Use of Principal Component Analysis in the Source Identification of PCDD/Fs

She, J.W.*, Hagenmaier, H.
Institute of Organic Chemistry, University of Tübingen, D-7400 Tübingen, F.R.G.

Principal Component Analysis (PCA) decomposes the original data into a model consisting of a signal part and an error (noise) part. The goal of the PCA is to define linear combinations of the original variables and thus to reduce the data space. Tools to interpret such models are primarily the score and loading-plots. The advantages of PCA are that PCA involves a minimum of statistical presumptions as to error distribution, and that PCA does not demand that the number of objects is at least three times the number of dimensions.
Application of PCA in environmental analysis is determined by the multivariate nature of environmental data. In many cases, for example dioxin analysis, more variables are available than samples. By using PCA, Czuczwa and Hites[1] found that both air particulate and sediment samples lie in the same region of the score plot of the first two components. To extract the maximum information from a PCDD/F data set, PCA was used by Pitea[2]; Stalling[3] used PCA to examine complex PCDD/F residues in environmental samples; Lindström[4] demonstrated that the congener profiles in the milk for some of the areas are unique; and Stephens et. al.[5] have used PCA to evaluate the PCDD/Fs data in human milk and blood from different laboratories. Most of the studies were carried out by using a commercial software package, such as SIMCA, SAS or SPSS. We believed that the commercial programs have their advantages, but not every laboratory can afford them due to the cost. In this paper, we will give a PASCAL program based on Nonlinear Iterative Partial Least Squares (NIPALS) algorithm and discuss some of the properties of the principal components. One application example, using PCA to aid in the identification of the source of PCDD/Fs in soil samples, will be given.

A The NIPALS algorithm and Program

The procedure for calculating the principal components involves a mathematical method called eigenanalysis, yielding eigenvalues and associated eigenvectors. The four most commonly used methods are the power method[6], the Jacobi method[6], Singular Value Decomposition (SVD)[6] and NIPALS[7]. NIPALS is an algorithm designed to extract eigenvalues and

* Current Address: California Public Health Foundation, California Department of Toxic Substances Control, Hazardous Materials Laboratory, 2151 Berkeley Way, Berkeley, CA 94704, U.S.A.

eigenvectors directly from the data without requiring premultiplication of the data matrix by its transpose. It computes one component at a time. In cases where memory is limited and the processor is slow (e.g. microcomputers), the NIPALS method is still able to get results. The NIPALS algorithm according to Martens[8] will be presented in the following section.
The matrix of samples (rows) and variables (columns), **X**, is column centered, so that the mean of each variable (or column) is 0. We will call this matrix $\mathbf{X}_0$. The column vectors, initially consisting of the values of variable i over N experiments, will be denoted by $\mathbf{x}_{i(0)}$. The row vectors, initially consisting of the values of i variables for experiment n will be denoted by $\mathbf{r}_{n(0)}$. We start with matrix $\mathbf{X}_t$ where t refers to the current residual matrix and t = 0 initially.The following steps are performed.

1) Select the column (factor) of $\mathbf{X}_t$ with the greatest sum of squares: this is a first guess of the principal component scores. Call this vector $\hat{x}_{\mathbf{max(t+1)}}$ and call the sum of squares of its elements $\hat{x}_{max}^2(t+1)$.

2) Calculate the row vector
$\hat{r}_{\mathbf{max(t+1)}} = \hat{x}_{\mathbf{max(t+1)}}'\mathbf{X}_t / \hat{x}_{max}^2(t+1)$
and call the sum of squares of the elements of this vector $r_{max}^2(t+1)$

3) Scale the new vector to unit length
$\hat{r}_{\mathbf{max(t+1)}} = \hat{r}_{\mathbf{max(t+1)}} / \hat{r}_{max}(t+1)$

4) Calculate a new estimate of the principal component scores
$\hat{x}_{\mathbf{max(t+1)}} = \mathbf{X}_t \hat{r}_{\mathbf{max(t+1)}}' / \hat{r}_{max}^2(t+1)$

5) Re-estimate the sum of squares for the new estimate of the principal component, i.e. calculate
$S = \hat{x}_{max}^2(t+1)$

6) Compare this sum of squares to that of the previous estimate. If the difference between these two estimates is small (perhaps 0.0001 S) then the algorithm has converged for this principal component, and we go to step 7. Otherwise return to step 2 until convergence is obtained. When converged, $\hat{x}_{\mathbf{max(t+1)}}$ equals the score and $\hat{r}_{max(t+1)}$ equals the loading for component t + 1.

7) Calculate the new value $\mathbf{X}_{t+1} = \mathbf{X}_t - \hat{x}_{\mathbf{max(t+1)}} \mathbf{r}_{\mathbf{max(t+1)}}$ and then repeat the algorithm, increasing t by 1 , as from step 2 in order to obtain further principal components. The program in PASCAL will be presented.

B The properties of the principal components

If we take out the statistical concept from PCA, the remaining problem becomes a pure linear algebra problem. The main problem is to get the eigenvectors and eigenvalues from a matrix and to make a linear transformation on the matrix. Then, a statistical problem regarding the correlation relationship between variables and objects can be solved through investigating the relative position of the vectors. Therefore, the principal components should possess all the properties of the eigenvectors. Some important properties are:

1) The eigenvectors are orthogonal, so the principal components represent jointly perpendicular directions through the space of the original variables.

2) The first principal component has the largest variance of any unit-length linear combination of the observed variables. The jth principal component has the largest variance of any unit-length linear combination orthogonal to the first j-1 principal components.

3) Loadings, a_{ij}, represent the correlation coefficients between the variable X_i and component F_j

4) a^2_{ij} represents the contribution of the variance of component F_j to the variable X_i. The bigger the sum of the squares of a^2_{ij} of the same column, the more important the component F_j.

The properties of PCA determine its application, as the first two properities guide us in selecting the variables in the data set. For example, we need not to worry about the collinearity among the variables, because the principal components are always orthogonal. The last two properties will tell us how to interpretate the results obtained from PCA. For example, if the first two principal components do not explain most of the variance in the data set, one must increase the number of principal components.

C Use of PCA in the identification of the sources of PCDD/Fs

In Rheinfelden, a south Germany city, PCDD/Fs were found around a no longer operating pentachlorophenol(PCP) plant. The deeper slag residue (mixed with soil) and soil samples show a special PCDD/Fs isomer pattern (Fig. 1), different from that of the contamination in PCP. The historic records show that this residue predates 1920. We analysed 17 samples from the area, and all of the samples showed a similar isomer pattern[9]. The score plot of the first two components show that the samples "grouped" very well with the samples from residue from the electrolysis of sodium chloride. This grouping is confirmed by cluster analysis and coincides with the fact that chlorine electrolysis was the only industrial application of chlorine chemistry at that time.

Acknowledgement

We would like to express our appreciation to Dr. M. McKinney for reviewing this manuscript.

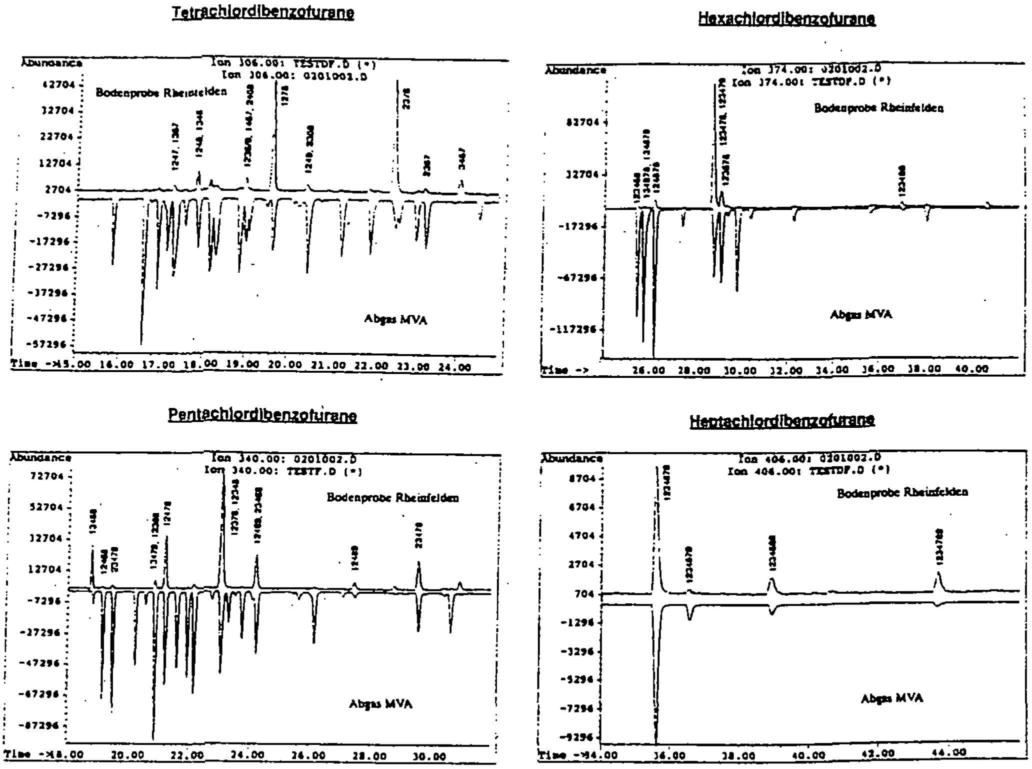

Fig. 1: Comparison of the isomer patterns of tetra to hepta-CDFs in the soil sample from Rheinfelden with that of a flue gas sample; positive peaks represent a soil sample, negative peaks represent a flue gas sample

References

1 Czucwa J.M., Hites R.A., Environ. Sci. Technol., 20, 195-200 (1986)

2 Pitea D., Bonati L., Lasagni M., Moro G., Todeschini R., Chiesa G., *Chemosphere* 18 (7-8), 1457-1464 (1989)

3 Stalling D. L., Peterman P.H., Smith L.M., Norstrom R. J., Simon M., *Chemosphere* 15 (9-12), 1435-1443 (1986)

4 Lindström G., Rappe C., Sjöström M., *Chemosphere* 19 (1-6), 745-750 (1989)

5 Stephens R. D., Rappe C., Hayward D. G., et. al. Analytical Chemistry 64 (24), 3109-3117 (1992)

6 Press W. H., Flannery B. P., Teukolsky S. A., and Vetterling W. T., *Numerical Recipes in PASCAL*, Cambridge University Press, Cambridge, p 375-418, 1989

7 H. Wold, in Krishnaish P. (Ed.), *Multivariate Analysis*, Academic Press, Orlando, 1966, P. 391

8 Martens H., Naes T., *Multivariate Calibration*, John Wiley & Sons, p111, 1991

9 She J.W., *Dissertation*, University Tübingen, 1992