NMF and PCA as applied to gearbox fault data

Anna M. Bartkowiak[1] and Radoslaw Zimroz[2]

[1] Wroclaw University, Inst. of Computer Science, 50-383 Wroclaw PL
[2] Wroclaw University of Technology, Diagnostics and Vibro-Acoustics Science Laboratory, 50-421 Wroclaw PL

Abstract. Both Non-negative matrix factorization (NMF) and Principal component analysis (PCA) are data reduction methods. Both of them act as approximation methods permitting to represent data by lower rank matrices. The two methods differ by their criteria how to obtain the approximation. We show that the main assumption of PCA demanding orthogonal principal components leads to a higher rank approximation as that established by NMF working without that assumption. This can be seen when analyzing a data matrix obtained from vibration signals emitted by a healthy and a faulty gearbox. To our knowledge this fact has not been clearly stated so far and no real example supporting our observation has been shown explicitly.

Keywords: Non-negative matrix factorization, Principal component analysis, Matrix approximation, Optimization with constraints, Gearbox fault detection.

1 Introduction

Nowadays, when a huge amount of data is collected anywhere, the problem of extracting the proper information hidden in the data and reduce their amount – without losing the hidden information – becomes crucial. One popular method of finding the intrinsic dimensionality of the gathered data and reduce them to a smaller representative amount is the method known widely as Principal Component Analysis (PCA) [8]. It permits to construct new independent variables called principal components (PCs), which are representative for the data matrix dealt with.

Nonnegative Matrix Factorization (NMF), launched as a bio-inspired method by Lee and Seug [9], is in its goals similar to PCA, however in its principles quite different. The NMF method was presented as an analogy to recognizing objects by the human brain which learns objects by learning parts of them [9]. The main idea of NMF is that a given observable data matrix $\mathbf{V}$, necessarily with nonnegative elements, is approximated by the product of two smaller rank matrices $\mathbf{W}$ and $\mathbf{H}$, also with nonnegative elements: $\mathbf{V} \approx \mathbf{W} * \mathbf{H}$. The derived approximation matrices $\mathbf{W}$ and $\mathbf{H}$ offer frequently contextual explanation on the structure of the data contained in $\mathbf{V}$. The explanation – because of the non-negativeness of the elements of $\mathbf{W}$ and $\mathbf{H}$ – is easier to perceive. A recent description of the NMF principles with its extensions and real data examples may be found the in the book by Cichocki *et al.* [3], also in Gillis [5] and in references therein.

The publication [9] gave beginning of large amount of discussion and analysis of experimental non-negative data. The method has found interesting applications in image analysis [6, 12], technical fault detection [10], analysis of satellite data [13], music data [4], analysis of of sound waves, EEG, etc., and became more and more popular.

On the other hand, the classic PCA is still much in use. It is developing by adapting a probabilistic approach [1].Our question is: What are the advantages and the disadvantages of the two methods? In particular: May NMF give something different as PCA? We elaborate these questions when looking at a particular set of data collected for a healthy and a faulty gearbox. Our main goal will be to compare the approximation of $\mathbf{V}$ obtained by NMF and PCA.

In the following we introduce in Sect. 2 both methods formally and indicate for similarities and differences between them.

In Sect. 3 we describe shortly the gearbox data (given as matrix $\mathbf{V}_{n\times m}$), which will serve as a test rig for our considerations. We will calculate for our matrix $\mathbf{V}$ its approximates of lower rank using the NMF and PCA methods. The obtained results for rank r=2 and r=3 are shown in Figs. 1 and 2. We consider specifically 3 types of results produces both by NMF and PCA.

Short discussion and closing remarks are presented in Sect. 4.

2 The NMF and PCA methods: definitions, similarities and dissimilarities

Firstly we describe here briefly the NMF and the PCA methods. Next we present a summarization of their main properties valuable for practitioners.

2.1 The NMF method.

After [9], we consider a data matrix $\mathbf{V}$ of size $n \times m$ (denoted also as $\mathbf{V}_{n\times m}$), with n rows and m columns. The authors [9] call this matrix 'visible' (hence denotation by $\mathbf{V}$), because this should be a real observed data matrix. The elements (v_{ij}) of $\mathbf{V}$ should be expressed as non-negative real numeric values satisfying the constraints: $v_{ij} >= 0, \quad i = 1, \dots n,\ j = 1, \dots, m.$

The authors [9] proposed to approximate the recorded data matrix $\mathbf{V}_{n\times m}$ by the product of two lower rank components $\mathbf{W}$ and $\mathbf{H}$ with non-negative elements:

$$\mathbf{V}_{n\times m} \simeq \mathbf{W}_{n\times r} * \mathbf{H}_{r\times m}, \tag{1}$$

where $0 < r <= min(m, n)$ is an integer denoting the rank of the sought component matrices. For practical reasons the following inequality should be satisfied: $(n + m) * r << n * m$.

The first derived approximation matrix $\mathbf{W}_{n\times r}$ is built from r columns vectors $\mathbf{w}_k,\ k = 1, \dots, r$, each of size $n \times 1$, which may be written as

$$\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_r],\ \text{with}\ \ r <= min(n, m).$$

The column vectors $\mathbf{w}_k$ $(k = 1, \ldots, r)$ play the role of representatives of all the m columns of $\mathbf{V}$. They are referred to as the *basis vectors* of the entire data [7]. They may be mutually dependent. Each of them has size $n \times 1$, similarly as each column of the data matrix $\mathbf{V}$. We call them *meta-features* (MFs) of the analyzed data. They may be also considered as building blocks for the observed data matrix $\mathbf{V}$ [7].

The second component matrix $\mathbf{H}$ appearing in (1) contains elements called *encoding coefficients*. Matrix $\mathbf{H}$ is frequently viewed as composed from r row vectors denoted as $\mathbf{h}^{(k)},\ k = 1, \ldots, r$. Obviously: $\mathbf{h}^{(k)} = [H_{k,1}, H_{k,2}, \ldots, H_{k,m}]$.

The original observed data matrix $\mathbf{V}$ may be approximated either directly by the product of $\mathbf{W}_{n\times r} * \mathbf{H}_{r\times m}$, as shown in (1), or in r cumulative steps as:

$$\mathbf{V}_{n\times m} \simeq \sum_{k=1}^{r} \mathbf{w}_k * \mathbf{h}^{(k)}. \tag{2}$$

The above formula (2) says that the approximation of the entire data matrix $\mathbf{V}$ may be done in r steps; each step uses the product matrix $[\mathbf{w}_k * \mathbf{h}^{(k)}]$, which provides a rank one approximation matrix of size $n \times m$.

Algorithms for finding the factorization matrices $\mathbf{W}$ and $\mathbf{H}$

The authors [9] launched three algorithms for three objective functions yielding the sought factorizing matrices (mean square error, Poissonian likelihood and Kullback-Leibler divergence). All algorithms work as iterative multiplicative algorithms and all they – being easily programmable in Matlab – contributed to the popularity of NMF.

The discussion od these algorithms may be found in [2]. The extensions of NMF went in the direction of using other optimization criteria, in particular, *differently defined divergences*, also in adding additional constraints suitable, for example, for discriminant analysis or cluster analysis [3, 5, 17, 16, 4, 7, 11, 12].

2.2 The PCA method.

Let $\mathbf{X}_{n\times m}$ denote the observed data matrix with real numeric values. The matrix $\mathbf{X}$ is centered and usually additionally standardized to have unit variances. The rows of $\mathbf{X}$ denote objects ('individuals'), each of them characterized by m variables, contained in the columns of $\mathbf{X}$. The PCA method seeks for new (constructed) variables which (i) are linear functions of the primary variables contained in $\mathbf{X}$, (ii) are mutually orthogonal, and (iii) explain as much as possible of the total sum of variances evaluated for all m variables included into $\mathbf{X}$. The sought new variables (their realizations $Y_1, Y_2, \ldots, Y_r$ called *Principal Components*) are then constructed as [8]:

$$Y_k = \mathbf{X} * \mathbf{a}_k, \quad k = 1, \ldots, r\ (r \leq m), \tag{3}$$

where $\mathbf{a}_k$ are eigenvectors of the covariance (or correlation) matrix of $\mathbf{X}$.

The reconstruction of $\mathbf{X}$ is done on the basis of the derived new variables (i.e. principal components Y_k) and connected with them eigenvectors $\mathbf{a}_k$:

$$\hat{\mathbf{X}}^{(r)} = \sum_{k=1}^{r} Y_k * \mathbf{a}_k^T, \quad \text{or equivalently} \quad \hat{\mathbf{X}}^{(r)} = [Y_1, \ldots, Y_r] * ([\mathbf{a}_1, \ldots, \mathbf{a}_r]^T). \quad (4)$$

Comparing the two methods

Similarity

• Both methods have the same goal: Reduction of dimensionality by constructing new meaningful features which for *r=2* and *r=3* may serve for a 2D- and a 3D-visualization of the original data vectors projected to the reduced space.

Dissimilarity

• Both methods use different optimization criteria how to obtain their goals, hence different algorithms for constructing the approximation matrices. NMF approximates the entire data matrix $\mathbf{X}$. PCA approximates the trace of $\mathbf{S} = \frac{1}{n-1} * (\mathbf{X}^T * \mathbf{X})$ of properly standardized data matrix $\mathbf{X}$.

• Different properties of the constructed features: PCA yields uncorrelated features, NMF yields features which may be correlated.

• NMF needs a data matrix with non-negative elements; PCA takes any real numeric values, however the work *in the classic setting* is done using centered data matrix (where the elements are *ex definitione* both positive and negative) and the proper calculations (evaluating eigen-values and eigen-vectors) are carried out using covariance or correlation matrix of the (rescaled) data.

• NMF is iterative with a random start; results may vary in different runs. Algorithm for PCA is stabile, and in a repeated runs one obtains the same solution.

3 The analyzed data matrix and its approximation by NMF and PCA

3.1 The data

We have used data described and analyzed partially in [14, 15]. For the present analysis, we have sampled a data matrix $\mathbf{V}$ of size 1000×15. All elements of this 'visible' matrix are positive. In rows (called segments) we have firstly 500 data vectors from the healthy, and next 500 data vectors from the faulty gearbox. Columns in $\mathbf{V}$ contain values of 15 variables called $pp1, \ldots, pp15$ evaluated for each row vector of $\mathbf{V}$. The pp variables are power spectra characteristics obtained from a PSD analysis of subsequent segments of the vibration signal emitted by the machines (their gearboxes).

To identify more easily graphical results, the 500-samples were sorted according to an external variable called *ZWE* denoting load state of the gearboxes.

3.2 Main Results from NMF and PCA applied to the gearbox data

We have investigated the approximation for *r=2* and *r=3*. For NMF we have used the matlab nmfmse function [7] working with the mean square error (*mse*) criterion applied to $\mathbf{V}$. For PCA we have used own script in Matlab which needed only the eig function from basic package, without any additional toolboxes.

Applying NMF we obtained for $r = 2$ and $r = 3$ the approximations: $\mathbf{V}_{1000\times 15} \simeq \mathbf{W2}_{1000\times 2} * \mathbf{H2}_{2\times 15}$ and $\mathbf{V}_{1000\times 15} \simeq \mathbf{W3}_{1000\times 3} * \mathbf{H3}_{3\times 15}$.
The results of our analysis are depicted in Fig. 1.

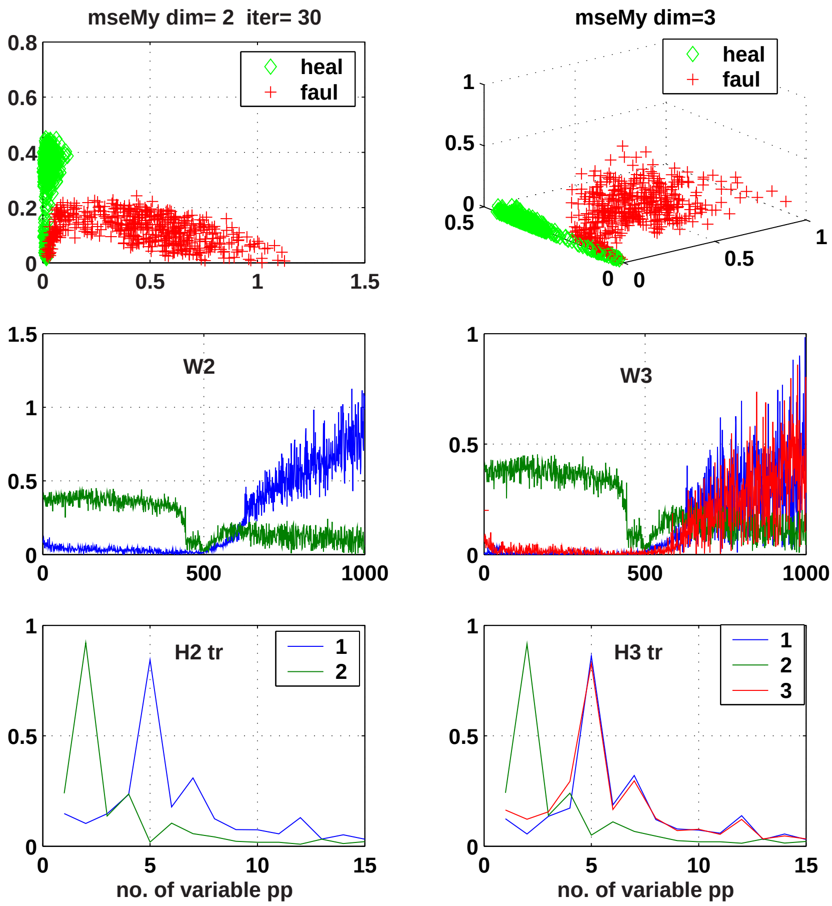

Fig. 1. Main results from NMF when using approximation of rank $dim = r = 2$ (left column of plots) and of rank $dim = r = 3$ (right column of plots). See text for explanation.

When **applying the PCA method** and using the correlation matrix for calculation of the eigenvectors, we obtained for r=*2* two eigenvectors $[\mathbf{a}_1, \mathbf{a}_2\ ]$ and two principal components $[PC1, PC2]$. When opting for dimension r=*3*, we needed additionally the eigenvector $\mathbf{a}_3$ and the additional principal component denoted as PC3. The results are depicted in Fig. 2.

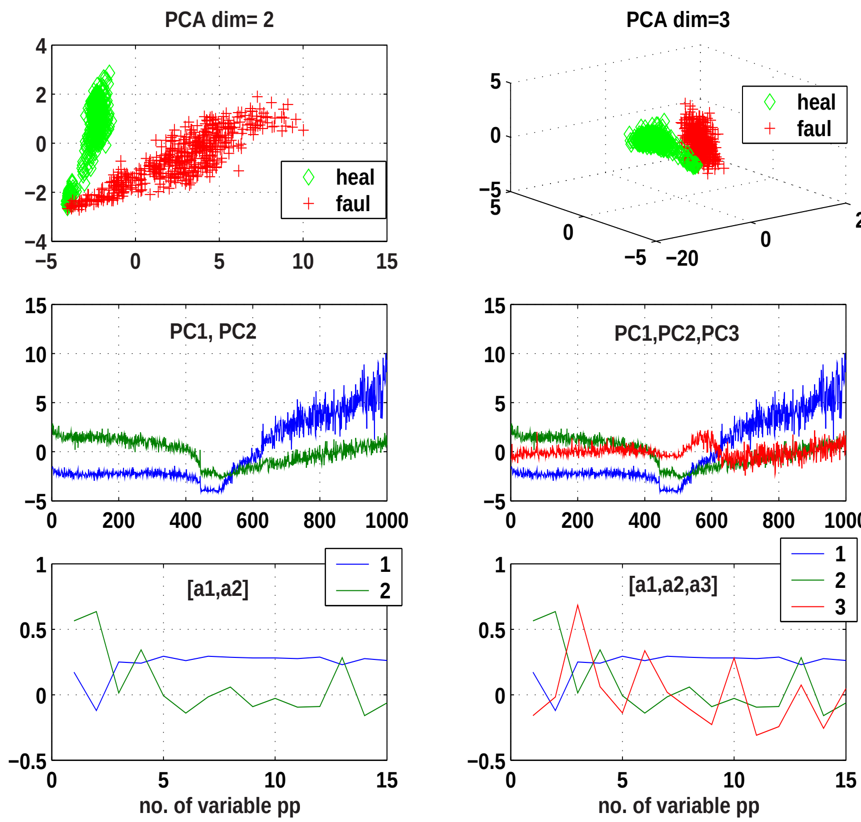

Fig. 2. As in Fig. 1, however depicting results from PCA. See text for explanation.

Both figures contain 3 types of graphs depicted for r=*2*(left) and r=*3* (right):

Type 1. Scatter plots of data points shown in the coordinate system of meta-features W2 or W3 (for NMF), or from the principal components [PC1,PC2] and [PC1,PC2,PC3] (for PCA). Shown in upper row of subplots in the respective figures.

Type 2. Profiles of the meta-features W2 and W3 (for NMF), and of the PC-s [PC1,PC2] or [PC1,PC2,PC3] (for PCA). Shown in middle row of subplots in the respective figures.

Type 3. Profiles of the encoding matrices H2 and H3 (for NMF), and of the corresponding eigenvectors $[\mathbf{a}_1, \mathbf{a}_2]$ and $[\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3]$ (for PCA). Shown in bottom row of subplots in the respective figures.

3.3 What is seen in the 3 types of graphs from from NMF and PCA

Graphs of the first type, located in subplots 1 and 2 of both figures, show scatter plots depicting subsequent data vectors (rows) of the analyzed data as represented by the MFs contained in W2 – for NMF, and similarly the representations by the first two or first three PC-s – for PCA. Both methods recognize properly that the data are composed from two different groups of data.

Graphs of the second type, located in subplots 3 and 4 of both figures, show profiles of the derived MFs contained in W2 and W3 (Fig.1), and of the respective PC-s (Fig.2).

Concerning the exhibits in subplot 3, based on W2 or [PC1,PC2], one might say that the exhibits show the same pattern for both methods. We may deduce here that the first MF (or PC1) expresses the *faultiness* of the gearbox, and that the second MP (or PC2) expresses the normal functioning of the gearbox. The dip in the middle, around items no.s 500, corresponds to the no/low load state of the gearboxes.

Concerning the exhibits in subplot 4 based on W3 or [PC1,PC2,PC3], it is obvious that the exhibits are essentially different. The exhibit from NMF depicts the three MFs contained in W3. We see here that the first and the third MF are overlapping (with a correlation coefficient r=0.92), moreover they have high variability for the faulty items. One may deduce: Dimension $r=3$ is too high for the analyzed data. Analogous exhibit from PCA looks stabilized. Segments no.s 1-500 have distinct values for all three PC-s. Only the PC2 for faulty segments is overshadowed by PC3 (however not so much as was observed on the graph obtained by NMF). One may deduce: The dimension $r=3$ is OK, 3 dimensions are needed, especially when considering the 'healthy ' part of the data coded as items 1-500.

Graphs of the third type, located in subplots 5 and 6 of both figures, show profiles of the transposed encoding matrix H2 or H3. The exhibits from NMF and PCA are completely different.

Subplot 5 and 6 obtained from NMF shows very clearly two profiles obtained from the first two columns of $(H2)^T$ and $(H3)^T$ appropriately. They are similar in the left and right exhibit. The third profile in exhibit 6 coincides practically with the 1st profile of that exhibit. Analogous Subplots 5 and 6 obtained from PCA show completely different profiles of $[\mathbf{a}_1, \mathbf{a}_2]^T$ and $[\mathbf{a}_1, \mathbf{a}_2, \mathbf{a}_3]^T$. The 3 profiles in subplot 6 are different and do not resemble those obtained by NMF, each of them is needed for encoding. The same can be said about profiles exhibited in subplot 5.

4 Discussion and closing remarks

We have compared results of NMF and PCA for a set of gearbox data (15 variables characterizing 1000 segments of recorded vibration signals from a healthy and a faulty gearbox). Our goal was to reduce dimensionality of the data. Both methods constructed new features representative for the observed variables. We found that both methods are effective for that purpose. We found also that PCA - due to the requirement of the orthogonality of the constructed new features - is perhaps more stabile, but also comes later to its goal, that is, it chooses higher dimensionality as the NMF does.

References

1. Bartkowiak A.M., Zimroz R.: Probabilistic principal components, how this works. Saeed K., E. Homenda E. (eds.): CISIM'2015, IFIP LNCS 9339, 24-35, (2015)
2. Berry M.W., et al.: Algorithms and applications for approximate nonnegative matrix factorization. Computational Statistics & Data Analysis 52, 155-173 (2007)
3. A. Cichocki, R. Zdunek, A.H. Phan, Sh. Amari: Nonnegative matrix and tensor factorizations. Applications to exploratory multi-way data analysis and blind source separation. Wiley, Chichester U.K. (2009)
4. Fevotte C., Bertin N., Durrieu J-L.: Nonnegative matrix factorization with the Itakura-Saito divergence. With application to music analysis. Neural Computation 21 (3), 793-830 (2009)
5. Gillis N.: The why and how of nonnegative matrix factorization. In: Suykens J.A.K., et al. (eds.), Regularization, Optimization, Kernels and Support Machines. Chapman&Hall/CRC, Chapter 1, pp. 3-39 (2014)
6. Guillamet D., Vitria J., Schiele B., Introducing a weighted non-negative matrix factorization for image classification. Pattern Recogn. Lett. 24, 2447-2454 (2003)
7. Hoyer P.O.: Non-negative matrix factorization with sparseness constraints. J. Mach. Learn. Res. 5, 1457-1469 (2004)
8. Jolliffe I.T.: Principal Component Analysis, 2nd edn. Springer, New York (2002)
9. Lee D.D., Seung H. S.: Learning the parts of objects by nonnegative matrix factorization, Nature 401, 788–791 (1999)
10. Li B., et al.: A new feature extraction and selection scheme for hybrid fault diagnosis of gearbox. Expert Syst. Appl. 38, 1000-1009 (2011)
11. Li Xi, Fukui K.: Fisher non-negative matrix factorization with pairwise weighting. IAPR Conf. on Machine Vision Application, MVA 2007, pp. 380-383, Tokyo, Japan, May 16-18 (2007)
12. Li S.Z. et al.: Learning spatially localized, parts-based representation. In: Proc. of IEEE International Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 207-2012, Havaii December 2001
13. Tsagaris V., Anastassopoulos V.: Feature extraction of Meris data. In: Proceedings of the 2004 Envisat & ERS Symposium, Salzburg, Austria, 6-10 September 2004 (ESA SP-572, April 2005), pp. 1-5 (2005)
14. Zimroz R., Bartkowiak A.: Two simple multivariate procedures for monitoring planetary gearboxes in non-stationary operating conditions. Mech. Syst. Signal Process. 38, 237-247 (2013)
15. Zimroz, R., Bartkowiak, A.: Investigation on spectral structure of gearbox vibration signals by principal component analysis for condition monitoring purposes. Journal of Physics: Conference Series 305 (1), art. no. 012075 (2011).
16. Zdunek R.: Trust-region algorithm for nonnegative matrix factorization with alpha- and beta-divergences. A. Pinz et al. (eds), DAGM/OAGM 2012, LNCS 7476, 226-235, Springer (2012).
17. Zurada J.M., Ensari T., Asi E.H., Chorowski J.: Nonnegative matrix factorization and its application to pattern recognition and text mining. Proc. of the 13th FedCSIS Conf. on Computer Science and Information Systems, Cracow, pp. 11-16 (2013)