DISCUSSION

of the John Mandel paper, The Regression Analysis of Collinear Data

R. W. Gerlach
Monsanto Agricultural Products Co.

I fully agree with Mandel in that one's model can (usually) only be assigned a degree of validity in the region spanned by the data used to generate the model. Though the range for all variables may be quite large, collinearity effectively restricts the model to a particular subregion. One should be aware of these restrictions so as not to misapply the model to those regions not represented by the data. I want to point out the need to carry out an additional initial operation; one should always examine the dataset for outliers. Otherwise the suggestion for using the largest and smallest values on each principal coordinate to examine the constraints may lead to overstating the region for a valid calibration. In fact, this could happen anyway if the shape formed by the data vectors was peculiar, perhaps occupying two disjoint regions for instance.

Additional constraints are frequently available to the analytical chemist. Minimum and maximum values along the original variables as well as conditions upon functions of these variables are frequently encountered. The location of the effective predictive domain within this potentially allowed domain could also be useful. A comparison might lead the researcher to conclude that more effort should be spent gathering additional data so that the calibration equation was valid over the desired region.

The variance factor (VF), resulting from the propagation of errors through the transformations, is a good method for observing how well characterized the model is at any location. Though principal components regression has been in the literature of analytical chemistry for some time [1,2][1] a paper dealing with the region of applicability for the model has only recently been published [3]. In this case the authors used as their criteria the expected mean square error. Hopefully, the propagation of error in this and related techniques will become more commonplace in analytical chemistry.

I think that the comparison of the measures advocated in this paper to the condition number is somewhat misdirected. The condition number can be used to provide a measure of how sensitive a model could be to variations in the data matrix. However, it would certainly not be appropriate to consider a condition number for the complete data matrix if one is dealing with only a subset of its dimensions in the principal component regression. The condition number assists one in interpreting the sensitivity of the model given all the original variables (or any orthogonal transformation). The condition number for the rotated coordinate system of the principal coordinates will be the same as for the original coordinate system. In the original coordinate system a large condition number signaled that the regression coefficients were not all well known. In the rotated eigenvector coordinate system this same condition number reflects the fact that coefficients for the eigenvectors with small eigenvalues will not be estimated accurately. However, since only the eigenvectors with significant eigenvalues will be considered in the principal component regression, the condition number for the entire matrix is not an appropriate parameter to consider. In fact, the only thing we can say is that one expects large condition numbers every time a principal component regression is the method of choice.

It should also be pointed out that other aspects of collinearity are frequently encountered by analytical chemists. While this paper deals with collinearity as it affects the region for applicability of the model in terms of predictability, it doesn't address questions as to the reliability of the model coefficients. Also, instead of generating a calibration or predictive equation, one might wish to evaluate possible models in which the independent factors behave somewhat similarly. What limitations are placed on the results of the traditional regression analysis? I want to mention that statisticians have already developed several appropriate techniques [4], such as methods to estimate confidence regions and the effective sample size. Hopefully, these and other measures to test the validity of the proposed model will be more widely used.

The propagation of errors through a constrained correlated regression would also be an appropriate technique for investigating the significance of the terms in a proposed model. As mentioned above, often there are known constraints, yet this information is commonly overlooked. A recent comparison of multivariate techniques applied to source apportionment of aerosols in which collinearity was an important factor showed that the known constraints were mostly ignored [5]. Mathematical techniques which deal with these extra conditions [6], though more complex numerically, should be investigated for their potential benefits to areas of analytical chemistry and brought into more common use.

[1]Figures in brackets indicate literature references.

References

[1] Bos, M., and G. Jasink, The Learning Machine in Quantitative Chemical Analysis, *Analytica Chimica Acta* **103**, pp 151–165 (1978).

[2] Martens, H., Factor Analysis of Chemical Mixtures, *Analytica Chimica Acta* **112**, pp 423–442 (1979).

[3] Fredericks, P. M.; Lee, J. B.; Osborn, P. R., and D. A. J. Swinkels, Materials Characterization Using Factor Analysis of FT-IR Spectra. Part 2: Mathematical and Statistical Considerations, *Applied Spectroscopy* **39**, pp 311–316 (1985).

[4] Willan, A. R., and D. G. Watts, Meaningful Multicollinearity Measures, *Technometrics* **20**, pp. 407–412 (1978).

[5] Currie, L. A.; Gerlach, R. W., and C. W. Lewis, *et al*, Interlaboratory Comparison of Source Apportionment Procedures: Results for Simulated Data Sets, *Atmospheric Environment* **18**, pp 1517–1537 (1984).

[6] Rust, B. W., and W. R. Burrus, Mathematical Programming and the Numerical Solution of Linear Equations, Elsevier (1972).