Metabolite fingerprinting: an ICA approach

M. Scholz[1], S. Gatzek[1], A. Sterling[2], O. Fiehn[1], and J. Selbig[1]

[1]Max Planck Institute of Molecular Plant Physiology, Germany

[2]Advion BioSciences Ltd., Norwich NR9 3DB, UK

Introduction

Metabolite fingerprinting is a technology for providing information from spectra of total compositions of metabolites. We will show, that *independent component analysis* (ICA) applied to such high dimensional data has a higher informative power than the classical *principal component analysis* (PCA).

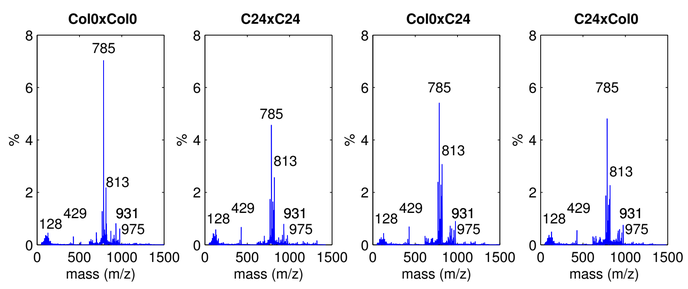

Figure 1: Mass spectra of *Arabidopsis thaliana* crosses are analysed by PCA and ICA to investigate how metabolite fingerprinting reflects the biological background.

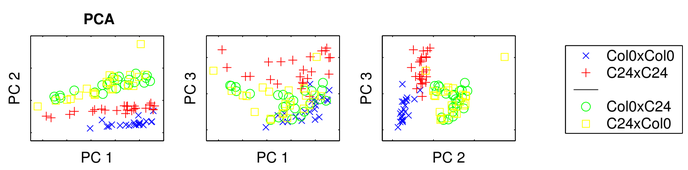

Figure 2: PCA fails to show optimal projections of the data with differences between the lines and crosses. The first principal components (PC 1), the component of highest variance, has no information about discriminating the lines or crosses. A better result is given by the components PC 2 and PC 3 of smaller variance, meaning that the required experimental information is not related to the highest variance in the data.

ICA versus PCA

In contrast to PCA, the components of ICA are constructed such as to minimise the statistical dependence and are therefore termed *independent components* (ICs).

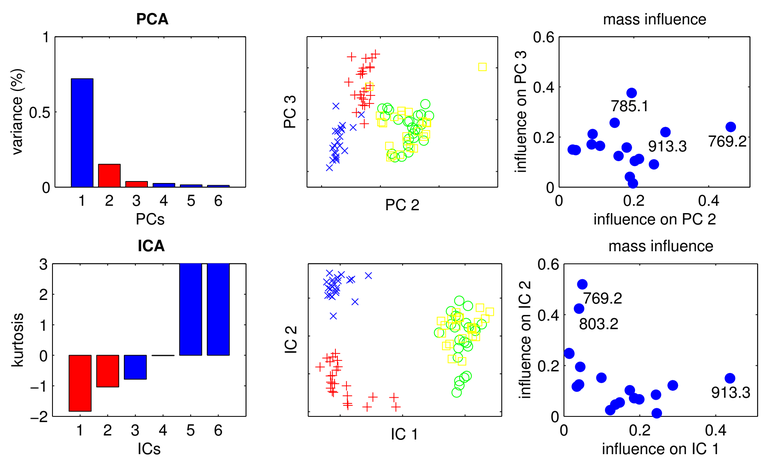

Figure 3: ICA gives a better projection result than PCA and this result is already given by the first two ICs (when ranked by the kurtosis measure). Also, in ICA the masses are more separated to different ICs, confirming that different ICs represent independent biological processes, where different metabolites are involved.

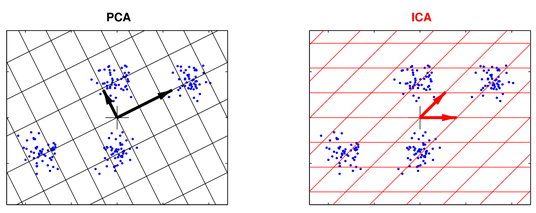

Figure 4: PCA and ICA applied to an artificial data set. The components of ICA are related to the cluster structure of the data and are not restricted to be orthogonal.

Kurtosis

The kurtosis measure is used to rank the extracted independent components to our interest. The kurtosis is a classical measure of non-Gaussianity, it indicates whether the data are peaked or flat relative to a Gaussian (normal) distribution.

$$kurtosis(z) = \frac{\sum_{i=1}^{n}(z_i - \mu)^4}{(n-1)\sigma^4} - 3$$

where $z = (z_1, z_2, ..., z_n)$ is representing a variable or component with mean μ and and standard deviation σ, n is the number of samples. The kurtosis is the fourth auto-cumulant after mean (first), variance (second), and skewness (third).

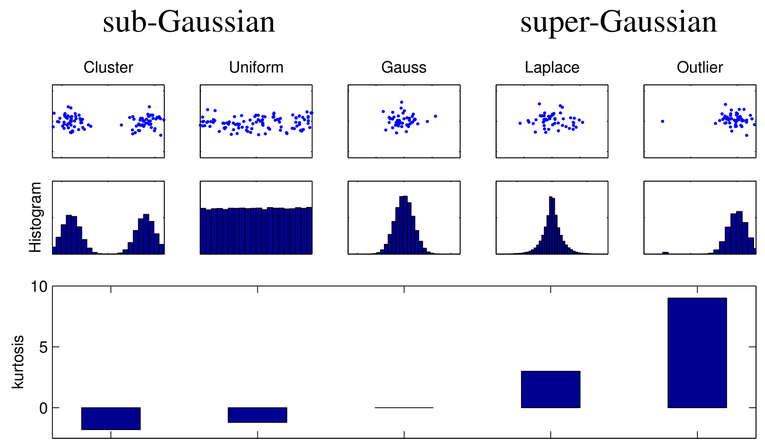

Figure 5: Negative kurtosis can indicate a cluster structure (different experimental conditions) or an uniformly distributed factor (temperature). Thus the components with the most negative kurtosis can give us the most relevant information.

PCA – pre-processing

The higher informative power of ICA is only achieved when ICA is combined with PCA, to reduce first the dimension of the data set. The number of principal components determines the quality of ICA significantly, therefore we propose the kurtosis as a criterion for estimating the optimal dimension automatically.

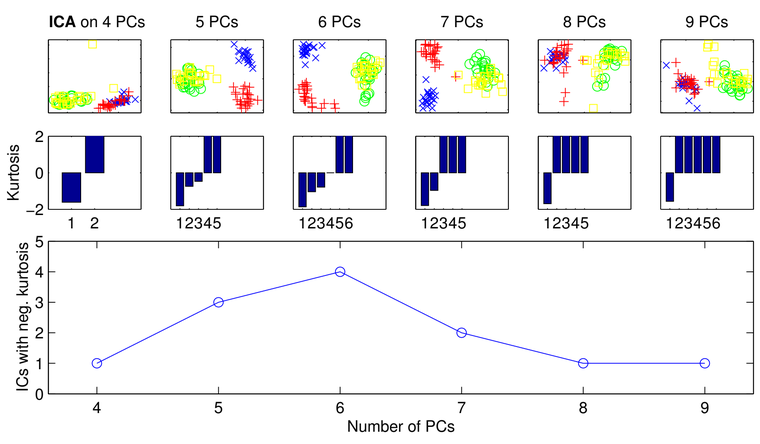

Figure 6: ICA is applied to reduced data sets with different numbers of PCs. At 6 components of PCA, ICA can extract the highest number of interesting ICs, i.e. ICs with negative kurtosis.

As a negative kurtosis indicates relevant components, the dimension, where we can extract the highest number of independent components with negative kurtosis is the optimal dimension.

Experimental artefact

We found that ICA could detect three relevant components. The first independent component is usable for separating the Arabidopsis crosses from the background parental lines, the second contains information for discriminating the two parental lines. The third component is not related to the biological experiment, but we could find a relation to the identifier of the samples, representing the order over time, measured in the mass spectrometer. Hence IC 3 is an experimental artefact due to increasing contamination of the QTOF skimmer along the analytical sequence. This technical factor could not be detected by PCA.

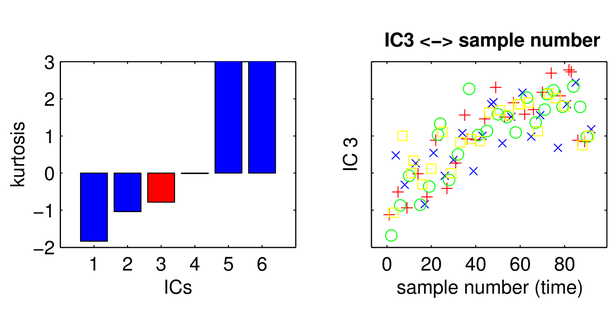

Figure 7: Three components with clearly negative kurtosis are detected. The third component (IC 3), an almost uniformly distributed factor, could be interpreted as an experimental artefact, related to the order over time, when the samples were measured.

Conclusion

ICA has a high informative power when it is combined with suitable pre-processing and evaluation criteria.

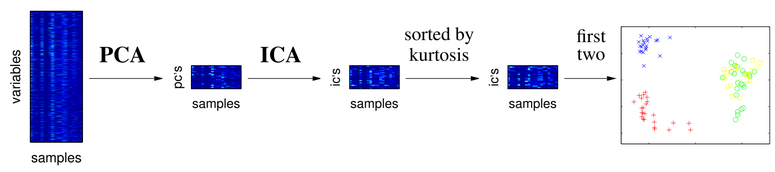

Figure 8: The proposed ICA procedure. First, the data set is reduced by PCA thereby maintaining all of the relevant variances. ICA is applied to this reduced data set and the extracted independent components are sorted by their kurtosis value.

The resulting independent components have been interpreted: The first component discriminates the *Arabidopsis* crosses from the background parental lines, and the second component discriminates the two parental lines. The third component could be interpreted as an experimental artefact. The described approach is available for public at the MetaGeneAlyse (`http://metagenealyse.mpimp-golm.mpg.de`), a web-based analysis tool for analysing biological data from metabolomics, proteomics and transcriptomics.

Acknowledgements

The authors thank Thomas Altmann and Rhonda Meyer for initiating and stimulating the *Arabidopsis hybrid vigour* (heterosis) project, which aims to use recombinant inbred and near isogenic lines for functional genomics.

References

[1] A. J. Bell and T. J. Sejnowski. An information-maximization approach to blind separation and blind deconvolution. *Neural Computation*, 7:1129–1159, 1995.

[2] P. Comon. Independent component analysis, a new concept? *Signal Processing*, 36(3):287–314, 1994.

[3] A. Hyvärinen, J. Karhunen, and E. Oja. *Independent Component Analysis*. J. Wiley, 2001.

[4] A. Hyvärinen and E. Oja. Independent component analysis: Algorithms and applications. *Neural Networks*, 4–5(13):411–430, 2000.

[5] W. Liebermeister. Linear modes of gene expression determined by independent component analysis. *Bioinformatics*, 18(1):51–60, 2002.

[6] M. Scholz, S. Gatzek, A. Sterling, O. Fiehn, and J. Selbig. Metabolite fingerprinting: detecting biological features by independent component analysis. *Bioinformatics*, Advance Access published on April 15, 2004.

Contact: Matthias Scholz, `scholz@mpimp-golm.mpg.de`