Rank Selection in Non-negative Matrix Factorization: systematic comparison and a new MAD metric

Laura Muzzarelli
Inst. of Neurosci. and Medicine
INM-7, Forschungszentrum
Jülich, and Inst. of Systems
Neurosci., HHU Düsseldorf
Germany
lau.muzzarelli@fz-juelich.de

Susanne Weis
Inst. of Neurosci. and Medicine
INM-7, Forschungszentrum
Jülich, and Inst. of Systems
Neurosci., HHU Düsseldorf
Germany
s.weis@fz-juelich.de

Simon B. Eickhoff
Inst. of Neurosci. and Medicine
INM-7, Forschungszentrum
Jülich, and Inst. of Systems
Neurosci., HHU Düsseldorf
Germany
s.eickhoff@fz-juelich.de

Kaustubh R. Patil
Inst. of Neurosci. and Medicine
INM-7, Forschungszentrum
Jülich, and Inst. of Systems
Neurosci., HHU Düsseldorf
Germany
k.patil@fz-juelich.de

Abstract—**Non-Negative Matrix Factorization (NMF) is a powerful dimensionality reduction and factorization method that provides a part-based representation of the data. In the absence of a priori knowledge about the latent dimensionality of the data, it is necessary to select a rank of the reduced representation. Several rank selection methods have been proposed, but no consensus exists on when a method is suitable to use. In this work, we propose a new metric for rank selection based on imputation cross-validation, and we systematically compare it against six other metrics while assessing the effects of data properties. Using synthetic datasets with different properties, our work critically evidences that most methods fail to identify the true rank. We show that properties of the data heavily impact the ability of different methods. Imputation-based metrics, including our new *MADimput*, provided the best accuracy irrespective of the data type, but no solution worked perfectly in all circumstances. One should therefore carefully assess characteristics of their dataset in order to identify the most suitable metric for rank selection.**

Keywords— non-negative matrix factorization, rank selection, cross-validation.

I. Introduction

Large multivariate datasets are commonplace in research as well as industry, and it is often crucial to identify latent components that underlie the data. For example, one may aim at creating new features that compress the original information and reduce noise, or one may aim at finding interpretable non-redundant factors to gain insights into the structure of the data. Unsupervised methods are often employed for obtaining such factors and several options exist, e.g. principal components analysis (PCA) and non-negative matrix factorization (NMF). There is usually no a priori knowledge available about the true dimensionality of the data. Therefore, independently of which factorization method is used, it is necessary to select the dimensionality (or rank) of the reduced data. We define this issue as rank selection problem, as it ultimately results in having to select the rank of the reduced representation that best captures structure and ignores noise.

Matrix factorization provides a unified framework for clustering, factorization and feature construction [1]. All these tasks, in fact, can be represented as a low-rank matrix approximation, in which the original data is represented as multiplication of two or more low-rank matrices. Singular value decomposition (SVD) and closely related principal components analysis (PCA) are popular methods that constrain the factors to be orthogonal. Non-negative Matrix Factorization (NMF) represents non-negative data V as a product of two matrices W and H that are constrained to be non-negative. NMF derives factors that group elements into parts that are additive, and are often notably sparser and more interpretable than the ones derived with other methods like SVD and PCA [2]. These features of NMF have made it increasingly popular for dimensionality reduction with important applications in text mining and natural language processing [3][4], image processing [2][5], signal decomposition and source separation [6][7], genetics [8][9], and neuroscience [10]–[13], among others.

However, for a given data how to optimally select the number of factors or rank in NMF is still an unsolved question. Several methods have been proposed for solving the rank selection problem (see section II), but the advantages and limitations of different methods are still unknown due to lack of systematic comparison. For instance, characteristics of the data like sparseness might affect rank selection methods. No previous study yet has explored such effects on rank selection. Another motivation for our work was the consideration that most NMF rank selection methods are based on the stability of the NMF solutions. However, stability does not necessarily imply accuracy. For instance, a trivial factorization method that returns some fixed factors would be highly stable across runs, but it would not yield a useful reduced representation of the data, as it would overfit. Nonetheless, stability based methods would favor such solutions. Although such extreme behavior is unlikely in real scenarios, it is not unlikely that similar biases could impact rank selection methods.

In this paper, we explicitly address these issues by systematically comparing several rank selection methods on a large number of datasets with two different data generation strategies. We also propose a new rank selection method that is based on the strategy commonly used for evaluating supervised methods. In supervised settings, the choice of the best model is generally done via cross-validation (CV) in which part of the data are held-out and used to validate the model induced on rest of the original data [14]. Evaluation of held-out data provides an estimate of the ability of the model to generalize. This feature is highly desirable in NMF applications in which one might want to identify a factor structure that is stable and

This study was partly supported by the Helmholtz Portfolio Theme "Supercomputing and Modeling for the Human Brain" and the European Union’s Horizon 2020 Research and Innovation Programme under Grant Agreement No. 785907 (HBP SGA2).

replicable on new samples. CV can be applied to factorization methods, such as PCA, if appropriately designed [15] and to date several such attempts have been carried out for NMF [16]–[18].

Taken together, in this work we make three main contributions: (1) we provide the first systematic comparison of several NMF rank selection methods; (2) we analyze for the first time the effect of different properties of the data on rank selection; (3) we propose a new CV-based rank selection metric that can potentially overcome limitations of previous methods.

The paper is organized as follows: first we will introduce NMF and previously proposed NMF rank selection methods (section II); then we will outline imputation CV-based metrics including our new metric (section III); we will then describe the experimental setup (section IV) and results of the experimental testing (V); finally, we will discuss the results (section VI) and conclude (section VII).

II. Non-negative Matrix Factorization

A. Algorithm

NMF is a low-rank approximation method that factorizes a non-negative matrix $V_{n\times m}$ with n observations and m features, into two non-negative matrices $W_{n\times k}$ and $H_{k\times m}$. This results in a part-based decomposition of V into a basis matrix H, representing how the m features are structured in k underlying factors, and a coefficient matrix W, representing the loadings for each of the n observations on the k factors.

Several algorithms exist for NMF, and numerous variants have introduced additional constraints on the factor structure [19]. For simplicity and increased comparability, we used Lee and Seung's [20] multiplicative update rule NMF algorithm that aims at minimizing the Euclidean distance between the original and reconstructed matrix. Under this formulation, NMF can be stated as the following optimization problem:

$$\|V - WH\|_2 \quad subject\ to\ W \geq 0,\ H \geq 0 \qquad (1)$$

And it is solved via the following iterative update rules:

$$H \leftarrow H \odot \frac{W^T V}{W^T W H} \qquad W \leftarrow H \odot \frac{V H^T}{W H H^T} \qquad (2)$$

where the $\odot$ symbol refers to element-wise multiplication, and where the division is a component-wise division between the two matrices. Note that in this paper, we factorize m features into k factors of interest using the basis matrix H.

As NMF is sensitive to initialization, for similar reasons of simplicity, we employ random initialization on the initial values of W and H together with SVD-based initializations [21].

B. Rank selection methods

Absence of prior knowledge makes it necessary to identify the optimal rank to factorize a given dataset. This is usually achieved by performing NMF for a range of ranks, and then choosing a rank based on some metric of goodness.

Currently available metrics for rank selection in NMF can be grouped into one of following general approaches: stability of NMF-derived clusters across multiple NMF runs [8][22]; similarity of NMF factors across halves in a split-half validation [12]; relative difference in reconstruction error between original matrix and permuted matrix [23]; minimization of reconstruction error of held-out values in a CV framework [16], [17]; estimators of the mean squared error (MSE) [24]; Bayesian methods [25], [26]; and information-theory based methods [27]. Here, we compare seven methods in total, six previously proposed methods and our new method.

1) Consensus methods

Consensus methods evaluate consistently with which pairs of features get clustered together across multiple NMF runs. NMF is computed at a given k and each of the m features are assigned to one of the k factors according to their highest loading in the H matrix. This yields a binary connectivity matrix for each NMF run. By repeating NMF with different initializations, a consensus matrix (i.e. average of connectivity matrices) is computed. This consensus matrix expresses the probability with which pairs of features cluster together. Two metrics utilizing the consensus matrix have been proposed as a criterion for rank selection:

1. *Cophenetic correlation coefficient* (*coph*) [8] of the consensus matrix is calculated, and the rank k at which the magnitude of *coph* "begins to fall" is selected. Here we operationalize this procedure by selecting the rank with maximum *coph*.
2. *Dispersion* [22] uses the same approach as above, but calculates the dispersion coefficient of the consensus matrix. The rank with the maximum *dispersion* is selected.

Note that consensus methods have been criticized in [23], because they do not evaluate the factors themselves, but rather they evaluate the ability of the NMF to identify stable clusters.

2) Similarity in split-half validation

In these methods, the sample is randomly divided in two halves and NMF is computed separately within each half. The factors identified in each half are then matched (e.g., via the Hungarian algorithm), because NMF factors are not ordered as in PCA. This procedure is repeated with different splits, and the rank at which the factors are maximally similar on average is selected. Similarity between the factors can be computed via:

1. *Inner product* [12] between the matched factors is calculated and then averaged.
2. *Adjusted Rand Index* (*aRI*). In this case, each of the m features is assigned to the factor on which it has highest loading. The similarity of the two resulting clusterings is then assessed via *aRI*.

3) Comparison with permuted matrix (perm) [23]

In this approach, a new matrix is constructed by permuting the columns (features) of each row independently. NMF is computed on this permuted matrix as well as on the original matrix on a range of ranks. The slopes of the corresponding reconstruction errors with respect to the ranks are compared. The rank k is then selected where the slope of the original matrix is lower than the slope of the permuted matrix. This metric is based on the rationale that when the original error slope is equal to that of the permuted matrix, no further information is gained.

4) *Reconstruction error of imputed held-out values*

Kanagal and Sindhwani [16] proposed to use cross-validation (CV) with weighted NMF for rank selection while presenting a NMF variant that can account for weights. By settings some of the weights to 0 these entries are effectively held out of the matrix *V*. Then, weighted NMF is computed on the data with these weights. NMF, however, reconstructs the entire matrix, thus essentially imputing the values the missing values. The reconstruction error is then calculated as the Frobenius distance between the original and imputed. The rank is selected at the minimum of imputation error. A similar approach is also suggested in the R package NNLM's vignette [18]. This provides a framework similar to CV in supervised settings.

Of note, we exclude from our comparison: Bayesian methods [25], [26], due to their complexity and to the required assumptions on the choice of the priors associated with the *W* and *H* matrices; Bi-Cross-Validation [17], as it has already been shown inaccurate [16][24]; and the Stein's unbiased risk estimator [24], as it has been criticized for providing optimistic estimates [28].

III. Imputation CV

In this section we formalize use of reconstruction error of imputed values as a method for rank selection. We also propose a new metric which expands on the imputation-based rank selection methods presented above (section II.D: [16], [18]).

A. *General principles*

Selecting a rank lower than the true rank will lead to poor representation of the structure within the data (underfitting), while selecting a higher rank would cause the factorization to model noise (overfitting). A possible solution to find the right balance between under- and overfitting is to adopt cross-validation, a resampling technique commonly used in supervised settings to assess the generalization of a model [14]. In a supervised setting, the dataset is partitioned into training and validation sets and the model learned from the training set is evaluated on the validation set. Higher accuracy on the validation set is indicative of a more generalizable model. It is fundamental, here, that the model is evaluated in predicting data it has not seen before.

However, entire observations (rows or columns) cannot be held-out in matrix factorization due to its unsupervised nature. [15]. We can still obtain an independent test sample by holding out some randomly selected entries (also referred to as a speckled pattern) in the matrix *V* as missing values. NMF is still able to learn structure in the data by considering only the available entries. Then, the approximation of the original matrix $\hat{V} = WH$ reconstructs the held-out entries, effectively imputing the missing values. The imputation error is then calculated by comparing the imputed values with the corresponding original values. To reduce the variance due to the randomness in the choice of missing value locations, it is necessary to repeat this procedure multiple times at each rank with different missing values. The minimum of the average imputation error across multiple runs at each rank is then used as criterion for rank selection, since it accounts for the trade-off between under- and overfitting. Thus, this method is similar to Kanagal and Sindhwani's approach described above.

B. *New metric: Median Absolute Deviation*

The averaged imputation error, however, does not provide the complete information about the reconstruction in different parts of the data. For instance, when the factors are representative of unequal portions of the data, the average imputation error can be low even though a good reconstruction is only achieved for a part of the data. This can mislead rank selection methods.

Variance in the imputation error due to different missing values can provide additional information, especially about relative reconstruction quality in different parts of the data. Lower variance indicates equal reconstruction quality in different parts of the data and thus the variance should be minimal at the true rank. This provides a new metric for rank selection which has not been explored yet. Here we estimated this variance using a robust estimator median absolute deviation (MAD) and call the resulting method *MADimput.*

IV. Experimental setup

A. *Datasets*

Our aim was to test if the ability of various methods to identify the true rank is affected by characteristics of the data. We, therefore, created two types of synthetic datasets with different structure of the *W* and *H* matrices (Figure 1);

- Dense datasets where the *W* and *H* matrices were uniformly sampled in the range [0, 1].
- Sparse datasets where the *W* and *H* matrices were block-diagonal and non-zero entries were sampled uniformly in the range [0, 1].

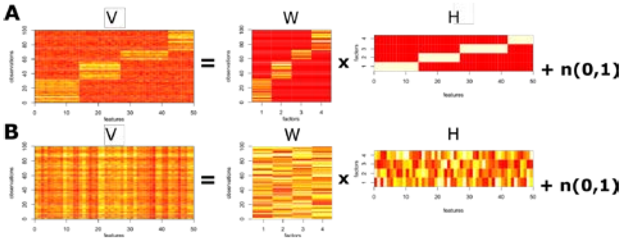

Fig. 1. Exemplar structure of sparse (A) and dense (B) synthetic datasets, at latent components $k = 4$.

In both cases the *W* matrix was scaled by 10 and normally distributed random noise with mean 0 and standard deviation 1 was added to the final *V* matrix.

For both types of data, we created 20 different synthetic datasets of size 100x50 (observations x features) for five values of latent dimensionality {2, 4, 6, 8, 10}, amounting to a total of 100 datasets per type.

Moreover, we evaluate the performance of all methods on three real datasets:

1. **MED5** (reduced MEDLINE) dataset is a database of medical abstracts available from the DTU:NMF Toolbox [29], having size 1159 terms by 124 abstracts, and 5 underlying human-labelled topics.

2. **Dig0246** dataset consists of a subset of the Optical Recognition of Handwritten Digits dataset from the University of California at Irvine (UCI) repository (training sample). It composed of 64 attributes and 1520 samples representing digits {0, 2, 4, 6}. The aim is to separate the digits, thus the true rank is four.

3. **ALL-AML** dataset [30] is a cancer gene expression dataset, containing the gene expression patterns of 5000 genes obtained in 38 samples (features of interest). The samples are either of acute lymphoblastic leukemia (ALL), or of acute myeloid leukemia (AML) – but further subtyping might be possible. The creators of the dataset found the rank to be 3 using cophenetic correlation [8].

B. *Imputation CV implementation*

We used a repeated cross-validation, in which at each repetition 10% of the values chosen at random were set to missing. The *nnmf* function in the NNLM package [18] was used as it can handle missing values by eliminating them before the estimation of *W* and *H*. For each rank *k*, the original matrix was reconstructed and the imputation error was computed as mean squared error (MSE). We then calculated the median (*MSEimput*) and MAD (*MADimput*) of the MSE values across 100 repetitions. Both metrics select a rank *k* at which their value is minimum.

C. *NMF and methods settings*

As NMF is sensitive to initialization, we ran NMF with 20 different initializations and selected the solution with the lowest reconstruction error. Ten initializations were based on uniformly sampled values while the other 10 were based on SVD [21]. This initialization strategy was used for all methods except for consensus methods as they rely on differences in NMF solutions due to initializations.

We repeated NMF with abovementioned initializations 100 times at each rank. The median of the corresponding metrics across repetitions was used to select the rank. For the permutation based method, a single permuted matrix was used. All methods were evaluated on ranks from 2 to 15.

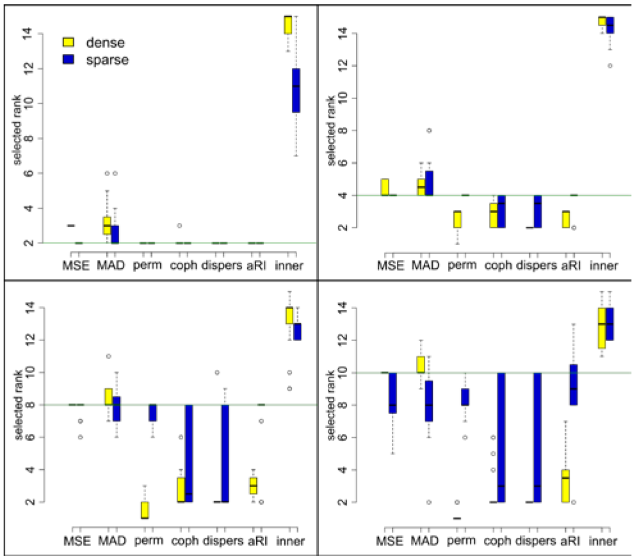

Fig. 2. Boxplots showing median selected rank on 20 synthetic datasets by the seven methods at four different latent components (2, 4, 8 and 10). For each method, performance in dense and sparse datasets is shown; latent k is shown as the green horizontal line. Results with true rank of 6 are not shown for brevity but they showed a similar pattern.

V. RESULTS

A. *Synthetic data*

Experimental results on synthetic data are presented in Fig. 2 and Table 1. We observed several systematic effects of data characteristics and the true rank on the performance of all rank selection methods: (1) firstly, it is clear that most of the methods tested here fail in the exact identification of the true rank; (2) remarkably, none of the methods other than imputation CV showed good performance on dense datasets, except at true rank of two; and (3) sparse datasets and higher true rank caused decreased accuracy for all methods.

TABLE I. ACCURACY OF TOP RANK SELECTION[a]

Type	True rank	MSE	MAD	perm	coph	disp	aRI
Dense	2	0	25	100	95	100	100
	4	70	50	0	25	0	0
	6	100	50	0	0	0	5
	8	100	60	0	0	0	0
	10	100	65	0	0	0	0
	Avg.	**74**	**50**	20	*24*	20	21
Sparse	2	100	65	100	100	100	100
	4	100	55	100	50	50	90
	6	100	55	100	35	45	95
	8	80	30	70	35	25	80
	10	30	15	20	30	30	15
	Avg.	**82**	44	**78**	50	50	*76*

[a] *inner* was excluded from the table as it consistently showed 0 accuracy

Both imputation CV based methods, *MSEimput* and *MADimput*, were highly accurate for both data types and over the whole range of true ranks (Fig. 2, Table I). However, these methods did not select the rank perfectly and notably showed increasing error at higher true ranks. Specifically, it appears that *MSEimput* is more accurate in selecting the true rank in dense datasets, but is less accurate for sparse datasets, especially at higher true ranks. *MADimput*, on the other hand, showed moderate accuracy in both sparse and dense datasets.

Split-half validation with inner product similarity (*inner*) was the overall worst performing method, and consistently overestimated the rank. While aRI similarity in split-half validation showed a remarkably good performance on sparse datasets, it also underestimated the rank of dense datasets with equal remarkability. The good performance of aRI can be attributed to the match between cluster-like property of the sparse data (see Fig. 1) which is then exploited by aRI through NMF's clustering ability. Permutation-based rank selection (*perm*) showed a similar pattern with slight underestimation for sparse datasets at higher true ranks. Consensus methods (*coph* and *dispersion*) tended to underestimate the rank with a much higher variability of the estimates on sparse datasets.

TABLE II. ACCURACY OF TOP-THREE RANK SELECTION[a]

Type	True rank	MSE	MAD	perm	coph	disp	aRI
Dense	2	100	85	100	100	100	100
	4	100	95	90	85	70	100
	6	100	100	0	20	25	15
	8	100	95	0	0	20	0
	10	100	100	0	5	60	0
	Avg.	**100**	**95**	38	42	55	43
Sparse	2	100	100	100	100	100	100
	4	100	95	100	100	100	100
	6	100	100	100	95	100	100
	8	95	75	100	85	95	95
	10	40	45	80	60	75	60
	Avg.	87	83	**96**	88	**94**	*91*

[a] *inner* was excluded from the table as it consistently showed 0 accuracy

We observed that sometimes local optima of the metrics tended to be closer to the true rank. In such cases a user can manually select the rank aided by the metrics. To assess the impact of local optima, we calculated the top-three accuracies considering if any of the top-three optimal ranks for each metric matched the true rank (Table II). As we can see, the top-three accuracy of all methods increased compared with the top-accuracy, indicating that local optima of all the metrics are meaningful and should be taken into account. Interestingly, *dispersion* showed a large increase in the top-three accuracy on sparse datasets surpassing many other methods.

B. Real data

Results on real datasets confirmed our observations in synthetic data and further suggested that more complex data properties might affect rank selection in real data. In fact, we can see (Figure 3) how for all three real datasets the inner product systematically overestimated the rank, while the other methods show more diverse and interesting patterns.

For the MED5 dataset, only *coph* identified the correct rank, *perm* had the highest error with 12 as the selected rank while other methods selected either 3 or 4.

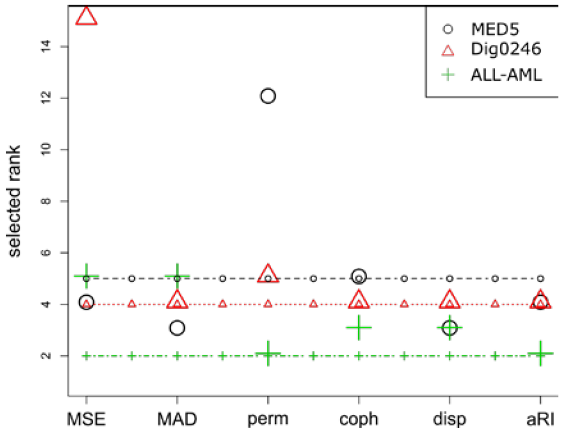

Fig. 3. Results on three real datasets. The method *inner* was excluded due to its bad performance. Each horizontal line shows the corresponding expected true rank.

For the Dig0246 dataset, *MADimput*, consensus methods and aRI converge on the expected 4 latent components; perm identified 5 components. The performance of *MSEimput* was puzzling as it selected the highest tested rank of 15. The *MSEimput* monotonously decreased for the whole range of ranks tested. More detailed analysis is needed to pinpoint reasons behind this bad performance but it is out of scope of this work.

For the ALL-AML dataset, perm and aRI identify the 2 expected latent components. Consensus methods converge on rank of 3, replicating the result of Brunet et al. [8], in which it was suggested that 3 components yield a more fine-grained interpretation of the data. *MADimput* and *MSEimput* both identified 5 components as the best representation.

It should be noted that all three real datasets can be expected to have a sparse latent structure, as they are composed of relatively distinct groups (even though no clear patterns were visible by plotting the data). This can explain the accuracy of *aRI*, and the relatively better performance of *MADimput* and *perm* as compared to *MSEimput*. However, none of the methods were accurate on all three real datasets. Given that consensus methods and *aRI* performed reasonably well on sparse datasets at low ranks (Table I), their apparently good performance on real data can be expected. However, their performance was more varied than other methods across runs and here we chose the best performing run.

VI. Discussion

In this paper one of our aims was to systematically compare effects of data properties on NMF rank selection methods. Towards this end, we tested effects of data characteristic of sparsity/orthogonality and a range of true ranks. We also proposed use of variance of MSE in imputation CV as a new rank selection method.

Overall, our results provide evidence that most of the currently available methods fail in identify the true rank, and no method works perfectly for all data types. However, metrics based on imputation CV appear to be the more reliable across data types, strongly supporting the use of CV for model selection in unsupervised settings [14, 15]. The results on synthetic datasets showed *MSEimput* to be more accurate, whereas our new method *MADimput* performed better than the rest on dense datasets while also showing good performance with sparse datasets. Interestingly *MADimput* performed better than *MSEimput* on the real datasets, especially the Dig0246 data. Our new method, therefore, seems to capture additional properties of the data, albeit with some caveats.

Our study, for the first time, shows that data characteristics can influence the ability of rank selection methods. Critical properties of the data identified in this work include sparsity/orthogonality and true rank of the underlying representation, but it is likely that other characteristics influence rank selection ability of all methods. This is evident in the performance of different methods in real datasets, in which we observed overall similar patterns as in synthetic data. However, the performance of many methods, including imputation-based methods, worsened in real-world data, which might be due to their increased complexity and noise.

The inability of most methods to correctly identify the rank in sparse datasets at higher true ranks is possibly due to the presence of dominating factors that influence large portions of the data along with some factors that influence only small parts of the data (see section IV). Specifically for the average imputation error *MSEimput*, such a data structure can bias the MSE estimates towards the dominating factors, effectively underestimating the rank (as observed in our results, see Fig. 2). In-fact, our current sparse data simulation process (adapted from the R package NMF, [31]) tends to create imbalanced dominance structure when the true rank approaches number of features. However, a more detailed analysis is required to establish such a relationship which is out of the scope of current work.

VII. Conclusion

In conclusion, even though imputation CV based methods seem to be the most promising option for rank selection, characteristics of the data should be carefully assessed prior to deciding which method to apply for NMF rank selection. We have also exposed several potential research directions along these lines. Our new method *MADimput* showed promise on both synthetic and real data and will be a valuable addition to the existing arsenal of NMF rank selection methods. We suggest that a combination of *MSEimput* and *MADimput* can provide clues about the true rank. Furthermore, exploring additional parameters such as number of held-out values is another potential future research direction.

A more thorough investigation of the effect of the above mentioned properties along with others like noise level and NMF variants is necessary before establishing definitive indications on which is the best rank selection method. Our analysis suggests that it is likely that different rank selection methods might emerge as being more effective pertaining to different properties of the data at hand. An example of such an effect can be seen for *aRI* (a method for evaluating similarity between two clusterings) based split-half validation, which works well only when the underlying representation is cluster-like. Such aspects can be further investigated via meta-learning approaches [32].

Acknowledgment

We are grateful to Ji Chen and Robert Langner for insightful discussions. The R project for statistical computing and associated packages made implementation and computations possible.

References

[1] A. P. Singh and G. J. Gordon, “A Unified View of Matrix Factorization Models,” in *Machine Learning and Knowledge Discovery in Databases*, 2008, pp. 358–373.

[2] D. D. Lee and H. S. Seung, “Learning the parts of objects by non-negative matrix factorization,” *Nature*, vol. 401, no. 6755, pp. 788–791, Oct. 1999.

[3] V. Pauca, F. Shahnaz, M. Berry, and R. Plemmons, “Text Mining using Non-Negative Matrix Factorizations,” in *Proceedings of the 2004 SIAM International Conference on Data Mining*, 0 vols., Society for Industrial and Applied Mathematics, 2004, pp. 452–456.

[4] F. Shahnaz, M. W. Berry, V. P. Pauca, and R. J. Plemmons, “Document Clustering Using Nonnegative Matrix Factorization,” *Inf Process Manage*, vol. 42, no. 2, pp. 373–386, Mar. 2006.

[5] D. Guillamet and J. Vitrià, “Non-negative Matrix Factorization for Face Recognition,” in *Topics in Artificial Intelligence*, 2002, pp. 336–344.

[6] C. Laroche, M. Kowalski, H. Papadopoulos, and G. Richard, “A structured nonnegative matrix factorization for source separation,” *2015 23rd Eur. Signal Process. Conf. EUSIPCO*, pp. 2033–2037, 2015.

[7] S. Wang and J. Ortiz, “Non-negative matrix factorization of signals with overlapping events for event detection applications,” in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2017, pp. 5960–5964.

[8] J.-P. Brunet, P. Tamayo, T. R. Golub, and J. P. Mesirov, “Metagenes and molecular pattern discovery using matrix factorization,” *Proc. Natl. Acad. Sci.*, vol. 101, no. 12, pp. 4164–4169, Mar. 2004.

[9] K. Devarajan, “Nonnegative Matrix Factorization: An Analytical and Interpretive Tool in Computational Biology,” *PLOS Comput. Biol.*, vol. 4, no. 7, p. e1000029, Jul. 2008.

[10] A. Sotiras, S. M. Resnick, and C. Davatzikos, “Finding imaging patterns of structural covariance via Non-Negative Matrix Factorization,” *NeuroImage*, vol. 108, pp. 1–16, Mar. 2015.

[11] E. L. Mackevicius *et al.*, “Unsupervised discovery of temporal sequences in high-dimensional datasets, with applications to neuroscience,” *bioRxiv*, p. 273128, Jun. 2018.

[12] A. Sotiras, J. B. Toledo, R. E. Gur, R. C. Gur, T. D. Satterthwaite, and C. Davatzikos, “Patterns of coordinated cortical remodeling during adolescence and their associations with functional specialization and evolutionary expansion,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 114, no. 13, pp. 3527–3532, 28 2017.

[13] D. P. Varikuti *et al.*, “Evaluation of non-negative matrix factorization of grey matter in age prediction,” *NeuroImage*, vol. 173, pp. 394–410, Jun. 2018.

[14] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*, 2nd ed. New York: Springer-Verlag, 2009.

[15] R. Bro, K. Kjeldahl, A. K. Smilde, and H. a. L. Kiers, “Cross-validation of component models: a critical look at current methods,” *Anal. Bioanal. Chem.*, vol. 390, no. 5, pp. 1241–1251, Mar. 2008.

[16] B. Kanagal and V. Sindhwani, “Rank Selection in Low-rank Matrix Approximations : A Study of Cross-Validation for NMFs,” 2010.

[17] A. B. Owen and P. O. Perry, “Bi-Cross-Validation of the SVD and the Nonnegative Matrix Factorization,” *Ann. Appl. Stat.*, vol. 3, no. 2, pp. 564–594, 2009.

[18] X. E. Lin and P. Boutros, “NNLM: A package For Fast And Versatile Nonnegative Matrix Factorization.” [Online]. Available: https://cran.r-project.org/web/packages/NNLM/vignettes/Fast-And-Versatile-NMF.html.

[19] Y. Zhang and Y. Wang, “Nonnegative Matrix Factorization: A Comprehensive Review,” *IEEE Trans. Knowl. Data Eng.*, vol. 25, no. 6, pp. 1336–1353, 2013.

[20] D. D. Lee and H. S. Seung, “Algorithms for Non-negative Matrix Factorization,” in *Proceedings of the 13th International Conference on Neural Information Processing Systems*, Cambridge, MA, USA, 2000, pp. 535–541.

[21] C. Boutsidis and E. Gallopoulos, "SVD Based Initialization: A Head Start for Nonnegative Matrix Factorization," *Pattern Recogn*, vol. 41, no. 4, pp. 1350–1362, Apr. 2008.

[22] H. Kim and H. Park, "Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis," *Bioinforma. Oxf. Engl.*, vol. 23, no. 12, pp. 1495–1502, Jun. 2007.

[23] A. Frigyesi and M. Höglund, "Non-negative matrix factorization for the analysis of complex gene expression data: identification of clinically relevant tumor subtypes," *Cancer Inform.*, vol. 6, pp. 275–292, 2008.

[24] M. O. Ulfarsson and V. Solo, "Tuning parameter selection for nonnegative matrix factorization," *2013 IEEE Int. Conf. Acoust. Speech Signal Process.*, pp. 6590–6594, 2013.

[25] A. T. Cemgil, "Bayesian Inference for Nonnegative Matrix Factorisation Models," *Comput. Intell. Neurosci.*, vol. 2009, pp. 788–791, 2009.

[26] V. Y. F. Tan and C. Févotte, "Automatic Relevance Determination in Nonnegative Matrix Factorization with the /spl beta/-Divergence," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 7, pp. 1592–1605, Jul. 2013.

[27] S. Squires, A. Prügel-Bennett, and M. Niranjan, "Rank Selection in Nonnegative Matrix Factorization using Minimum Description Length," *Neural Comput.*, vol. 29, no. 8, pp. 2164–2176, Aug. 2017.

[28] R. J. Tibshirani and S. Rosset, "Excess Optimism: How Biased is the Apparent Error of an Estimator Tuned by SURE?," *J. Am. Stat. Assoc.*, pp. 1–16, Feb. 2018.

[29] L.K. Hansen, *NMF:DTU Toolbox*. 2006. Available: http//cogsys.imm.dtu.dk/toolbox/nmf/

[30] T. R. Golub *et al.*, "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," *Science*, vol. 286, no. 5439, pp. 531–537, Oct. 1999.

[31] R. Gaujoux and C. Seoighe, "A flexible R package for nonnegative matrix factorization," BMC Bioinformatics, vol. 11, no. 1, p. 367, Jul. 2010.

[32] P. Brazdil, C. G. Carrier, C. Soares, and R. Vilalta, Metalearning: Applications to Data Mining. Berlin Heidelberg: Springer-Verlag, 2009.