Analytica Chimica Acta 704 (2011) 1–15

Contents lists available at ScienceDirect

Analytica Chimica Acta

journal homepage: www.elsevier.com/locate/aca

# Fast and simple methods for the optimization of kurtosis used as a projection pursuit index

S. Hou, P.D. Wentzell*

*Department of Chemistry, Dalhousie University, Halifax, NS, B3H 4J3 Canada*

ARTICLE INFO

*Article history:*
Received 25 April 2011
Received in revised form 21 July 2011
Accepted 4 August 2011
Available online 11 August 2011

*Keywords:*
Optimization
Quasi-power method
Univariate kurtosis
Multivariate kurtosis
Projection pursuit
Independent component analysis

ABSTRACT

As a powerful method for exploratory data analysis, projection pursuit (PP) often outperforms principal component analysis (PCA) to discover important data structure. PP was proposed in 1970s but has not been widely used in chemistry largely because of the difficulty in the optimization of projection indices. In this work, new algorithms, referred as "quasi-power methods", are proposed to optimize kurtosis as a projection index. The new algorithms are simple, fast, and stable, which makes the search for the global optimum more efficient in the presence of multiple local optima. Maximization of kurtosis is helpful in the detection of outliers, while minimization tends to reveal clusters in the data, so the ability to search separately for the maximum and minimum of kurtosis is desirable. The proposed algorithms can search for either with only minor changes. Unlike other methods, no optimization of step size is required and sphering or whitening of the data is not necessary. Both univariate and multivariate kurtosis can be optimized by the proposed algorithms. The performance of the algorithms is evaluated using three simulated data sets and its utility is demonstrated with three experimental data sets relevant to analytical chemistry.



## 1. Introduction

Exploratory data analysis and classification methods have always been important tools of multivariate data analysis in chemistry. The application of these methods has expanded in recent years due to, among other things, an increased emphasis on high throughput biological analysis, where researchers are often interested in differentiating among different biological states of organisms. Discriminant methods are widely used for classification purposes in these applications, but because the number of variables is typically high and the number of samples is limited, careful validation is required to ensure that there is a meaningful separation of classes. On the other hand, exploratory methods, such as principal components analysis (PCA) or hierarchical cluster analysis (HCA), are unsupervised, so any class separation that is observed is likely to be real. PCA has dominated as a method to visualize high-dimensional data in lower dimensional spaces, but suffers from the drawback that it is based on maximizing the variance along the projection vectors, which is not always the best way to separate classes. This problem can be circumvented through the use of projection pursuit (PP) analysis, which uses different criteria to identify projection vectors. While there are examples of the application of this technique to chemistry (see for example Refs. [1–3]), it is not nearly as widely applied as PCA and HCA, probably because the algorithms are fairly complex and not readily accessible in many standard packages. In this work, algorithms are presented to carry out PP analysis that are straightforward and efficient, allowing this important tool to be readily adapted to any application.

The term "projection pursuit" was firstly coined by Friedman and Tukey [4], but the concept of PP can be tracked back to the work of Kruskal [5,6] who proposed the term "index of condensation". PP generally refers to an unsupervised technique for exploratory data analysis, but some researchers have used this term for discriminant analysis [7]. The primary purpose of PP is to look for "interesting" projections in a low-dimensional subspace that can reveal the natural structure of the data. The notion of "interestingness" may have different interpretations in different applications, but in the present context, interesting projections are those where the data projected in the low dimensional space can reveal clusters or outliers.

Because the description of PP does not unambiguously define how to determine what is interesting, any linear projection method, including PCA, could be regarded as a special case of PP. PCA is perhaps the most widely used method in exploratory data analysis, but in many cases PP can outperform PCA. This is because the directions of the greatest variance in the data set (determined by PCA) do not necessarily show the most useful information, but PP may find directions that reveal "interesting" data structure.

An objective function that characterizes the "interestingness" is called a "projection index". In the literature, various projection

* Corresponding author. Tel.: +1 902 494 3708; fax: +1 902 494 1310.
*E-mail address:* Peter.Wentzell@Dal.Ca (P.D. Wentzell).

0003-2670/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.aca.2011.08.006

indices have been developed, leading to many PP variants. The original projection index was proposed by Friedman and Tukey [4], but this was followed by proposals for other projection indices in the literature [8–15]. Most of the projection indices are designed to measure the non-normality of a distribution. Deviations from normality in the projected data are considered interesting because, for multivariate data, the observed variables are often the linear combinations of a small number of latent variables. By the central limit theorem, even if the latent variables reveal important elements of data structure, such as clusters or outliers, the observed variables often cannot directly disclose meaningful information because they tend towards normality. The latent variables that reveal useful information deviate from a normal distribution, so projections that deviate strongly from normality may uncover this structure.

In theory, any function that relates directly to the normality of a distribution can be used as a projection index, but a good index should be a simple measure and easy to optimize. Several functions have been used, with entropy and kurtosis being the most familiar. Kurtosis was one of the early functions proposed [8] and has the advantage of conceptual simplicity. Peña and Prieto showed that maximization of the kurtosis can be used to detect outliers [16], although this is not always effective and other methods may be preferred [17]. On the other hand, projections with bimodality tend to have a small kurtosis, and minimization of kurtosis can therefore be used as a criterion to search for clusters [15]. Kurtosis is also used to measure the non-normality in independent component analysis (ICA) [18,19], which is a technique closely related to PP. In univariate statistics, a normal distribution has a kurtosis of 3. A super-gaussian (peaked, or leptokurtic) distribution has a larger kurtosis, while a sub-gausssian (flat, or platykurtic) distribution has a smaller kurtosis. Either maximization or minimization of kurtosis can give useful information. Kurtosis satisfies the condition of the Class III objection functions set by Huber [8] for good projection indices; that is, scaling and translation do not change the values of the functions. One more appealing property of kurtosis is that the univariate case can be easily generalized to multivariate kurtosis, which not only has the useful properties of univariate kurtosis, but also is independent of the choice of the basis for a subspace. Therefore, kurtosis appears to be an ideal statistic for the projection index.

The projection index acts as the heart of PP, but its utility is mostly dependent on computational aspects. Optimization of the projection index, which greatly determines whether a projection index is successful, plays a crucial role in PP. Because of the quartic nature of kurtosis, optimization is a difficult problem. Kurtosis can have multiple local maxima and minima, and commonly used optimization algorithms cannot guarantee the global extrema. Therefore, it is generally necessary to start from different initial guesses to search for the global optimum, or better local optima, and therefore the speed of an optimization algorithm is critical. Gradient descent or ascent methods are ubiquitous in optimization problems, but gradient methods have the well-known shortcoming of slow convergence rates and the choice of optimal step size is difficult. Gradient methods have been used for the optimization of kurtosis [20,21], but other algorithms have also been developed in the literature. Peña and Prieto [15] proposed iterative methods for optimization of kurtosis by applying a modified Newton's method, which is complicated, or by solving first-order optimality conditions, similar to one method proposed in this work (differences are noted in the Supplementary Information). Croux's algorithm [22] has also been used for the optimization of kurtosis as a projection index [3]. This algorithm calculates the objective function for many projections based on the sample space and works well when the number of variables is relatively small, but will perform poorly if the dimensionality of the data becomes too high. Hyvärinen et al. proposed a fast fixed-point algorithm to optimize the kurtosis [19,23] based on sphered data. Sphering, which differs from autoscaling, is a transformation that ensures the data have unit variance when projected in any direction [24]. This algorithm is one of the most widely used because of its fast convergence. It has several variants [25–29] and can be viewed to be a continuum between gradient methods and Newton's method. As with other such methods, the determination of the optimal step size for the fixed-point algorithm is computationally involved, but this has been described [27,28].

In the present work, new algorithms, referred as "quasi-power methods", to optimize kurtosis are proposed. The algorithms use the well-known conclusion in calculus that if all the partial derivatives are zeros at a point, the point may be a maximum or a minimum. By setting all the derivatives of kurtosis to be zero followed by re-arrangements, equations emerge that allow the principle of the power method and its variants (used to solve eigenvalue problems) to be employed. Because the algorithms are developed from the perspective of the power method instead of gradient methods, they are simple, fast, and stable. Commonly required preprocessing steps, such as sphering or whitening of the data, are not necessary. The algorithms can search for maxima or minima according to user's requirements, without the need to optimize step size, and they can be used for both univariate and multivariate kurtosis with little modification.

## 2. Theory

### 2.1. Univariate kurtosis

For univariate data, the sample kurtosis ($K$) is defined as

$$K = \frac{1/n\sum_{i=1}^{n}(z_i - \bar{z})^4}{\left(1/n\sum_{i=1}^{n}(z_i - \bar{z})^2\right)^2} \tag{1}$$

where $n$ is the number of samples, $z_i$ is the individual sample value, and $\bar{z}$ is the sample mean. The numerator is the fourth central moment and denominator is the square of the second central moment or the biased sample variance (as opposed to the unbiased variance which has $n-1$ degree of freedom). For the purpose of optimization, the offset of "−3" that is included in some definitions of kurtosis to give the normal distribution a kurtosis of zero is not included. The current definition ensures that the kurtosis is always positive.

For multivariate data, if there are $n$ samples measured on $p$ variables, the entire data can be arranged in a $n \times p$ matrix:

$$\underset{(n\times p)}{\mathbf{X}} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix} = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_n^T \end{pmatrix} \tag{2}$$

Each column of $\mathbf{X}$ represents a set of samples measured on a single variable and each row contains the measurements on different variables for a single sample, denoted by the notation $\mathbf{x}_i^T$, where the subscript "$i$" is the sample index. In the following, the data matrix $\mathbf{X}$ is assumed to have been column mean-centered to simplify the derivation. PP tries to search for a unit length projection vector $\mathbf{v} = \begin{bmatrix} v_1 & v_2 & \cdots & v_p \end{bmatrix}^T$ such that, when the $p$-dimensional data $\mathbf{X}$ are projected onto this projection vector, the kurtosis of the projected data reaches a maximum or a minimum. If a projected data