Available online at www.sciencedirect.com
SCIENCE DIRECT
Analytica Chimica Acta 552 (2005) 1–12
ANALYTICA CHIMICA ACTA
www.elsevier.com/locate/aca

# Chemometrical exploration of an isotopic ratio data set of acetylsalicylic acid

I. Stanimirova [a], M. Daszykowski [a,1], E. Van Gyseghem [b], F.F. Bensaid [b], M. Lees [b], J. Smeyers-Verbeke [a], D.L. Massart [a], Y. Vander Heyden [a,*]

[a] *ChemoAC, FABI, Analytical Chemistry and Pharmaceutical Technology, Vrije Universiteit Brussel-VUB, Laarbeeklaan 103, B-1090 Brussels, Belgium*
[b] *Eurofins Scientific Analytics, Rue Pierre Adolphe Bobierre, 44323 Nantes Cedex 3, France*

Received 7 April 2005; received in revised form 12 July 2005; accepted 21 July 2005
Available online 22 August 2005

**Abstract**

A data set consisting of fourteen isotopic ratios or quantities derived from such ratios for samples of acetylsalicylic acid (aspirin), commercialized by various pharmaceutical companies from different countries, was analyzed. The goal of the data analysis was to explore whether results can be linked to geographical origin or other features such as different manufacturing processes, of the samples. The methods of data analysis used were principal component analysis (PCA), robust principal component analysis (RPCA), projection pursuit (PP) and multiple factor analysis (MFA). The results do not seem to depend on geographic origin, except for some samples from India. They do depend on the pharmaceutical companies. Moreover, it seems that the samples from certain pharmaceutical companies form clusters of similar samples, suggesting that there is some common feature between those pharmaceutical companies. Variable selection performed by means of MFA showed that the number of variables can be reduced to five without loss of information.


*Keywords:* Chemometrics; Principal component analysis; Projection pursuit; Multiple factor analysis; Acetylsalicylic acid; Isotopic ratios

## 1. Introduction

Counterfeiting of pharmaceutical products is a serious global problem. Recent statistics published by the European Commission have shown that seizures of food products and medicines at the EU's external borders increased by 77% between 2002 and 2003. In addition to the obvious threat to human health posed by fraudulently produced medicines, the pharmaceutical industry itself sustains substantial financial losses due to the availability of counterfeit products or cases of patent infringement. Through an EU-funded research project, work has been proceeding in order to provide a solution to the problem of counterfeit pharmaceuticals and patent protection by developing a reliable methodology based on a combination of trace impurity profiles, isotopic fingerprinting and chemometric classification. This paper describes part of the preliminary work carried in this project out on a model compound, aspirin, which is widely available over the counter. A set of acetylsalicylic acid samples has been collected from around the world and analyzed using stable isotope ratio analysis. Stable isotope ratios are defined as the relative amounts of heavy and light forms of an element in a given sample (e.g. $^{2}H/^{1}H$, $^{13}C/^{12}C$ and $^{18}O/^{16}O$), which can be measured at natural abundance with a very high precision, either by Site specific Natural Isotopic Fractionation-Nuclear Magnetic Resonance (SNIF-NMR) or by Isotopic Ratio Mass Spectrometry (IRMS). Their performance, particularly for assessing the authenticity of compounds used in the flavor and perfume industries, has been well documented. These naturally occurring chemical tracers can provide an unfalsifiable fingerprint for a given pharmaceutical product that can

* Corresponding author. Tel.: +32 2 4774734; fax: +32 2 4774735.
*E-mail address:* yvanvdh@vub.ac.be (Y. Vander Heyden).
[1] On leave from the University of Silesia, Department of Chemometrics, Institute of Chemistry, 9 Szkolna Street, 40-006 Katowice, Poland.

0003-2670/$ – see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.aca.2005.07.041

be used to trace its origin and to distinguish it from other chemically identical products available on the market from other manufacturing sources.

To extract information from the isotopic data generated by the analyses, multivariate analysis is used. Some of the methods used are standard, such as principal component analysis (PCA) [1,2], but most of them are less usual. Robust principal component analysis (RPCA) [3–5] and projection pursuit (PP) with the entropy criterion [6] are applied to find outliers or samples with extreme characteristics, while PP with the kurtosis criterion [7] to verify the existence of clusters. Among the variables, some blocks can be recognized and these blocks might have a too large weight during data analysis. Multiple factor analysis (MFA) [8–10] avoids this problem or at least should make it less important.

In exploratory analysis, it is recommended [1] to use several methods, in order to uncover the data structure. The set of methods used here could be considered as a minimal number of approaches giving complementary information about the data and together helping to discover data characteristics.

The aim of the data exploration was to see whether classes of samples could be discovered that can be linked to different manufacturing processes (e.g. different synthetic pathways, raw materials used) of the acetylsalicylic acid.

Another question to be resolved was whether all variables measured are necessary. We applied two methods for variable selection, which were proposed for this aim for PCA, but adapted them for application with MFA. These are the methods of Krzanowski [11] and Guo et al. [12].

From a chemometric point of view an additional aim of this study was to compare the results of these relatively new or less frequently used methods both with PCA and with each other to have a better idea of their potential in exploratory data analysis of chemical data.

## 2. Theory

### 2.1. Principal component analysis

PCA is a projection method [1,2] revealing the data structure. It projects the data from a high dimensional to a lower dimensional space spanned by a few latent factors, so-called principal components (PCs). The uncorrelated principal components are linear combinations of the original variables obtained by maximizing the data variance. Therefore, PCA is considered a dimensionality reduction tool. The projection of an object on the PC is called score and the projection of a variable on the PC is called loading. Consequently, the scores give information about the similarity between objects, while loadings give information about the contribution of each original data variable to the certain PC. The original data, $\mathbf{X}$ ($m \times n$), can then be reconstructed, using the PCA model, as follows:

$$\mathbf{X} = \mathbf{T}\mathbf{D}^{\mathrm{T}}, \tag{1}$$

where $\mathbf{T}$ is a matrix containing in columns orthogonal vectors (principal components) for $m$ objects, $\mathbf{D}$ contains in columns orthonormal vectors (loadings) for $n$ variables and $\mathbf{D}^{\mathrm{T}}$ is its transposed matrix.

### 2.2. Projection pursuit

Like PCA, PP is also a projection method [2,13–15]. It is considered a special case of PCA, where the new latent factors, called projection pursuit features (PPFs), are obtained by maximizing a certain projection index that describes the inhomogeneity of the data, instead of their variance. The PP algorithm used in this study is proposed by Croux and Ruiz-Gazen [3] and is described in detail in [3,5]. Firstly, the data are sphered, i.e. each variable is of zero mean and unit variance. All objects are then projected onto all possible normalized directions passing trough the objects and the data origin, contrary to PCA, where the directions are not restricted to pass via objects. The projection index for all projections is calculated and the direction with the highest projection index value is selected. The next direction with the highest projection index value is found in the residual space, i.e. the space reduced by one projection from the current data space. This continues until the desired number of orthogonal directions is obtained. At the end of the algorithm, PPFs are obtained by projecting all objects onto the directions found. Many projection pursuit indices have been described in the literature [14,16]. In this study, kurtosis [7,17] and entropy [6,18] were used. Kurtosis of a projection, $y$, is defined as the fourth central moment, $y(4)$, divided by the fourth power of the standard deviation, $\sigma$.

$$\text{Kurtosis}\,(y) = \frac{y(4)}{\sigma^4} - 3 \tag{2}$$

Kurtosis is equal to zero for normally distributed (non-interesting) projection. A normally distributed projection is non-interesting from the PP point of view, because the technique looks for inhomogeneities. A positive or a negative kurtosis value is a measure of deviation of a projection from the normal distribution. Kurtosis is also called “peakedness”. A negative kurtosis indicates less peaked distributions, whereas a positive kurtosis indicates more peaked distributions. When the data contain clusters the distribution becomes multi-modal and less peaked. It has been shown in [17] that kurtosis has a minimum for two clusters containing the same number of objects.

A larger value of the entropy corresponds to a larger inhomogeneity in the data. In this way, extreme observations (possible outliers) are highlighted. In this application of PP the approximation of the entropy, used in independent component analysis (ICA) [6], is applied.

The use of specially designed projection indices should make the visual detection of clusters and outliers easier and more evident than by PCA.