Discussion of NMF Papers

Jon R. Kettenring[1]
[1]Research Institute for Scientists Emeriti (RISE), Drew University, Madison, NJ 07940

Abstract
This is a summary of comments I made about four papers presented in the JSM Session on nonnegative matrix factorization (NMF).

Key Words: principal components analysis, cluster analysis, latent semantic indexing, singular value decomposition, canonical correlation analysis, contingency tables, robust estimation

1. Principal Components Analysis (PCA)

Hotelling's famous paper on PCA was published 80 years ago this year. (Hotelling 1933). I will begin my discussion by recalling some of the key properties of PCA because of its close ties to nonnegative matrix factorization (NMF) and its importance in multivariate statistical analyses.

PCA is based on elegant well understood mathematics. It possesses useful theoretical properties. The computations are straightforward and are usually based on the spectral decomposition of the covariance or correlation matrix or on the singular value decomposition (SVD) of a data matrix, $X\ (n \times p)$, after appropriate centering and scaling of the p variables. Data-based PCs are unique apart from sign, except in special cases. They come nicely ordered in terms of importance. Users usually focus on a few of the leading PCs. Fortunately, these do not change if it is decided to consider additional ones later.

PCA has its awkward points. PCs based on the covariance matrix typically have nothing to do with those based on the correlation matrix. This is a reminder that scaling really does matter with PCA. While there are various ad hoc methods for determining how many PCs to retain, none of them is fail-safe. Furthermore, those that are retained are often very hard to interpret. Attempts to make them understandable can lead to charges of reification. Unless the number of observations, n, is large enough relative to p, the PCs can be unstable, which will only increase the possibility of reading too much into them.

PCA is basically a least squares method and as such is well known to be highly sensitive to outliers. Indeed, the sensitivity is great enough that it can be used as an effective vehicle for detecting multivariate outliers. Robustly centering and scaling the data prior to PCA can help to establish the core of the data and enhance the detection process.
Hotelling developed PCA against the backdrop of multivariate normal distributions or data. The PCs were designed to capture orthogonal directions of greatest variability. However, the most common usage of PCA is to find clusters, a consideration that Hotelling does not even mention. The fuzzy logic for doing this is that directions of

largest variability may also correspond to directions where clusters reveal themselves most clearly. Fortunately for those who take this cheap road to cluster analysis, it often works. Little attention is paid to the fact that clear-cut, albeit more subtle, cluster structure can easily be missed, and might only show up among the smaller PCs if at all.

There have been numerous attempts to improve PCA. A very simple idea is to reduce the number of variables in the analysis using subject-matter knowledge or purely statistical considerations, as discussed in Jolliffe (1972, 1973). Some of his empirical work with real data even suggests that the number of variables can sometimes be cut in half without loss. The CUR matrix decomposition approach, developed in Mahoney and Drineas (2009), uses a leverage statistic to decide which variables to retain. Sparse PCA (Zou, Hastie, and Tibshirani, 2006) is a lasso-based methodology for picking modified PCs with hopefully many zero coefficients to ease interpretation. Several approaches to robust PCA are mentioned by Candes et al. (2009), who also offer a recent perspective on the topic.

While Hotelling assumed that the data for PCA are multivariate normal, the modern assumption is murkier. A scan of the applied literature would reveal that the (usually unstated) assumptions are only that the data are multivariate, "reasonably" continuous, and possibly grouped.

2. Nonnegative Matrix Factorization (NMF)

There are several very nice overview papers on NMF. They can help to introduce the topic to statisticians, which is one of the goals of this session. The opening paper presented by Luta et al. is the most current of the batch and contains helpful examples and insights. I would also recommend Lee and Seung (1999) and Devarajan (2008). The paper by Brunet et al. (2004) shows how NMF can be used to reduce the dimensionality and cluster gene expression data.

The assumptions behind NMF are about as fuzzy as those for PCA. The basic one is that the data to be modeled are nonnegative (apart possibly from some contamination or random noise).

NMF is performed by factoring X without any centering. The factorization is into the product of two non-negative rank k matrices, with an error term left over. The model is often described as consisting of "additive parts" plus error in contrast to the less constrained SVD-based model.
Whether fitting the additive parts by least squares or another criterion, iterations are required to search for a local optimum. Checking whether the solution is also global will be necessary. In brief, there will be a lot of computing to do, and this may prove quite a challenge for large or massive datasets.

Luta et al. mention the lack of uniqueness of the basic NMF decomposition. This is a very important point with practical implications. At best the decomposition is determined only up to an arbitrary $k \times k$ scaling matrix, where k is the rank of the factoring matrices. Since subsequent computations, e.g., for cluster analysis, may be based on the individual matrices in the decomposition, it is essential that a rational way be found for settling the indeterminancy. (This point seems not to have received much attention.) Donoho and Stodden (2004) ask: under what conditions is NMF well defined

and correct. Rules of thumb to guide practitioners considering the use of NMF would be very helpful.

The choice of the rank parameter *k* is a critical step in NMF. Unlike PCA, the additive parts of the NMF can vary fundamentally with *k.* Luta et al. argue that picking *k* correctly is not as important for NMF as it is for PCA or SVD. I am not so sure about that!

3. Robust NMF

Keeping in mind the sensitivity of PCA to outliers, it is not surprising that NMF has the same issue. The question is how to effectively robustify the fitting process. The presentation by Sun et al. offers some promising ideas on how to do this. My suggestion for this work is to connect it as much as possible to existing ideas, methods, algorithms, and theory from the field of robust estimation. Ideas such as influence functions and breakdown points should be explored, for example.

4. Contingency Tables

The presentation by Das et al. offers an NMF approach to two-way count data. They title their paper "Contingency Table Analysis via Matrix Factorization", but not all two-way count data is of this type. Specifically, contingency tables capture the association between two variables. The usual goals are to measure the strength of this association and to test for independence.

Das et al. use an idea of I. J. Good's (1969) for testing the independence of the rows and columns in such a table. Independence implies the contingency table matrix has rank equal to one. If it is not of rank one, then maybe it is of rank 2, etc. Good proposed to use chi-squared tests for these hypotheses which are based on the SVD of the contingency table matrix. Das et al. suggest using an NMF factorization instead of the SVD. This may be helpful in some types of two-way data, but it is not clearly so for standard contingency tables.

Another of Hotelling's great contributions is relevant here: canonical correlation analysis (CCA). If each row-column pair of observations involved in the contingency table is represented by a pair of 0-1 indicator vectors, then the table can be analyzed by CCA. This is a well known fact (see, e.g., Kendall and Stuart, 1961). Testing for independence is usually done using a chi-squared test statistic which is proportional to the sum of squares of the canonical correlations. These correlations in turn are the singular values not of the raw contingency table but a modified version of it, which may have negative values. In my opinion this makes more sense than Good's approach. One last note: the CCA could be modified to force the coefficients which define the canonical variables to be nonnegative, in the spirit of NMF.

5. Nonnegative Data

Both Luta et al. and Das et al. assert that when the data matrices are constrained to be nonnegative, then the matrix factors should "arguably" be nonnegative as well. This need not be the case. The standard contingency table discussed above is a prominent example.

Another example is latent semantic indexing (LSI) or analysis (Deerwester et al., 1990). The data for LSI is a term-by-document matrix of word counts for an electronic collection of documents. Hence all of the entries are nonnegative. Nevertheless, while NMF may be a logical method to apply, LSI utilizes the SVD to decompose the matrix. The leading terms, which may be large in number, are used in the document searching. LSI has proved to be remarkably successful even for very large collections. One reason is no doubt its speed and efficiency. Another is that users don't typically pause to analyze the unwieldy singular vectors involved. Success, instead, is measured by the end quality of the search process—whether or not the documents identified as most relevant turn out to be just that. Indeed, black box approaches do have their place!

Taking the LSI discussion one step further, it may make sense to consider both "additive" and "subtractive" parts in the modeling of term-by-document matrices. For example, I might want to retrieve from a collection of multivariate analysis papers all those pertaining to NMF but omit those written by Stan Young because I already have those. Or maybe I want to search a library of travel documents for information about Miami, Florida but not Miami, Ohio.

6. Cluster Analysis

Luta et al. point out that "NMF factors can be used in much the same way as those coming from PCA" for cluster analysis. There is related literature on this including the paper by Brunet et al, 2004. The number of clusters is assumed equal to the rank number, k. One issue that looms large is how to scale the relevant factoring matrix, given its indeterminancy, and the potential effect this may have on the clustering. Special care seems warranted given that the clustering is being performed on approximating factor matrices rather than the raw data or their exact derivatives such as interpoint distances.

7. Theoretical Developments

There has been excellent theoretical progress on the NMF topic but more is needed. The paper by Devarajan et al. (2011) develops a unified algorithm for NMF and provides an elegant treatment of divergence measures to assess goodness-of-fit. They offer a probabilistic LSI model invoking the assumptions that word counts are independent and follow a Poisson distribution. The independence assumption is convenient although it may be hard to justify in practice. The question is how much impact it has on the analysis.

8. Conclusion

The papers in this session have been excellent for introducing NMF to the statistical community. Clearly, NMF has appealing advantages for certain types of problems involving nonnegative data. This is important because nuances and complexities associated with NMF suggest that there is more to worry about with this methodology than with traditional methods such as PCA.

As more is learned about NMF, it will be crucial to obtain a better understanding of its strengths, weaknesses, and limitations. The review paper by Devarajan mentions that "normalization of the observed data prior to NMF analysis is an important problem and

one that has not been systematically studied." This is a reminder that NMF has some of the same issues as PCA. Another, no doubt, is the impact of small samples sizes, when $n << p$, which can cause havoc with PCA. As more evidence is accumulated we should reach a better understanding of both the *pros* and the *cons* of NMF.

Acknowledgements

Many thanks to Stan Young of NISS and the authors of the papers presented in this session for shining light on this important and timely topic.

References

Brunet, J-P., Tamayo, P. Golub, T. R., and Mesirov, J. P. (2004. Metagenes and molecular pattern discovery using matrix factorization. P NATL ACAD SCI USA 101, 4164-4169.

Candes, E. J., Li, X., Ma, Y., and Wright, J., (2009). Robust principal component analysis? J ASSOC COMP MACH **58**, 1-37.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., Harshman, R. (1990). Indexing by latent semantic analysis. J AM SOC INFORM SCI **41**, 391-407.

Devarajan, K. (2008). Nonnegative matrix factorization: an analytical and interpretive tool in computational biology. PLOS COMPUT BIOL **4**, 1-12.

Devarajan, K., Wang, G., and Ebrahimi, N. (2011). A unified approach to non-negative matrix factorization and probabilistic latent semantic indexing. *COBRA Preprint Series*, Working Paper 80.

Donoho, D. and Stodden, V. (2003). When does non-negative matrix factorization give a correct decomposition into parts? ADV NEUR IN.

Good, I. J. (1969). Some applications of the singular decomposition of a matrix. TECHNOMETRICS **11**, 823-831.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. J EDUC PSYCHOL **24**, 417-441, 498-520.

Hotelling, H. (1936). Relations between two sets of variates. BIOMETRIKA **28**, 321-377.

Huber, P. J. (1981). *Robust Statistics*. Wiley, New York.

Jolliffe, I. T. (1972). Discarding variables in a principal component analysis. I. Artificial Data. J ROY STAT SOC C-APP **21**, 160-173.

Jolliffe, I. T. (1973). Discarding variables in a principal component analysis. II. Real Data. J ROY STAT SOC C-APP **22**, 21-31.

Kendall, M. G. and Stuart, A. (1961). *The Advanced Theory of Statistics, Vol. 2*. Hafner, New York.

Lee, D. D. and Seung, H. S. (1999). Learning the parts of objects by non-negative matrix factorization. NATURE **401**, 788-791

Mahoney, M. W. and Drineas, P. (2009). CUR matrix decompositions for improved data analysis. P NATL ACAD SCI USA **106**, 697-702.

Zou, H., Hastie, T., and Tibshirani, R. (2006). Sparse principal component analysis. J COMPUT GRAPH STAT **15**, 265-286.