Independent Components Analysis in multivariate regression : application to the prediction of concentration and temperature of sugar solutions using Near Infrared spectra

Douglas N. Rutledge

INRA, UMR 214 INRA/INA P-G, 16, rue Claude Bernard, 75231 Paris cedex 05 France

Keywords : Independent Components Analysis, regression, Near Infrared spectrsocopy

The objective of Principal Components Analysis (PCA) is to find vectors that describe the maximum of the variance within a dataset, with the constraint of orthogonality, thus de-correlating the resulting "signals" under the assumption that all of the sources follow a Gaussian probability distribution. For a complex dataset where many different signal sources contribute, the Gaussian assumption may not be valid. Consequently, PCA cannot then be expected to work well for the extraction of pure and statistically independent signals from a dataset of such mixtures

Numerous other bilinear decomposition methods are available, depending on which objective function is chosen. If the objective function maximises the variance in the extracted components, this decomposition leads to Principal Component Analysis; if it maximises the correlation between the variables, a form of Factor Analysis (FA) is the result. The space spanned by the directions of the new variables may not yield the optimal interpretation in relation to the application at hand (either for the shape of the extracted signal-vectors or for the distribution of the samples), so that some sort of rotation may have to be employed as a second step in both these analyses.

Independent Components Analysis (ICA) is the decomposition of a set of vectors into linear components that are "*as independent as possible*". Here, "independence" should be understood in its strong statistical sense: it goes beyond (second-order) decorrelation and thus involves the non Gaussianity of the data.

ICA attempts to recover the original signals by estimating a linear transformation, using a criterion related to information theory and entropy, that provides statistical independence between the sources, under the assumption that the data does not follow a Gaussian distribution. This may be achieved by the use of higher-order information that can be extracted from the densities of the data.

The implicit objective of ICA is usually to find "physically significant" components. This is its strength and its weakness compared to PCA, which does not really to look for (and usually does not find) components with physical reality. Whereas a PCA can always be done, a successful ICA is not possible if there do not exist any underlying components that can be represented by the model :

X=A*S

where :

X is the matrix of observed spectra ;

S is the matrix of pure spectra ;

A is the mixing matrix of coefficients, related to the corresponding concentrations.

The assumptions made on the independent components are that they are mutually statistically independent and have non-Gaussian distribution. The ICA model constructs a demixing matrix, **W**, approximating the inverse mixing matrix, $\mathbf{A}^{-1}$, so that pure component spectra can be recovered from the measured mixed spectra by :

S=W*X

W can be approximated by maximizing the non-Gaussianity of all $\mathbf{w}^T\mathbf{x}$, where **w** denotes one row of **W**. With this approach, the observed data are linearly transformed into mutually statistically independent components. A thorough description of the mathematical basis for ICA can be found in Hyvaerinen and Oja [1, 2] and De Lathauwer et al. [3]. The widely used Jade (Joint Approximate Diagonalization of Eigenmatrices) algorithm for ICA used in this study was obtained from [4]. This method [5, 6] for the blind separation of independent non-Gaussian sources in Gaussian noise is based on the optimisation of the second and fourth order cumulants from the data.

In this paper, Independent Components Analysis was applied to the analysis of second-derivative Near Infrared (NIR) spectra of water-sugar solutions at different temperatures and sugar concentrations with the objective of developing IC-based regression models, in a way similar to the use of PCs to create PCR models.

In ICA, different results may be produced as a function of the number of Independent Components (ICs) being extracted. For this reason, it is first necessary to determine the optimal number of ICs to extract from the dataset. One can then go on to determine which ICs are most correlated with the properties to be predicted, and which spectral regions are optimal.

This was done using a Matlab procedure to perform the ICA with varying numbers of ICs (between 1 and 10), within a variable-sized window sweeping across the spectra. The squared correlation coefficients (R^2) were calculated between the property to be predicted (temperatures and concentration) and the scores of each IC obtained for the different sets of ICs. The maximum R^2 values and corresponding IC numbers and total number of ICs for each window were plotted as a function of the starting and ending frequencies of the corresponding spectral window.
A leave-3-out Cross-Validation PLS regression was also performed for each set of extracted ICs and the minimal RMSECV values were calculated. These minimal RMSECV values and the corresponding IC numbers and total number of ICs were similarly plotted as a function of the starting and ending frequencies of the corresponding spectral window.

The optimal number of ICs to extract was determined as being 6, and the region starting anywhere between 750 and 850 nm and ending at 1050nm gives an R^2 above 0.95 for the regression between IC3 and sugar concentration. It was also shown that IC6 gives a very high R^2 (above 0.95) for the spectral region beginning anywhere between 750 and 820nm and ending at about 950nm. At the same time, IC4 gives a high R^2 (above 0.90) for the spectral region that was optimal for sugar (beginning at 750 nm and ending at 1050nm).

---

1 A. Hyvaerinen, E. Oja, Neural Computation'. 9 (1997) 1483–1492.
2 A. Hyvaerinen, E. Oja, Neural Networks 13 (2000) 411–430.
3 L. De Lathauwer, B. De Moor, J. Vanderwalle, J. Chemometr. 14 (2000) 123–149.
4 http://www.tsi.enst.fr/icacentral/Algos/cardoso/
5 J-F. Cardoso, Neural Computation 11 (1999):157-192
6 J-F. Cardoso, A. Souloumiac, IEE Proceedings-F, 6, 140, (1993) 362-370