MULTIVARIATE STATISTICAL ANALYSIS OF SPECTROSCOPIC DATA

Haisheng Lin, Ognjen Marjanovic, Barry Lennox

Control Systems Centre, School of Electrical and Electronic Engineering, University of Manchester

Abstract: This paper focuses on the application and comparison of Principal Component Analysis (PCA) and Independent Component Analysis (ICA) using two generic artificially created datasets. PCA and ICA are assessed in terms of their abilities to infer reference spectra and to estimate relative concentrations of the constituent compounds present in the analysed samples. The results show that ICA outperforms PCA and is able to identify the reference spectra of all the constituent compounds. On the other hand, PCA fails to identify one of the constituent compounds for the first dataset. Also, ICA estimates relative concentrations of all the constituent compounds present in both datasets much more accurately than PCA.

Keywords: principal component analysis, independent component analysis, spectroscopic data analysis

1. INTRODUCTION

In recent years, the pharmaceutical industry has been obliged to comply with increasingly stringent regulatory requirements enforced by Food and Drugs Administration. These requirements have been introduced to ensure that the risks associated with pharmaceutical products to public health are minimised. Many of these regulatory obligations translate into the requirement that the chemical composition of a given therapeutic drug, i.e. pharmaceutical product, satisfies certain conditions imposed on the concentrations of its constituent compounds.

To meet these requirements, many pharmaceutical companies have started to explore the use of various vibrational spectroscopic techniques, such as near infrared (NIR) and Raman, in order to obtain spatial and spectral information for a given sample. Specific applications in the pharmaceutical industry include producing chemical images of the constituent compounds on the surfaces of tablets, determination of content uniformity and the examination of the association of spatial distribution as a function of dissolution properties. In particular, these techniques are being explored as a method for the estimation of the relative concentrations of constituent compounds present in a given sample (Zhang et al., 2005).

Particularly important application area for spectroscopic techniques is the estimation of the relative concentrations of constituent compounds present in a given sample (Zhang et al., 2005). This is a non-trivial task and requires the application of sophisticated multivariate statistical analysis techniques (Windig, 1991; Sasic, 2007). This task is simplified if the reference spectra for the constituent compounds exist. In those cases, it is possible to use the so-called Direct Classical Least Squares (DCLS) that estimate concentrations by regressing the data onto the reference spectra’ vectors (Zhang, et al., 2005).

However, for more general cases where the reference spectra of the constituent compounds are not available, multivariate data analysis methods, such as Principal Component Analysis (PCA) and Independent Component Analysis (ICA), have been shown to be effective in their ability to extract chemical information from the imaging data (Zhang et al., 2005; Shashilov et al., 2006). The application of these methods to spectroscopic data, therefore, offers the potential to provide an important tool in ensuring that pharmaceutical products conform to the stringent regulatory requirements.

Both PCA and ICA belong to the class of so-called *unsupervised* analysis methods. In the context of spectroscopic data analysis, this means that neither PCA or ICA require any *a priori* knowledge regarding the spectra of the individual chemical compounds present in the sample in order to compute the corresponding models and, thereupon, to estimate relative concentrations of the constituent compounds in a given sample.

While both PCA and ICA have been shown to be effective in terms of their abilities to extract chemical information from the imaging data (Zhang et al., 2005; Shashilov et al., 2006), they have rarely been compared with each other. This paper attempts to address this issue by providing a comparison between PCA and ICA using two generic spectral datasets. Comparison is performed in terms of the abilities of PCA and ICA to infer reference spectra and then to estimate the relative concentrations of the constituent compounds present in two two constituent compounds.

This paper is organised as follows. Section 2 describes the datasets used as well as the analysis methods of PCA and ICA. The results from the subsequent multivariate analysis are provided in section 3. Finally, the conclusions and future work plan are discussed in section 4.

2. MATERIALS AND METHODS

In this section, method of creating datasets used in the comparison is described. Also, the brief review of the Principal Component Analysis and the Independent Component Analysis are provided in this section.

2.1 Creating Simulated Dataset

Both of the artificially created datasets were created using a specially constructed MATLAB function, the main features of which are outlined below.

First of all, reference spectra were created for each of the 4 constituent compounds. Each of these spectra is comprised of 200 spectral channels or wavenumbers. Also, each reference spectrum contains a random number of peaks, centred at random wavenumbers and having random widths. The number of peaks, locations of their centres and their width sizes, were generated using "rand" command in MATLAB, which creates uniformly distributed random numbers. Peak shapes were modelled using Gaussian contours. Resulting reference spectra for the dataset 1 and dataset 2 are shown in Figure 1 and Figure 2, respectively.

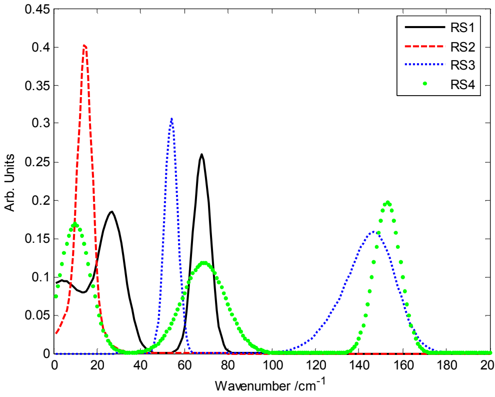

Fig. 1. Reference spectra of the constituent compounds of the dataset 1.

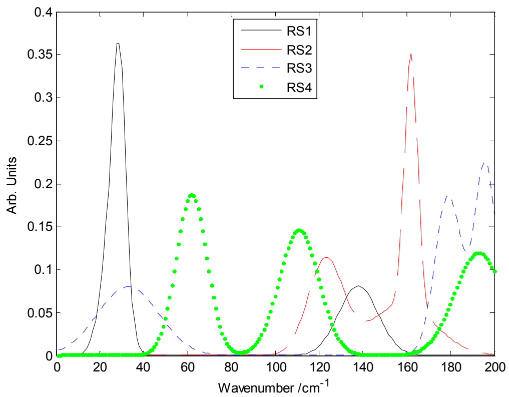

Fig. 2. Reference spectra of the constituent compounds of the dataset 2.

After the reference spectra were created, concentrations of the 4 constituent compounds were generated for 500 samples as uniformly distributed random numbers ranging between 0.0 and 1.0. Individual concentration values were made to conform to the "closure constraint", which states that all concentrations of all the constituent compounds sum up to 1 for every sample.

Finally, the 'measured' spectra were created by multiplying reference spectra of the constituent compounds with the corresponding concentrations.

2.2 Principal Component Analysis (PCA)

The general objective of the Principal Component Analysis (PCA) is to capture majority of variation present in data using a minimal number of composite variables, namely principal components (PCs). This dimensionality reduction is performed by exploiting the inter-dependence between original variables present in a given dataset. PCA has been applied in many areas of science and engineering with many excellent references that describe this technique in detail, see (Jackson, 1991) for further details. For details regarding the application of PCA to spectral data see (Geladi, 1997) for a good overview.

In the context of spectroscopic data analysis, the power of PCA lies in its ability to condense the correlated information from hundreds of spectral channels (wavenumbers) into a small number of orthogonal principal components (PCs), which can then be exploited for data visualization and feature extraction. These PCs are defined as follows:

$$\mathbf{Z} = \sum_{k=1}^{nc<n} \mathbf{t}_k \mathbf{p}_k^T + \mathbf{E} \qquad (1)$$

Where $\mathbf{Z}$ is a zero-mean, $m \times n$ matrix containing the spectral data with m samples that are made at n wavenumbers. $\mathbf{t}_k$ are referred to as score vectors and $\mathbf{p}_k$ are the loading vectors. $\mathbf{E}$ represents the information contained within the matrix $\mathbf{Z}$ that is not represented in the first *nc* principal components (a single PC being the combined $\mathbf{t}_k$ and $\mathbf{p}_k$ pair).

If a particular principal component (PC) identifies the reference spectrum of a constituent compound, then the score value associated with that PC and a given sample will reveal the relative concentration of the corresponding constituent compound for that sample. Therefore, provided the individual PCs are able to identify the reference spectra of the individual compounds, PCA provides a convenient tool for estimating relative concentrations of the constituent compounds in a given sample.

However, each PC is calculated by maximising the amount of variance it can explain. The PCs will, therefore, not necessarily correspond to specific chemical components (Zhang et al., 2005; Vrabie et al., 2007). In fact, principal components' loadings vectors will typically correspond to linear combinations of the reference spectra. This phenomenon is known as "rotational ambiguity" and is particularly evident when several reference spectra overlap significantly (Zhang et al. 2005).

2.3 Independent Component Analysis (ICA)

Similarly to PCA, Independent Component Analysis (ICA) separates a multivariate signal, such as a spectroscopic measurement, into its constituent subcomponents, e.g. constituent compounds. However, ICA identifies mutually statistically independent components, rather than simply uncorrelated ones, as is the case with PCA (De Lathauwer, et al., 2000). In probabilistic theory, independence is a high-order statistic and it is a much stronger condition than orthogonality. Also, the condition of independence in the components allows ICA to overcome rotational ambiguity problem encountered with PCA. It is important to note, however, that PCA and ICA are closely related. In fact, ICA can be considered as a 'fine-tuning' form of PCA, since it rotates the principal components in order to remove high-order dependencies between source signals (De Lathauwer, et al., 2000; Hyvarinen and Oja, 2000). Also, PCA and ICA are related from a computational perspective, since they both rely on an eigenvalue decomposition for the identification of their corresponding models (De Lathauwer, et al., 2000; Hyvarinen and Oja, 2000).

As in the case of PCA, ICA assumes that spectroscopic measurements are linear combinations of the constituent compounds' reference spectra:

$$\mathbf{Z} = \sum_{k=1}^{nc<n} \mathbf{c}_k \mathbf{s}_k^T + \mathbf{F} \qquad (2)$$

Where $\mathbf{Z}$ is a zero-mean, $m \times n$ spectroscopic data matrix with m samples recorded at n wavenumbers. Similarly to PCA, $\mathbf{c}_k$ represent score vectors while $\mathbf{s}_k$ represent loading vectors associated with independent components (a single IC being the combined $\mathbf{c}_k$ and $\mathbf{s}_k$ pair). Finally, $\mathbf{F}$ represents additive noise present in the original data that cannot be explained by *na* independent components.

3. RESULTS

Both PCA and ICA models were developed with four retained principal/independent components for each of the two spectral datasets. Firstly, loading vectors, corresponding to particular PCs and ICs, were matched to the reference spectra that they most closely resemble. Concentrations of the constituent compounds were then estimated by observing the values of scores associated with the loading vectors.

3.1 Results for Dataset 1

Plots of the loading vectors, corresponding to PCA and ICA models, and the reference spectra related to the constituent compounds present in dataset 1 that the loading vectors most closely resemble are shown in Figures 3-6.

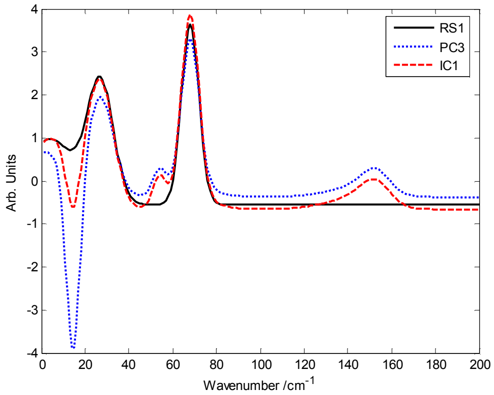

Fig. 3. Plot of the 1st reference spectrum (RS1), 3rd PC and 1st IC.

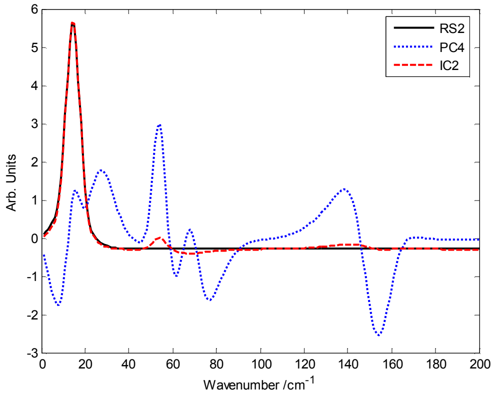

Fig. 4. Plot of the 2nd reference spectrum (RS2), 4th PC and 2nd IC.

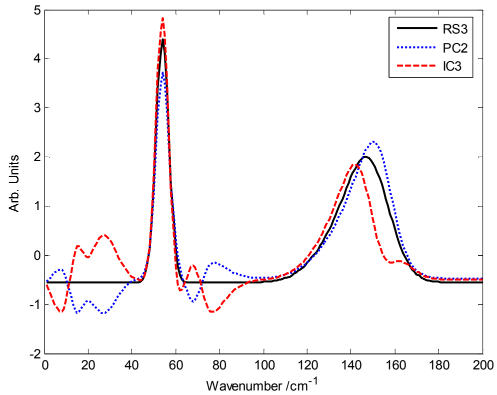

Fig. 5. Plot of the 3rd reference spectrum (RS3), 2nd PC and 3rd IC.

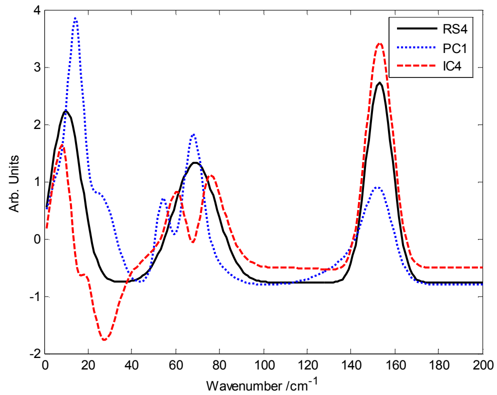

Fig. 6. Plot of the 4th reference spectrum (RS4), 1st PC and 4th IC.

Figures 3, 5 and 6 show that both PCA and ICA managed to identify 1st, 3rd and 4th reference spectrum, respectively. However, it is shown in Figure 4 that, while the 4th IC did manage to infer 2nd reference spectrum, PCA model failed in identifying this reference spectrum. Hence, ICA managed to clearly identify reference spectra of all the constituent compounds present in dataset 1. On the other hand, PCA failed to identify one of these reference spectra. Hence, in terms of their abilities to infer reference spectra, ICA clearly outperforms PCA.

Next, the comparison between PCA and ICA was made in terms of their abilities to estimate relative concentrations of the constituent compounds present in analysed samples. Comparison was conducted by performing the correlation analysis between the estimated and the actual concentrations, results of which are shown in Table 1.

Table 1 shows that ICA is much more accurate when compared to PCA in terms of its ability to accurately estimate relative concentration of each of the constituent compounds. In particular, note that correlation between the true concentrations and those estimated using ICA model is greater than or equal to 0.81.

Table 1 Cross-correlation coefficients between the actual concentrations of the constituent compounds and their estimates obtained using PCA and ICA models for dataset 1.

	RS1	RS2	RS3	RS4
PCA	0.93	0.94	0.99	0.81
ICA	0.78	0.19	0.90	0.16

Also, note that the cross-correlation between ICA estimates and the true concentrations are clearly stronger, i.e. higher in value, than those between the actual concentrations and the PCA estimates for each of the constituent compounds. Finally, it is observed in Table 1 that PCA model clearly fails to provide reasonable estimates of the concentrations associated with the 2nd and the 4th constituent compounds. Since PCA failed to accurately identify RS2, as shown in Figure 4, it is unsurprising that PCA model failed to provide satisfactory estimates of the relative concentration of the second constituent compound. However, somewhat surprising result is that the PCA model clearly failed to accurately estimate concentrations of the 4th constituent compound, even though Figure 6 demonstrates that it managed to capture main features of the corresponding reference spectrum (RS4). This finding demonstrates that even if a particular reference spectrum is successfully inferred using a particular IC or PC it does not necessarily follow that the corresponding score values will truthfully reflect concentration values of the corresponding constituent compound.

3.1 Results for Dataset 2

Ability of the PCA and ICA models to infer reference spectra of the constituent compounds present in dataset 2 is demonstrated in Figures 7-10.

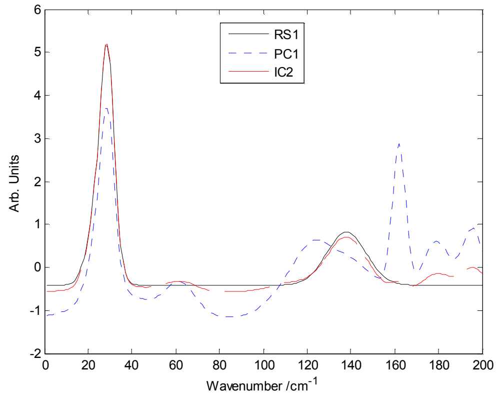

Fig. 7. Plot of the 1st reference spectrum (RS1), 1st PC and 2nd IC.

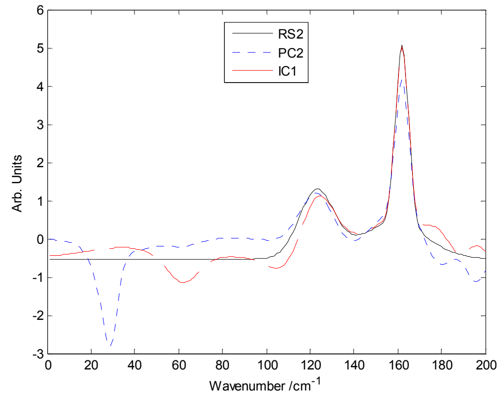

Fig. 8. Plot of the 2nd reference spectrum (RS2), 2nd PC and 1st IC.

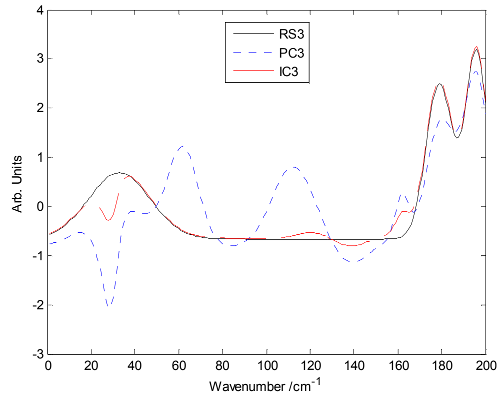

Fig. 9. Plot of the 3rd reference spectrum (RS3), 3rd PC and 3rd IC.

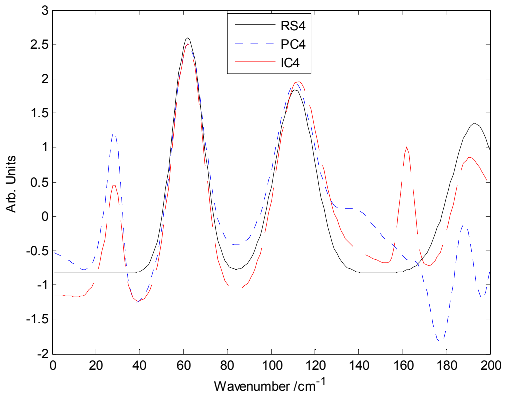

Fig. 10. Plot of the 4th reference spectrum (RS4), 4th PC and 4th IC.

Figures 7-10 demonstrate that both PCA and ICA managed to identify reference spectra of all the constituent compounds present in dataset 2.

However, additional observation made from Figures 7-10 is that the PCA was found to be less selective than ICA with individual PCs containing features of several different reference spectra. In particular, PC1, shown in Figure 7, identifies features related to three different reference spectra (RS1, RS2 and RS3). Also, PC3, shown in Figure 9, captures reference spectrum RS3 as well as the two peaks contained in the reference spectrum RS4.

Next, the ability to estimate relative concentrations of the constituent compounds present in dataset 2 was assessed for both PCA and ICA. Correlation analysis between the estimated and actual concentrations related to the constituent compounds was then performed and the resulting cross-correlation coefficients are provided in Table 2.

Table 2 Cross-correlation coefficients between the actual concentrations of the constituent compounds and their estimates obtained using PCA and ICA models for dataset 2.

	RS1	RS2	RS3	RS4
PCA	0.97	0.98	0.96	0.94
ICA	0.67	0.96	0.54	0.78

First of all, it is observed in Table 2 that ICA model is capable of estimating concentrations of all the constituent compounds with a high level of accuracy. In fact, the cross-correlation between the actual concentrations and those estimated using ICA is greater than or equal to 0.94. On the other hand, in the case of PCA model the cross-correlation between actual and estimated concentrations is great than 0.9 for only one out of four constituent compounds. Also, PCA failed to provide satisfactory estimates of the concentrations related to the first and the third constituent compounds, with the cross-correlations coefficients being equal to 0.67 and 0.54, respectively. This poor performance is likely to be due to the lack of selectiveness observed for loadings vectors of PC1 and PC3, which contained features from several different reference

4. CONCLUSIONS

This paper compared the relative abilities of Principal Component Analysis (PCA) and Independent Component Analysis (ICA) to infer reference spectra and relative concentrations of the constituent compounds present in two generic artificially created spectral datasets. Although the reference spectra and relative concentrations were known, this information was not used during the computation of either PCA or ICA. Hence, it was assumed that the user is unaware of the reference spectra and relative concentrations of the constituent compounds present in the analysed samples. The results showed that ICA clearly outperformed PCA in terms of its ability to accurately infer the reference spectra of the constituent compounds. In particular, ICA successfully identified the reference spectra of all the constituent compounds while PCA failed to identify one of the constituent compounds. Also, PCA model was found to be less sensitive than ICA with individual PCs containing features of several different reference spectra. In terms of its ability to infer concentrations of the constituent compounds, ICA more clearly identified reference spectra when compared to PCA. In particular, it was found that the cross-correlation coefficients between the true concentrations and those estimated using ICA were all greater than 0.8. On the other hand, cross-correlation coefficients calculated for the actual concentrations and the estimates made by PCA model were greater than 0.8 for only 25% of the constituent compounds.

REFERENCES

De Lathauwer, L., B. De Moor and J. Vandewalle (2000). An introduction to independent component analysis. *Journal of Chemometrics*, Vol. **14**, pp. 123-149.

Geladi, P. and H. Grahn (1997). *Multivariate Image Analysis*. John Wiley & Son Ltd, New York.

Hyvarinen, A. and E. Oja (2000). Independent component analysis: algorithms and applications. *Neural Networks*, Vol. **13**, pp. 411-430.

Sasic, S. (2007). An in-depth analysis of raman and near-infrared chemical images of common pharmaceutical tablets. *Applied Spectroscopy,* Vol. **61,** pp. 239-250.

Shashilov, V.A., M. Xu, V.E. Ermolenkov and I.K. Lednev (2006). Latent variable analysis of Raman spectra for structural characterization of proteins. *Journal of Quantitative Spectroscopy & Radiative Transfer,* Vol. **102,** pp. 46-61.

Vrabie, V., C. Gobinet, O. Piot, A. Tfayli, P. Bernard, R. Huez and M. Manfait (2007). Independent component analysis of Raman spectra: Application on paraffin-embedded skin biopsies. *Biomedical Signal Processing and Control,* Vol. **2,** pp. 40-50.

Windig, W. and J. Guilment (1991). Interactive self-modelling mixture analysis. *Analytical Chemistry,* Vol. **63,** pp. 1425-1432.

Zhang, L., M.J. Henson and S.S. Sekulic (2005). Multivariate data analysis for Raman imaging of a model pharmaceutical tablet. *Analytica Chimica Acta,* Vol. **545,** pp. 262-278.