Step : Project onto the Current Cone.
    H = $\arg\min_{B \geq 0} ||X - X_A B||_2^2$ (ADMM)
  4. Update Residual Matrix : $R = X - X_A H$
end

step finds a column(s) of X to be added as an anchor, and (ii) a projection step: In which all data points are projected onto the current cone to get the residuals. Projection is done by solving simultaneous nonnegative least squares problem. The least square problem is solved in a parallel mode using cvxpy software. Every residual vector R_i obtained after the projection step is normal to one of the faces of the current cone. In the selection step, we pick a face of the current cone (identified by its normal R_i), normalize all the data points to lie on the hyper plane $p^T x = 1$ $Y = \frac{Xj}{p^T X_j}$ for a strictly positive vector p. In this report I selected to choose $p^T = [1, 1, 1, ...]$, and expand the current cone by selecting an extreme ray that maximizes the inner product $R_i^T Y_j$. In the selection step, the choice of i can be implemented in various ways. For this project I choose two approaches (a) max : i = $\arg\max_j ||R_k||_2$, choosing the maximum residual vector (b) dist: i = $\arg\max_j ||(R_k^T X)_+||_2$. So, Rest of the report uses these two approaches to solve the problem.

Experiments & Results

Three Variants for choosing the i in selection were implemented at code level (max, dist, rand). But for the purpose of results only two are shown namely max, dist. The experiments were done both in synthetic Data and Real world Datasets. The next section talks about the experimental Setup and Results.

Synthetic Experiments:

Synthetic experiments were carried out by constructing W $\in \mathbb{R}_{+}^{\mathrm{210xr}}$ and H $\in \mathbb{R}_{+}^{\mathrm{rx200}}$ matrices with varying inner dimension r $\in$(10, 20, 30). Each entry of matrix W is generated by i.i.d uniform distribution between 0 and 5. The matrix H is decomposed into two parts $H_1 \in I_{rXr}$ (Identity Matrix) and $H_2 \in \mathbb{R}_{+}^{\mathrm{rx200-r}}$. Each column of H_2 is generated according to i.i.d uniform distribution between 0 and 1. The data Matrix is set to X = WH + N, where N is the controlled noise. Each Entry in N is a i.i.d Gaussian with mean zero and standard deviation δ, the range of δ is chosen from 0 to 1.4. Table 1 shows the anchor recovery rate for inner dimension of r = 10, 20, 30. Both variants of the algorithm shows noise - robustness in terms of anchor recovery, For most of noise level δ the algorithms were able to recover anchor vectors. As the noise level increased to higher level, there is a slight performance reduction in anchor recovery. Table 2 shows the run time in seconds for the algorithms, On average for recovering 10 anchors the algorithm took 4 seconds compared to 20 which is 15 seconds and anchors 30 45 seconds. This suggests that the algorithm could be scalable to higher dimension data.

Real Data Experiments

I applied the algorithm to 4 new real datasets to recover the topics in the corpus, these datasets were not used in original paper. This can be considered as major contribution to this report. Table 3 provides the details of the dataset for topic modeling. The value of r in the Table 3 indicates the number of topics present in the corpus. So, this value of r is chosen for running the algorithm. The datasets have been preprocessed by removing stop-word and low term frequency filtering (count < 20), then log TF-IDF and L2 document length normalization. For the four datasets, a document -term matrix were constructed. Table 5 6 7 8 shows the ranking of leading words of the mined topics for bbc, bbcsport, guardian, irishtimes data set. It is clear from these tables that XRAY algorithm could not able to recover all the anchor vectors. For example in case of irishtimes article table 8, the topics like politics are repeated more than once, I suspect this can be due to (a) both of these documents are on the same extreme ray (b) The documents are noisy. This kind of effect is even seen for all the real world datasets. Its is evident that XRAY method fails to give clear topics with real world noisy data. This experimental results suggest that a robust anchor algorithm has to be designed.

I applied the algorithm to 4 new real datasets to recover the topics in the corpus, these datasets were not used in original paper. This can be considered as major contribution to this report. Table 3 provides the details of the dataset for topic modeling. The value of r in the Table 3 indicates the number of topics present in the corpus. So, this value of

Table 1. Show the Anchor Recovery rate for data matrix X $\in \mathbb{R}_{+}^{210\times200}$ using two variants of XRAY algorithm (max,dist)

Noise Level δ	r = 10		r = 20		r = 30	
	Max	Dist	Max	Dist	Max	Dist
	Anchor Recovery	Anchor Recovery	Anchor Recovery	Anchor Recovery	Anchor Recovery	Anchor Recovery
0	1.000	1.000	1.000	1.000	1.000	1.000
0.2	1.000	1.000	1.000	1.000	1.000	1.000
0.4	1.000	1.000	1.000	1.000	1.000	1.000
0.5	1.000	1.000	1.000	1.000	1.000	1.000
0.6	1.000	1.000	1.000	1.000	1.000	1.000
0.8	1.000	1.000	0.998	1.000	0.998	0.998
1	0.995	0.995	0.995	0.998	0.993	0.990
1.2	0.995	1.000	0.990	0.993	0.978	0.975
1.4	0.980	0.985	0.971	0.958	0.967	0.955

Table 2. XRAY Algorithm run Time for noise injected data matrix X $\in \mathbb{R}_{+}^{210\times200}$

Noise Level δ	r =10		r =20		r=30	
	Max	Dist	Max	Dist	Max	Dist
	Run Time(Sec)	Run Time(Sec)	Run Time(Sec)	Run Time(Sec)	Run Time(Sec)	Run Time(Sec)
0	4.101	3.903	15.034	17.332	42.366	43.806
0.2	3.934	3.811	16.480	16.978	42.140	42.925
0.4	3.890	3.731	15.082	15.120	38.209	39.118
0.5	3.704	3.495	13.913	14.745	36.705	41.346
0.6	3.592	3.491	16.092	14.238	58.629	36.560
0.8	3.473	3.414	14.610	14.077	34.489	35.025
1	3.551	3.299	13.357	14.782	35.263	36.559
1.2	3.520	3.441	13.245	14.611	39.143	42.159
1.4	3.661	3.545	15.359	14.234	59.311	54.684

r is chosen for running the algorithm. The datasets have been preprocessed by removing stop-word and low term frequency filtering (count $<$ 20), then log TF-IDF and L2 document length normalization. For the four datasets, a document -term matrix were constructed. Table 5 6 7 8 shows the ranking of leading words of the mined topics for bbc, bbcsport, guardian, irishtimes data set. It is clear from these tables that XRAY algorithm could not able to recover all the anchor vectors. For example in case of irishtimes article table 8, the topics like politics are repeated more than once, I suspect this can be due to (a) both of these documents are on the same extreme ray (b) The documents are noisy. This kind of effect is even seen for all the real world datasets. Its is evident that XRAY method fails to give clear topics with real world noisy data. This experimental results suggest that a robust anchor algorithm has to be designed.

Conclusion

XRAY algorithm with two variants (max, dist) were implemented for synthetic and real world data set . XRAY shows good noise robustness for synthetic datasets with good anchor recovery rates. XRAY algorithm was used to identify topics in bbc, bbcsport, guardian, irishtimes real world datasets. Experiment on these datasets shows that the algorithm couldn' t able to recover all the anchor vectors, some vectors have same topic. This suggest that XRAY approach has draw backs while it comes to noisy real world dataset.

References

[1] P. K. Abhishek Kumar, Vikas Sindhwani. Fast conical hull algorithms for near-separable non-negative matrix factorization. *ICML*, pages 1735–1780, 2013.

[2] H. T. C. Papadimitriou, P. Raghavan and S. Vempala. Probabilistic latent semantic analysis. *UAI*, page 289296, 1999.

[3] H. T. C. Papadimitriou, P. Raghavan and S. Vempala. Latent semantic indexing: a probabilistic analysis. *JCSS*, page 217235, 2000.

[4] A. N. D. Blei and M. Jordan. Probabilistic latent semantic analysis. *JMLR*, page 9931022, 2003.

[5] D. Donoho and V. W. Stodden. does non-negative matrix factorization give a correct decomposition into parts. *NIPS*, 2003.

[6] R. K. S. Arora, R. Ge and A. Moitra. Computing a nonnegative matrix factorization provably. *STOC*, 2012.

Table 3. Details of the corpora used in the experiments, including the total number of documents n, words m, and number of anchors r.

Corpus	**n**	**m**	**r**	**Desciption**
bbc	2,225	3,121	5	Genearal News Articles from the BBC
bbc-Sport	737	969	5	Sports News Articles from the BBC
guardian-2013	6520	10801	6	News Articles published by The Guardian
irishtimes-2013	3,246	4,832	6	News Articles published by Irish Times

Table 5. Top 10 terms for reference ranking sets generated by XRAY Algorithm on **bbc Dataset** with **m=3121, n=2225 r = 5**, . The Topic names are found with the help of words for each topic

	Max					Dist				
Topic	**Sports**	**Sports**	**Business**	**Politics**	**Entertainment**	**Sports**	**Politics**	**Politics**	**Entertiment**	**Tech**
1	leicester	serve	complaints	indonesia	wonder	france	yukos	blair	best	music
2	bath	australian	id	relief	character	game	oil	labour	film	mobile
3	england	break	fraud	tsunami	woman	french	bankruptcy	brown	award	phones
4	sale	open	identity	offer	film	ireland	russian	election	british	mobiles
5	robinson	melbourne	theft	effort	female	important	auction	prime	actress	sales
6	newcastle	win	credit	offered	series	wales	control	cabinet	category	replacement
7	lock	assessment	consumers	government	silver	impressed	court	party	finding	design
8	squad	leap	total	royal	write	nations	russia	government	mind	phone
9	andy	guys	internet	downing	feature	focused	state	mps	actor	markets
10	row	ease	someone	operation	produced	half	firm	way	movie	people

Table 6. Top 10 terms for reference ranking sets generated by XRAY Algorithm on **bbcsport** Dataset with **m=969, n=737, r = 5**. The Topic names are found with the help of words for each topic

	max					dist				
Topic	**Cricket**	**athletics**	**Football**	**Football**	**Football**	**UnKnown**	**Cricket**	**Tennis**	**Athletics**	**FootBall**
1	series	seed	penalty	fa	southampton	nations	south	open	indoor	chelsea
2	sri	felt	williams	sports	club	victory	africa	quarter	record	league
3	chris	broke	net	federation	linked	australia	de	round	world	united
4	match	strong	subs	suspended	manager	france	england	seed	ran	champions
5	australia	take	james	body	jones	england	vaughan	federer	olympic	season
6	squad	favourite	minutes	move	tuesday	world	series	finals	champion	arsenal
7	day	thomas	penalties	football	boss	wales	andrew	france	ireland	manchester
8	craig	taylor	scott	pay	return	coach	jacques	meet	mark	premiership
9	scott	reached	corner	committee	former	ireland	runs	williams	time	west
10	shoulder	suffered	hosts	remains	player	title	jones	roddick	excellent	cup

Table 7. Top 10 terms for reference ranking sets generated by XRAY Algorithm on **guardian Dataset** with **m= 10801, n= 6520, r = 6**. The Topic names are found with the help of words for each topic

	max						**dist**					
Topic	**Music**	**Music**	**Music**	**Flights**	**Politics**	**Sports**	**Business**	**Politics**	**Fashion**	**Books**	**FootBall**	**FootBall**
1	steve	orchestra	album	flights	bowie	wives	profit	thatcher	fashion	novel	manchester	charles
2	band	event	entirely	airlines	vuitton	sua	growth	margaret	capturing	james	universally	leeds
3	jorge	events	requisite	boeing	muse	rez	ultimately	speaker	footwear	book	middlesbrough	player
4	swallow	free	comprised	houston	invitation	sport	serco	political	somerset	says	bayer	universally
5	saxophonist	productions	hurts	flight	advert	sports	centrica	tributes	heels	awesome	leverkusen	cup
6	cheek	ensemble	runner	incidents	louis	behaviour	price	disagreed	london	write	portsmouth	rangers
7	harmony	quartet	stevie	plane	campaign	field	shares	directness	bag	burning	watford	season
8	organ	symphony	compulsory	flying	singing	man	warmly	herself	pink	hunters	city	transfer
9	remarkably	opera	engineer	grounded	model	cup	analysts	politics	week	light	norwich	village
10	rhythms	scotland	commodity	tokyo	modelling	handball	government	clegg	shoes	flew	rooney	football

Table 8. Top 10 terms for reference ranking sets generated by XRAY Algorithm on **irishtimes Dataset** with **m=4832 , n= 3246 r = 6**. The Topic names are found with the help of words for each topic

	max						**dist**					
Topic	**Health**	**Marketing**	**Sports**	**Politics**	**Politics**	**Politics**	**Health**	**Sports**	**Sports**	**Music**	**Politics**	**Politics**
1	illness	consumers	scarlets	jobs	european	stone	health	keane	goal	leinster	votes	exit
2	disease	ways	connacht	shatter	institutions	armagh	sources	achilles	league	ulster	seanad	decision
3	heart	day	december	welcomed	ombudsman	discovered	maintained	neill	visitors	mcgrath	referendum	precautionary
4	mental	getting	saturday	plant	citizens	trial	initiatives	robbie	eto	schmidt	dublin	euro
5	exercise	practical	edinburgh	garda	reilly	brain	service	poland	premier	heaslip	referendums	ireland
6	rates	lunchtime	ulster	senator	union	walked	budget	donall	side	jackson	castle	european
7	medication	lots	ravenhill	leader	voice	heard	maternity	republic	arsenal	connell	campaign	imf
8	smoking	offers	glasgow	business	luxembourg	court	insurance	get	hazard	samoa	donaill	programme
9	death	suggestions	dragons	announcement	officially	injuries	pressures	captain	home	hip	count	zone
10	factors	consumer	murrayfield	minister	renewed	women	cuts	surgery	gave	mcfadden	cent	bailout