Detection of Specific Sugars in Dairy Process Samples Using Multivariate Curve Resolution

P. W. HANSEN,*,[1] A. S. VAN BRAKEL,† J. GARMAN,† and L. NØRGAARD‡
*Foss Electric A/S, R&D Greenhouse, Slangerupgade 69, DK-3400 Hillerød, Denmark
†New Zealand Dairy Research Institute, Private Bag 11 029, Palmerston North, New Zealand
‡The Royal Veterinary and Agricultural University, Department of Dairy and Food Science, Rolighedsvej 30, DK-1958 Frederiksberg C, Denmark

ABSTRACT

Dairy process monitoring by application of multivariate curve resolution using alternating least squares is presented. Alternating least squares was used for resolving Fourier transform infrared spectral data from a dairy batch process in which lactose is enzymatically hydrolyzed to glucose and galactose. It was possible to extract four compounds (fat, lactose, and two other sugar components) from the spectral data obtained from nine process runs. Subsequently, the pure spectra obtained in this way were used to monitor the content of these compounds in two new process runs. In this way, alternating least squares made it possible to follow the hydrolysis process by Fourier transform infrared spectroscopy without the need for reference analyses. When the results were correlated to reference results for lactose, the accuracy was similar to that obtained when a partial least squares regression was performed on the same data; lactose correlation was 0.980 when alternating least squares was used and was 0.987 when partial least squares was used.

(**Key words**: lactose hydrolysis, multivariate curve resolution, process control, reference-independent estimation)

Abbreviation key: **ALS** = alternating least squares, **FTIR** = Fourier transform infrared, **MIR** = mid-infrared, **NIR** = near-infrared reflection, **NIT** = near-infrared transmission, **PARAFAC** = parallel factor analysis, **PLS** = partial least squares, **SSE** = sum of squared errors.

INTRODUCTION

Process control of industrial processes is increasing in importance as online analytical equipment providing fast and reliable results becomes available. Near-infrared reflection (**NIR**) and near-infrared transmission (**NIT**) spectroscopies are the most frequently used methods in many branches of industry, and mid-infrared (**MIR**) spectroscopy has proved very useful for process milk analysis. Milk analysis using MIR equipment is generally more accurate than the corresponding NIR or NIT method because MIR contains more specific information (fundamental absorptions) and has stronger signals than NIR or NIT, which detect derived information (overtones and combination bands). In addition, full spectrum instruments based on Fourier transform infrared (**FTIR**) spectroscopy for dairy product analysis in the laboratory are showing promising results with regard to the number of components [e.g., specific sugars (17), casein (9, 11), and urea (7)] that can be measured.

Hydrolysis of lactose in milk is of interest because a large number of racial groups suffer from lactose intolerance (i.e., they are not able to cleave lactose into glucose and galactose). Therefore, low lactose milk products produced by the action of the enzyme β-galactosidase are of commercial interest. The process is sensitive to the initial conditions such as temperature and β-galactosidase concentration (3, 6). Therefore, the concentrations of the sugars need to be monitored during the course of the reaction to control the process. The reaction is typically completed within a few hours, thus it requires a fast analytical method such as MIR.

The general approach when analyzing spectral data with the intention of generating future predictions of milk constituents is to use one of several multivariate methods relating the data to wet chemistry results. These methods include partial least squares (**PLS**) regression, which is described elsewhere (12). Ordinary multivariate methods require an accurate and reproducible reference method to obtain a reliable calibration, which tends to be resource demanding—especially when the typical number of calibration samples (15 to hundreds) is taken into account. In addition, during a process,

Received September 10, 1998.
Accepted March 3, 1999.
[1]Corresponding author.

some intermediate species might only exist for a limited period of time (i.e., they are both produced and consumed during the reaction). In such a case, it might be difficult to isolate the intermediates and to measure them using the reference methods.

Such problems can be solved using a regression method of higher order, such as parallel factor analysis (**PARAFAC**) (1). A PLS regression handles second order data, and data can be arranged in a matrix. The PARAFAC method requires data to be of an order higher than two, which occurs when a spectral landscape is obtained for each sample, and the whole data set can be arranged in a cube. The landscape could be obtained by measuring a reacting sample at fixed time intervals during the process. The individual spectra constituting the landscape will be related, and these relationships implicitly contain information on the concentrations of all compounds in the sample absorbing infrared light.

The PARAFAC method is able to resolve these variations and to produce concentration profiles and pure spectra corresponding to the absorbing species present in the sample. The concentrations will be arbitrary but proportional to the true concentrations. If correlated species are present, the concentration profiles will be the sums of such correlated compounds. Usually, PARAFAC is the most useful method for multivariate curve resolution; because it can handle more than one sample at a time, the solutions to the mathematical problem are unique (i.e., only one solution to each problem exists), and they might resemble real spectra and concentrations. Qualitative results have been obtained on resolving absorption and emission profiles from fluorescence spectra of sugar samples with PARAFAC (1).

In the present case, PARAFAC would not work because the actual shape of the concentration profiles were strongly dependent on the initial conditions of the process. The PARAFAC would be able to analyze only one landscape (sample) at a time or the unfolded data set. When a landscape is unfolded (e.g., when the spectra from the individual runs are appended to each other), one of the directions in the three-dimensional structure is lost.

Alternating least squares (**ALS**), sometimes referred to as alternating regression (10), is a two-way method that handles one landscape or unfolded data set at a time. Tauler et al. (19) reported how such curve resolution methods work. The use of ALS produces pure spectra and concentration profiles in a way similar to that of PARAFAC performed on the unfolded data set. The ALS method has been used for resolution of infrared process data with excellent results (4, 18).

The aim of the present work was to investigate whether ALS is capable of resolving the changes occurring in the FTIR spectra during the course of the lactose hydrolysis process and thus to obtain concentration profiles and pure spectra for the involved compounds. This procedure was done without the use of reference analyses to show the resolving power of the ALS method. It should be possible to monitor the concentrations of the components in new process runs because pure spectra were obtained. The results were compared with results from an ordinary PLS regression performed on the same data set.

MATERIALS AND METHODS

ALS

The ALS method relies on the assumption that the Beer-Lambert law is obeyed perfectly [i.e., that a spectrum (the row vector **x**) of a given sample can be seen as a linear combination of the pure constituent spectra (contained in the matrix **A**)], thus

$$\mathbf{x} = \mathbf{cA} \qquad [1]$$

where **c** = row vector containing the concentrations of the constituents corresponding to the pure spectra in **A**. When more than one spectrum is measured, the general expression becomes

$$\mathbf{X} = \mathbf{CA} \qquad [2]$$

where **X** = landscape containing the spectra in its rows, and **C** = matrix containing the concentrations corresponding to each spectrum. In this context, one sample is **X** (i.e., a collection of spectra from one process run).

A typical landscape from one lactose hydrolysis run with seven FTIR spectral recordings is presented in Figure 1. Most of the variation of the spectra was between 1000 and 1200 cm^{-1}. This result was expected because the only compounds affected by the reaction are sugars, which show strong absorptions because of stretching of the sugar C-O bonds in this range. The ALS method calculates the pure spectra **A** from the input spectra in **X** (the landscape) and an estimate of **C** (e.g., random numbers) using a rearranged form of [2]:

$$\mathbf{A} = \mathbf{C}^{+}\mathbf{X} \qquad [3]$$

where $\mathbf{C}^{+}$ = some pseudoinverse of **C**, followed by a calculation of a better estimation of **C** from this **A**:

$$\mathbf{C} = \mathbf{XA}^{+} \qquad [4]$$

where $\mathbf{A}^+$ = pseudoinverse of **A**. Equations [3] and [4] are repeated until convergence (or a maximal number of iterations) has been reached.

Various constraints can be applied to the spectra and concentration profiles in **A** and **C** to avoid physically meaningless solutions (1). For example, the concentrations of **C** cannot possibly become negative, so a non-negativity constraint would be reasonable. In addition, when we examined compounds produced and consumed during a chemical process, the concentration profiles were expected to have only one maximum during the course of the reaction. This observation led us to apply the unimodality constraint, which limits solutions to smooth concentration curves with only one maximum each. It could be argued that a non-negativity constraint would be appropriate for the pure spectra in **A** as well, but in this specific application it is not. The FTIR absorbance spectra were calculated using a water background, which causes slightly negative absorbances. Thus, constraint of **A** to non-negativity would restrict the algorithm too much.

Non-negativity can be applied in various ways. The most straightforward approach is to force negative values to zero (e.g., in **C**) after each iteration. This method is very simple and does not necessarily lead to the optimal description of **X** (i.e., the least squares solution). The approach employed here [adopted from reference (2)] forces only one concentration profile (or spectrum) at a time to zero and is followed by a correction of the pure spectrum matrix, **A** (or concentration matrix, **C**). This modification leads to the optimal result.

If more process runs were contained in the same data matrix **X**, the unimodality constraint of **C** would not work. In such a case, only the parts of **C** originating from the same process should be constrained. This approach was used in the present work.

After the concentration profiles and pure spectra have been obtained, the same principle can be used for prediction of the constituents of an unknown sample by applying the vector form of Equation [4] to the spectrum, **x**, of the sample:

$$\mathbf{c} = \mathbf{x}\mathbf{A}^+ \qquad [5]$$

The concentration row vector **c** will be in arbitrary units but will be linearly related to the actual concentrations.

Calculations

The data analysis and calibration were performed on a computer using Matlab® software (13). The pseudoinverse in Equations [3] and [4] were calculated using the built-in functions of Matlab®. The calibration routines were either programmed by the authors or taken from the PLS_Toolbox (14).

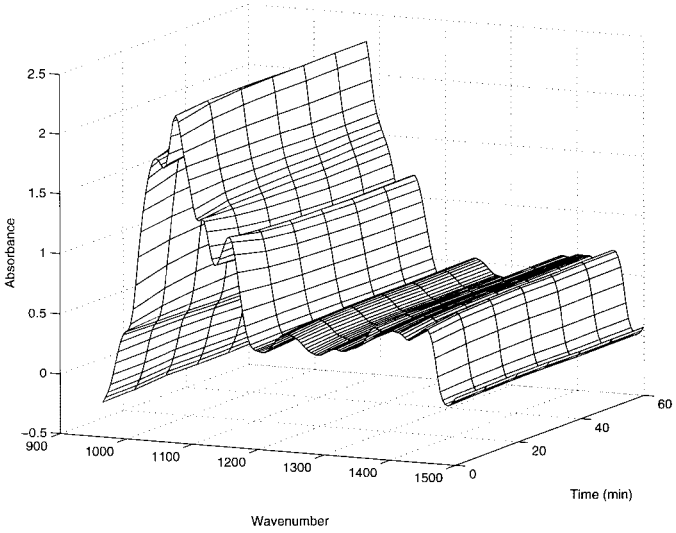

Figure 1. Infrared landscape obtained during 60 min (with sampling every 10 min) of a lactose hydrolysis process. Major changes are in the range from 1000 to 1200 cm^{-1} where sugars generally show strong absorptions.

Repeatability is expressed as a mean standard deviation (s_r) of multiple determinations performed under identical conditions and is calculated as

$$s_r = \sqrt{\frac{1}{q(n-1)} \sum_{j=1}^{q} \sum_{i=1}^{n} (x_{j,i} - \bar{x}_j)^2}$$

where q = number of samples, n = number of replicates, $x_{j,i}$ = result of replicate i of sample j, and $\bar{x}_j$ = average result of the sample j.

Accuracy is expressed as the root mean square error of prediction (RMSEP) and is calculated as

$$\text{RMSEP} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (x_{i,\text{reference}} - x_{i,\text{predicted}})^2}$$

where N = number of determinations [number of samples (q) × number of replicates (n) from above], and $x_{i,\text{reference}}$ and $x_{i,\text{predicted}}$ = reference and predicted values corresponding to determination i, respectively.

When a bias (mean difference between reference results and predictions) is observed, the standard error of prediction (SEP) is used. It is calculated as

$$\text{SEP} = \sqrt{\frac{1}{N-1} \sum_{i=1}^{N} (x_{i,\text{reference}} - x_{i,\text{predicted}} - \text{bias})^2}$$

If two variables are related by performing a univariate linear regression (slope a, intercept b) between instrumental responses ($x_{i,instrumental}$) and reference results ($x_{i,reference}$), the accuracy of future predictions can be estimated by the use of the standard error of calibration (SEC), calculated as

$$SEC = \sqrt{\frac{1}{N-2}\sum_{i=1}^{N}(x_{i,reference} - (ax_{i,instrumental} + b))^2}$$

Correlation is calculated as

$$R^2 = \left[\frac{\frac{1}{N}\sum_{i=1}^{N}(x_{i,reference} - \bar{x}_{reference})(x_{i,predicted} - \bar{x}_{predicted})}{s_{reference}s_{predicted}}\right]^2$$

where N, $x_{i,reference}$, and $x_{i,predicted}$ are defined above, and $\bar{x}_{reference}$, $s_{reference}$, $\bar{x}_{predicted}$, and $s_{predicted}$ = mean and standard deviations of the reference or predicted results, respectively.

Finally, the fit of the model to **X** (Equation [2]) is expressed as the sum of squared errors (**SSE**):

$$SSE = \sum_{i=1}^{N}\sum_{j=1}^{M}(x_{i,j} - \mathbf{c}_i\mathbf{a}_j)^2$$

where $x_{i,j}$ = element in **X**, $\mathbf{c}_i$ = row vector containing the concentrations of sample i, $\mathbf{a}_j$ = column vector containing the absorbencies of the wavelength j, and M = number of wavelengths in the spectra. Note that the reference results have not been used in the calculation of the SSE.

Experiments

Sample sets. The samples obtained for this work were from New Zealand and were divided into calibration samples or test samples.

Calibration samples. This set contained 124 samples. They were collected from nine process runs (five based on skim milk, four based on whole milk) carried out in May 1997 using an experimental set-up in the laboratory. Lactozym 3000 (Novo-Nordisk, Bagsværd, Denmark) was the enzyme used. Samples were taken from the reaction mixture at various time points over a 3-h period, and they were immediately heated to 80°C in a 750-W microwave oven to deactivate the enzyme. Duplicate samples were taken, and the subsequent reference analyses and spectral measurements were carried out independently. Thus, the set of 124 samples composed two very similar sets of 62 samples.

Test samples. This set contained 23 samples obtained from two process runs carried out in the laboratory in November and December 1997 using the same experimental set-up. Samples were taken at various intervals, and only in the subsample used for reference analysis was the enzyme deactivated. The spectral measurement was carried out on the non-deactivated sample immediately (i.e., less than 1 min) after sampling to make the FTIR measurements as close to an online application as possible.

Reference measurements. Lactose was determined on 90 of the calibration samples and on 23 test samples using the following HPLC set-up.

Equipment. Lactose was analyzed by HPLC using a Waters Maxima 820 Workstation (Millipore Corp., Milford, MA). One hundred microliters of sample extract were injected with a Waters WISP automatic injector (Millipore Corp.) into an Alltima NH_2 column (250 × 4.6 mm, 5-μm particles; Alltech Associates, Auckland, New Zealand) protected by a 10-mm adsorbosphere NH_2 guard column cartridge (Alltech Associates). The column and guard column were kept at 28°C in an electrical column heater (Jones Chromatography Ltd., Hengoed, United Kingdom). To protect the column against deterioration, a silica saturation column (New Zealand Dairy Research Institute, Palmerston North, New Zealand) was placed between the pump (Waters model 510; Millipore Corp.) and the injector. Lactose was eluted by 80% acetonitrile (2 ml/min) and detected by a Shimadzu RID 6A reflective index detector (Shimadzu Corp., Kyoto, Japan).

Sample preparation. Liquid milk containing less than 1.0 g of total solids was diluted to 20 ml with water. The proteins were precipitated with barium hydroxide and zinc sulfate. The volume was increased to 40 ml before centrifugation for the unhydrolyzed lactose samples and to 30 ml for the hydrolyzed lactose samples. After centrifugation at 2000 rpm for 10 min, 2 ml of the clear supernatant were diluted to 10 ml with acetonitrile and filtered through a 0.45-μm syringe filter (Whatman, Singapore, Singapore).

Spectral measurements. The FTIR spectral measurements were carried out using a MilkoScan FT 120 (Foss Electric A/S, Hillerød, Denmark). The infrared spectrum from 925 to 5000 cm^{-1} was recorded. The calibration samples were measured in duplicate, and the test samples were measured in triplicate.

In the data analysis, only the ranges 964 to 1542, 1724 to 1847, and 2699 to 2965 cm^{-1} were used, because these areas contain useful chemical information.

All measurements were determined against a water background and were log-transformed to give

TABLE 1. Partial least squares results from the cross-validated calibrations for the determination of lactose in the calibration samples.[1]

	2 CVS[2]	4 CVS	6 CVS	8 CVS	10 CVS
Number of factors[3]	5	5	5	5	4
R^2	0.997	0.996	0.996	0.996	0.996
RMSEP[4]	0.82	0.90	0.88	0.93	0.88
s_r[5]	0.22	0.23	0.23	0.23	0.24

[1]Forty-four samples ranging from 0 to 45% dry base lactose were used.

[2]Cross-validation segments.

[3]Optimal number of partial least squares factors.

[4]Root mean square error of prediction.

[5]Repeatability.

absorbance spectra. A typical time resolved landscape for the first hour of an experiment is shown in Figure 1. No spectral preprocessing was performed prior to data analysis because only minor improvements are obtained when using full spectrum data of the present type (7).

RESULTS AND DISCUSSION

PLS Results

For comparative purposes, a PLS calibration against the lactose reference results was performed.

Because the calibration set comprised two dependent sample sets, only the first set (62 samples of which 44 had been reference analyzed) was used for finding the optimal model (regarding the number of PLS factors). To this end, a PLS calibration was cross-validated using the calibration samples. The procedure was repeated using 2, 4, 6, 8, and 10 cross-validation segments to check the stability of the result. The results are shown in Table 1, and a reference versus a measured plot for lactose is shown in Figure 2. A model using five PLS factors was the most accurate.

All samples analyzed using the reference method in the calibration set (a total of 90 samples) were used to build the final model using five PLS factors. This model was used for predicting lactose in the 23 test samples. The result is shown in Figure 3. The result was not as accurate as that indicated by cross-validation, but most of the error was due to a substantial bias. Thus, the calibration set was not representative of the test set. The reason could be

1. The time difference (more than 6 mo) between the measurement of calibration samples and test samples.
2. A difference in chemical composition of the test samples caused by seasonal variations in the milk. Seasonal variations are of great importance in New Zealand, where most dairy cattle calve simultaneously, as farming is pasture based.
3. From the spectral data (not shown), it is evident that the protein content of the calibration samples is virtually the same for all. Therefore, the calibration cannot account for protein variations in the test samples.

Despite this lack of reproducibility, the reference results and lactose predictions correlate well, which is the main issue in this context.

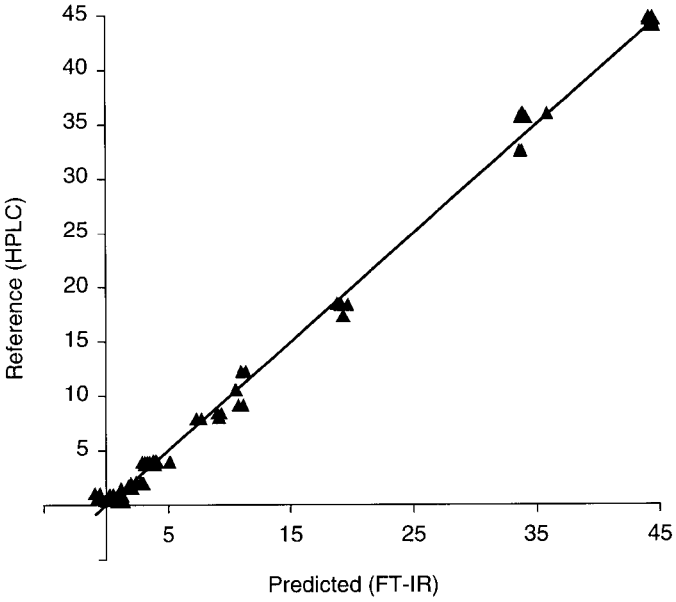

Figure 2. Reference versus predicted plot for lactose showing 44 calibration samples (two replicates for each sample) predicted using cross-validation (with six cross-validation segments) against the reference results. The model uses five partial least squares factors with $R^2 = 0.996$, root mean square error of prediction = 0.88, and repeatability = 0.23. The reference results range from 0 to 45% dry base lactose. FT-IR = Fourier transform infrared.

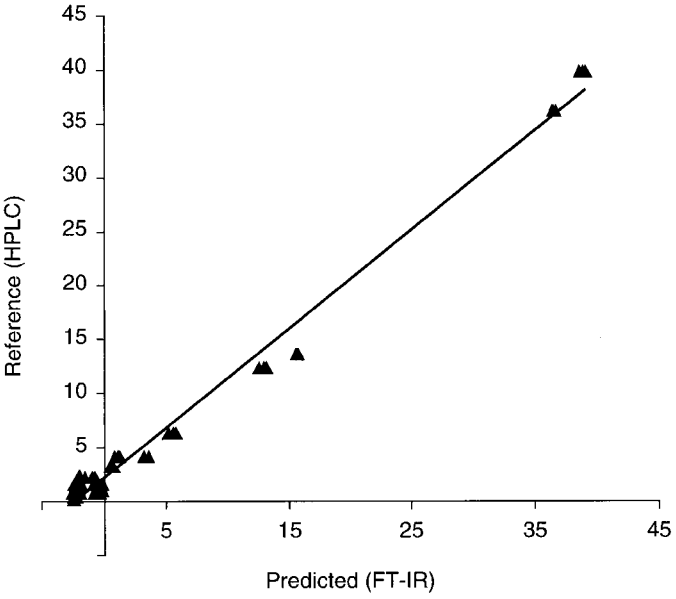

Figure 3. Reference versus predicted plot for lactose showing the 23 test samples (three replicates for each sample) predicted using a partial least squares model with five factors. R^2 = 0.987, root mean square error of prediction = 2.49, standard error of prediction = 1.55, bias = –1.96, and repeatability = 0.23. The reference results range from 0 to 40% dry base lactose. FT-IR = Fourier transform infrared.

ALS Results

The PLS results suggest that five independent factors are necessary for predicting the lactose content of the samples. Thus, a similar number of components in the ALS is expected.

To obtain reasonable solutions, two constraints were applied during the regression: 1) non-negativity of the concentration profiles, and 2) unimodality of the concentration profiles (except for fat) in each process run.

Although the lactose concentrations of most of the samples were already known, they were not used as start guesses, because the purpose of this study was to see whether it was possible to obtain good results having only a limited knowledge of the process. Therefore, the following—very simple—start guesses were used: 1) the skim and whole milk samples had a start guess for fat of 0 or 1, respectively; 2) because lactose is known to be the only sugar present at the beginning of the process, the lactose concentration of the first sample of each run was set to 1, and all others were set to 0; and 3) the remaining components were given random numbers as start guesses (uniformly distributed numbers between 0 and 1).

Because random numbers were used, many different solutions were possible. For this reason, the ALS run was carried out 100 times with new start guesses for each number of components. From three to six components (corresponding to the number of pure spectra in **A**, Equation [2]) were tried on both (almost identical) calibration sets of 62 samples. When six components were used, the concentration profiles and pure spectra became noisy and highly correlated, so they will not be discussed in the following text.

The best models, in terms of how well the **X** matrix is described (measured as SSE) and how well the lactose profile correlates with the reference results, are shown in Table 2 (A and B parts). The results for three to five components were as follows.

With three components, the same pure spectra and concentration profiles were reached every time for both calibration sets—at the least, the differences were insignificant. The solutions with the lowest SSE are shown in Figures 4 and 5. Note that both pure spectra and concentration profiles have been normalized to make a presentation on the same scale possible. The pure spectra describing fat (having strong absorptions near 2900 cm^{-1} due to C-H stretching vibrations) generally have very negative contributions in the areas where the sugars absorb because high-fat samples normally contain less lactose as a result of the displacement of the water phase by fat. This phenomenon is unfortunate, as the ALS model will predict samples with a high fat content to contain less sugars, but it cannot be avoided with the present data set.

With four components, many different solutions were reached. They belonged to a limited number of groups of solutions inside which the variations were small. Some results are shown in Figures 6 and 7. Of the four components, at least two were sugars (having strong absorptions in the 1000 to 1200 cm^{-1} range), one was fat (strong absorptions between 2800 and 3000 cm^{-1}), and one was difficult to assign. Of the sugars, the component decreasing rapidly through each batch is lactose.

With five components, the problem of finding the optimal solution becomes more difficult. But, as is evident from Table 2, good lactose correlations were still obtained. No major improvement was observed when using five components instead of four. The three- and four-component solutions were thus chosen for further examination.

An ALS with three components gives the most stable result and the most reasonably pure spectra, but the correlation to the actual lactose concentration is relatively poor. In Figure 8, the lactose concentration profiles are plotted against the reference results,

TABLE 2. Alternating least squares results of the two calibration sets (parts A and B) and the independent test set (parts C and D) using three, four, or five components.[1]

		3 Components		4 Components		5 Components	
Set[2]		Set 1	Set 2	Set 1	Set 2	Set 1	Set 2
A	SSE[3]	0.56	1.01	0.38	0.44	0.26	0.47
	R^2	0.928	0.942	0.942	0.971	0.978	0.977
	SEC[4]	3.47	3.15	3.12	2.23	1.92	2.00
B	SSE	0.71	1.06	0.44	51.63	0.40	10.50
	R^2	0.928	0.942	0.971	0.986	0.991	0.992
	SEC	3.47	3.15	2.22	1.54	1.26	1.16
C	R^2	0.894	0.891	0.959	0.980	0.959	0.940
	SEC	3.46	3.50	2.16	1.51	2.14	2.61
	s_r[5]	0.15	0.15	0.16	0.27	0.16	0.13
D	R^2	0.894	0.891	0.959	0.953	0.978	0.973
	SEC	3.46	3.50	2.15	2.29	1.59	1.78
	s_r	0.15	0.15	0.16	0.14	0.19	0.19

[1]Reference results range from 0 to 45% dry base lactose.

[2]A = Results obtained from the calibration sets (62 samples each). The models were selected on the basis of the lowest sum of squared errors. B = Results obtained from the calibration sets (62 samples each). The models were selected on the basis of the highest correlation to the lactose reference results. C = Results obtained from the test set (23 samples) using the models (selected by use of the sum of squared errors) calculated from the individual calibration sets. D = Results obtained from the test set (23 samples) using the models (selected by use of the correlation) calculated from the individual calibration sets.

[3]Sum of squared errors.

[4]Standard error of calibration.

[5]Repeatability.

and the resulting plot is highly nonlinear. The four-component model (Figure 9) gives a more linear relationship to the lactose reference results. It cannot be due to overfitting (i.e., a too optimistic estimate of the error) as the reference results were not involved in the optimization. In addition, the concentration profiles of the four-component model agree with previous observations (5, 15, 16) that not only the monosaccharides (galactose and glucose) but also various other sugars (containing two or more monosaccharide units), generally known as the oligosaccharides, are formed during the process. The shapes of the concentration profiles are very similar to these previous observations. The way in which

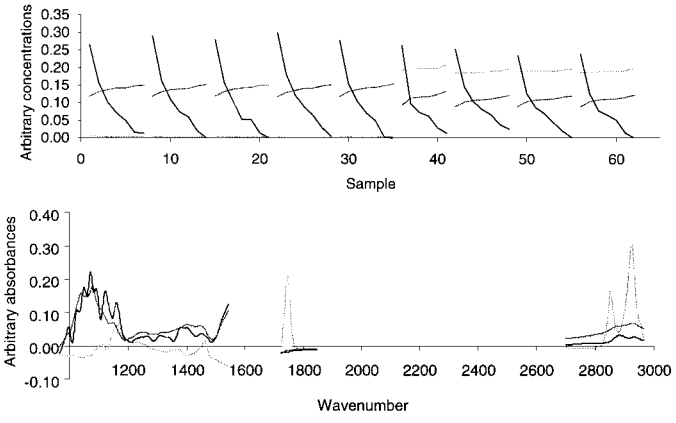

Figure 4. The three-component alternating least squares solution with the lowest sum of squared errors (out of 100 runs) for set 1. Upper part shows the concentration profiles, lower part is the pure spectra.

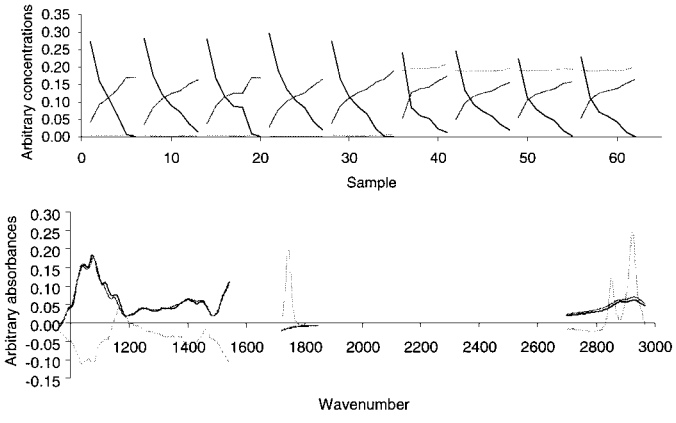

Figure 5. The three-component alternating least squares solution with the lowest sum of squared errors (out of 100 runs) for set 2. Upper part shows the concentration profiles, lower part is the pure spectra.

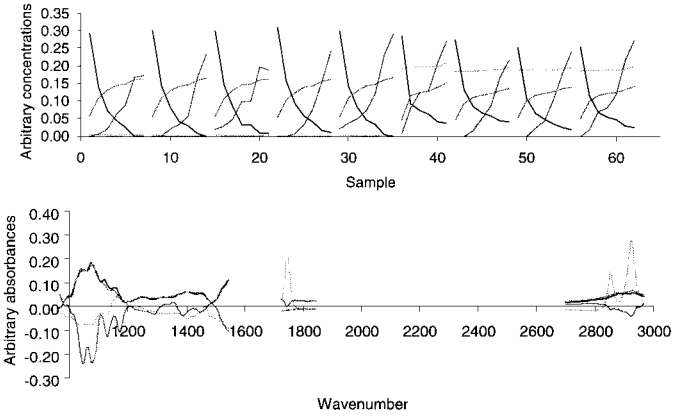

Figure 6. The four-component alternating least squares solution with the lowest sum of squared errors (out of 100 runs) for set 1. Upper part shows the concentration profiles, lower part is the pure spectra.

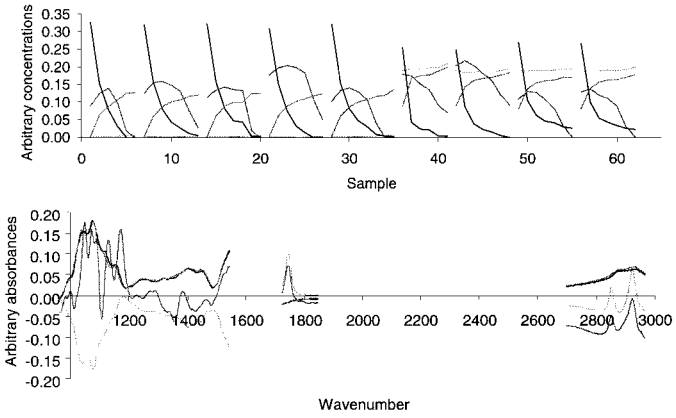

Figure 7. The four-component alternating least squares solution with the lowest sum of squared errors (out of 100 runs) for set 2. Upper part shows the concentration profiles, lower part is the pure spectra.

these sugars are distributed among the last two components (Figures 6 and 7) can vary between separate ALS runs, which is why many different solutions are observed when four (and five) components were tried.

Whether the SSE or the correlation should be used as the selection criterion is not clear. When three components were used, both criteria gave almost the same result, whereas the SSE criterion gave a somewhat higher prediction error for four or five components. In both cases, a significant improvement over the three-component result was obtained.

The final test of the hypothesis used the ALS models on the test set obtained from two new process runs. The results are shown in Table 2 (parts C and D). A large improvement in the standard error of calibration when increasing the number of components from three to four was observed. It is almost independent of the method (correlation or SSE) used for selecting the optimal model. These results led us to the conclusion that nothing is gained by selecting the optimal model by use of the lactose reference results (i.e., by looking at the correlation). The best result is obtained by using the SSE. This is very promising, since reference analyses are not needed when process data of the present type are analyzed. The ALS alone can be used for generating a model that can be used for future process monitoring. Note that the ALS concentrations are in arbitrary units, so only relative process changes can be detected.

The results in Table 2 should be compared with the PLS result shown in Figure 3. The ALS predictions of lactose (using four components) were almost as good as the PLS predictions—in some cases the results were the same.

The normalized predictions of three of the four components (using the model with lowest SSE based on calibration set 2) for each of the two processes included in the test set are shown in Figure 10 with the corresponding results from the PLS model and the reference method. Fat was omitted because it was constant during each run. The ALS lactose predic-

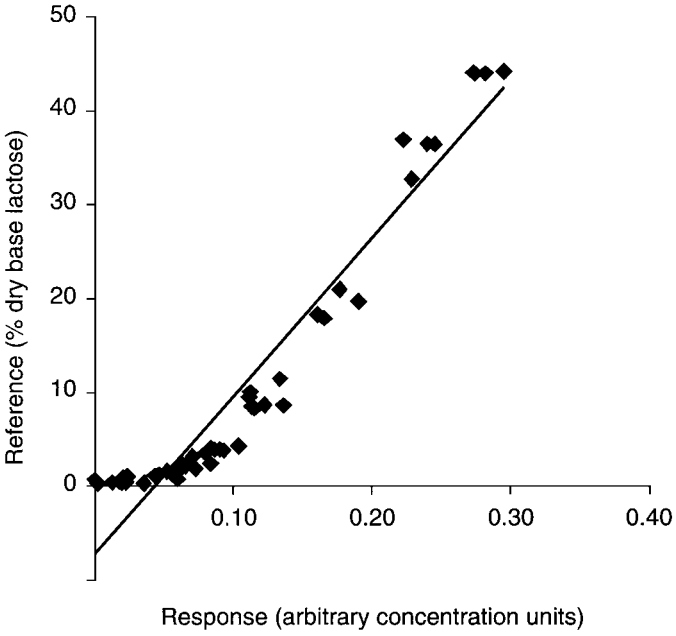

Figure 8. Lactose profile and reference lactose results plotted against each other. The profile originates from an alternating least squares regression on calibration set 2 with three components. The relationship is strongly nonlinear. $R^2 = 0.942$, and standard error of calibration = 3.15.

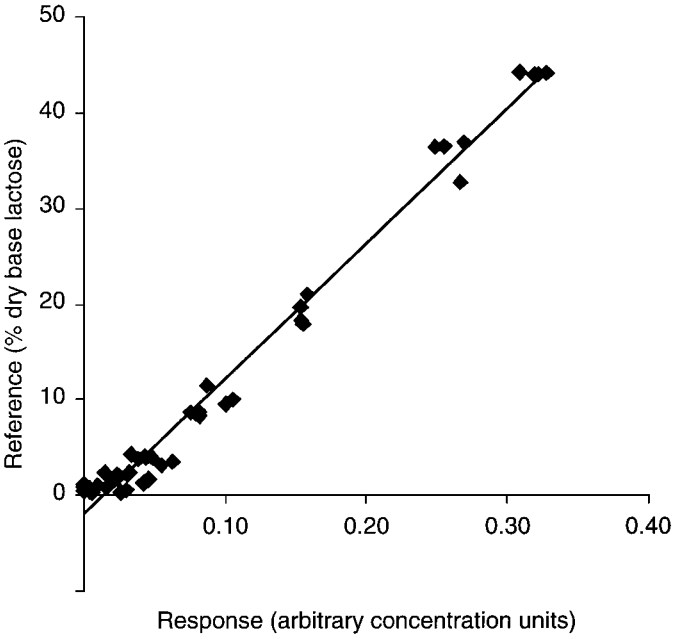

Figure 9. Lactose profile and reference lactose results plotted against each other. The profile originates from an alternating least squares regression on calibration set 2 with four components. The linearity was more accurate than that of only three components (Figure 8). R^2 = 0.971, and standard error of calibration = 2.23.

tions do not agree perfectly, neither with the reference nor the PLS results based on the same spectra, but they follow roughly the same curve. The main reason for the disagreement between the ALS and PLS results was the negative PLS lactose predictions, which were due to the earlier discussed bias of the PLS calibration. Both PLS and ALS concentration profiles follow a smooth curve, which should be expected when dealing with a chemical reaction. Thus, the fluctuations in the lactose reference results are likely to be caused by the lack of reproducibility of the reference method rather than real variations in the lactose content.

The two test runs (Figure 10) gave the same shapes of the concentration profiles of the third and fourth components as those observed during calibration. Therefore, component 3 was assigned to the sum of galactose and glucose, and component 4 was likely to be caused by oligosaccharides formed during the reaction. Another data set with reliable reference results from sugars other than lactose is required to confirm this result.

The remaining problem of allowing implementation of ALS for practical use in dairy process monitoring is selection of the optimal number of components in the ALS model, which corresponds to the problem of selecting factors in PLS, but for ALS, no prediction error (e.g., RMSEP) can be minimized. In the present case, the obvious choice would have been three components, which gave the most stable result. Only the comparison of the profiles to actual lactose results indicated that four components were optimal. Methods for determining the number of independently varying species present in the samples are, therefore, required.

Scores obtained through principal component analysis (12) could solve the problem. After a principal component analysis of all calibration samples, the scores (not shown) indicated structures originating from the batch structure of the data and revealed up to four or five components. Thus, four or five components would be expected to be optimum in ALS, which supports the actual findings shown above.

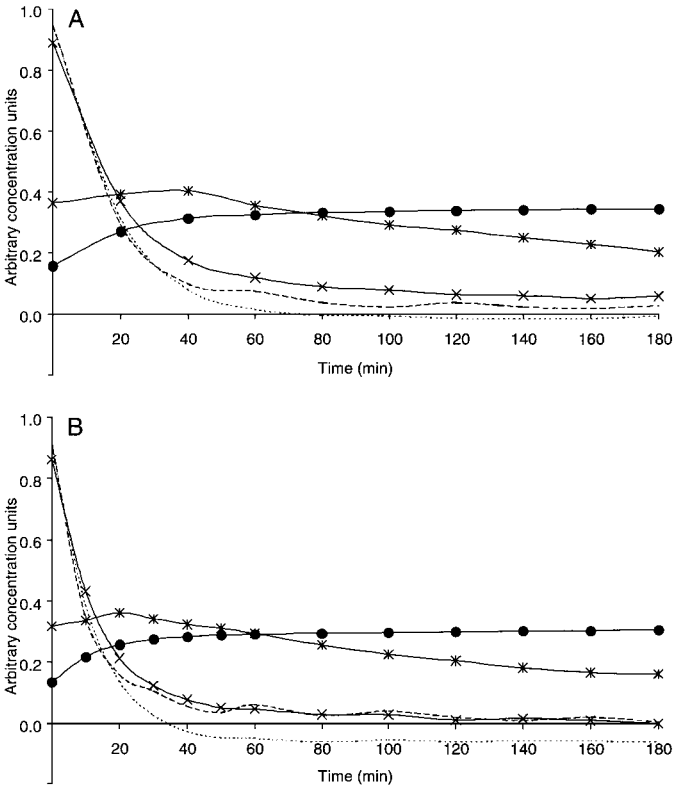

Figure 10. Concentration profiles from the reference results (lactose only), partial least squares results (lactose only), and the four components obtained from alternating least squares for the first (A) and second (B) experiments present in the test set. All profiles are normalized to unity to make them comparable. Only the three sugar components are shown. Component 1 (fat) was constant throughout the process run. Lactose predicted using alternating least squares (—X—), lactose predicted using partial least squares (· · ·), lactose reference results (– – –), component 3 predicted using alternating least squares (–●–), and component 4 predicted using alternating least squares (–*–).

In this experiment, we were successful in extracting concentrations from FTIR data from a lactose hydrolysis process and in monitoring new process runs with this knowledge. The present data set must be considered to be worst case because the infrared spectra of reactants, intermediates, and products were very similar (i.e., they are all sugars, which gave almost the same absorption peaks). Resolution of spectra from processes in which the compounds are less similar should therefore be easier; the problem of determining the number of components especially should be less difficult.

The method described here has been patented (8).

CONCLUSIONS

The present study has shown that ALS is a promising method for use in dairy process optimization. Without the need for reference analyses, it was possible to extract four components from the lactose hydrolysis process data (fat, lactose, and two other sugar components) and to obtain a lactose prediction error similar to the one obtained from an ordinary PLS regression. Such use of ALS for a reference-independent prediction of process parameters is not limited to dairy products only but is likely to be useful for process monitoring and identification of intermediates in all branches of the food and beverage industry.

By use of ALS combined with FTIR, it becomes possible to obtain quick information on compounds present during the process including intermediates produced and consumed in the course of the reaction. A further advantage (in many cases the most important) is that the pure spectra obtained by ALS makes it possible to generate predictions of process parameters without the need for expensive and time consuming reference procedures.

ACKNOWLEDGMENTS

The authors thank Foss Electric A/S, Hillerød, Denmark, and The Danish Academy of Technical Sciences, Lyngby, Denmark, for providing the funds for the work of P. W. Hansen. E. Conaghan is thanked for his help in relation to the practical measurements. R. Bro and C. Ridder are thanked for useful discussions and help in the implementation of the mathematical routines.

REFERENCES

1 Bro, R. 1997. PARAFAC. Tutorial and applications. Chemometrics Intelligent Lab. Syst. 38:149–171.
2 Bro, R., and C. A. Andersson. 1998. The n-way toolbox for Matlab®. http://newton.foodsci.kvl.dk/Matlab/nwaytoolbox/index.html. Accessed Feb. 4, 1999.
3 Forsman, E., M. Heikonen, L. Kiviniemi, M. Kreula, and P. Linko. 1979. Kinetic investigations of the hydrolysis of milk lactose with soluble *Kluyveromyces lactis* β-galactosidase. Milchwissenschaft 34:618–621.
4 Furusjö, E., L.-G. Danielsson, E. Könberg, M. Rentsch-Jonas, and B. Skagerberg. 1998. Evaluation techniques for two-way data from in situ Fourier transform mid-infrared reaction monitoring in aqueous solution. Anal. Chem. 70:1726–1734.
5 Greenberg, N. A., and R. R. Mahoney. 1983. Formation of oligosaccharides by β-galactosidase from *Streptococcus thermophilus*. Food Chem. 10:195–204.
6 Guy, E., and E. Bingham. 1978. Properties of β-galactosidase of *Saccharomyces lactis* in milk and milk products. J. Dairy Sci. 61:147–151.
7 Hansen, P. W. 1998. Urea determination in milk using Fourier transform infrared spectroscopy and multivariate calibration. Milchwissenschaft 53:251–255.
8 Hansen, P. W., inventor. 1998. Evaluation of spectroscopic data. Foss Electric A/S, assignee. Danish Pat. Appl. No. PA 1998 01177.
9 Hewavitharana, A. K., and B. van Brakel. 1997. Fourier transform infrared spectrometric method for the rapid determination of casein in raw milk. Analyst 122:701–704.
10 Karjalainen, E. J. 1989. The spectrum reconstitution problem. Use of alternating regression for unexpected spectral components in two-dimensional spectroscopies. Chemometrics Intelligent Lab. Syst. 7:31–38.
11 Kjær, L. 1997. Say cheese—and think of direct casein determination. Scand. Dairy Inform. Feb.:28–30.
12 Martens, H., and T. Næs. 1989. Multivariate calibration. Wiley, Chichester, England.
13 Matlab®, Version 5.2.1. 1998. The MathWorks, Inc., Natick, MA.
14 PLS_Toolbox, Version 1.5.1. 1995. Eigenvector Technologies, Manson, WA.
15 Prenosil, J. E., E. Stuker, and J. R. Bourne. 1987. Formation of oligosaccharides during enzymatic lactose hydrolysis and their importance in a whey hydrolysis process. Part I. State of art. Biotechnol. Bioeng. 30:1019–1025.
16 Prenosil, J. E., E. Stuker, and J. R. Bourne. 1987. Formation of oligosaccharides during enzymatic lactose hydrolysis and their importance in a whey hydrolysis process. Part II. Experimental. Biotechnol. Bioeng. 30:1026–1031.
17 Ridder, C., and L. Kjær. 1995. Sweet dreams come true. Applied FTIR technology in ice cream mix analysis. Scand. Dairy Inform. Sept.:34–36.
18 Tauler, R., B. Kowalski, and S. Fleming. 1993. Multivariate curve resolution applied to spectral data from multiple runs of an industrial process. Anal. Chem. 65:2040–2047.
19 Tauler, R., A. Smilde, and B. Kowalski. 1995. Selectivity, local rank, three-way data analysis and ambiguity in multivariate curve resolution. J. Chemometrics 9:31–58.