RESEARCH PAPER

Chemometrics Approach to the Determination of Polymorphism of a Drug Compound by Infrared Spectroscopy

Kwan R. Lee,[1,*] Gary Zuber,[2] and Lee Katrincic[2]

[1]*Statistical Sciences Department and* [2]*Analytical Sciences Department, SmithKline Beecham Pharmaceuticals, 709 Swedeland Road, King of Prussia, PA 19406-0939*

ABSTRACT

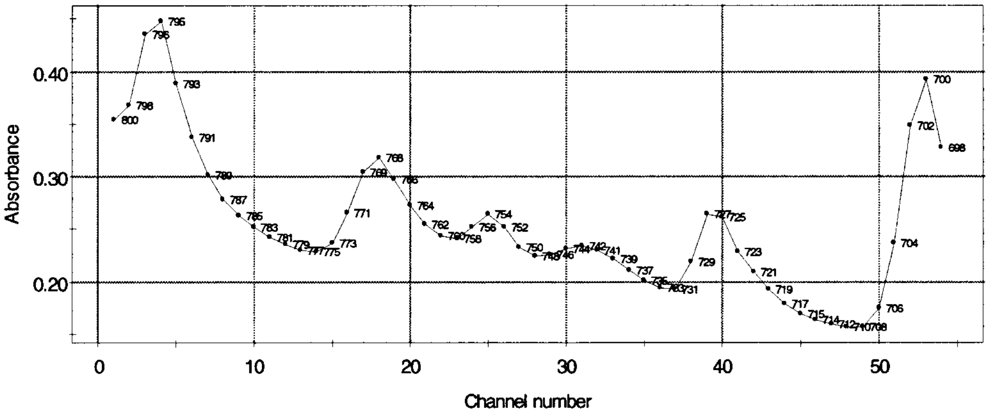
A chemometrics approach, multivariate calibration in particular, was used to determine the polymorphism of a drug compound based on Fourier transform infrared (FTIR) spectroscopy. The partial least-squares projection to latent structure makes use of all of the data, and the latent variables created by the method make use of hidden or partially separated peaks for quantitation. This paper illustrates the usefulness of the partial least-squares multivariate calibration method as an efficient tool to determine the polymorphism of a drug. Also, the analysis suggests the use of information from the modeling as diagnostic tools to gain more insight from the data. In particular, the diagnostic tools allow an analyst to assess design characteristics and any shortcomings of a calibration experiment for the polymorphism of a drug compound.

Key Words: *Infrared spectroscopy; Multivariate calibration; Partial least squares; Polymorphism.*

INTRODUCTION

Chemometrics can be described as the marriage of statistics with chemistry. It is an approach to analytical and measurement science that is based on the use of indirect observation. Currently, most of the qualitative and quantitative Fourier transform infrared (FTIR) spectroscopic methods developed within our laboratories make use of direct measurements, with only one variable considered at a time. These traditional approaches are often unsuc-

[*] To whom correspondence should be addressed.

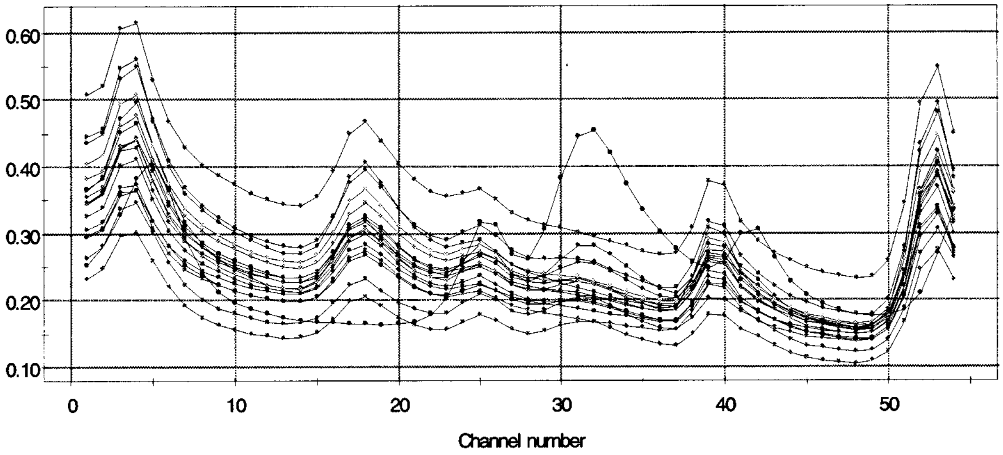


Simca-P 7.0 by Umetri AB 1999-01-14 10:55

Figure 1. Infrared spectroscopic profiles of a polymorphism for a drug compound.

cessful or limited and may require additional methods of data manipulation to visualize and isolate each data point. These univariate approaches may also be complicated and time consuming. Chemometrics allows the spectroscopist to use all the data or variables at the same time. These multivariate approaches typically make use of even the hidden information while utilizing data redundancies, which help to increase the efficiency of a method. By using these multivariate approaches, we may now develop quantitative methods previously considered impossible and have the potential to simplify analytical methods already in existence.

During the past few years, chemometrics techniques have proved their power in the quantitative data analysis of components in complex systems (1–5). Partial least-squares (PLS) projection to latent structure (6) is a proven multivariate calibration method in quantitative analysis using infrared spectroscopic profiles (1,2). It is called multivariate calibration since it utilizes all the channels of data (7). Its main idea is to make latent variables of original matrix *X* (predictor variable) and matrix *Y* (dependent variable). Latent variables are formed as a linear combination of all the original variables in *X* in such a way that most of the association with *Y* variables can be



Simca-P 7.0 by Umetri AB 1999-04-16 10:56

Figure 2. All 22 spectras overlayed.

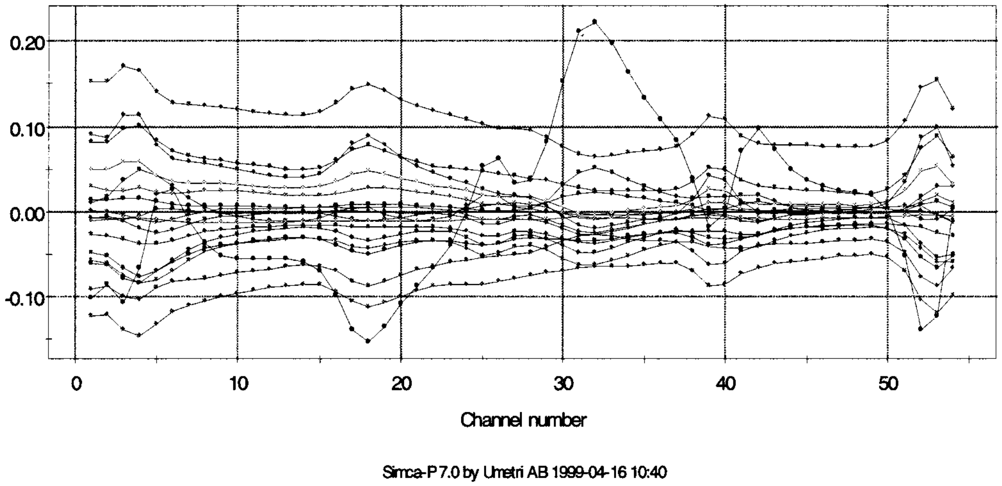


Figure 3. Centered spectras overlayed.

explained. The weights of the linear combination are called *loadings*, and the resultant linear combinations are called *scores*. Next, the main function of PLS is dimensional reduction. As in the case of principal component analysis (PCA), for which dimensional reduction is achieved by explaining most of the variation in the *X* matrix, PLS can achieve dimensional reduction when the first few linear combinations of the *X* matrix can explain most of the variation in the *Y* matrix. As in an ordinary least-squares (OLS) regression, PLS eventually comes up with equations that relate all the variables in the *X* matrix with the *Y* matrix. However, PLS is better for the highly multicollinear *X* matrix since it can overcome the undesirable properties of OLS. With multicollinearities in the *X* matrix, the regression coefficient of OLS will be highly unstable and will have large standard errors, which makes prediction from the model less useful. Also, by examining the first few factors (the loadings and scores) of PLS, we can gain valuable insights into the data analysis.

In the case of spectroscopy calibration for polymorphism of a drug compound, the matrix of spectra formed from the several samples will make up an *X* matrix. In our example below, it is a matrix of 22 rows and 54 columns. The 22 rows derive from the number of different samples, and there are 54 different wavelengths (channels). The *Y* matrix has the same 22 rows, but only a single column since it contains percentage concentration of a form II of the drug compound. We call the response ‘‘low melt.’’

EXPERIMENTAL

Infrared spectroscopy deals with the study of vibrational modes of a system. Infrared spectral profiles arise from the intensities of absorption of the fundamental vibrations and describe the chemical structure. In the case of a mixture of chemical components arising from a reaction, the spectroscopic profiles represent the combination of the spectral profiles of the products formed during the reactions in the system. Thus, changes in the infrared profiles of the containing mixtures formed during a reaction

Table 1

Model Summary Statistics for Partial Least-Squares Calibration

Compound	R^2X	R^2X (cumulative)	R^2Y	R^2Y (cumulative)	Q^2	Q^2 (cumulative)
1	0.470	0.4706	0.811	0.8118	0.114	0.114
2	0.524	0.9948	0.181	0.9929	0.953	0.959
3	0.004	0.9990	0.006	0.9995	0.914	0.996

X = spectra; *Y* = % concentration of low melt.

Table 2

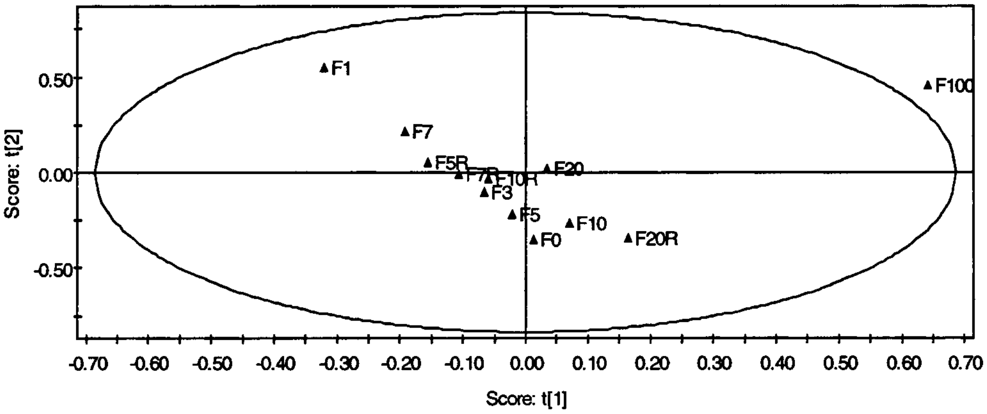
Scores for the First Two Components

Observation	Name	*t*[1]	*t*[2]	*u*[1]	*u*[2]	Low Melt	*u*[1] − *t*[1]
1	F100	0.641	0.455	0.818	0.456	100	0.177
2	F20	0.034	0.025	0.042	0.020	20	0.008
3	F20R	0.164	−0.343	0.042	−0.314	20	−0.122
4	F10	0.070	−0.263	−0.055	−0.322	10	−0.125
5	F10R	−0.059	−0.032	−0.055	0.011	10	0.004
6	F7	−0.192	0.220	−0.084	0.279	7	0.108
7	F7R	−0.106	−0.004	−0.084	0.057	7	0.022
8	F5	−0.021	−0.214	−0.103	−0.212	5	−0.082
9	F5R	−0.156	0.056	−0.103	0.135	5	0.052
10	F3	−0.067	−0.103	−0.123	−0.144	3	−0.056
11	F1	−0.320	0.552	−0.142	0.458	1	0.178
12	F0	0.013	−0.347	−0.152	−0.425	0	−0.165

Table 3

Correlation Matrix of the First Two Scores

	t[1]	*t*[2]	*u*[1]	*u*[2]	Low Melt	*u*[1] − *t*[1]
t[1]	1.000					
t[2]	0.000	1.000				
u[1]	0.901	0.425	1.000			
u[2]	0.000	0.981	0.434	1.000		
Low melt	0.901	0.425	1.000	0.434	1.000	
u[1] − *t*[1]	0.000	0.981	0.434	1.000	0.434	1.000



Ellipse: Hotelling T2 (0.05)
Simca-P 7.0 by Umetri AB 1999-01-14 11:35

Figure 4. Two-dimensional projection of the spectral profiles.

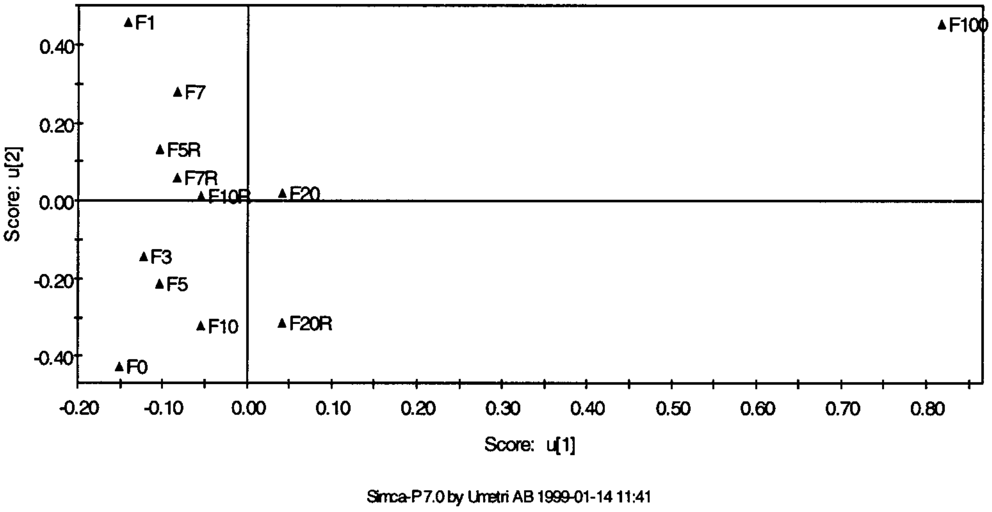


Figure 5. The scores plot of response *Y*.

imply quantitative variation in the reaction products formed in the system.

An FTIR method has been developed to detect the presence of the low-melting polymorphic form (form III; differential scanning calorimetry [DSC] endotherm at 156°C) of a drug compound within the high-melting polymorphic form (form II; DSC endotherm at 177°C). The normal pretreatment would make use of Fourier deconvolution (resolution enhancement) across the aromatic region of the IR spectrum (780–740 cm^{-1}). A typical spectroscopic profile of a drug compound is shown in Fig. 1.

However, we have used this nonpretreated area of the IR spectra and PLS multivariate calibration to predict each polymorphic form successfully, which resulted in simpler procedures. Concentrations ranged from 0% low melt to 20% low melt in the presence of high melt and made use of 12 samples as a work set (calibration set), with the remaining 10 samples used as a test set (validation set). Figure 2 shows all 22 spectra overlayed, and

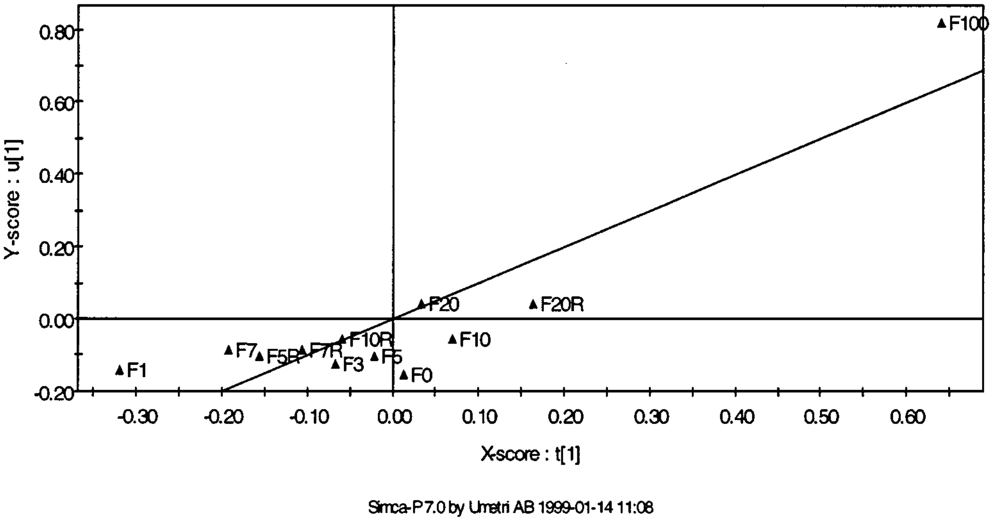
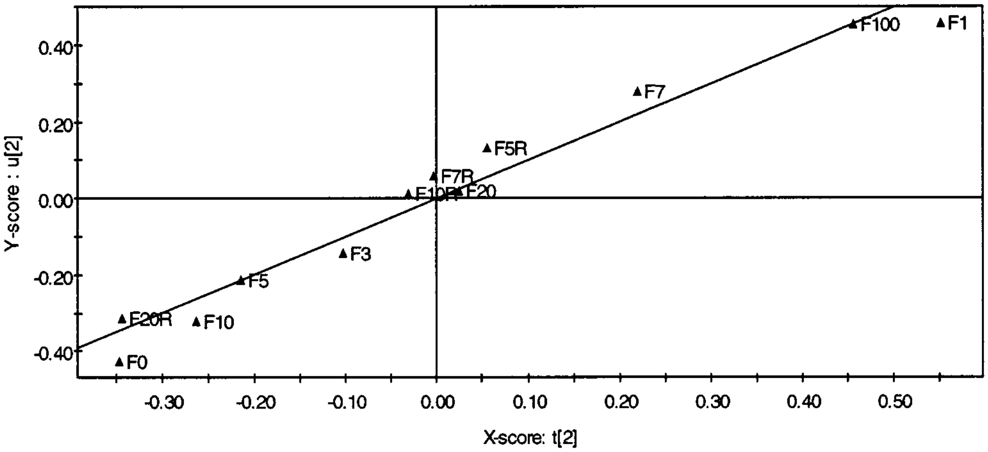


Figure 6. The first *X* scores versus *Y* scores.



Simca-P 7.0 by Umetri AB 1999-01-14 11:32

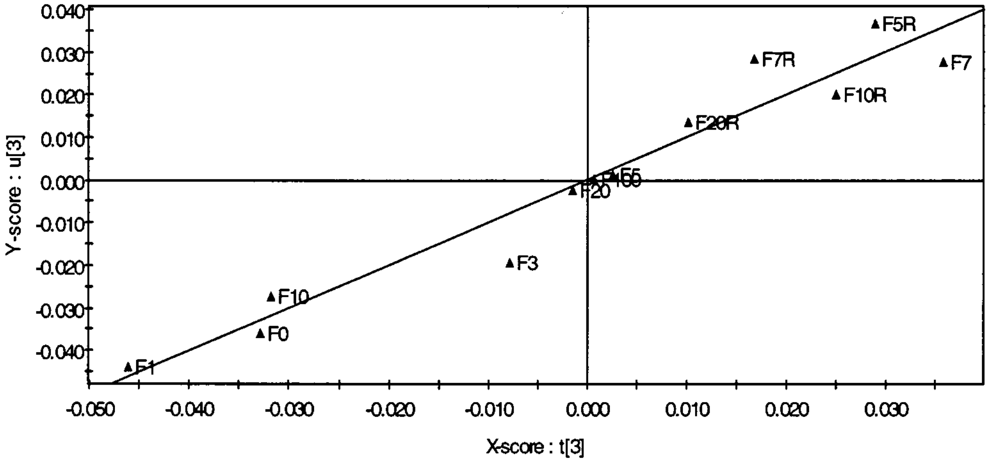
Figure 7. The second *X* scores versus *Y* scores.

Fig. 3 shows a centered (mean was subtracted) version of the same spectra. The distinctly different spectrum is F100 for low melt 100% (i.e., it is the spectrum for 100% concentration of form II of the drug compound).

MULTIVARIATE CALIBRATION

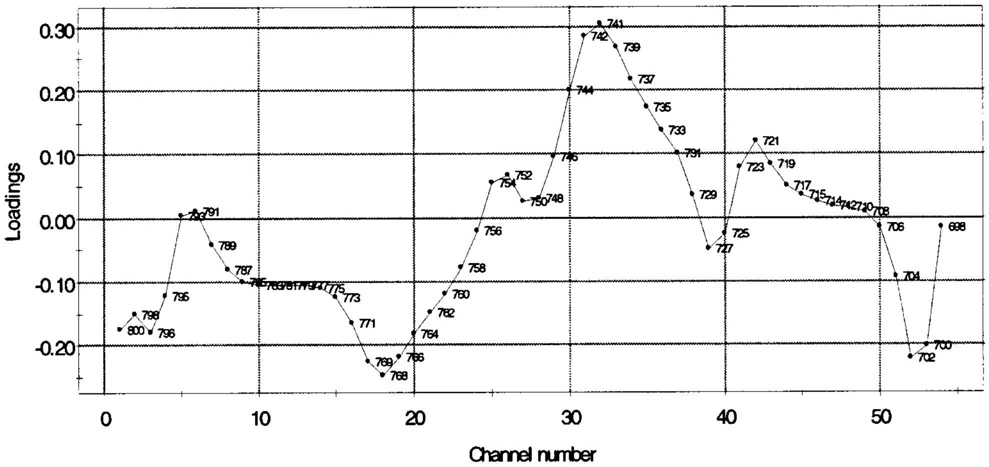
PLS Multivariate Calibration

The chemometrics software SIMCA® (8) was used to perform the PLS analysis and to generate graphics. The first three statistically significant components explained most of the variation in the data. R^2 is the fraction of the sum of squares of all the variables explained by the current component. Q^2 is similar to R^2, but was computed using cross validation. The predicted error sum of squares (PRESS) for the cross validation is the squared differences between observed and predicted values for the data kept out of model fitting. Parts of the data are kept out of the model development, then predicted by the model, and compared with the actual values. This procedure is repeated several times until every data element has been



Simca-P 7.0 by Umetri AB 1999-01-14 13:30

Figure 8. The third *X* scores versus *Y* scores.

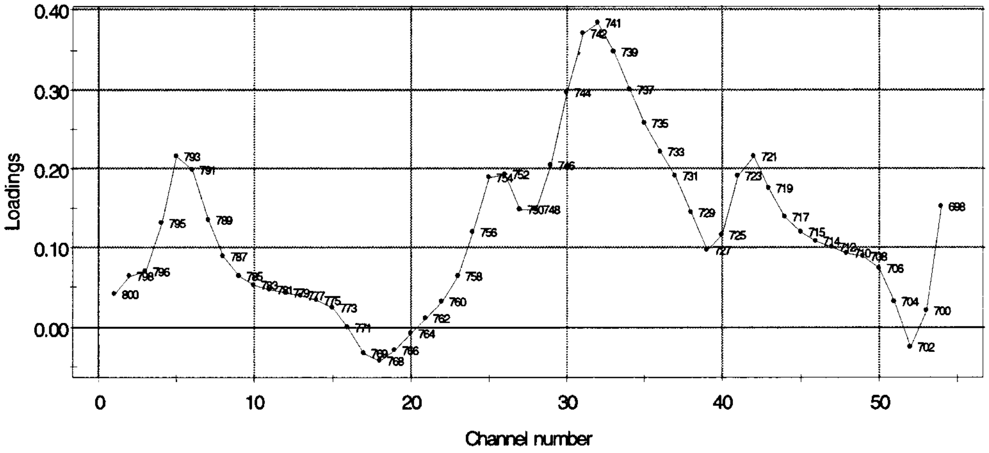


Simca-P 7.0 by Umetri AB 1999-01-14 13:54

Figure 9. The first loadings of spectral profiles.

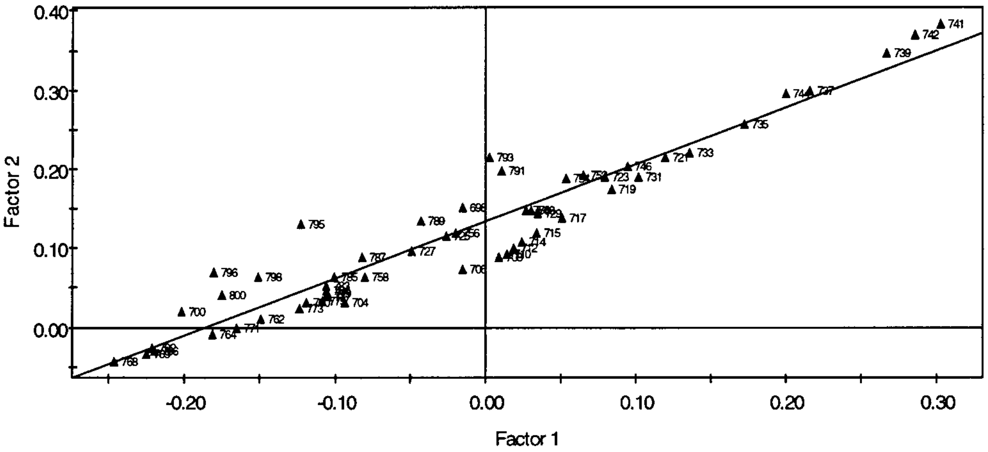
kept out once and only once. The final PRESS then has contributions from all the data. The Q^2 is then computed by 1 − PRESS/SS, where SS is the overall sums of squares. Q^2 is a more realistic measure of the goodness of approximation than R^2, but Q^2 is not additive, unlike R^2. As we can see in Table 1, the first two factors explain 99.5% of the variation in the X matrix and 99.3% of the variation in the Y part. The corresponding cross-validated measure is 95.9%. A two-component model may have been adequate, but a third component was added to the final model since it had some contribution to Q^2.

The next step is to examine scores and loadings of the first two dimensions. The actual values of the X scores (t's) and Y scores (u's) are listed in Table 2. As you can see in the correlation matrix in Table 3, the values of t are not correlated to each other, making the contribution of each dimension independent from others. Since we have only one response (low melt), the first Y score u[1] is simply statistically equivalent to low melt itself. In other words, u[1] is a linear combination of low melt, and they have perfect correlation. The second Y score u[2] is leftover from u[1] after subtracting the effect of



Simca-P 7.0 by Umetri AB 1999-01-14 13:56

Figure 10. The second loadings of spectral profiles.

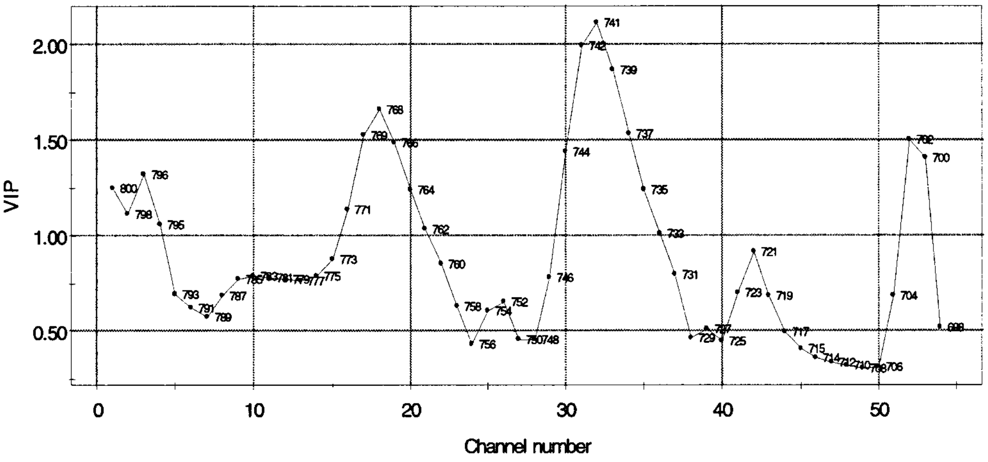


Simca-P 7.0 by Umetri AB 1999-01-14 13:58

Figure 11. The first two loadings of spectral profiles.

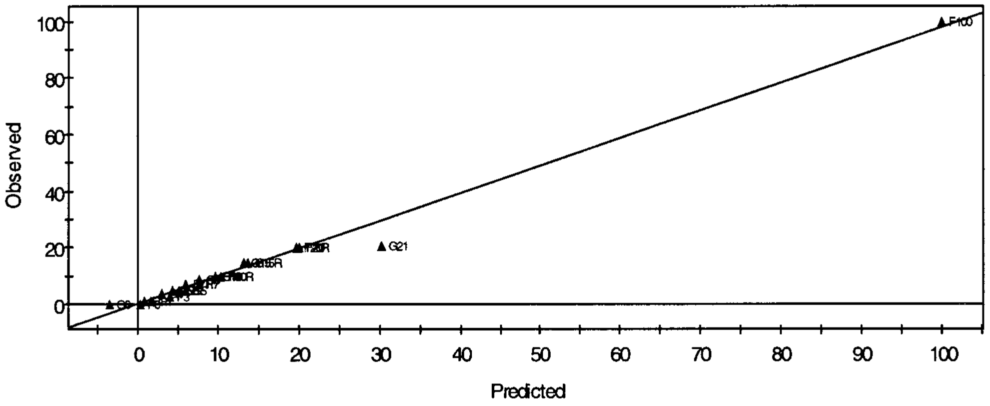
$t[1]$. In other words, $u[2]$ is an unexplained part of the response by the first X score, which is to be explained, hopefully, by next scores. In this case, we have only three components each since the three factors could adequately describe the variation in both the spectra and the response low melt.

The plot of two-dimensional scores in Fig. 4 can be interpreted as the best projection of the spectral data into two dimensions. The ellipsoid defines the 95% confidence region, and all 11 points except F100 are comfortably inside the region. The further away a point is from the center of the model, the more ''leverage'' it exerts for prediction of the response. In that sense, F1 and F100 have the most influence, but most of the points are close to the center. Ideally, it is desirable to have sample points scattered uniformly over the ellipsoidal region. Also, the replicates were not measured precisely. The replicates F5R, F7R, and F20R were not close to corresponding to F5, F7, and F20, which added uncertainty to the calibration of polymorphism. Ideally, replicate-to-replicate variation should be smaller than the sample-to-sample variation.



Simca-P 7.0 by Umetri AB 1999-01-14 14:01

Figure 12. Variable importance in projection (VIP) for spectral profiles.



RMSEP=2.25484
Simca-P 7.0 by Umetri AB 1999-01-14 14:03

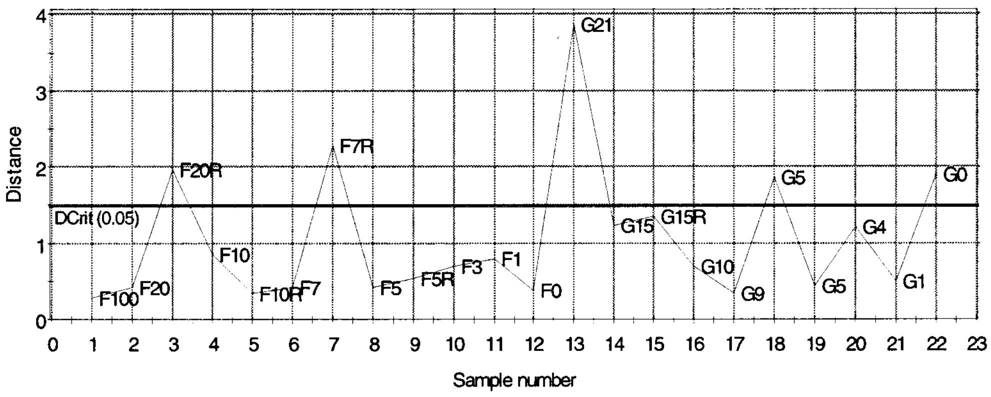
Figure 13. Predicted versus observed values for both work and test sets.

Table 4

Prediction of Test Set (ts) Based on Work Set (ws) Model

Observation	Name	Set	Low Melt (Observed)	Low Melt (Predicted)	SE	Residual
1	F100	ws	100.00	100.03	0.643	−0.033
2	F20	ws	20.000	20.074	0.193	−0.074
3	F20R	ws	20.000	19.694	0.339	0.306
4	F10	ws	10.000	9.609	0.360	0.391
5	F10R	ws	10.000	10.408	0.274	−0.408
6	F7	ws	7.000	7.679	0.397	−0.679
7	F7R	ws	7.000	5.984	0.245	1.016
8	F5	ws	5.000	5.126	0.241	−0.126
9	F5R	ws	5.000	4.334	0.320	0.666
10	F3	ws	3.000	3.980	0.218	−0.980
11	F1	ws	1.000	0.808	0.609	0.192
12	F0	ws	0.000	0.271	0.393	−0.271
13	G21	ts	21.000	30.281	0.645	−9.281
14	G15	ts	15.000	13.187	0.261	1.813
15	G15R	ts	15.000	13.733	0.442	1.267
16	G10	ts	10.000	9.609	0.360	0.391
17	G9	ts	9.000	7.679	0.397	1.321
18	G5	ts	5.000	5.984	0.245	−0.984
19	G5R	ts	5.000	4.334	0.320	0.666
20	G4	ts	4.000	2.919	0.285	1.081
21	G1	ts	1.000	1.690	0.493	−0.690
22	G0	ts	0.000	−3.526	0.731	3.526

Fxx are work set (calibration) samples, and *Gxx* are test set (validation) samples.



Dcrit [3] = 1.4801 Normalized distances, Non weighted residuals

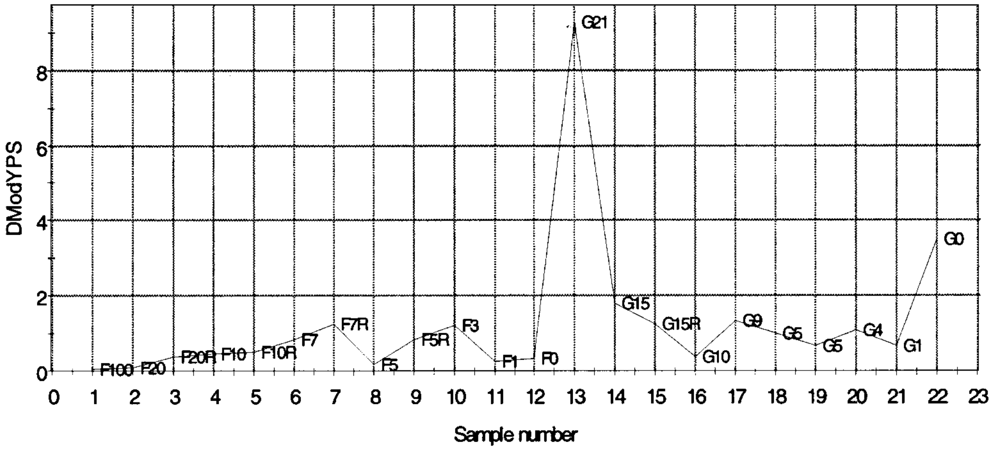
Simca-P 7.0 by Umetri AB 1999-01-15 15:05

Figure 14. Residual distance to the model in spectral space.

The same conclusion concerning the experimental design of samples can be made by looking at the plot of scores for the *Y* space (Fig. 5). Ideally, all the sample points should be uniformly scattered around the center. In the plot below, we can observe that the sample F100 is alone in the right corner. We could have a better experimental design by having more samples between F20 and F100 or, rather, make the samples more evenly distributed between the minimum and maximum value of low melt.

By plotting *X* and *Y* scores together (Figs. 6–8), we can assess the strength of linear relationships accounted for by each component. The first scores can predict most of the samples reasonably (81%), but model the sample of midlevel concentration (Fig. 6). The second scores add accuracy to more extreme samples, which are not well accounted for in the first dimension. Those samples include F0, F1, F3, F5, F7, F20R, and F100 (Fig. 7). The accuracy improvement by adding the third dimension is very small (Fig. 8). However, by the third dimension, most of the variation in the data was well modeled, and adding more components to the model may not improve the prediction ability of the model in the future.

Loadings are useful information for finding which channels are more important than others in explaining the variation in the spectral data and also for predicting the



Simca-P 7.0 by Umetri AB 1999-01-15 15:09

Figure 15. Residual distance to the model in response space.

response (Figs. 9–11). The two line plots of the loadings in the first two dimensions show that the two loadings are very similar except for some scale difference (Figs. 9 and 10). The scatter plot of the two loadings shows almost perfect correlation (Fig. 11). It is interesting to note that 739, 741, and 742 are important wavenumbers (Figs. 9 and 10), and F100 was an ''outlier'' since it had large values in those channels (Fig. 3). In fact, after close examination, we found that the first and the second loadings have a shape very similar to the F100 spectra. What this means is that sample F100 has most of the influence in explaining the variation of the data and making the calibration model for low melt. Again, it would have been better if we had designed the samples uniformly over the range of the concentration instead of making most samples in the range of 0–20 and then adding 100. One possible design of the samples could be concentrations of 0%, 10%, 20%, 30%, 40%, and 50% with replicates at each point. If application of the model is toward a lower concentration, say 0–20%, then again we can spread the samples uniformly over the range with possible replicates.

The variable importance in projection (VIP) takes account of the influence of each channel on the response and is a summary measure for assessing the relevance of the channel for explaining the response (Fig. 12). VIP is a function of the loading weights for all three components used in the modeling. The VIP plot shows that the wavenumbers 741, 768, 702, and 796 are important for the prediction of the response and again are mostly influenced by the F100 sample.

Prediction of the Test Set

Although the experimental design of the samples could have been better, the calibration model predicted the samples in the test (validation) set adequately except for samples G21 and G0, which are sample points further from the midlevel concentrations (Fig. 13). Actual predicted values and their standard errors are given in Table 4. Approximate prediction intervals for the predicted values can be found by the formula prediction ± t ∗ standard error. The degrees of freedom of the t value should be 9 ($= 12 - 3$), and the tabled t value is 2.26 for 95% confidence.

The root-mean-squared error of prediction (RMSEP) was computed as 2.25, about 4.5% of the midlevel concentration (50%). The samples F7R and F20R can be considered possible outliers in spectral space, but were included in the calibration model. In the test (validation) set, G21, G5, and G0 had larger residual distances in the X space, so their predictions were poor (Fig. 14). Removing those two possible outliers did not improve the quality of the model appreciably, however. For the test set, G21 and G0 were not predicted well (Fig. 15). It appears that 1% or below cannot be predicted very precisely.

Final Calibration Equation

The final equation is of the following form:

$$\begin{aligned}\%\ \text{Low melt concentration} = 3.405 &- 6.361 * (w800)\\ &- 2.240 * (w798)\\ &+ 6.29 * (w796)\ldots\end{aligned}$$

where $w800$ is the absorbance at wavelength 800, for example.

The raw coefficients together with the VIP values of the first 30 channels are listed in Table 5, but actual use

Table 5

Coefficients and Variable Importance in Projection (VIP) Data

Channel Number	Wavelength	Coefficient	VIP
0	Const	3.405	—
2	800	−6.361	1.246
3	798	−2.240	1.112
4	796	6.292	1.323
5	795	25.533	1.055
6	793	36.825	0.695
7	791	27.294	0.619
8	789	9.284	0.575
9	787	−3.463	0.687
10	785	−10.569	0.769
11	783	−13.082	0.785
12	781	−15.680	0.779
13	779	−17.453	0.768
14	777	−19.831	0.767
15	775	−20.678	0.785
16	773	−22.573	0.875
17	771	−25.320	1.140
18	769	−28.119	1.524
19	768	−27.885	1.665
20	766	−26.459	1.489
21	764	−24.826	1.241
22	762	−22.118	1.037
23	760	−19.080	0.850
24	758	−13.213	0.631
25	756	0.138	0.431
26	754	15.446	0.609
27	752	15.470	0.652
28	750	4.157	0.456
29	748	3.904	0.458
30	746	16.206	0.784

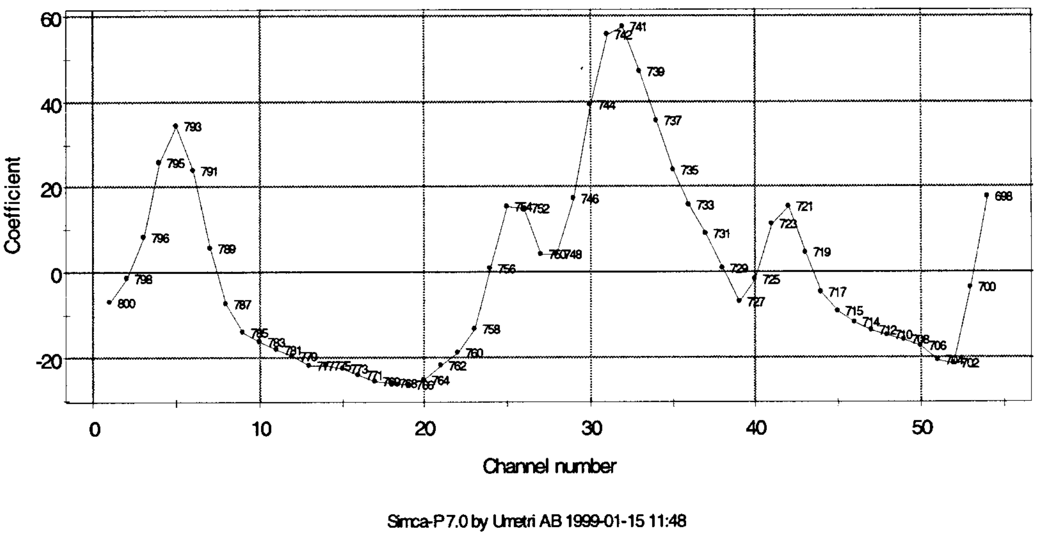


Figure 16. Coefficients of the final calibration equation.

of the coefficient and prediction for future determination of polymorphism will be made using software instead of manual computation from the equation above. The magnitude of the coefficient is related to VIP; actually, they have a correlation of about 70% in this case. But, VIP values are more reliable for assessing the importance of a specific wavelength.

Figure 16 shows the coefficients as a line plot; again, its shape is very close to the first and second loadings. We might say the coefficients were largely determined by the influence of the F100 sample, which again emphasizes the importance of sample design in calibration.

To validate further that the calibration model was not due to pure chance, we used 10 random permutations of

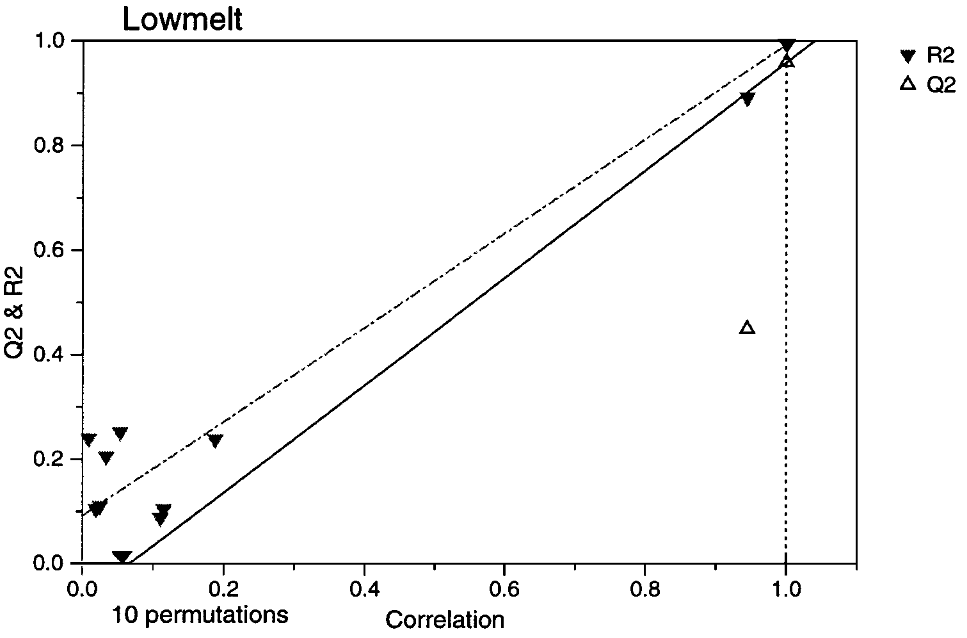


Figure 17. Validation based on random permutation of response.

the data and tried to come up with a calibration model for each permutation. As can be seen in Fig. 17, most of the permutations had negligible prediction ability as measured by R^2 and Q^2. One permutation had an R^2 of more than 80% by chance, but its Q^2 is again less than 50%. The point at correlation 1.0 is that of the original data.

CONCLUSION

A chemometrics approach to the use of FTIR spectroscopy in the qualitative and quantitative analysis of drug substance polymorphism mixtures has a great deal of potential. Since a multivariate method makes use of all the data contained within an information-rich technique such as infrared spectroscopy, our ability to design and implement quick and efficient spectroscopic methods may greatly increase. This has already been observed in the areas of near IR and react IR, for which information and utility would be very limited if it were not for the use of chemometrics.

This paper illustrated the usefulness of a PLS multivariate calibration method as an efficient tool to determine the polymorphism of a drug, as well as the suggestion of using information from the modeling (scores and loadings) as diagnostic tools to gain more insight from the data. Although PLS can do a good job of utilizing multivariate information and of predicting polymorphism of a drug compound, it should be noted that careful attention should be paid to the experimental design of mixtures. The concentration range of the mixture should uniformly cover the range of possible applications to come up with more accurate and useful calibration models. Consideration of experimental design will become more important when we are dealing with polymorphism of more than two-component mixtures (7).

REFERENCES

1. A. A. Kristy, O. M. Kvalheim, F. O. Libnau, G. Aksnes, and J. Toft, Vib. Spectrosc., 6, 1–14 (1992).
2. J. Toft, O. M. Kvalheim, T. V. Karstang, A. A. Christy, K. Kleveland, and A. Henriksen, Appl. Spectrosc., 46, 1002–1008 (1992).
3. O. M. Kvalheim and Y. Z. Liang, Anal. Chem., 64, 936–945 (1992).
4. Y. Z. Liang, O. M. Kvalheim, H. R. Keller, D. L. Massart, P. Kiechle, and F. Emi, Anal. Chem., 64, 946–953 (1992).
5. H. R. Keller, Y. Z. Liang, O. M. Kvalheim, and D. L. Massart, Anal. Chem. Acta, 267, 63–71 (1992).
6. S. Wold, A. Ruhe, H. Wold, and W. J. Dunn III, SIAM J. Sci. Stat. Comput., 5, 735–743 (1984).
7. H. Martens and T. Naes, *Multivariate Calibration*, John Wiley and Sons, New York, 1989.
8. SIMCA, SIMCA-P for Windows, version 7.0, Umetri AB, Umea, Sweden, 1998.