NONNEGATIVE MATRIX FACTORIZATION USING A ROBUST ERROR FUNCTION

Chris Ding and Deguang Kong

Department of Computer Science & Engineering, University of Texas at Arlington, Arlington, TX, 76013

ABSTRACT

Nonnegative matrix factorization (NMF) is widely used in image analysis. However, most images contain noises and outliers. Thus a robust version of NMF is needed. We propose a novel NMF using a robust error function which smoothly interpolates between the least squares at small errors and L_1-norm at large errors. An efficient computational algorithm is derived with rigorous convergence analysis. Extensive experiments are made on six image datasets to show the effectiveness of proposed approach. Robust NMF consistently provides better reconstructed images, and better clustering results as compared to standard NMF.

Index Terms— NMF, robust, error function, clustering

1. INTRODUCTION

Nonnegative Matrix Factorization (NMF) has been popularly studied in data mining and machine learning areas since the initial work of Lee and Seung [1]. As originally proposed method for finding matrix factors with parts-of-whole interpretations, NMF has been applied to a number of applied areas, environmetrics, chemometrics [2], pattern recognition, multimedia data analysis and text mining. Algorithmic extensions of NMF have been developed to accommodate a variety of objective functions and a variety of data analysis problems, including classification, collaborative filtering, etc. One of the key features of NMF is its clustering capabilities. It is shown [3] that NMF essentially solves a matrix clustering problem.

Standard NMF uses the least square error function which is well-known to be non-robust w.r.t. noises and outliers [4, 5]. On the other hand, many real life data contain noises and outliers [6]. For this reason, a robust NMF model is needed. We analyze the error models and propose a characterization of desired error function, which interpolates between the least squares for smaller white noises and L_1 norm (absolute value) for large errors. We propose a specific error function with the desired behaviors at both limits. We use this robust error function for NMF and derive an efficient computational algorithm and provide rigorous analysis on its convergence. Extensive experiments on six image datasets demonstrate the usefulness of robust NMF on image analysis and image clustering.

2. STANDARD NMF REVISIT

Given input data vectors $X = (x_1, \cdots, x_n)$, where $x_i \in \Re^p$ represents an image (a vector of image features or linearized pixels). The standard NMF is defined as

$$\min_{F,G} \ \|X - FG\|_F^2, \ \ s.t. \ F \geq 0, \ G \geq 0, \tag{1}$$

where $\|X\|_F^2 = \sum_{ij} X_{ij}^2$ is the Frobenious form of a matrix.

One of the most important drawback of the standard NMF is that it is prone to large derivations (outliers and noises at image level), because the error for both data features (index j) and data vector (index i) are **squared**. Thus a few noisy features or a few outliers with large errors easily dominate the objection function because the errors are squared. Below, we first discuss Gaussian white noise and large deviations and then propose a error function which has the correct behaviors in both sides.

3. STATISTICAL DISTRIBUTIONS FOR WHITE NOISE AND LARGE DERIVATIONS

For simplicity, we first consider the case where the observations $(x_1, \cdots, x_n)$ are **scalars**. Observed value x_i can be contaminated by noisy features and outliers, i.e., $x_i = \theta + \varepsilon_i$, where θ is the unobservable true value, and ε_i is the additive noise. Different distributions of ε_i define different error models.

The most common noise is white noise, which follows the zero-mean normal distribution with standard deviation of σ,

$$p(x_i) \sim \exp\{-\frac{(x_i - \theta)^2}{2\sigma^2}\}, \tag{2}$$

and the error function (negative log-likelihood) is

$$-\log[\Pi_{i=1}^n p(x_i)] \propto \sum_{i=1}^n (x_i - \theta)^2. \tag{3}$$

White noises generally have small magnitudes. Outliers and large deviations are usually modeled by the Laplacian distribution with zero mean,

$$p(x_i) \sim \exp\{-\frac{|x_i - \theta|}{\sigma}\}, \tag{4}$$

and the error function (negative log-likelihood) is

$$-\log[\Pi_{i=1}^n p(x)] \propto \sum_{i=1}^n |x_i - \theta|. \tag{5}$$

In real life data, smaller and frequent white noises are usually mixed with infrequent but large rare deviations. This suggests a good error function should cover both the smaller white noise and the larger Laplacian-type deviations, i.e., $f(\cdot)$ should have the following desired behaviors on residue $r = |x - \theta|$,

$$f(r) \to \begin{cases} r^2 & \text{if } r \ll \sigma, \\ r & \text{if } r \gg \sigma, \end{cases} \tag{6}$$

where σ is the fixed quantity close to the variance of x. This characterization of the error function is a key point of this paper: the error function should smoothly interpolate between the least square function for small errors and L_1-norm for large errors.

3.1. A smooth robust error function

In this paper, we propose a novel robust error function $\Gamma(\cdot)$ on residue r which has the correct characteristics in both limits,

$$\Gamma(r) = \sigma\sqrt{r^2 + \sigma^2} - \sigma^2. \tag{7}$$

It is easy to check that this function has the right asymptotic behaviors:

$$\Gamma(r) \to \begin{cases} \frac{1}{2}r^2 & \text{if } r \ll \sigma, \\ \sigma r - \sigma^2 & \text{if } r \gg \sigma. \end{cases} \tag{8}$$

We note there exist other error functions which take account of these considerations. The most well-know one is Huber M-estimator [7]:

$$H(r) = \begin{cases} \frac{1}{2}r^2 & \text{if } r \le \sigma, \\ \sigma r - \frac{1}{2}\sigma^2 & \text{if } r > \sigma, \end{cases} \tag{9}$$

where σ is a positive number. Huber function is fairly close to $\Gamma(\cdot)$ function, except that $H(\cdot)$ consists of **two** smooth function fused together in an *ad hoc* way while $\Gamma(\cdot)$ is a **single** smooth twice differentiable function. Computationally, $H(\cdot)$ must test the range condition; such a *branching* operation typically takes hundreds of CPU clock cycles, while *add, multiply* take only 1-2 clock cycles. Thus computationally, $\Gamma(\cdot)$ function is much faster than $H(\cdot)$. Another robust function is the Beaton-Tukey function

$$F_{BT}(r) = \begin{cases} \frac{1}{6}\sigma^2[1 - \left(1 - \left(\frac{r}{\sigma}\right)^2\right)^3] & \text{if } r \le \sigma, \\ \frac{1}{6}\sigma^2 & \text{if } r > \sigma. \end{cases} \tag{10}$$

A somehow similar error function is the Cauchy function:

$$C(r) = \frac{\sigma^2}{2}\log[1 + (\frac{r}{\sigma})^2]. \tag{11}$$

At small r, both Beaton-Tukey and Cauchy functions approach $r^2/2$, same as $\Gamma(\cdot), H(\cdot)$. At large r, however, both Beaton-Tukey and Cauchy functions grow much slower than linear growth of $\Gamma(\cdot), H(\cdot)$.

3.2. Error function for vector data

In above discussions, x_i are assumed to be scalar observations. Many data come in as vector data, say, a p-dimensional vector data set $(\mathbf{x}_1, \cdots, \mathbf{x}_n)$. Let $\mathbf{r}_i = \mathbf{x}_i - \theta$ be the residue of $\mathbf{x}_i$. We propose two forms of robust error function.
(1) The element-wise error function

$$\Gamma(\mathbf{r}) = \sum_{j=1}^{p} \sigma(\sqrt{r_j^2 + \sigma^2} - \sigma), \tag{12}$$

where r_j is the j-th component (element) of vector $\mathbf{r}$.
(2) The q-norm robust function

$$\Gamma_q(\mathbf{r}) = \Gamma(\|\mathbf{r}\|_q) = \sigma\sqrt{\|\mathbf{r}\|_q^2 + \sigma^2} - \sigma^2, \tag{13}$$

where $\|\mathbf{r}\|_q = (\sum_j |r_j|^q)^{1/q}$ is the L_q-norm of vector $\mathbf{r}$.

Using the element-wise definition of Eq.(12), the robust error function of the entire input data X is, $R_{ij} = |X_{ij} - \theta_{ij}|$,

$$J = \Gamma(R) = \sum_{i=1}^{n}\sum_{j=1}^{p} \sigma(\sqrt{R_{ji}^2 + \sigma^2} - \sigma). \tag{14}$$

We mention here that for matrix variables, besides the standard L_2 or Frobenius norm error function $\|R\|_F = \sqrt{\sum_{i=1}^{n}\sum_{j=1}^{p} R_{ji}^2}$, there are two robust error functions: (1) the L_1 error function $\|R\|_1 = \sum_{i=1}^{n}\sum_{j=1}^{p} R_{ji}$ and (2) the $L_{2,1}$ error function $\|R\|_{2,1} = \sum_{i=1}^{n}\sqrt{\sum_{j=1}^{p} R_{ji}^2}$ [8, 9].

4. ROBUST NMF USING THE SMOOTH ERROR FUNCTION

The robust error function Eq.(14) leads to a new formulation of NMF,

$$J(F, G) = \sum_{i=1}^{n}\sum_{j=1}^{p} \sqrt{(X - FG)_{ji}^2 + \sigma^2}, \tag{15}$$

where the constant term $\sum_{ij} \sigma^2$ and the proportional constant σ are both ignored. Formally, this novel NMF is formulated as

$$\min_{F,G} \; J(F, G) \;\; s.t. \;\; F \ge 0, \; G \ge 0. \tag{16}$$

A main contribution of this paper is to derive the following updating algorithms

$$F_{jk} \Leftarrow F_{jk}\frac{[X \circ WG^T]_{jk}}{[(FG) \circ WG^T]_{jk}}, \tag{17}$$

$$G_{ki} \Leftarrow G_{ki}\frac{[F^T X \circ W]_{ki}}{[F^T(FG) \circ W]_{ki}}, \tag{18}$$

where W is a matrix given by

$$W_{ij} = \left((X - FG)_{ij}^2 + \sigma^2\right)^{-1/2}, \tag{19}$$

and $\circ$ is the Hadamard product, i.e., elementwise product between two matrices. Here we assume Hadamard product has higher operator precedence over regular matrix product, i.e., $AB \circ CD = A(B \circ C)D$.

Generally it is harder to solve this robust NMF due to its complicated formulation. Thus it is a bit surprising that our algorithm updating rules of Eqs.(17,18) for robust NMF is very similar to the updating rules for standard NMF of Eq.(1). [In fact, if we set $W_{ij} = 1$, these two algorithms are identical.] Both are very simple to implement and have nearly identical computational complexity.

4.1. Illustration on image reconstruction

As discussed above, robust NMF is most useful for noises within each images. This is best illustrated by considering occlusion as noises. On AT&T dataset, we add occlusions to 5 of the 10 images of each person (details are given in Section 7). The occluded images are shown in Fig. 1.

We run robust NMF and standard NMF on the entire dataset (400 images). We use the computed F and G to reconstruct the corresponding original images. Due to space limit, we show only 20 images for two persons in Fig. 1. Clearly, robust NMF results are generally better than those of standard NMF. Many of the occluded blocks are removed in reconstructed images from robust NMF while they remain on the constructed images from standard NMF. For example, on the images of second person, "glasses" in the original images are preserved in robust NMF results while most of them are lost in standard NMF results. Furthermore, the robust NMF has some corrective effects. On images of the first person, two reconstructed images (4th and 6th images from left) using robust NMF are corrected to the proper (vertical) orientation.

5. CONVERGENCE OF THE ALGORITHM

Here we present main analysis results (detailed proofs are skipped due to space limit). The main result is

Theorem 1. *(A) Updating G using the rule of Eq.(18) while fixing F, the objective function of Eq.(15) monotonically decreases. (B) Updating F using the rule of Eq.(17) while fixing G, the objective function of Eq.(15) monotonically decreases.*

Here we list two key lemmas in proving Theorem 1A for updating G while fixing F.

Lemma 1. *Let G^t be the old G [on the RHS of Eq.(18)] and G^{t+1} be the new G [on the LHS of Eq.(18)]. Under the updating rule of Eq.(18), the following holds*

$$\|X - FG^{t+1}\|_W^2 \leq \|X - FG^t\|_W^2,$$

where W_{ij} is defined in Eq.(19) and $\|A\|_W^2 = \sum_{ij} A_{ij}^2 W_{ij}$.

Lemma 2. *Under the updating rule of Eq.(18), the following holds*

$$J(F, G^{t+1}) - J(F, G^t) \leq \frac{1}{2}\Big(\|X - FG^{t+1}\|_W^2 - \|X - FG^{t+1}\|_W^2\Big). \quad (20)$$

Fig. 1: Reconstruction of occluded AT&T dataset. Experiment is done for all 400 images of 40 subjects. 20 images for 2 persons are shown as 2 panels. On each panel, upper images are original occluded images, middle images are reconstructed from standard NMF, and lower images are reconstructed from robust NMF.

Table 1: Detailed Description of different Data sets

Dataset	# Size	# Dimension	# Class
AT&T	400	2576	40
MNIST	150	784	10
CMUPIE	680	1024	68
UMIST	360	644	20
YALE	1984	2016	31
Bin-alpha	1404	320	36

We can similarly prove Theorem 1B.

6. CORRECTNESS OF THE ALGORITHM

We can prove that the converged solution is the correct optimal solution, i.e., the converged solution satisfies the Karush-Kohn-Tucker condition of constrained optimization theory.

Theorem 2. *At convergence, the converged solution F^* of the updating rule of Eq.(17) satisfies the KKT condition of optimization theory.*

Theorem 3. *At convergence, the converged solution G^* of the updating rule of Eq.(18) satisfies the KKT condition of optimization theory.*

7. EXPERIMENT

We use six widely used image data sets. Table 1 summarizes the characteristics of them. For all image data sets, we use the same raw gray level values as features in the original space without making any changes. All of them only have non-negative values as features. To construct the occluded data

Table 2: Clustering Results of robust-NMF(rNMF) with NMF and K-means on six *original* data sets.

Dataset	Metric	Approaches		
		rNMF	NMF	K-means
AT&T	ACC	**0.6792**	0.6496	0.6519
	NMI	**0.8294**	0.7945	0.8134
	PUR	**0.7600**	0.6822	0.7021
MNIST	ACC	**0.7523**	0.7297	0.6872
	NMI	**0.7298**	0.6966	0.6788
	PUR	**0.7687**	0.7461	0.7068
UMIST	ACC	**0.4973**	0.4861	0.4744
	NMI	**0.6123**	0.5869	0.6030
	PUR	**0.5032**	0.5029	0.5185
CMUPIE	ACC	**0.4278**	0.4138	0.2227
	NMI	**0.6912**	0.6557	0.5386
	PUR	**0.4307**	0.4286	0.2429
YALE	ACC	**0.2419**	0.1950	0.0870
	NMI	**0.3179**	0.2882	0.0933
	PUR	**0.2495**	0.2082	0.0943
Bin-alpha	ACC	**0.3889**	0.2183	0.3342
	NMI	**0.4763**	0.3287	0.5072
	PUR	**0.3946**	0.2091	0.3897

Table 3: Clustering Results of robust-NMF(rNMF) with NMF and K-means on six *occluded* data sets

Dataset	Metric	Approaches		
		rNMF	NMF	K-means
AT&T	ACC	**0.6325**	0.5000	0.6237
	NMI	**0.7972**	0.6652	0.7670
	PUR	**0.6650**	0.6298	0.5250
MNIST	ACC	**0.7613**	0.7400	0.7160
	NMI	**0.7540**	0.7307	0.6907
	PUR	**0.7980**	0.7629	0.7253
UMIST	ACC	**0.4194**	0.3917	0.3623
	NMI	**0.5538**	0.4955	0.4872
	PUR	**0.4417**	0.4056	0.3957
CMUPIE	ACC	**0.3647**	0.3500	0.2097
	NMI	**0.6201**	0.5966	0.5211
	PUR	**0.3838**	0.3676	0.2293
YALE	ACC	**0.1976**	0.1598	0.0912
	NMI	**0.2767**	0.2389	0.0970
	PUR	**0.2072**	0.1704	0.0981
Bin-alpha	ACC	**0.4295**	0.1624	0.3858
	NMI	**0.5530**	0.2891	0.5328
	PUR	**0.4544**	0.1695	0.4163

sets, we first randomly select half of the images from each category of each data set, and then occlude a square block with wxw pixels(e.g., $w = 10$) on the selected ones. The locations of the occlusions are randomly generated. Through this procedure, six occlusion image datasets are generated corresponding to six original data sets.

7.1. Reconstruction for Image Analysis

Nonnegative matrices F and G can be used to reconstruct the original images. As is shown in Fig. 1 on occluded AT&T data set, we compare the reconstructed images by using F and G from both robust NMF and NMF. It is clear to see the robust NMF reconstructed images are much better than those from NMF on occluded data set AT&T. We can get the same results on other data sets no matter whether images are occluded or not. Due to space limit, we did not show all here.

7.2. Clustering Results

We report clustering results by making a comparison with K-means clustering and standard NMF approach. The evaluation metrics we used here are clustering accuracy(ACC), normalized mutual information(NMI), purity(PUR). These measurements are widely used in the evaluation of different clustering approaches. (Higher values of these quantities indicate better clustering results.)

In experiments, σ is set to the median of residues computed from standard PCA. We use K-means as initialization as suggested by theoretical analysis [3]. We average 100 iterations of NMF, robust NMF, and K-means results to get the average of the three metrics(ACC, NMI, PUR) for each approach and show them in Table 2(original data sets) and Table 3(occlusion data sets). We can see that robust NMF performs consistently better than standard NMF and K-means on all data sets, including both original data sets and occluded image data sets.

8. CONCLUSION

We propose a novel NMF using a robust error function which smoothly interpolates between the least squares at small errors and L_1-norm at large errors. We derive an efficient computational algorithm with rigorous convergence analysis. We demonstrate the effectiveness of proposed approach on six image datasets. Robust NMF consistently provides much better reconstructed images, and also better clustering results.

Acknowledgements. This work is partially supported by NSF-CCF-0939187, NSF-CCF-0917274, NSF-DMS-15228.

9. REFERENCES

[1] D.D. Lee and H.S. Seung, "Algorithms for non-negative matrix factorization," in *Advances in Neural Information Processing Systems 13*, Cambridge, MA, 2001, MIT Press.

[2] Y.-L. Xie, P.K. Hopke, and P. Paatero, "Positive matrix factorization applied to a curve resolution problem," *Journal of Chemometrics*, vol. 12, no. 6, pp. 357–364, 1999.

[3] C. Ding, X. He, and H.D. Simon, "On the equivalence of nonnegative matrix factorization and spectral clustering," *Proc. SIAM Data Mining Conf*, 2005.

[4] Q. Ke and T. Kanade, "Robust l_1 norm factorization in the presence of outliers and missing data by alternative convex programming," in *CVPR (1)*, 2005, pp. 739–746.

[5] D. Kong, C. Ding, and H. Huang, "Robust nonnegative matrix factorization using l_{21}-norm," in *CIKM*, 2011, pp. 673–682.

[6] H. Huang and C. Ding, "Robust tensor factorization using r_1 norm," in *CVPR*, 2008.

[7] W. Li and J.J.Swetits, "The linear l_1 estimator and the huber m-estimator," *SIAM Journal on Optimization*, vol. 8, 1998.

[8] C. Ding, D. Zhou, X. He, and H. Zha, "R_1-pca: rotational invariant l_1-norm principal component analysis for robust subspace factorization," in *ICML*, 2006, pp. 281–288.

[9] F. Nie, H. Huang, X. Cai, and C. Ding, "Efficient and robust feature selection via joint $l_{2,1}$-norms minimization," in *NIPS*, 2010, pp. 1813–1821.