Organic Process Research & Development **2004**, *8*, 495−503

Online Spectroscopy and Multivariate Data Analysis as a Combined Tool for Process Monitoring and Reaction Optimization

Hans-René Bjørsvik*

Department of Chemistry, University of Bergen, Allégaten 41, N-5007 Bergen, Norway

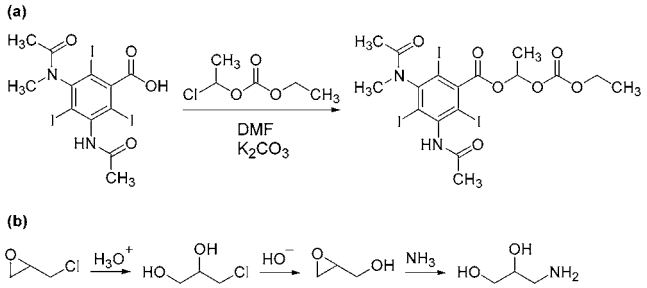
Abstract:

Multivariate modeling and spectroscopy taken together constitute a powerful tool applicable to both process monitoring and to process and reaction optimizing. The present article establishes how near-IR spectroscopy utilizing a fiber-optical transmission probe can be combined with principal component analysis to extract concentration profiles from a set of spectra recorded during the course of a reaction. Moreover, it is demonstrated how to easily determine the end-point (optimized reaction time, that is, the minimized time) of a reaction, and a simple calibration for quantitative analysis is demonstrated.

Introduction

Experimental design[1−3] and multivariate modeling[4] have during recent years become established research tools in organic process research and development. Increased utilization of computerized machines for parallel synthesis, dedicated and automated equipment for reaction workup combined with robotized chromatographic instrumentation for the quantification of reaction mixtures has resulted in extensive automation and speeding up of process development and optimization. Even though online FT-IR spectroscopy is currently utilized to some extent in research and process development,[5−10] a huge unused potential still exists also to combine it with advanced computational techniques for multivariate modeling and analysis. The combined approach constituted by spectroscopy and multivariate mathematical and statistical methods[11] represents an extremely powerful tool that can provide new and highly useful insight into the organic reaction or process under study. Near-infrared (NIR) spectroscopy[12,13] offers a vast unexploited potential for monitoring organic reactions. Although the combined methodology using NIR instrumentation and multivariate mathematical methods gives good predictive power for quantitative analysis,[14] it has not become especially widespread in the research and development of organic processes and synthesis. This is most probably due to (i) the difficulties in interpreting spectra recorded in the near-infrared region, hence difficulties in extracting direct chemical information, and (ii) the apparent complexity of understanding the multivariate methods necessary for the multivariate data analysis.

Scheme 1



However, I have previously demonstrated how the combined technique of NIR spectroscopy and multivariate methods can be successfully used to monitor synthetic organic processes and to determine the shortest reaction time ("the end point") for the synthesis of 1-(ethoxycarbonyloxy)-ethyl 5-acetylamido-3-(*N*-methylacetamido)-2,4,6-triiodobenzoate[15] (Scheme 1a) and in the synthesis of 3-aminopropan-1,2-diol[16] (Scheme 1b). The later work demonstrated, moreover, how the results from the multivariate calculations based on only one quantitative analysis might be used as an estimate for the production of the time−concentration profile for the reaction.

Figure 1 shows schematically a fiber-optical near-IR instrument connected to a transmission cell probe for "directly-in-reactor" spectral recording. Such a setup was used for the spectral recording in the present report as well

* To whom correspondence should be addressed. E-mail: Hans.Bjorsvik@kj.uib.no.

(1) Box G. E. P.; Hunter, J. S. *Technometrics* **1961**, *3*, 311−351.
(2) Box G. E. P.; Hunter, J. S. *Technometrics* **1961**, *3*, 449−458.
(3) Box, G. E. P.; Hunter, W. G.; Hunter, J. S. *Statistics for Experimenters, An Introduction to Design, Data Analysis, and Model Building*; Wiley: New York, 1978; pp 1−653.
(4) Draper, N. R.; Smith, H. *Applied Regression Analysis, 3rd ed.*; Wiley: New York, 1998; pp 1−709.
(5) Tschaen, R.; Valante, G.; Smith, G.; Riseman, S.; Shinkai, I. *J. Org. Chem.* **1989**, *54*, 3793−3797.
(6) Dozeman, G. J.; Fiore, P. J.; Puls, T. P.; Walker, J. C. *Org. Process Res. Dev.* **1997**, *1*, 137−148.
(7) Hoyt, S. B.; Overman, L. E. *Org. Lett* **2000**, *2*, 3241−3244.
(8) Pasquale, L.; Modena, G.; Contarca, L.; Delogu, P.; Mantovania, S. *J. Org. Chem.* **2000**, *65*, 8224−8228.
(9) Sun, X.; Collum, D. B. *J. Am. Chem. Soc* **2000**, *122*, 2452−2458.
(10) Kaba, M. S.; Barteau, M. A.; Lee, W. Y.; Song, I. K. *Appl. Catal., A* **2000**, *194−195*, 129−136.
(11) Bjørsvik, H.-R.; Bye, E. *Appl. Spectrosc.* **1991**, *45*, 771−778.
(12) *Handbook of Near-Infrared Analysis*, 2nd ed.; Burns, D. A., Ciurczak, E. W., Eds.; Marcel Dekker: New York, 2001; Vol. 27.
(13) *Near-Infrared Spectroscopy. Principles, Instrumentation, Applications*; Siesler, H. W., Ozaki, Y., Kawata, S., Heise, H. M., Eds.; Wiley-VCH: Weinheim, 2002.
(14) Bjørsvik, H.-R.; Martens, H. Data Analysis. Calibration of NIR Instruments by PLS Regression. In *Handbook of Near-Infrared Analysis*, 2nd ed.; Burns, D. A., Ciurczak, E. W., Eds.; Marcel Dekker: New York, 2001; Vol. 27, Chapter 8, pp 185−207.
(15) Bjørsvik, H.-R. *Acta Chem. Scand.* **1994**, *48*, 445−452.
(16) Bjørsvik, H.-R. *Appl. Spectrosc.* **1996**, *50*, 1541−1544.

10.1021/op049971f CCC: $27.50
Published on Web 04/23/2004

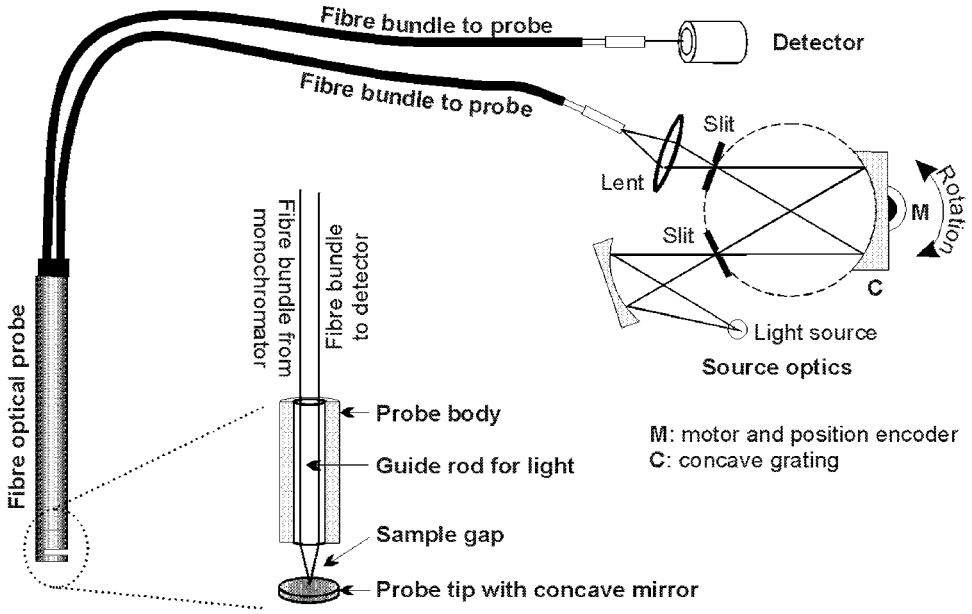
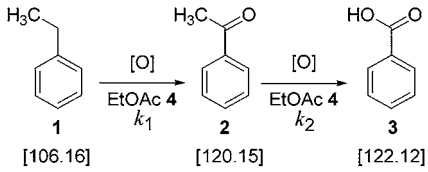


Figure 1. **General outline for a fiber-optical NIR spectrophotometer with a transmission probe (cell).**

Scheme 2



as during the study of the processes shown in Scheme 1 and previously reported by the author.[15,16]

This article will present and discuss a method using fiber-optical near-infrared spectroscopy as the analytical technique, multivariate data analysis using principal component analysis as a monitoring and analyzing tool for organic reactions, and some new ideas concerning how results from multivariate data analysis can be used for elucidation of mechanistic features of synthetic reactions and processes.

Methods and Results

Many new NIR instruments designed for application in the chemical process industry have become commercially available in recent years. Most of these instruments are built for remote measurement using fiber optics in combination with an assorted collection of transmission measurement probes or cells, see Figure 1. Such an instrumental setup is an excellent tool for nearly real-time monitoring of any organic process, and permits remote measurements in, e.g. hazardous environments or inside high-pressure equipment.

The combined method of multivariate modeling, using principal component analysis, and online spectroscopy will be illustrated by means of an imagined oxidation process. This process is constituted by the partial oxidation of ethylbenzene **1** to an intermediate, acetophenone **2**, which in a subsequent and last step is further oxidized to the terminal product, benzoic acid **3**. The oxidation reaction is conducted with ethyl acetate **4** as reaction medium. The imagined oxidation process is outlined in Scheme 2.

Let the imagined oxidation process be a consecutive first-order reaction, where the rate constants are $k_1 = 50.0 \times 10^{-3}$ and $k_2 = 40.0 \times 10^{-2}$. The numerical values of k_1 and k_2 for the imagined process were selected for experimental convenience, as will be demonstrated later. The kinetic rate model that describes a consecutive first-order reaction is given by the eqs 1−3.

$$-\frac{\mathrm{d}[\mathbf{1}]}{\mathrm{d}t} = k_1[\mathbf{1}] \quad (1)$$

$$-\frac{\mathrm{d}[\mathbf{2}]}{\mathrm{d}t} = k_2[\mathbf{2}] - k_1[\mathbf{1}] \quad (2)$$

$$\frac{\mathrm{d}[\mathbf{3}]}{\mathrm{d}t} = k_2[\mathbf{3}] \quad (3)$$

The integrated form of eq 1 yields:

$$[\mathbf{1}]_\tau = [\mathbf{1}]_\mathrm{o}\, \mathrm{e}^{-k_1\tau} \quad (4)$$

The initial (time $\tau = 0$) concentration of ethylbenzene **1** is $[\mathbf{1}]_\mathrm{o}$. At any reaction time $\tau = \theta$ is the concentration of the substrate $[\mathbf{1}]_\theta$. Substitution of eq 4 in eq 2 and then integrating yields the rate equation for the intermediate acetophenone **2**, thus:

$$-\frac{\mathrm{d}[\mathbf{2}]}{\mathrm{d}t} = k_2[\mathbf{2}] - k_1[\mathbf{1}]_\mathrm{o} e^{-k_1\tau} \Rightarrow$$

$$[\mathbf{2}]_\tau = \frac{k_1}{(k_2 - k_1)}[\mathbf{1}]_\mathrm{o}(\mathrm{e}^{-k_1\tau} - \mathrm{e}^{-k_2\tau}) \quad (5)$$

For the final product, benzoic acid **3**, the expression for the variation of concentration over the reaction time $[\mathbf{3}]_\tau$ can easily be found, since $[\mathbf{1}]_\tau + [\mathbf{2}]_\tau + [\mathbf{3}]_\tau = [\mathbf{1}]_\mathrm{o}$ at any reaction time τ during the course of the reaction. Hence, for the imagined oxidation process, eq 6 gives the concentration−time profile for the final product benzoic acid **3**.

$$[\mathbf{3}]_\tau = [\mathbf{1}]_\mathrm{o} - [\mathbf{1}]_\tau - [\mathbf{2}]_\tau \quad (6)$$

Table 1. **Calculated and actual masses of the three analytes, ethylbenzene 1, acetophenone 2, and benzoic acid 3**

	calculated model values [g]			measured values [g]			score values calculated by means of PCA[c]		
t	$\mathbf{1}_{(c)}$	$\mathbf{2}_{(c)}$	$\mathbf{3}_{(c)}$	$\mathbf{1}_{(m)}$	$\mathbf{2}_{(m)}$	$\mathbf{3}_{(m)}$	t_1 99.8716%	t_2 0.1276%	t_3 0.0006%
0[a]	10.6150	0.0000	0.0000	10.6136	0.0000	0.0000	−6.1806	−0.4338	−0.0317
1[a]	10.0973	0.4822	0.1055	10.1233	0.4842	0.1052	−6.2099	−0.4138	−0.0361
2[a]	9.6048	0.7818	0.3675	9.6081	0.7826	0.3678	−6.2181	−0.3965	−0.0160
3[a]	9.1364	0.9604	0.7249	9.1718	0.9610	0.7254	−6.2538	−0.3682	−0.0112
4[a]	8.6908	1.0588	1.1375	8.6921	1.0545	1.1375	−6.2869	−0.3384	−0.0032
5[b]	8.2670	1.1045	1.5787	8.2771	1.1048	1.5791	–	–	–
6[a]	7.8638	1.1159	2.0310	7.8938	1.1146	2.0319	−6.3961	−0.2645	0.0069
7	7.4803	1.1052	2.4831	7.4882	1.1051	2.4845	−6.4145	−0.2409	0.0143
8	7.1154	1.0806	2.9277	7.1155	1.0827	2.9279	−6.4570	−0.2094	0.0187
9	6.7684	1.0475	3.3606	6.7758	1.0460	3.3600	−6.5018	−0.1785	0.0222
10[b]	6.4383	1.0096	3.7789	6.4383	1.0105	3.7783	–	–	–
11[b]	6.1243	0.9692	4.1812	6.1540	0.9691	4.1810	–	–	–
12[a]	5.8256	0.9279	4.5668	5.8317	0.9294	4.5675	−6.6240	−0.0877	0.0247
13	5.5415	0.8866	4.9357	5.5478	0.8862	4.9360	−6.6517	−0.0637	0.0253
14	5.2713	0.8460	5.2878	5.2777	0.8507	5.2879	−6.6989	−0.0283	0.0088
15[b]	5.0142	0.8065	5.6237	5.0166	0.8075	5.6232	–	–	–
16	4.7696	0.7684	5.9438	4.8175	0.7663	5.9434	−6.7687	0.0174	0.0152
17	4.5370	0.7317	6.2487	4.5363	0.7310	6.2491	−6.8011	0.0384	0.0125
18[a]	4.3157	0.6966	6.5390	4.3222	0.6978	6.5392	−6.8233	0.0600	0.0145
19	4.1053	0.6630	6.8153	4.1160	0.6618	6.8147	−6.8642	0.0891	−0.0001
20	3.9050	0.6309	7.0782	3.9096	0.6303	7.0776	−6.8744	0.0990	0.0091
21	3.7146	0.6003	7.3285	3.7135	0.5995	7.3280	−6.9180	0.1254	−0.0020
22	3.5334	0.5711	7.5665	3.5481	0.5706	7.5660	−6.9378	0.1418	0.0019
23	3.3611	0.5433	7.7930	3.3610	0.5494	7.7932	−6.9724	0.1610	−0.0044
24[a]	3.1972	0.5169	8.0085	3.1969	0.5185	8.0085	−6.9771	0.1710	0.0007
25	3.0412	0.4917	8.2135	3.0442	0.4935	8.2134	−6.9907	0.1842	−0.0002
26	2.8929	0.4677	8.4084	2.9112	0.4700	8.4091	−7.0310	0.2068	−0.0107
27	2.7518	0.4449	8.5939	2.7624	0.4440	8.5946	−7.0297	0.2123	0.0004
28	2.6176	0.4232	8.7704	2.6428	0.4238	8.7703	−7.0493	0.2275	−0.0001
29	2.4900	0.4026	8.9382	2.4928	0.4007	8.9388	−7.0868	0.2501	−0.0130
30[a]	2.3685	0.3830	9.0979	2.3916	0.3850	9.0975	−7.0553	0.2301	−0.0115
60[a]	0.5285	0.0855	11.5171	0.0000	0.0000	12.2100	−7.4036	0.4877	−0.0348

[a] Values used for the curve fitting in Figures 5, 6, and 7, indicated by black-filled circles. [b] During the introductory PCA, the object was determined to be an outlier. The object was thus removed from the data matrix before the final PCA was carried out. [c] The score values were estimated in a PC model constituted by three principal components: $a = 1$ (99.8716%), $a = 2$ (0.1276%), and $a = 3$ (0.0006%) explain 99.9997% of the total variance.

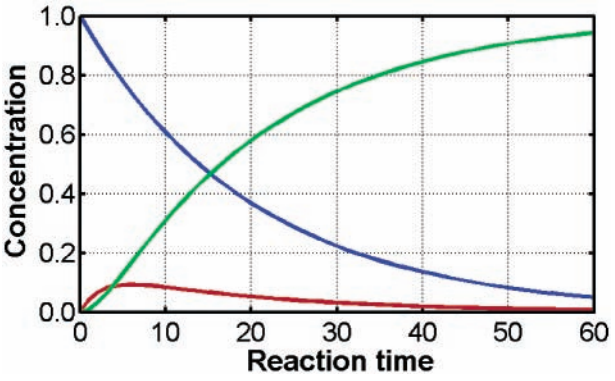


Figure 2. **Reaction profile. The blue curve shows how the concentration of the substrate ethylbenzene 1 changes over time. The red curve shows how the concentration of the intermediate product acetophenone 2 increases, and then decreases during the course of the reaction. The green curve shows how the concentration of the final product benzoic acid 3 increases over time.**

By using eqs 4−6, the numerical values selected for k_1 and k_2, and an initial concentration of the substrate **1** of $[\mathbf{1}]_o = 1$ M one obtains the reaction profile outlined in Figure 2. The blue curve shows how the concentration of the substrate ethylbenzene **1** decreases over the course of the reaction, while the red curve shows how the concentration of the intermediate acetophenone **2** first increases and then decreases over the course of the reaction. The green curve shows the concentration profile of the final product, benzoic acid **3,** over the course of the reaction.

Equations 4−6 were used to calculate the mass weight for each of the substances **1**, **2**, and **3** for a reaction volume of 100 mL when the start concentration of ethylbenzene **1** is $[\mathbf{1}]_o = 1$ M. The calculated quantities for the reaction times $\tau = 0, 1, 2, ..., 30, 60$, are shown in the three columns designated as $\mathbf{1}_{(c)}$, $\mathbf{2}_{(c)}$, and $\mathbf{3}_{(c)}$ in Table 1. The three columns next to these shows the actual quantities of the compound indicated as $\mathbf{1}_{(m)}$, $\mathbf{2}_{(m)}$, and $\mathbf{3}_{(m)}$ as measured on an analytical laboratory balance. These mixtures were diluted with ethyl acetate until the 100 mL mark of a measuring flask. The NIR-spectra for all 32 solutions were recorded on a NIR instrument equipped with a fibre optical transmission measurement probe, see an outline of the instrumentation in Figure 1. The $I = 32$ NIR spectra recorded from the "reaction mixtures" listed in Table 1 were organized in such a way that the first NIR spectrum, that corresponding to time $\tau = 0$ was placed in the first row of the data matrix $\boldsymbol{X}_{raw}$. The next row was constituted by the NIR spectrum corresponding to time $\tau = 1$, the third row by the spectrum corresponding to time $\tau = 2$, and so on until time $\tau = 30$. In the very last

row of the data table, $\boldsymbol{X}_{raw}$, was placed the NIR spectrum of a sample constituted only by the final product, benzoic acid **3**, dissolved in the reaction medium ethyl acetate. This final sample corresponds approximately[17] to a reaction time of $\tau = 60$. The final raw spectral data matrix $\boldsymbol{X}_{raw}$ is constituted by 32 × 700 matrix elements (lines × variables) = 22 400 absorption measurements. The upper regions of the spectra contain considerable noise due to light absorption from the optical fibers. Hence, the spectral region λ 2222−2498 nm was removed from the raw NIR spectra data matrix, $\boldsymbol{X}_{raw}$, to leave the final spectra data matrix $\boldsymbol{X}$. The final spectra data matrix $\boldsymbol{X}$ thus contains the absorbances in the range 1100−2220 nm, corresponding to a 32 × 561 = 17 952 element spectra matrix $\boldsymbol{X}$. The spectra matrix $\boldsymbol{X}$ is shown as a 3D plot in Figure 3a, with the wavelength on the *x*-axis, the reaction time on the *y*-axis, and the absorbance along the *z*-axis. From this figure, it is almost impossible to distinguish between the spectra by visual inspection alone. The lower section of Figure 3 shows the NIR spectra of the three compounds **1**, **2**, and **3** as 1 M solutions in ethyl acetate. Moreover the NIR spectrum of the solvent ethyl acetate is included for the purpose of visual comparison.

Multivariate data analysis techniques[18] can successfully extract systematic differences from such a data matrix. In this contribution, the principal component analysis (PCA) method will be used for the purpose of analyzing the spectral matrix $\boldsymbol{X}$. A brief description of the PCA method will be given below.

Principal Component Analysis (PCA).[19,20] Every data table displays two types of variation. The within-object variation in the variables is displayed horizontally, while the between-object variation in the variables is displayed vertically. These features may be analyzed individually by means of (PCA). The essence of the PCA method is that the systematic variation can be represented by fewer variables, the principal components, than the number of descriptor variables present in the original data table. The number of principal components $a = 1, ..., A$, is usually much lower than the number of original variables, $k = 1, ..., K$, ($A \ll K$). In this article the number of variables (that is wavelengths) is $K = 561$, whereas the number of principal components is expected to be in the range $A = 2-4$.

Figure 4 shows the matrix $\boldsymbol{X}$ that is composed of I rows (samples, often referred to as objects) and K columns (variables, here the wavelengths in the spectrum). This matrix describes a swarm of I points in a K-dimensional space. Figure 4 illustrates an example where the matrix $\boldsymbol{X}$ is constituted by $K = 3$ variables and $I = 17$ objects (samples). The essence of PCA is to fit a hyper-plane to the data of matrix $\boldsymbol{X}$. The hyper-plane can be of low dimension, e.g.

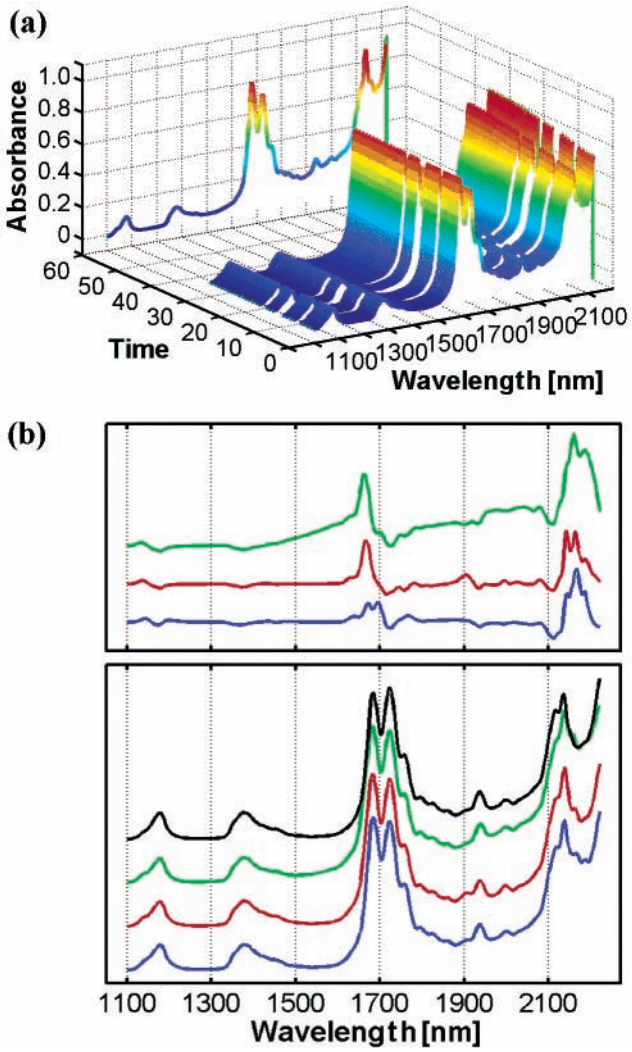


Figure 3. **(a) NIR spectra recorded in the region $\Delta\lambda = 1100-2220$ nm, over the course of the imagined reaction 1 → 2 → 3, $\tau = 0, 1, 2, ..., 30$ and $\tau = 60$. (b) The lower part of the plot shows the NIR spectra of 1 M solutions of ethylbenzene (blue line), acetophenone (red line), benzoic acid (green line), and the spectrum of the pure solvent ethyl acetate (black line), plotted to view the aspect and differences among the pure components. The upper part of plot (b) shows the NIR spectra of ethylbenzene (blue line), acetophenone (red line), and benzoic aid (green line) when the contribution from the ethyl acetate is removed, that is the difference spectra.**

$A = 1$, which implies that the data are described by a line in the multidimensional space spanned by the K-variables in matrix $\boldsymbol{X}$. Figure 4 shows, however, a model with $A = 2$ principal components. A dimensional reduction is thus carried out from the original three variables x_1, x_2, and x_3 to two principal component scores t_1, and t_2. The PC scores t_1 and t_2 are the projections of the points onto the plane, see Figure 4.

Mathematically, PCA involves a factorization of the original data matrix $\boldsymbol{X}$, into means ($\bar{x}_k$), the principal component scores (t_{ia}) which show the between-sample variation, the principal component loadings (p_{ak}), which describe the within-sample variation in the wavelengths, and finally the residuals (e_{ik}), the noise in the data. The mathematical description is shown in eq 7, where the

(17) Actually, at $\tau = 60$, both ethylbenzene **1** as well as a very small quantity of the intermediate acetophenone **2** should remain according to the kinetic models of eqs 4−6: the quantities are estimated to be 0.5285 g (0.0498 mol/L) (**1**) and 0.0855 g (0.0071 mol/L) (**2**), respectively.

(18) For an overview of methods used in chemistry see: Malinowski, E. R. *Factor Analysis in Chemistry*, 3rd ed.; Wiley: New York, 2002.

(19) Jolliffe, I. T. *Principal Component Analysis*; Springer-Verlag: New York, 1986; pp 1−271.

(20) Jackson, J. E. *A User's Guide to Principal Components;* Wiley: New York, 1991; pp 1−569.

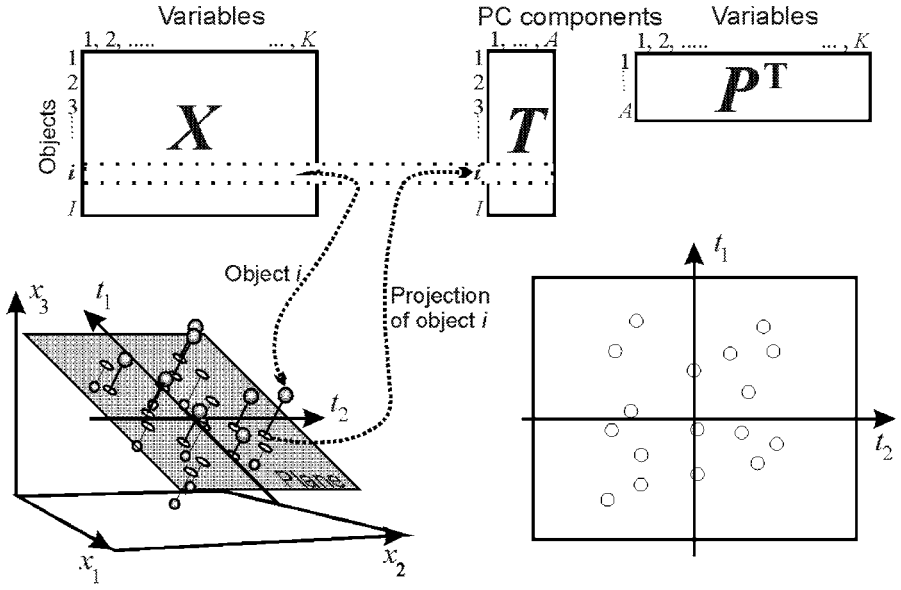


Figure 4. **Schematic explanation of the PCA method with three descriptor variables (x_1, x_2, x_3) and two principal components (t_1, t_2).**

parameter A denotes the number of significant principal components determined according to, for example, the cross-validation method.[21] The absolute value of the loading, p_{ak}, explains how much the variable k contributes to the ath principal component, whereas the sign gives information as to whether the variable is negatively or positively correlated with the principal component.

$$x_{ik} = \bar{x}_k + \sum_{a=1}^{A} t_{ia} p_{ak}^{T} + \epsilon_{ik} \qquad (7)$$

PCA of the NIR Data Table. The final data matrix $\boldsymbol{X}$ (28 rows × 561 variables)[22] containing the NIR spectra (in the range $\lambda = 1100-2220$ nm) was submitted for PCA. The PCA afforded a principal component model constituted by three ($a = 3$) principal components. The variance explained by the first principal component was 99.8716%. Even though the second and the third principal components constituted only 0.1276% (PC #2) and 0.0006% (PC #3) of the total variance, respectively, the additional principal components (PCs #2−3) show both clear systematic variations. An additional principal component (PC #4) showed no such systematic information. The three principal components, PCs #1−3 were plotted against the time scale; the three time−score plots are shown in Figures 5, 6, and 7, respectively. The time−score values for the objects at reaction times $\tau =$ 0, 1, 2, 3, 4, 6, 9, 12, 18, 24, 30, and 60 were used to estimate smooth curves of the time−score (proportional with time−concentration) profiles. Polynomial fit[23] and cubic spline fit[24] can be used for the purpose of estimating smooth curves. The time−score profiles (Figures 5−7) were all estimated using polynomial fitting in the present work. The time−score

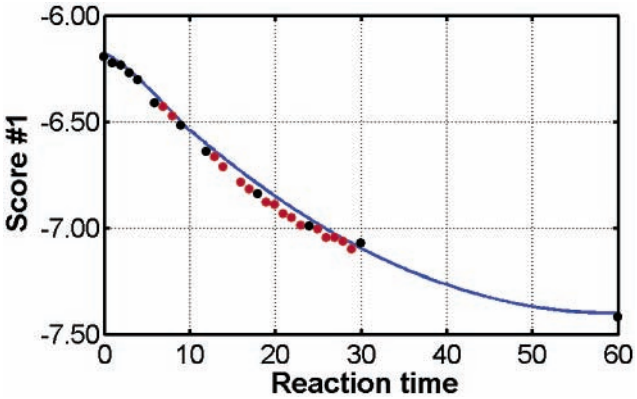


Figure 5. **Polynomial curve fit of the PC #1 score values versus reaction time. For the polynomial curve fitting of the smoothly drawn curve, a selection of time−score values were used, namely the objects at reaction times $\tau = 0$, 1, 2, 3, 4, 6, 9, 12, 18, 24, 30, and 60, and these are indicated as black-filled circles on the smoothed curve. For the time range $\tau = 0-10$ a polynomial fit of fifth degree was used. For the time range $\tau = 10-60$ a polynomial fit of third degree was used. This principal component represents changes in the concentration of ethylbenzene 1 during the course of the reaction.**

profile of principal component #1 was described by a polynomial fit of fifth degree for the interval $\tau = 0-10$, while for the rest of the cuve a third-degree fit was applied. A similar fit was used for the time−score fit for the principal component #2. In the case of the principal component #3, a polynomial fit of fifth degree was used for the interval $\tau =$ 0−10. To achieve a smooth connection between first and second part of the curve a small part of the estimated values of the first interval ($\tau = 0-10$) were used as input values

(21) Wold, S. *Technometrics* **1978**, *20*, 397.

(22) During the introductory principal component analysis, some few abnormal (outlier) NIR spectra (objects) were determined: namely at "reaction time" $\tau = 5$ (object #6), $\tau = 10$ (object #10), $\tau = 11$ (object #11), $\tau = 15$ (object #15). Those spectra were thus removed from the data matrix ($\boldsymbol{X}_{32\times561} \rightarrow \boldsymbol{X}_{28\times561}$) before the final principal component analysis was performed.

(23) (a) Gerald, C. F.; Wheatley, P. O. *Applied Numerical Analysis*, 6th ed.; Addison-Wesley: Reading, MA, 1999; pp 238−248. (b) *Using Matlab*, version 6; The MathWorks Inc.: Natick, MA, 2000; pp 13.17−13.21 and 13.31−13.39.

(24) (a) Gerald, C. F.; Wheatley, P. O. *Applied Numerical Analysis*, 6th ed.; Addison-Wesley: Reading, MA, 1999; pp 264−274. (b) *Using Matlab*, version 6; The MathWorks Inc.: Natick, MA, 2000; pp 12.11−12.13, and 13.31−13.39.

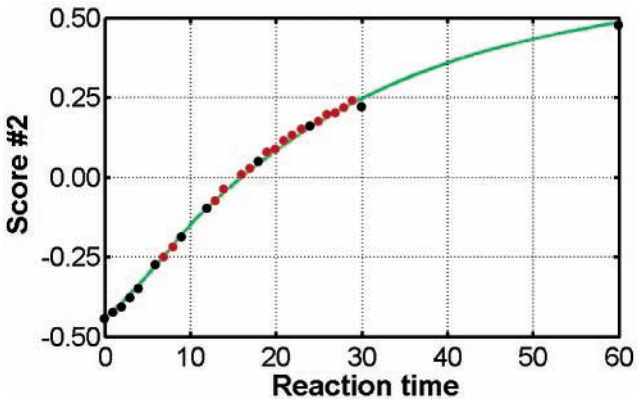


Figure 6. **Polynomial curve fit of the PC #2 score values versus reaction time. For the polynomial curve fitting of the smoothly drawn curve, a selection of time−score values were used, namely the objects at reaction times τ = 0, 1, 2, 3, 4, 6, 9, 12, 18, 24, 30, and 60, and these are indicated as black-filled circles on the smoothed curve. For the time range τ = 0−10 a polynomial fit of fifth degree was used. For the time range τ = 10−60 a polynomial fit of third degree was used This principal component represents changes in the concentration of benzoic acid 3 during the course of the reaction.**

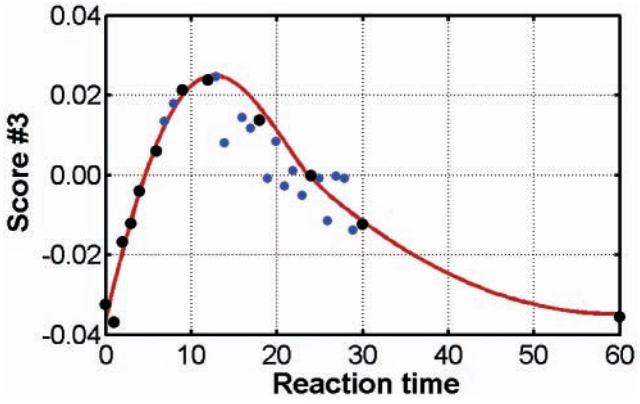


Figure 7. **Polynomial curve fit of the PC #3 score values versus reaction time. For the polynomial curve fitting of the smoothly drawn curve, a selection of time−score values were used, namely the objects at reaction times τ = 0, 1, 2, 3, 4, 6, 9, 12, 18, 24, 30, and 60, and these are indicated as black-filled circles on the smoothed curve. For the time range τ = 0−10 a polynomial fit of fifth degree was used. For the time range τ = 10−60 a polynomial fit of third degree was used. This principal component represents changes in the concentration of acetophenone 2 during the course of the reaction.**

together with the actual values in the region τ = 10−60. For this region a third-degree polynomial curve fit was applied.

The idea behind this data treatment and fitting may be summarized by the statement that each of the principal components should portray one chemical characteristic, which is different from the information portrayed in the other principal components. Nevertheless, each of the various principal components may contain further chemical or physical information, however, that covariates with the chemical composition of the reaction mixture.

At this stage of the data analysis, it is necessary to make an attempt to interpret the loadings from PCA. Figure 8 displays the principal component loading spectra (magenta line) for PC #1 (bottom row of subplots), PC #2 (center row of subplots) and PC #3 (top row of sub plots). The loading spectra PCs #1−3 are plotted together with the NIR spectra of ethylbenzene **1** (blue line), acetophenone **2** (red line), benzoic acid **3** (green line), and ethyl acetate **4** (black line). The NIR spectra of compounds **1**−**3** were recorded as solutions (1 M) in ethyl acetate **4**. The contribution of ethyl acetate **4** was removed from each of the three compound spectra. To simplify the comparison of the peaks of the loading spectrum and the NIR spectrum of each compound, the loading spectra were scaled as $sf \times \mathbf{p}_a$. The scaling factor sf is estimated according to eq 8 for each of the twelve subplot of Figure 8.

$$sf = \frac{\max(\mathbf{Abs}_j) - \min(\boldsymbol{Abs}_j)}{\max(\mathbf{p}_a) - \min(\mathbf{p}_a)}, \quad \begin{cases} j = 1, ..., 4 \text{ (compounds } \mathbf{1}-\mathbf{4}) \\ a = 1, ..., 3 \text{ (principal components)} \end{cases} \tag{8}$$

The loadings can be utilized to discern the various chemical or physical effects that are present in the different principal components. A monitored reaction, such as the imagined reaction of Scheme 1, can be considered as a closed system. The major variation that appears in such systems is the mutual concentration variations of the various reacting species and the products thereof.

Despite the complexity of NIR spectra it is possible by means of the wavelength-loading spectra achieved from the PCA to discern the wavelengths that can be assigned to the different functional groups found in the substrate **1**, the intermediate **2**, and in the final product **3**. The substrate, **1**, is expected to show specific absorption bands for CH_3 and CH_2 that are not present in the other two compounds. Likewise, the ketone carbonyl C═O group is only present in the intermediate **2**, and the carboxylic acid −COOH group is only present in the target product **3** of the oxidation reaction. In addition, the ester carbonyl present in the solvent is expected to be found in the principal component, explaining the majority of the variance of the spectral data matrix $\boldsymbol{X}$. Examining the loading spectra, some of the specific absorption bands of the functional groups can be assigned, if with some difficulty.

The PC #1 loading spectrum reveals an inverted NIR spectrum of ethyl acetate **4**, see subplot in the lower right corner, Figure 8. This is not very surprising, since ethyl acetate **4** is the compound that is present in the largest quantity. Interestingly, plotting the time−PC #1 score reveals a profile similar to the consumption of ethyl benzene **1** (Figure 5). This observation is difficult to explain but can reflect some sorts of solvatization or hydrogen bonds toward ethyl benzene (or benzoic acid?). In the following we will try to use this principal component score−time plot in an attempt to describe the consumption of ethyl benzene. Ethyl benzene can moreover be calculated by means of eq 6 if the two other principal components (PC #2 and PC #3) can be assigned to the two compounds benzoic acid **3** and acetophenone **2**, respectively.

The PC #2 loading spectrum clearly possesses features that assign the principal component to benzoic acid **3**, see

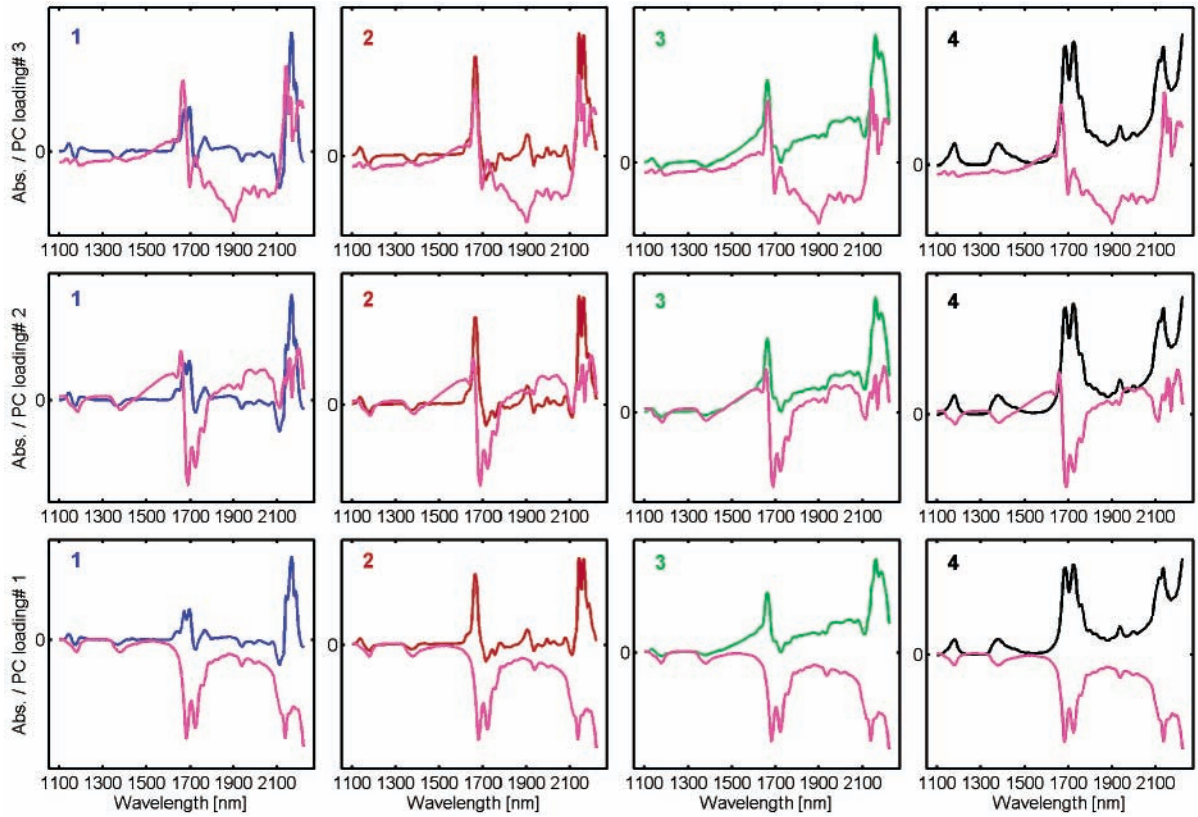


Figure 8. **The graphic shows the NIR spectra of ethylbenzene 1 (blue line), aceophenone 2 (red line), benzoic acid 3 (green line), and ethyl acetate 4 (black line). The NIR spectra of the compounds 1−3 were recorded as solutions (1 M) in ethyl acetate 4. The contribution of ethyl acetate was removed from each of the three compound spectra. Together with the compound and solvent spectra 1−4, the principal component loading spectra (magenta line) for PC #1 (row 1 from the bottom), PC #2 (row 2), and PC #3 (row 3) are plotted.**

the third subplot from left of the center row of Figure 8. The carboxylic acid group, −C(═O)OH, has a second overtone absorption band at 1890−1920 nm. The ridge at the wavelength region (1300 nm) 1500−1700 nm, the specific peak at 1680 nm, and spectrum profile in the region 1700−2100 nm fit well with the NIR spectrum of benzoic acid.

Comparison of the PC #3 loading spectrum with the NIR spectrum for each of the compound **1**−**4** reveals several similar absorption bands between PC #3 loading and the NIR spectrum of acetophenone **2**, see second subplot from left of the top row, Figure 8. The region 1700−2100 of the PC #3 loading spectrum corresponds however to the inverted acetophenone spectrum region. The time−score PC #3 plot (Figure 7) shows a profile very similar to the concentration profile of acetophenone in the reaction of Scheme 2.

Converting the Time−Score Curve Profile to a Time−Concentration Reaction Profile. The author has previously disclosed methods for (i) how the reaction time−score curve profile can be used to determine the endpoint of a reaction,[15] and (ii) how to estimate a plot of the reaction time−concentration profile for the final product on the basis of the time−scores.[16] These methods will for the sake of completeness be applied here to indicate an approximate end-point and to estimate the concentration profiles for the substrate, the final product, and for the intermediate.

The "end-point" of any reaction monitored using NIR spectroscopy can be easily determined by only one experiment. The reaction is conducted over an extended reaction time, to ensure that the reaction is brought to completeness. The spectral data matrix $\boldsymbol{X}$ is submitted for PCA, to estimate the scores and loadings. The score values are represented in a time−score plot. The endpoint is determined in the region where the time−score profile approaches a horizontal line. The end-point is more exactly determined at the point on the time−score plot where the differences among several successive score values becomes insignificant. Applying this to the imagined oxidation process, the approximated reaction time is determined to be at reaction time $\tau = 60$.

Using the start point for score-concentrations $t_{1(o)}$, $[\mathbf{1}]_o$ and the end point $t_{1(60)}$, $[\mathbf{1}]_{60}$, a simple regression model, eq 9, describing the concentration of ethylbenzene **1** as a function of the score value can be achieved. The scores t_1 are given by PC #1. The coefficient α is the regression coefficient for the regression line and α_o is the intercept. The model, eq 9, can then be used to estimate the reaction time−concentration profile based on the reaction time−score profile of Figure 5.

In the same way, a model for the product benzoic acid **3** can be achieved, eq 10. For this model development, the start concentration of benzoic acid is $[\mathbf{3}]_o = 0$ with a final concentration of $[\mathbf{3}]_{60} \approx 1$. The score values are for PC #

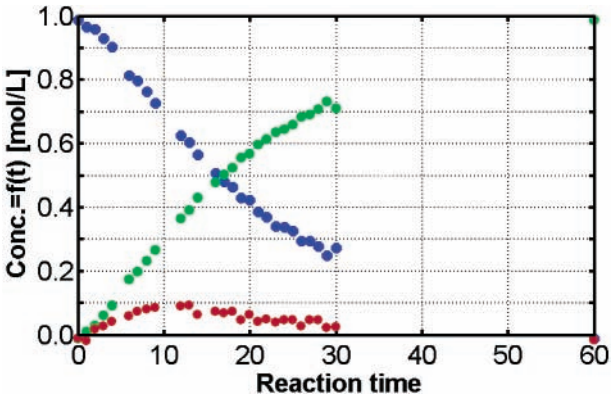


Figure 9. **Concentration−time profiles for each of the compounds ethylbenzene 1, acetophenone 2, and benzoic acid 3. The profiles are plotted on the basis of the scores estimated by PCA of the near-IR data matrix. Simple regression models are then used to transform the score values into concentration values, that in the present plot are plotted versus the reaction times.**

$$[\mathbf{1}]_\tau = \alpha_0 + \alpha \times t_1 \quad \begin{cases} \tau & t_1 & [\mathbf{1}] \\ 0 & -6.1806 & 1.00 \\ 60 & -7.4036 & 0.00 \end{cases} \Rightarrow$$

$$[\mathbf{1}]_\tau = 6.0536 + 0.8177 \times t_1 \quad (9)$$

$$[\mathbf{3}]_\tau = \beta_0 + \beta \times t_2 \quad \begin{cases} \tau & t_2 & [\mathbf{3}] \\ 0 & -0.4338 & 0.00 \\ 60 & 0.4877 & 1.00 \end{cases} \Rightarrow$$

$$[\mathbf{3}]_\tau = 0.4708 + 1.0852 \times t_2 \quad (10)$$

$$[\mathbf{2}]_\tau = \gamma_0 + \gamma \times t_3 \quad \begin{cases} \tau & t_3 & [\mathbf{2}] \\ 0 & -0.0317 & 0.0000 \\ 8 & 0.0187 & 0.0901 \end{cases} \Rightarrow$$

$$[\mathbf{2}]_\tau = 0.0567 + 1.7905 \times t_3 \quad (11)$$

To construct a graphical representation of the time−concentration profile for the intermediate acetophenone **2** on the basis of the time−score plot, two similar reactions must be carried out. The first one is conducted and analyzed as described above. From the time−score plot, the time and score pair for where the intermediate shows a maximum is determined. This will be used as the "end-point." The second experimental run is "quenched" at approximately the reaction time where the concentration of the intermediate shows a maximum, and the concentration of the intermediate is determined, for example, by means of a chromatographic method. For the oxidation process (Scheme 2), the intermediate (score#3) shows a maximum at approximately time $\tau = 8$. For the intermediate **2**, the two data points $t_{3(o)}$, $[\mathbf{2}]_o = [-0.0317, 0]$ and $t_{3(\tau=8)}$, $[\mathbf{2}]_8 = [0.0187, 0.0901]$ is used in eq 11.

Using the eq 9−11, the concentration profiles of the three compounds **1**, **2**, and **3** can be viewed. The intermediate acetophenone **2** is shown by the by red-filled circles, ethyl benzene **1** by the blue-filled circles, and benzoic acid **3** by the green-filled circles.

The predictive capacities of the models of eqs 9−11 are shown in Figure 10. In this figure, the measured values are plotted against the predicted ones for (a) the substrate ethyl benzene **1**, (b) the intermediate acetophenone **2**, and (c) the final product benzoic acid **3** of the imagined oxidation reaction.

As Figure 10a and c shows, the estimated simple linear regression explains the concentration profiles for ethyl benzene **1** and benzoic acid **3** superbly. For the intermediate acetophenone **2**, the model is somewhat aberrant for the region where the concentration increases. For that region (indicated with black circles around the red filled circles), the model predicts a too-low concentration. For the rest of the samples throughout the course of the reaction, the model shows a very good predictive ability.

Discussion and Conclusions

It has been shown that by means of a series of near-IR spectra, recorded during the course of a (imagined) synthetic reaction, it is possible to predict the current reaction profiles, that is, the 2D plot of reaction time versus concentration, of the substrate, the intermediate, and of the final product, respectively. The near-IR spectra were analyzed using PCA to achieve the scores and the loadings. By means of the score values for each of the significant principal components, it was possible to portray the reaction profile for each of the compounds that participates in the reaction. A simple linear regression model was established from only two samples using their scores (as x's) with their corresponding quanti-

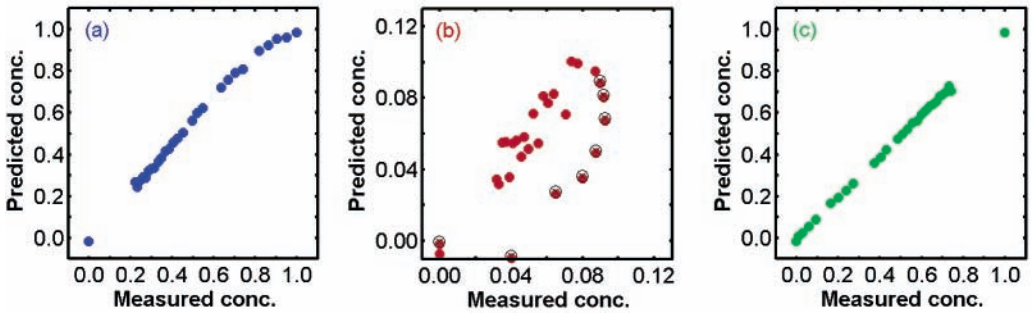


Figure 10. **Correlation of actual values and predicted values for models describing the concentration of: (a) the substrate ethyl benzene 1, (b) the intermediate acetophenone 2, and (c) the final product benzoic acid 3 of the imagined oxidation reaction. The product statistics: $R^2 = 0.9950$ and slope = 0.9044 for ethyl benzene; $R^2 = 0.7080$ and slope = 0.8964 for acetophenone, and $R^2 = 0.9997$ and slope = 0.9973 for benzoic acid shows a good predictive ability of the derived models. Regarding the plot (b), the filled red circles marked with a black circle show a marked deviation from the other points. These points represent the first eight NIR recordings of the reaction. The model predicts a lower concentration than the actual values for these points. For the rest of the measurements, an excellent agreement between the measured and the predicted values is observed.**

tatively measured content of product (as y's) in a small "calibration set". This model was used to recalculate the reaction time−score reaction profile into a reaction time−concentration reaction profile.

PCA was used in this study for the analysis of the NIR spectral data matrix $\boldsymbol{X}$. As an alternative, the singular value decomposition (SVD) method was utilized for the analysis of the NIR data. SVD[25] function released with Matlab 6.5.1,[26] provided similar results to those acquired by the PCA method.

Experimental Section

Preparation of Samples. The samples for the imagined oxidation process of Scheme 2 were prepared from commercial samples of ethylbenzene, acetophenone, and benzoic acid, respectively. Commercial samples were used directly without any purification prior to use. The samples were weighed on an analytical laboratory balance, the quantities given in Table 1. The samples were diluted with ethyl acetate (100 mL) in measurement flasks.

NIR Spectral Recording. The near-infrared (NIR) spectra (λ 1100−2498 nm, $\Delta\lambda = 2$ nm) were recorded by means of a NIR System model 6500 spectrometer equipped with a Optiprobe system fiber optical probe (a transmittance probe) with total path length of $L = 2$ mm. The upper region of each spectrum was removed due to noise and absorption by the fiber optics. The final spectrum from each recording is constituted by the wavelength range λ 1100−2220 nm, with a resolution of $\Delta\lambda = 2$ nm, that gives 561 absorption values in total for each spectrum.

Multivariate Calculations, Curve Fitting, and Graphics. The calculations of the principal components by means of PCA, the curve fitting using polynomial regression[23] and spline,[24] and the graphical representations[27] of the calculated results were performed by using in-house developed routines for MATLAB, version 6.1.

(25) Anderson, E.; Bai, Z.; Bischof, C.; Blackford, S.; Demmel, J.; Dongarra, J.; Du Croz, J.; Greenbaum, A.; Hammarling, S.; McKenney, A.; Sorensen, D. *LAPACK User's Guide*, 3rd ed.; SIAM: Philadelphia, 1999 (http://www.netlib.org/lapack/lug/lapack_lug.html).

(26) *MATLAB*, version 6.5.1, release 13, service pack 1; The MathWorks Inc.: Natick, MA, August 4, 2003.

(27) *Using Matlab Graphics*, version 6; The MathWorks Inc.: Natick, MA, 2000.

Acknowledgment

Economical support from the University of Bergen and Optimum Accipe is gratefully acknowledged. Professor Rolf Manne is acknowledged for discussions during the initial period of this project. An unknown reviewer is acknowledged for important criticism and corrections concerning the first version of this paper. Professor George Francis is acknowledged for linguistic assistance. Dr. Egil Nodland is acknowledged for assistance with the setup of the NIR instrument. Misses Heidi Blokhus and Monica Hyland are acknowledged for technical assistance preparing the samples and recording the NIR spectra.

Received for review January 16, 2004.

OP049971F