Clustering Scotch Whiskies using Non-Negative Matrix Factorization

S. Stanley Young, National Institute of Statistical Sciences Paul Fogel, Consultant, France
Douglas Hawkins, University of Minnesota

Most likely, you are familiar with the singular value decomposition, SVD, of a matrix whereby a matrix is decomposed into three parts, X = L Λ R', the left eigenvectors, a diagonal matrix of singular values and a matrix of right eigenvectors. If X is *n x p*, it is often useful to approximate X using only k<min(n,p) singular values, right and left eigenvectors, so that you have $X = L_k \Lambda_k R_k + E$, where E is an error or residual matrix. For an insightful paper on the use of SVD in statistics, Good (1969) is worth a careful reading. More recently, the non-negative matrix factorization, NMF, of a matrix with positive elements into two positive element matrices, Lee and Seung (1999), Hoyer (2004), with X=WP+E, has attracted much deserved attention. W can be taken to be a weight matrix and P a profile matrix. Usually only *k* components are fit, so E is an error or residual matrix. Very remarkably, when Lee and Seung decomposed pictures of human faces, NMF appeared to assign an eigenvector to each of the face parts. For the same data set, SVD produced right eigenvectors with no obvious visual interpretation. Donoho and Stodden (2003) examine the question of NMF uniquely (up to a multiplication factor) isolating out "parts/mechanisms." A good example of the use of NMF is Brunet et al. (2003), who examine a micro array data set using NMF. Whereas SVD can be solved easily with alternating least squares (see Good (1969), Gabriel and Zamir (1979), and Liu et al. (2003)) non-negative matrix factorization involves complex optimization. One strategy is to load the W and P matrices with random numbers and use an updating strategy to minimize the squared differences of the elements of X and WP. Lee and Seung (1999) minimize a Divergence criterion derived from a maximum likelihood approach assuming a Poisson distribution for the matrix cells. In this note, we show the utility of NMF using it to cluster a data set of 86 Scotch Whiskies.

Scientists often encounter two-way data tables. In our case, we have 86 Scotch whiskies that have been rated on a five-point scale for 12 flavor characteristics: Body, sweetness, smoky, medicinal, tobacco, honey, spicy, winey, nutty, malty, fruity, and floral. This data set comes from a wonderful book on the classification of Scotch whisky based on flavors by David Wishart (2002). The first comment is that good classification usually involves subject knowledge. The Wishart book provides considerable background knowledge of Scotch whisky. In particular, production methods are described in some depth so that factors that contribute to ultimate flavor are made clear. Wishart's (Wishart (2002)) premise is that people are interested in single malt whiskies for their different flavors and can benefit from a refined analysis of the factors of flavors (as opposed to a simple single "quality" scale). If the consumer understands the different dimensions of flavor then exploring different flavors and finding additional single malts of interest is possible. However, a 12-dimensional world is very big. How the 86 readily available single malts cluster is a major topic of Wishart's book. He provides 10, 6, and 4 level clusterings of the single malts.

How can NMF be used for clustering this data set? In many ways the data is ideal to display some of the potential advantages of NMF. The X matrix is positive with each flavor scaled from 0 (not present) to 4 (pronounced). Water and grain neutral alcohol have no flavor so all the flavors of Scotch whisky are designed or engineered in. For example, the still master controls the "cut" of the distillation process to give fewer or more fermentation secondary compounds, e.g. aldehydes, esters, ketones, acids, etc. The whiskey is aged in a wide variety of oak casts used to impart flavors, e.g. old, young, European or American, previously used for wine, port, other whiskies, etc. It is not unreasonable to think of the flavor components being "layered onto" the starting unflavored product, water and alcohol. Some of the production methods might add several flavor components in essentially fixed ratios. Might we find some prototypical flavor patterns and that the individual single malts are combinations of these flavor patterns? So our left eigenvectors will be the mixing levels, W, and our right eigenvectors will be the prototypical flavor patterns, P. Let's see how it works out. First, how many flavor patterns are present? For NMF a Scree plot can be computed and is shown in Figure 1.

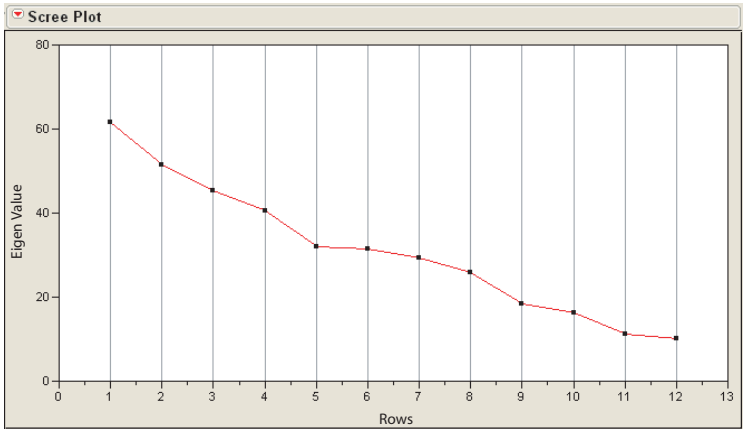

Figure 1. NMF Scree Plot

Obviously considerable care has been given to the various flavors chosen to characterize Scotch whisky as there is no dramatic clustering of the flavors. Even so, there appear to be "jumps" in the Scree plot when going from 9 to 8 factors and from 5 to 4. Zhu and Ghodsi (in press) give a method defined as profile likelihood, for evaluation of a Scree plot that gives the likelihood of a mixture distribution – where

the Scree plot can be cut so that noise components are to the right and signal components are to the left. The Scree plot for the profile likelihood method is shown in Figure 2.

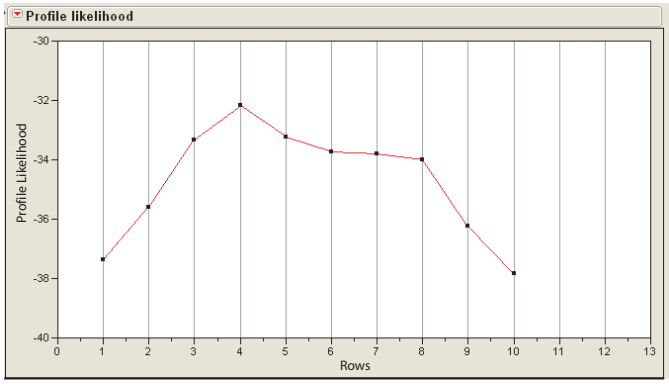

Figure 2. Profile Likelihood Scree Plot.

We choose to go with four major flavor factors. A look at the right eigenvectors is instructive and shown in Table 1. Component 1 mainly contains fruity, sweetness and floral characteristics. To a lesser extent Component 1 contains malty and nutty flavors. (We normalized each component so that the largest element is one.) Component 2 contains spicy, sweetness and floral flavors and to a lesser extent honey and malty flavors. Note that these components are not mathematically orthogonal. Pure flavors are not expected to be available to the designer of a single malt; it seems weird that a mathematical detail like orthogonality should intrude into the data analysis process! Component 3 has the winey flavor isolated from the other components. This suggests that winey flavor might be added specifically to the single malt, perhaps through the use of oak barrels previously used for wine aging. Medicinal, smoky, body, and tobacco flavors are captured by Component 4.

Are there single malts that appear to be relatively pure embodiments of these four flavor profiles? Results for eight single malts are given in Table 2. These eight single malts, two for each cluster, have flavor profiles very close to the flavor profile for the four flavor profile vectors that can be used to reconstruct all the single malt flavor profiles. (Here we normalized each component so that the sum of weights for each weighting component is one and assigned each single malt to its closest prototypical component. More specifically, we assigned each row of the matrix (single malt) to the number of the component with the highest element.)

Table 2. Profiles for Eight single malts.

	RowID	Cluster	Distillery	Con1	Con2	Con3	Con4
1	03	1	AnCnoc	0.0288	0.0000	0.0000	0.0000
2	65	1	Miltonduff	0.0250	0.0000	0.0000	0.0000
3	83	2	Tomatin	0.0000	0.0303	0.0156	0.0000
4	73	2	Speyburn	0.0108	0.0257	0.0000	0.0000
5	43	3	Glendronach	0.0000	0.0081	0.0569	0.0000
6	63	3	Macallan	0.0089	0.0085	0.0446	0.0000
7	04	4	Ardbeg	0.0006	0.0000	0.0000	0.0947
8	24	4	Clynelish	0.0052	0.0000	0.0000	0.0775

Some wild ideas: these eight single malts might form a small, cost-effective inventory to keep your single malt drinking friends happy; one of each pair might serve as the basis for making Scotch blends; it might be instructive to map these four flavor profiles back to the manufacturing methods.

Some final comments: SVD is mathematically driven to maximize the coverage of the variance of the measurements and have orthogonal components whereas non-negative matrix factorization aims to decompose the matrix into profile vectors and the weighting of those vectors to reconstruct the observed, positive data. Here and in other examples, not shown, the resulting right eigenvectors appear to point to underlying parts of a mixture or mechanisms. We think NMF can offer easier insight into the problem behind the non-negative data matrix.

Table 1. Right Eigenvectors for Four Major Flavor Factors.

	Variable	Comp 1		Variable	Comp 2		Variable	Comp 3		Variable	Comp 4
1	Fruity	1.0000	1	Spicy	1.0000	1	Winey	1.0000	1	Medicinal	1.0000
2	Sweetness	0.8833	2	Sweetness	0.6178	2	Body	0.7269	2	Smoky	0.8703
3	Floral	0.8019	3	Floral	0.5664	3	Nutty	0.5683	3	Body	0.7454
4	Malty	0.6341	4	Honey	0.4794	4	Fruity	0.4306	4	Spicy	0.3965
5	Nutty	0.6004	5	Malty	0.4408	5	Honey	0.3781	5	Malty	0.2830
6	Body	0.4538	6	Smoky	0.3361	6	Smoky	0.2926	6	Nutty	0.2575
7	Smoky	0.2985	7	Body	0.3031	7	Sweetness	0.2862	7	Tobacco	0.2128
8	Honey	0.2474	8	Fruity	0.0000	8	Malty	0.2531	8	Sweetness	0.1683
9	Medicinal	0.0000	9	Nutty	0.0000	9	Spicy	0.0000	9	Fruity	0.1646
10	Tobacco	0.0000	10	Medicinal	0.0000	10	Floral	0.0000	10	Winey	0.0000
11	Spicy	0.0000	11	Tobacco	0.0000	11	Medicinal	0.0000	11	Honey	0.0000
12	Winey	0.0000	12	Winey	0.0000	12	Tobacco	0.0000	12	Floral	0.0000

Code for SVD and NMF

The computations done for this paper were done with a special script written for SAS JMP. There are public programs available for SVD, *http://www.niss.org/PowerMV,* and non-negative matrix factorization, *http://www.simonshepherd.supanet.com/nnmf.htm*

Data

The data set used in this analysis can be obtained from the first author at *young@niss.org*.

References

Brunet J.P.; Tamayo P.; Golub T.R.; and Mesirov J.P. (2004). "Metagenes and Molecular Pattern Discovery Using Matrix Factorization". *Proceedings of the National Academy of Sciences* 101, pp. 4164–4169.

Donoho, D. and Stodden, V. (2003). "When Does Non-Negative Matrix Factorization Give a Correct Decomposition Into Parts?" *Advances in Neural Information Processing Systems 16.* MIT Press, Cambridge, MA.

Gabriel, K.R. and Zamir, S. (1979). Lower Rank Approximation of Matrices by Least Squares with any Choice of Weights. *Technometrics* 21, pp. 489–498.

Good, I.J. (1969). "Some Applications of the Singular Decomposition of a Matrix". *Technometrics* 11, pp. 823-831.

Hoyer, P.O. (2004). "Non-negative Matrix Factorization with Sparseness Constraints". *Journal of Machine Learning Research* 5, pp. 1457–1469.

Lee, D.D.and Seung H.S. (1999). "Learning the Parts of Objects by Non-Negative Matrix Factorization". *Nature* 401, pp. 788-791.

Liu, L.; Hawkins, D.M.; Ghosh, S. and Young, S.S. (2003). "Robust Singular Value Decomposition Analysis of Microarray Data". Proceedings of the National Academy of Sciences 100, pp. 13167-13172.

Wishart, D. (2002). *Whisky Classified, Choosing Single Malts by Flavor.* Pavilon, London.

Zhu, M. and Ghodsi, A. (2006). "Automatic Dimensionality Selection from the Scree Plot via the Use of Profile Likelihood". *Computational Statistics and Data Analysis,* (in press and available online at *Science Direct* since 25 October 2005).