Using t-SNE and KDE to assess the conformity of powdered products

F. Stevens[1], B. Carrasco[2], T. Fearn[3], V. Baeten[1], J.A. Fernández Pierna[1]

[1] Walloon Agricultural Research Centre (CRA-W), Valorisation of Agricultural Products Department, Food and Feed Quality Unit, Henseval Building, 24, chaussée de Namur, 5030 Gembloux, Belgium

e-mail : FoodFeedQuality@cra.wallonie.be

[2] Chemometric Brain, S.L.U.

C/ Cardenal Belluga, p.24/23, Pol. Ind. Oeste P.O. Box 487, 30169 San Ginés, Murcia, Spain

e-mail : beatriz.carrasco@chemometricbrain.io

[3] University College of London, Statistical Sciences Department, United Kingdom.

Keywords: food blends, quality control, *t-SNE, kernel density estimate*

1 Introduction

In the food industry, powdered products are about 40% of the human consumption. Moreover, pure or blend of ingredients are bought and used by the companies in the production of their final recipes. The quality control of the process for these type of ingredients (or even for the final powdered products) is a challenge for most of the producer companies as there are not specific methods giving a confident result. There is a real need of a method to perform quality control as a clear process, independent on the technician, reproducible with almost no time consuming and reliable. Furthermore, issues as food frauds are difficult to detect in this kind of products.

NIR technology combined with Chemometrics have proven to be a good alternative to perform in-line and on-line quality control. With this technology, it is possible to perform the assessment of any powder product determining also the uncertainty present in a new sample. So, this methodology provides to the industry an objective and reliable tool to ensure the composition and functionality of powdered products.

In this work, we propose the use of a new method to assess the conformity of a new sample in a known product. It is based on two techniques of data analysis, t-distributed stochastic neighbor embedding (t-SNE) [1] and kernel density estimate (KDE) [2].

2 Material and methods

The dataset used in this work consists on Near Infrared (NIR) spectra of different pure and blend products from the global Chemometric Brain dataset. Each product containing different number of ingredients and specific compositions.

t-SNE is a visualization technique used to project high-dimensional data into a two or three-dimensional map that reveals structure at many different scales. The basic idea is to create a probability distribution that dictates the relationships between various neighboring in the high-dimensional space and then to recreate a low dimensional space that follows that probability distribution as best as possible.

KDE is a statistical tool used to create a smooth curve/surface given a set of points. The basic idea is to place a peak function (the kernel function) at each data point and to average over all the points to obtain an estimate of the probability distribution.

In this application, all the (qualitative) samples of the product available in the database are first joined with the tested sample. Next, a t-SNE algorithm is applied to this newly created dataset, using all the spectral features. Moreover, a classical PCA is applied in parallel. Depending on user choices (dimensionality of resulting map for t-SNE and selected components for PCA), multiple two- or three-dimensional mappings are produced. A KDE is then applied on each of these mappings. Finally, probability contours of these distributions are used to set a boundary for judging future samples as conforming or not conforming.

3 Results and discussion

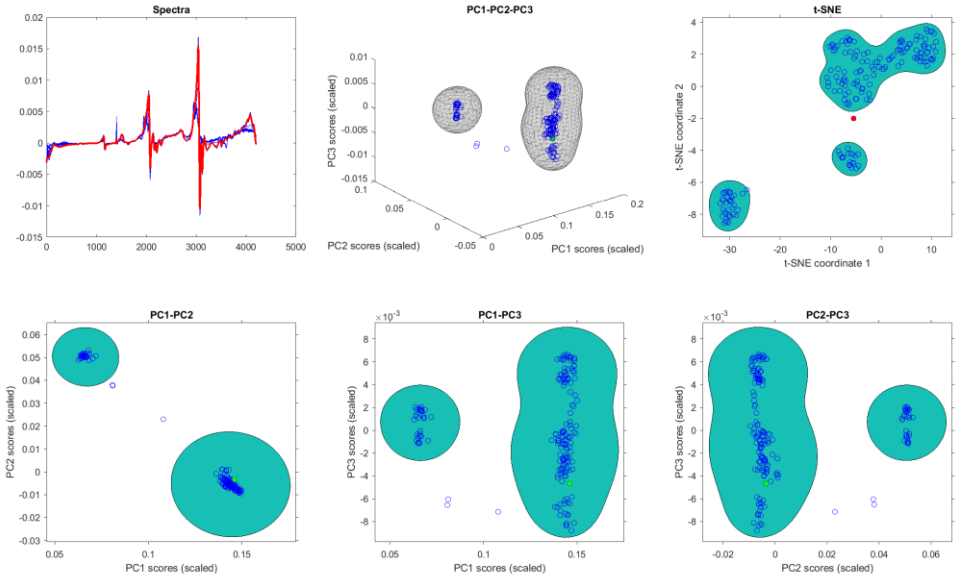

Figure 1– Comparison of a new spectrum (in red) to other spectra of the same product. Green areas and grey meshes have been defined using KDE. PCA analysis (up to the third component) indicates that the sample is conforming. However, analysis of the two t-SNE coordinates issues a warning (red point).

4 Conclusion

The advantages of t-SNE and KDE are exploited within a decision support tool allowing the user to assess the conformity of a new sample of a known product. By considering all the spectral information, t-SNE circumvents the problem of the choice of the best component faced with the use of PCA and takes advantage of all the spectral information. Besides, in contrast with ellipsoid or convex hull, KDE allows creating conformity boundaries in the low-dimensional space which respects the fact that samples are often distributed in a non-normal and/or multivariate way in this space.

5 References

[1] L. J. P. van der Maaten, & G. E Hinton, Visualizing data using t-SNE. Journal of Machine Learning Research, 9, 2431–2456, 2008

[2] M. Rosenblatt, Remarks on Some Nonparametric Estimates of a Density Function,. The Annals of Mathematical Statistics. 27 (3): 832–837. doi:10.1214/aoms/1177728190, 1956.