Latent Outlier Detection and the Low Precision Problem

Fei Wang
University of Sydney
Sydney, Australia
fwan7957@it.usyd.edu.au

Sanjay Chawla
University of Sydney and NICTA
Sydney, Australia
sanjay.chawla@sydney.edu.au

Didi Surian
University of Sydney and NICTA
Sydney, Australia
didi.surian@sydney.edu.au

ABSTRACT

The identification of outliers is an intrinsic component of knowledge discovery. However, most outlier detection techniques operate in the observational space, which is often associated with information redundancy and noise. Also, due to the usually high dimensionality of the observational space, the anomalies detected are difficult to comprehend. In this paper we claim that algorithms for discovery of outliers in a latent space will not only lead to more accurate results but potentially provide a natural medium to explain and describe outliers. Specifically, we propose combining Non-Negative Matrix Factorization (NMF) with subspace analysis to discover and interpret outliers. We report on preliminary work towards such an approach.

1. INTRODUCTION

It is well known that new scientific discoveries or "paradigm shifts" are often triggered by the need to explain outliers [11]. The availability of large and ever increasing data sets, across a wide spectrum of domains, provides an opportunity to actively identify outliers with the hope of making new discoveries.

The obvious dilemma in outlier detection is whether the discovered outliers are an artifact of the measurement device or indicative of something more fundamental. Thus the need is not only to design algorithms to identify complex outliers but also provide a framework where they can be described and explained. Sometimes it is easy to explain outliers. For example, we applied the recently introduced k-means-- algorithm [4] on the 2012 season NBA player data set[1]. k-means-- extends the standard kmeans algorithm to simultaneously identify clusters and outliers. The result of the Top-5 outliers are shown in Table 1 and matches with the top players in the NBA "All Star" team. An NBA star is an outlier and given the highly competitive nature of NBA, *an outlier is most likely a star.* Or in other words there are no bad players in the NBA but some players are very good! However, in many other applications it is not at all clear how to proceed to explain outliers. This can be termed as the "Low Precision Problem (LPP)" of outlier detection.

[1]www.basketball-reference.com

ODD'13, August 11th, 2013, Chicago, IL, USA.

Table 1: Given the highly competitive nature of the NBA, not only are stars outliers, but outliers are stars! All the top five outliers are well known leading players of NBA.

Outlier Rank	Player Name	All Star Team (Y/N)
1	Kevin Durant	Y
2	Kobe Bryant	Y
3	LeBron James	Y
4	Kevin Love	N
5	Russell Westbrook	Y

PROBLEM 1. *The Low Precision Problem (LPP) in outlier detection is that*

$$P(\textit{genuine outlier}|\textit{predicted outlier}) \approx \textit{low}$$

LPP occurs because it is hard to disambiguate genuine outliers from errors occurring in the measurement device.

2. THE MULTIPLE SUBSPACE VIEW

A starting point towards addressing LPP and explaining and sifting genuine outliers from measurement errors is to view data from multiple perspectives [12]. In the context where data entities are described by a vector of features, examining an entity in all possible feature subspaces can potentially lead to isolating genuine outliers. This is especially true in high dimensional settings. For example assume that each entity is described by a feature vector of size m. Furthermore, assume that the probability of each feature being recorded incorrectly is p and is independent of other features. Then if m is large, the probability that at least one feature value has been recorded incorrectly is $1-(1-p)^m$ and this can be close to 1 when m is large. Thus having at least one feature value which is corrupted due to measurement error is high. However if we can view the data in multiple subspaces then a genuine outliers will consistently stand out.

A limitation of the multiple subspace approach is that there are exponentially many subspaces leading to intractable algorithms. However the problem can be ameliorated if we notice that in real data sets, the *intrinsic dimensionality* of the data is much lower than the *ambient dimensionality* as we now explain.

3. HIGH-DIMENSIONAL ANOMALIES

It is now part of the data mining folklore that in real data sets, the "degrees of freedom" which actually generate the data is small, albeit unknown. This can be illustrated using examples from computer vision. For example, consider a subset of the Yale Face data shown in Figure 1. Each image is very high-dimensional ($64 \times 64 = 4,096$), however the set of images together live on a three dimensional manifold where the degree of freedom are governed by the rotation of the camera and the lighting. The bottom right hand image (transpose of the top left image) is an outlier as it lives outside the manifold [5].

Thus given a high-dimensional space, if we can project data into a lower-dimension space which preserves the intrinsic structure of the data, then not only can we identify outliers efficiently but potentially explain the discovered outliers. An example of manifold-preserving projection are the family of random projections which preserve pairwise distances with high probability [5]. However, while random projections can lead to improvements in efficiency, by their very nature they make it nearly impossible to interpret the outliers. Thus we need a set of projections to which we can also ascribe some meaning. We next describe matrix factorization methods which are projections of data into lower dimensional space where each dimension aggregates a group of correlated features.

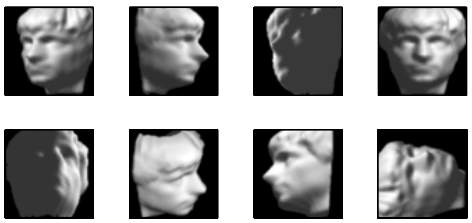

Figure 1: An example to explain the difference between intrinsic and ambient dimension. Samples from the 698-image Yale face data. Each 64 x 64 is a point in a 4,096 dimensional space. However the set of images live in a three dimension set. The bottom right image is added as the transpose of the top left image and is an outlier.

4. MATRIX FACTORIZATION

As we have noted, the challenge in outlier detection is the difficulty to separate true outliers from those data points that are caused because of measurement errors. We have also noted that in high-dimensional space most of the features tend to be correlated. Thus if a data point is a true outlier that fact should be visible in several features. Thus if we take a subspace approach then a genuine outlier will show up as an outlier in more subspaces than an accidental outlier. The challenge in pursuing a subspace approach is that the *space of subspaces* is exponential in the number of features and thus intractable to explore for most practical problems.

One way to address the intractability is to reduce the dimensionality of the original space. This can be carried out using matrix factorization approaches. Factorization is a principled approach of simultaneously aggregating correlated features into a reduced number of "meta-features" which in turn can be imbued with semantics related to the application domain. While Singular Value Decomposition (SVD) and Principal Component Analysis (PCA) have been around for a long time, the recent surge in new methods like Non-Negative Matrix Factorization (NMF) and Bayesian factorization have enhanced the reach of these methods [13]. The key advantage of NMF, say over SVD, is the enhanced interpretation that these methods afford. For example, if X is non-negative document-word matrix or data from a micro-array experiment and $X = UV$ is a non-negative factorization (i.e., both U and V are also non-negative) then the factors can be ascribed a meaning as shown in Table 2.

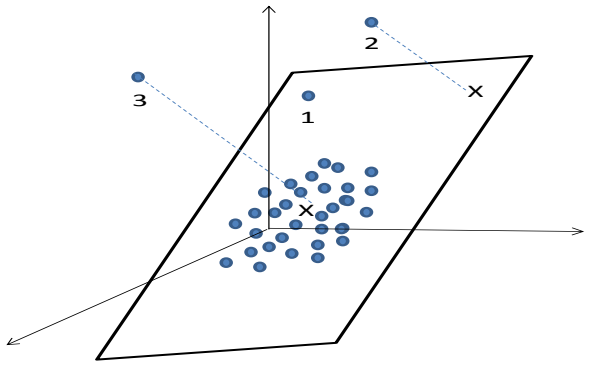

Figure 2: The figure shows the impact of projections of outliers in a lower dimensional space. Data points 1 and 2 remain outliers after projection, while data point 3 is mixed with normal after the projection [8].

4.1 The impact of Projections

Outliers can potentially be impacted in different ways depending upon the nature of outliers. For example, consider the projection shown in Figure 2. The projection shown will have no impact on data point 1, will force data point 3 into a cluster and data point 2 will continue to remain an outlier even though it is far away from the projection plane. Now, which one of these points are genuine outliers is potentially application dependent. However, if we take a subspace perspective, then data point 1 is more likely a genuine outlier. This is because it preserves the correlation between its components but each component is far removed from the main cluster.

4.2 Sensitivity to Outliers

While techniques like NMF provide a promising way to address the combinatorial explosion problem associated with multiple subspace viewing, like SVD, they are highly sensitive to outliers. Thus if our aim is to find outliers, then our method of discovering outliers should not in turn be affected by them. For example, it is well known that both mean and the variance-covariance matrix are extremely sensitive to the presence of even one extreme value and their use for outlier detection will often mask the discovery of genuine outliers. Thus we first have to modify NMF to make them more robust against outliers. Thus we define the following problem:

Table 2: Non-Negative Factorization provides enhanced interpretation of the meta-features. In text processing, the meta-features can be interpreted as topics, while in micro-array analysis, the meta-features are group of correlated genes.

X	U	V
Document-Word	Document-Topic	Topic-Word
Exp-Gene	(Exp,Functional Group)	(Functional Group, Gene)

PROBLEM 2. *[NMF(k,ℓ)] Given a non-negative matrix* $\mathbf{X} \in \mathbb{R}_+^{m\times n}$*, fixed integers* k *and* ℓ*, find matrices* $\mathbf{U} \in \mathbb{R}_+^{m\times k}$*,* $\mathbf{V} \in \mathbb{R}_+^{k\times n}$ *and a subset* $L \subset N$*,* $|L| = \ell$*, which minimizes* $\|X_{-\ell} - UV_{-\ell}\|_F$*, where* $X_{-\ell}$ *is a submatrix consisting of all columns except those from the set* L*.*

To solve the $NMF(k,\ell)$ problem we present the R-NMF algorithm shown in Algorithm 1. The algorithm belong to the class of alternating minimization methods and is very similar to the standard NMF algorithm except for a few caveats. We begin by initializing U in Line 1. In Line 4, we solve for V which minimizes the Frobenius norm of $\|X - U^{i-1}V\|_F$. In Line 5, we compute the residual between X and the current estimate of the product $U^{i-1}V$. In Line 6, we rank the residuals based on the norm of their column values, and L is the index vector of the ranking. We then generate new matrices $X_{-\ell}$ and $V_{-\ell}$ by removing the first ℓ values of the set X and V in Line 7 and 8. In Line 9, we estimate U by minimizing the Frobenius norm of $X_{-\ell}$ and $UV_{-\ell}^i$. We iterate until the convergence criterion is met.

We also propose algorithm O-NMF which is simply using classical NMF algorithm to identify anomaly. The anomalies are calculated by taking ℓ data points which correspond to the top ℓ residual of the final matrices X and UV that is calculated identical to Line 5 of Algorithm 1.

The R-NMF algorithm is an analogous extension of the recently proposed proposed k-means-- algorithm [4]. We should note that another extension for NMF to find outliers has been proposed by Xiong et. al. [14] introduced the method of Direct Robust Matrix Factorization (DMRF). The DMRF method first assumes the existence of a small outlier set S and then infers the low-rank factorization UV by removing S from the data set. It then updates S by using the inferred factorization. In the experiment section we will compare R-NMF with DNMF.

Algorithm 1 [R-NMF Algorithm]

Input: A matrix X of size $m \times n$, m number of features, n number of samples
k the size of the latent space
Output: An $m \times k$ matrix U and $k \times n$ matrix V
$R \approx UV$
1: $U^0 \leftarrow$ random $m \times k$ matrix
2: $i \leftarrow 1$
3: **while** (no convergence achieved) **do**
4: $V^i = \arg\min_V \|X - U^{i-1}V\|_F$
5: $R = X - U^{i-1}V^i$ \\R is a residual matrix
6: Let $L = \{1, 2, \ldots, n\}$ be a new ordering of the columns of R such
$\|R(:,1)\| \geq \|R(:,2)\| \ldots \geq \|R(:,n)\|$
7: $X_{-\ell} \leftarrow X(:, L \setminus L(1:\ell))$
8: $V_{-\ell} \leftarrow V(:, L \setminus L(1:\ell))$
9: $U^i = \arg\min_U \|X_{-\ell} - UV_{-\ell}^i\|$
10: $i \leftarrow i + 1$

The R-NMF algorithm forms the kernel of the subspace algorithm, SR-NMF shown in Algorithm 2 which combines subspace enumeration with R-NMF. Note we only take subspace of the "meta-features." The intuition is that genuine outliers will emerge as outliers in the latent subspaces.

Here we design algorithm that incorporate both the concept of multi subspace view and matrix factorization. As we mentioned before the shortage in [12] is that due to the high dimensionality nature in most of the data set, one simply can not brute force and traversal each and every subspaces. We solve this problem by investigate the problem in a latent space where data are confined in a much small dimensionality.

Algorithm 2 [SR-NMF]

Input: A matrix X of size $m \times n$, m number of features, n number of samples, k the size of the latent space, ℓ number of outliers
Output: A vector R represent the ranking of anomalies with a score in descending order
1: Using $R - NMF$ algorithm we get U and V such that $X \approx UV$
$(U, V) = R - NMF(k,\ell)$
2: $j \leftarrow 0; RANKS \leftarrow$ empty matrix;
3: **STEP1** generate ranks for each subspace
4: **for** $i = 1 \rightarrow k$ **do**
5: generate all set of combinations AS from (k choose i)
6: **for** each $S \in$AS **do**
7: $Residual = X - U(:,S)V(S,:)$
8: $RNorm = columnNorm(Residual)$
9: $[-, RANK] = sort(RNorm, \text{'descend'})$
10: $RANKS = [RANKS; RANK]$
11: $j++$
12: **STEP2** merge ranks into one rank
13: $R \leftarrow$ vector of size n;
14: **for** $i = 1 \rightarrow j$ **do**
15: **for** $p = 1 \rightarrow n$ **do**
16: $R(RANKS(i,p)) = R(RANKS(i,p)) + i$
17: sort R in descending order
$[-, R] = sort(R, \text{'descend'})$ (Note: Matlab Notation)

5. EXPERIMENTS AND RESULTS

In this section we evalute both R-NMF and SR-NMF on several data sets. Our ultimate objective is to verify if SR-NMF can be used to address the **LPP** problem. All our experiments were carried out on a PC with following configurations. Intel(R) Core(TM) i5-2400 CPU @3.1GHz 4GB RAM running on 64-bit Microsoft Windows 7 Enterprise Edition.

5.1 Data Sets

We used three data sets from different application domains which we now describe.

NBA 2012

The NBA 2012 data set consists of nearly two hundred players with each player characterized by twenty features. Example of features include number of points scored, rebounds etc. The data set is freely available from basketball-reference.com.

Spam Email

'Spambase' is a spam email data set [6] consisting of 4,601 emails out of which 1,813 (39%) are spam. The spam e-mails came from their postmaster and individuals who had filed spam and non-spam e-mails from work and personal e-mails. Most of the features (48 out of 57) are frequency of key words.

Research Abstracts

We took around one thousand computer science paper titles from DBLP and also a thousand physics research paper abstracts. We created two data sets. In the first we kept the thousand CS titles and merged them with one hundred physics abstracts. For the second data set, we kept the thousand physics abstracts and merged them with a random subset of one hundred computer science titles. We call the former CSet and the latter PSet.

5.2 Results

We report results on robustness, convergence, runtime and accuracy on the three aforementioned data sets.

Results:Robustness of R-NMF

Here we report on results about the sensitivity of the R-NMF against the classical NMF algorithm used for outlier detection, the O-NMF. We applied both R-NMF and O-NMF algorithm on the NBA 2012 data set but modified one entry in the matrix as a multiple of the mean value. This is shown on the x-axis of Figure 3. For each different value on the x-axis we computed the U matrix and computed the difference in the norm of the new U matrix and the original U matrix. The U matrix is the base matrix and stores the meta-features in terms of the original features.

Figure 3 shows that R-NMF is more robust against perturbations while the U matrix using O-NMF increases without bound. This clearly demonstrates that the traditional NMF algorithm should not be used for any serious applications as it is extremely sensitive to data perturbations.

Results:Convergence Analysis

Here we investigate the convergence properties of the R-NMF algorithms. From Algorithm 1 we know that for each iteration R-NMF will reconstruct U with a given number of outliers excluded. However, each iteration the algorithm may exclude different data points as outliers, this could potentially make the algorithm unstable. Thus, it is necessary to study whether this new algorithm will converge properly.

We conduct the experiments as follows. We use the Spambase data set, and set the number of outliers for R-NMF as the number of spam emails. We vary k and present the results for k=9,12,15, and 18.

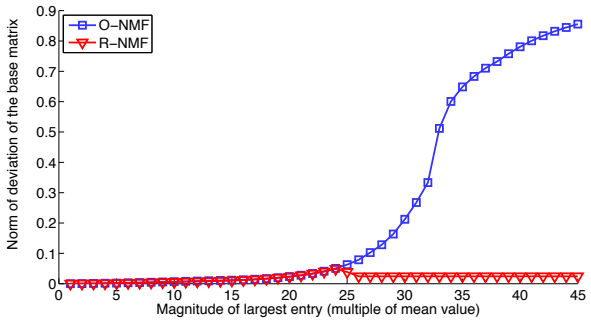

Figure 3: R-NMF is substantially more robust against the presence of outliers in the data compared to standard O-NMF.

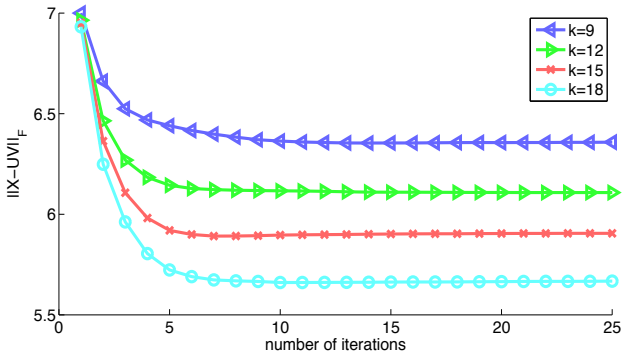

Figure 4: R-NMF converges with all given settings of k. As the dimension of the subspace (k) increases, residual of R-NMF algorithm goes down.

As can be seen from Figure 4, the first thing one can notice is that with bigger k, the residual of the algorithm goes down. This is because with bigger k, the decomposed matrices UV can better reconstruct the original X. Most importantly, the algorithm converge at all given settings of k within 20 repetitions.

Results:Runtime

We present the run time results of R-NMF algorithm for the Spambase data sets in Figure 5 respectively. As expected, we observe that the run time of R-NMF decreases as the number of outliers is increased. This trend follows the intuition of R-NMF algorithm that the construction of base matrix U is based on the data X without the anomalous points (Algorithm 1 line 5-8).

Results:Precision and Recall

We compute precision and recall on the Spambase, PSet and the CSet data sets. The outliers are considered as *positives*. The experiments are conducted as follows. We vary the two variables: k and ℓ, We compared the two proposed algorithms: R-NMF and SR-NMF against the Direct Robust Matrix Factorization (DMRF) approach proposed by [14]. The results for different values of k and different sizes of the outliers specified are show from Table 3-8. At the moment it is hard to draw conclusions from the results. Futher work is

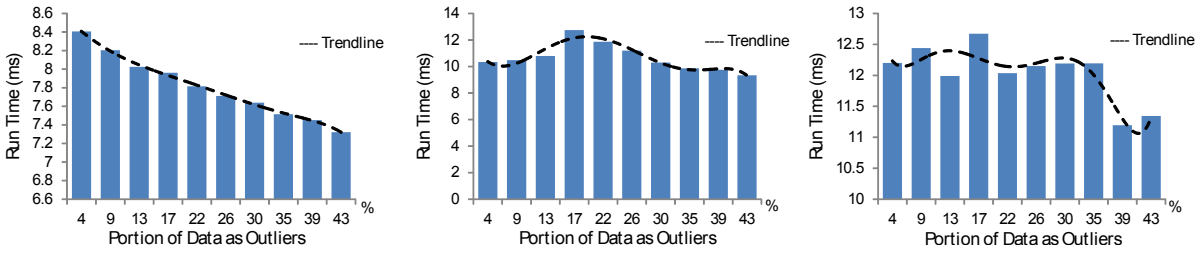

Figure 5: Average Run time R-NMF on Spambase data set: (Left) k = 1, (Middle) k = 2, (Right) k = 3. As the number of outliers increases, the run time for R-NMF decreases. The values here are the average values for all iterations.

Table 3: Precision on CSet: DRMF, SR-NMF and R-NMF.

k	Portion of data as outliers								
	35%			40%			45%		
	DRMF	SR-NMF	R-NMF	DRMF	SR-NMF	R-NMF	DRMF	SR-NMF	R-NMF
6	0.10	**0.13**	**0.13**	0.10	**0.12**	**0.12**	0.09	**0.11**	**0.11**
9	0.09	0.12	**0.14**	0.10	**0.12**	**0.12**	0.09	**0.11**	**0.11**
12	0.10	**0.12**	**0.12**	0.09	0.12	**0.13**	0.09	**0.11**	**0.11**

Table 4: Recall on CSet: DRMF, SR-NMF and R-NMF

k	Portion of data as outliers								
	35%			40%			45%		
	DRMF	SR-NMF	R-NMF	DRMF	SR-NMF	R-NMF	DRMF	SR-NMF	R-NMF
6	0.39	0.49	**0.50**	0.45	0.51	**0.54**	0.47	**0.56**	0.55
9	0.36	0.47	**0.52**	0.45	0.52	**0.54**	0.46	**0.56**	**0.56**
12	0.39	**0.48**	0.47	0.40	0.53	**0.55**	0.45	**0.56**	0.52

Table 5: Precision on PSet: DRMF, SR-NMF and R-NMF.

k	Portion of data as outliers								
	35%			40%			45%		
	DRMF	SR-NMF	R-NMF	DRMF	SR-NMF	R-NMF	DRMF	SR-NMF	R-NMF
6	0.13	**0.15**	0.15	**0.16**	0.15	**0.16**	**0.15**	**0.15**	0.14
9	0.16	0.16	**0.18**	0.16	0.16	0.16	0.15	0.15	0.15
12	0.17	0.16	**0.18**	0.16	0.16	0.16	0.15	0.15	0.15

Table 6: Recall on PSet: DRMF, SR-NMF and R-NMF.

k	Portion of data as outliers								
	35%			40%			45%		
	DRMF	SR-NMF	R-NMF	DRMF	SR-NMF	R-NMF	DRMF	SR-NMF	R-NMF
6	0.49	0.56	**0.58**	**0.72**	0.66	0.70	**0.72**	0.72	0.70
9	0.60	0.60	**0.69**	0.70	0.69	0.70	0.72	**0.73**	0.72
12	0.65	0.60	**0.69**	0.70	**0.72**	0.70	0.72	0.72	**0.73**

Table 7: Precision on Spambase: DRMF, SR-NMF and R-NMF. Best values are highlighted.

k	Portion of data as outliers								
	7%			10%			13%		
	DRMF	SR-NMF	R-NMF	DRMF	SR-NMF	R-NMF	DRMF	SR-NMF	R-NMF
6	0.27	**0.30**	0.29	**0.32**	0.26	0.29	**0.37**	0.32	0.36
9	0.26	0.26	**0.30**	0.28	**0.31**	0.28	0.31	**0.35**	**0.35**
12	0.25	**0.32**	0.30	0.30	**0.33**	0.29	0.30	0.32	**0.36**

required to analyse the results and determine the root cause of the outliers.

6. SUMMARY AND CONCLUSION

Outlier Detection is a core task in data mining. In fact as the size and complexity of data sets increases the need

Table 8: Recall on Spambase: DRMF, SR-NMF and R-NMF. Best values are highlighted.

k	Portion of data as outliers								
	7%			10%			13%		
	DRMF	SR-NMF	R-NMF	DRMF	SR-NMF	R-NMF	DRMF	SR-NMF	R-NMF
6	0.05	**0.06**	0.05	**0.08**	0.07	0.07	**0.12**	0.10	**0.12**
9	0.05	0.05	**0.06**	0.07	**0.08**	0.07	0.10	**0.12**	**0.12**
12	0.04	**0.06**	0.05	**0.08**	**0.08**	0.07	0.10	0.10	**0.12**

to identify meaningful and genuine outliers will only grow. Almost all major applications ranging from health analytic to network data management to bio-informatics require analytical tools which can identify and explain genuine outliers.

The core challenge in outlier detection is to distinguish between genuine and noise outliers. The former are indicative of a new, previously unknown process while the latter is often a result of error in the measurement device. The difficulty to distinguish between genuine and noise outliers leads to the Low Precision Problem (LPP). Our claim is that LPP is the fundamental problem in outlier detection and algorithmic approaches to solve LPP are urgently needed.

One approach to distinguish between genuine and noise outliers is to take a multiple subspace viewpoint. A genuine outlier will stand out in multiple subspaces while a noise outlier will be separated from the core data in much fewer subspaces. However the problem in subspace exploration is that current methods are unlikely to scale to high dimensions.

Matrix factorization methods provide a balanced compromise between full subspace exploration in the feature space versus exploration in the meta-feature or latent space. The advantage of working in the latent space is that many of the features are aggregated into a correlated meta-feature. Often these features in the latent space can be imbued with a semantic meaning relevant to the problem domain. For example, in the case of text mining, the features correspond to words while meta-features correspond to topics.

The challenge with matrix factorization methods is that they are highly sensitive to outliers. This can be a serious problem whenever there is a mismatch between the data and the proposed model. One way to ameliorate the problem is to use an alternate minimization approach to estimate both the matrix decomposition and the outlier set. This is the basis of the NMF(k,ℓ) problem and the R-NMF algorithm. Preliminary results show that R-NMF is substantially more robust compared to a standard NMF approach in the presence of data noise. This opens up a promising avenue for further exploration and address the LPP.

7. RELATED WORK

The task of extracting genuine and meaningful outliers has been extensively investigated in Data Mining, Machine Learning, Database Management and Statistics [3, 1]. Much of the focus, so far, has been on designing algorithms for outlier detection. However the trend moving forward seems to be on detection and interpretation.

While the definition of what constitutes an outlier is application dependent, there are two methods which gained fairly wide traction. These are distance-based outlier techniques which are useful for discovering *global* outliers and density-based approaches for *local* outliers [9, 2].

Recently there has been a growing interest in applying matrix factorization in many different areas, *e.g.* [7],[10]. To the best of our knowledge, probably the most closest work to ours is by Xiong *et al.* [14]. Xiong *et al.* have proposed a method called Direct Robust Matrix Factorization (DRMF) which is based on matrix factorization. DRMF is conceptually based on Singular Value Decomposition (SVD) and error thresholding.

The main algorithm proposed in this paper extends the work on k-means-- proposed in *et al.* [4] which unifies clustering and outlier detection. Furthermore we have taken inspiration from a body of work on multiple subspace outlier detection to distinguish between genuine and accidental outliers [12].

8. REFERENCES

[1] V. Barnett and T. Lewis. *Outliers in Statistical Data*. John Wiley and Sons, 1994.

[2] M. Breunig, H. Kriegel, R. Ng, and J. Sander. Lof: Identifying density-based local outliers. In *IEEE International Conference on Data Mining (ICDM)*, 2000.

[3] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection: A survey. *ACM Comput. Surv.*, 41(3), 2009.

[4] S. Chawla and A. Gionis. kmeans–: A unified approach to clustering and outlier detection. In *SIAM International Conference on Data Mining (SDM SIAM)*, 2013.

[5] T. de Vries, S. Chawla, and M. E. Houle. Density-preserving projections for large-scale local anomaly detection. *Knowledge and Information Systems (KAIS)*, 32(1):25–52, 2012.

[6] A. Frank and A. Asuncion. UCI machine learning repository, 2010.

[7] D. M. Hawkins, L. Liu, and S. S. Young. Robust singular value decomposition. *Technical Report, National Institute of Statistical Sciences*, 2001.

[8] M. Hubert, P. Rousseeuw, and Vanden. ROBPCA: a new approach to robust principal component analysis. *Technometrics*, 47:64–79, 2005.

[9] E. Knorr and R. Ng. Algorithms for mining distance-based outliers in large datasets. In *International Conference on Very Large Data Bases (VLDB)*, 1998.

[10] H.-P. Kriegel, P. Krogel, E. Schubert, and A. Zimek. A general framework for increasing the robustness of pca-based correlation clustering algorithm. In *Scientific and Statistical Database Management Conference (SSDBM)*, 2008.

[11] T. Kuhn. *The Structure of Scientific Revolutions*. University of Chicago, 1962.

[12] E. Müller, I. Assent, P. Iglesias, Y. Mülle, and K. Böhm. Outlier ranking via subspace analysis in

multiple views of the data. In *IEEE International Conference on Data Mining*, pages 529–538, 2012.

[13] A. Singh and G. Gordon. A unified view of matrix factorization models. In *Machine Learning and Knowledge Discovery in Databases*, volume 5212 of *Lecture Notes in Computer Science*, pages 358–373. Springer Berlin Heidelberg, 2008.

[14] L. Xiong, X. Chen, and J. G. Schneider. Direct robust matrix factorizatoin for anomaly detection. In *IEEE International Conference on Data Mining (ICDM)*, pages 844–853, 2011.