Available online at www.sciencedirect.com

SCIENCE DIRECT®

Chemometrics and Intelligent Laboratory Systems 78 (2005) 41–50

Chemometrics and intelligent laboratory systems

www.elsevier.com/locate/chemolab

Estimation of missing data using latent variable methods with auxiliary information

Koji Muteki[a], John F. MacGregor[a,*], Toshihiro Ueda[b]

[a]*Department of Chemical Engineering, McMaster University, Hamilton, Ontario, Canada L8S 4L7*
[b]*Mitsubishi Chemical Corporation, Yokkaichi, Mie, Japan*

Received 13 May 2004; received in revised form 25 October 2004; accepted 8 December 2004
Available online 9 February 2005

Abstract

Estimating missing data in a matrix is often done with methods, such as the EM algorithm, using the existing data in that matrix. However, if auxiliary data that is related to the missing measurements is available, it can help to estimate the missing values. This paper presents latent variable approaches that exploit an auxiliary data information matrix, as well as the data matrix itself, for estimating missing data on raw material properties of rubbers used in formulating industrial polymer blends. The use of auxiliary information is most useful when the percentage of missing data is high, and when there exist certain combinations of missing data that show little correlation with the data that are present. Two approaches to incorporating the auxiliary information are presented: a multi-block approach and a novel two-stage projection approach. The latter approach is shown to be more flexible and to provide slightly better estimates in two industrial polymer blending problems. However, in both cases, the addition of the auxiliary information is shown to significantly improve the estimates.

Keywords: Missing data; Principal component analysis (PCA); EM algorithm; Partial least square (PLS); Mixture designs; Polymer blends; Multi-block methods

1. Introduction

Missing measurements occur in various industrial settings. In industrial operations, data loss occurs periodically when sensors fail or are taken offline for routine maintenance, when measurements are removed from a data set because of gross measurement errors, or when measurements of certain variables are simply not collected at the same time as those of other variables [3]. In research and development, data loss occurs when some measurements fall outside the instrument range or when chemical property data is only partially available from the literature. In many instances, it is impractical to obtain all the property measurements for economic reasons. An insistence on using only complete data would entail throwing away large amounts of data. Therefore, effective methods to estimate missing data points are needed.

There are basically two issues in the missing data problem: building the model with missing data, and using an existing model with missing data. For the latter case, Nelson et al. [2,3] and Arteaga et al. [1] performed comparative studies among several methods, and Nelson et al. [4] studied the impact of missing data in prediction and in multivariate statistical process control (MSPC) applications. In most cases, the use of the expectation maximization (EM) algorithm provides the best estimation results, but several other approaches are comparable. The EM algorithm estimates the missing data Y_{miss} by maximizing the likelihood $l(Y_{\text{miss}}|Y_{\text{obs}})$ with respect to Y_{miss} for fixed Y_{obs} [9]. In terms of building the model with missing data, the use of single component projection (SCP) in the nonlinear iterative partial least squares

* Corresponding author. Tel.: +1 905 525 9140; fax: +1 905 521 1350.
E-mail address: macgreg@mcmaster.ca (J.F. MacGregor).

0169-7439/$ - see front matter © 2004 Published by Elsevier B.V.
doi:10.1016/j.chemolab.2004.12.004

(NIPALS) algorithm and the EM algorithm with latent variable methods such as PCA and PLS are well known [3,6–8]. This paper studies the problem of incorporating auxiliary information into the estimation of missing data during model building. It will use the EM algorithm with PCA modeling.

There are several important issues when building the model with missing data.

(1) It is important to understand the mechanism that leads to the missing data [8]. The important question is whether the variables that are missing are missing because they are related to the underlying values of the variables in the data set. Rubin and Little [8] define the following three categories: missing completely at random (MCAR), missing at random (MAR), and not missing at random (NMAR). Measurements are said to be MCAR when there is no relationship between values of the variables and their probability of being missing. Note that this assumption does not mean that the pattern itself is random, but rather that being missing does not depend on the true underlying value. Measurements are said to be MAR when the probability that an element is missing depends on the observed data only. In another word, the missingness does not depend upon the true underlying value of the missing data, and the observed data includes the necessary information to estimate the missing data. For example, data missing because of maintenance of a sensor in process operation would correspond to MAR. Measurements are said to be NMAR when the probability that an element is missing depends on the unobserved value of the missing elements. For example, NMAR occurs when some measurements are below a lower detection limit or over an upper detection limit. Problems arise when missing data belong to NMAR. In such a case, the distribution of missing measurements for a set of variables does not follow that of the observed measurements for that same set of variables in other objects. Therefore, the important assumption required to use PCA with EM is that missing measurements belong to MCAR or MAR, and not to NMAR.

(2) Data matrices which do not have overlapping objects or variables (Fig. 1) cannot be treated by any missing data method based only on these data because there exists no information on the joint distribution of the variables. Even if overlapping objects exist, when these are very few, as shown at Fig. 2a, the EM algorithm converges very slowly, and the estimates are poor. With overlapping variables (Fig. 2b) the estimation of the missing data is entirely dependent upon the strength of the correlation of the overlapping variables with the variables in the missing blocks.

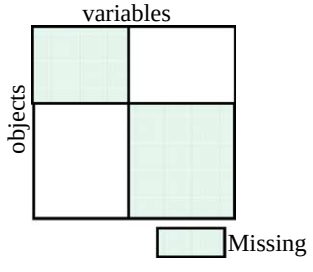


Fig. 1. No overlapping samples data structure.

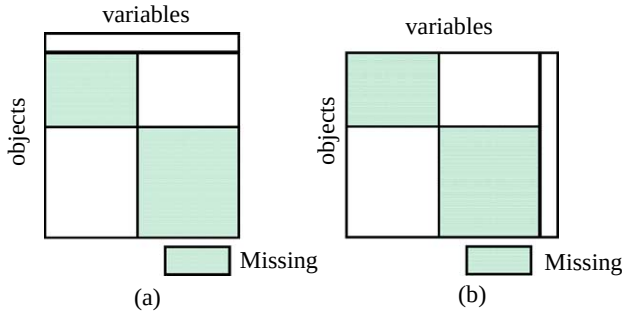


Fig. 2. Data structures: (a) where there are overlapping objects, (b) where some variables are always available.

(3) Certain critical combinations of missing measurements can give rise to large errors. When measurements on any variables that contribute significantly to capturing one or more of the latent variable dimensions are missing, one should exercise caution. Nelson et al. [3,4] presents diagnostic analyses to test for highly leveraged missing data combinations.

This paper discusses the use of auxiliary data sets that are influenced by the missing data to help in the estimation of the missing data. These methods will always bring more information to the problem, and hence lead to improved estimates. However, they are of particular benefit when the proportion of missing data is large, when critical combinations of missing measurements are present, or when there is a poorly overlapping data structures as in Figs. 1 and 2. In all the following we assume that the missing data are MCAR or MAR.

2. An industrial polymer blending problem

We consider the following example throughout this paper. It concerns the estimation of missing measurements of raw material property data in polymer blending. The raw materials consist of rubber, polypropylene, and oil. The mixture properties are mainly affected by the rubber properties, and the rubber property data is critical for design of any blended product. However, the rubber properties are difficult to measure, and the many different suppliers of the rubbers often provide measurements of different variables and omit measurements of others. This is often done by supplier to emphasize the features of their own specific materials. Even the same suppliers provide measurements on different properties for different product grades. Other measurements are omitted because they are costly to measure. Therefore, estimating the critical missing rubber property data is essential for the blending problem and for future product design.

The data structure for the industrial blending problem is represented in Fig. 3. X_{raw} matrix refers to the ($K \times N$) rubber material property data matrix. K is the number of rubber materials and N is the number of rubber material properties. The other data matrices X_{ratio} and $Y_{property}$ contain auxiliary information from blending studies in which one or more of these rubbers have been used. The $(M \times J)X_{ratio}$ contains the ratios of all components used in the formation of a blend. M is the number of blends, and J is the number of all the raw materials used in the polymer blends ($J > K$). The data in X_{ratio} contains the fraction of each raw material used in the blend ($0\% \leq x_i \leq 100\%$, and $\sum x_i = 100\%$). The $(M \times L)Y_{property}$ matrix contains the L properties measured on the final blends. The rubber property matrix X_{raw} and the mixture property matrix $Y_{property}$ usually contain measurements on different variables because the important properties of the final blends that a customer is interested in are quite different from those properties measured on the raw rubbers. Even if some of the same properties are measured (such as hardness and melt flow ratio (MFR)), they can be handled as different variables because their measurement ranges are so different.

To estimate the missing rubber property, the simplest approach would be to build a PCA model using the EM algorithm directly on X_{raw}. However, X_{raw} has a small number of property variables (N:small), a high proportion of missing measurements, and often the missing measurements are critical combinations [3] that lead to large estimation error. But in this problem the mixture property data $Y_{property}$ is clearly affected by the properties of the rubber and other materials in X_{raw}, as well as the ratios that were used (i.e., X_{ratio}). Therefore, this auxiliary information can be used to help estimate the missing values. Furthermore, the number of polymer blends (M) is usually much larger than the number of rubber materials (K), i.e., the same rubbers are often used in many blends, and the number of variables measured in $Y_{property}$ is usually larger than the number in X_{raw} since the blended product is sold for commercial end uses. Therefore, we will utilize the auxiliary X_{ratio} and $Y_{property}$ data as well as X_{raw} to obtain better estimates of the missing measurements in X_{raw}. However, a difficulty arises because of the different number of dimensions on each data matrix. There is no common dimension between X_{raw} and the auxiliary data matrices, a problem that would be common in most cases of auxiliary data. The objective of this paper is to propose some solutions to this missing data problem.

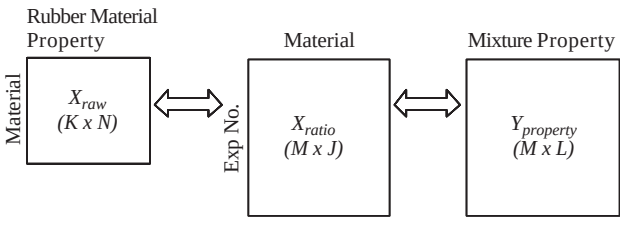


Fig. 3. Data structure in mixture design.

The paper is organized as follows. In Section 3, the EM technique for estimating missing data using PCA models is described briefly. In Section 4, three approaches are presented for estimating the missing data in X_{raw}. The first is the direct application of EM to X_{raw}. The second involves the use of a multi-block approach to incorporate the auxiliary data, and then using the EM algorithm on the multi-block matrix. The third is a new two-stage approach in which the X_{ratio} and $Y_{property}$ data are projected to a dimension compatible with X_{raw} and then EM is applied to the combined data matrices. The limitations of each approach are discussed. In the Section 5, two industrial examples are used for verifying the effectiveness of the auxiliary variable approaches, and comparing them with each other and with the standard approach without auxiliary data. Section 6 considers a diagnostic analysis to assess why certain combinations of missing data bring about large error; the effectiveness of the auxiliary information in reducing these errors; and insights that can be gained from these approaches on relationships between raw material properties and blended product properties.

3. Estimation of missing data

Before discussing the main analytical procedures, the common method to estimate the missing data in all the approaches is described briefly.

3.1. PCA with EM

PCA with EM refers to the iterative replacement method which uses the PCA modeling technique to estimate missing data. Fig. 4 shows the schematic procedure as follows:

(1) Fill in the missing elements with initial values, which is often a mean of the column or the value obtained by Single Component Projection (SCP) based on NIPALS;
(2) Build the PCA model using the full data matrix;
(3) Use the PCA model to predict the missing data ($\hat{\mathbf{X}}=\mathbf{TP}^{\mathbf{T}}$);
(4) Check for convergence;
(5) If not yet converged, replace the original missing data by the new predicted values and return to step (2).

Upon convergence, a PCA model and the predicted values of missing data are obtained.

The above procedure of iterative replacement can be considered to be equivalent to EM algorithms when they are formulated as missing data problems. The maximization (M)-step in the EM algorithm corresponds to PCA modeling at step (2) after filling in the missing data. The expectation (E)-step in the EM algorithm finds the conditional expectation of the missing data (step (3)) given the observed data

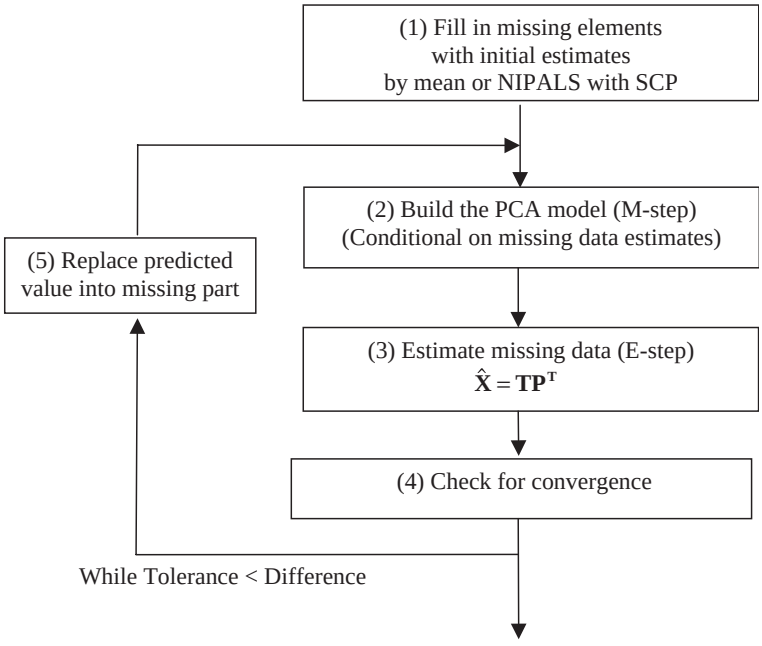


Fig. 4. PCA with EM.

and current model parameters **P**, and then substitutes these expectations for the missing data.

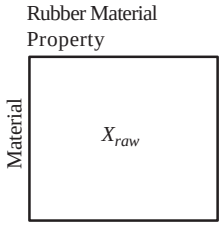
4. Strategies for estimating missing data on rubber material properties

This section discusses three different approaches to estimate the missing data in the raw material property matrix X_{raw}.

4.1. Direct approach using X_{raw} data only

This direct approach refers to directly applying PCA with EM to the rubber material property matrix X_{raw}, as shown in Fig. 5.

This approach will generally provide good estimates of the missing data if the number of objects is large and representative of the population, and if there are reasonably large number of highly correlated measured variables, and if the proportion of missing data is not too large. However, as mentioned earlier, for our problem the number of variables in X_{raw} is small, the number of objects (rubbers) is small (5 or 6), and there are often critical combination of missing data that will lead to large estimation errors [3]. Therefore, the approach was unsatisfactory for the industrial problems mentioned here (see results in Section 5).

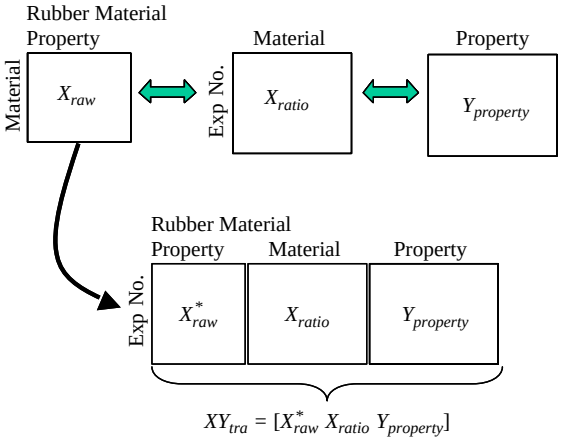


Estimation on missing data = PCA with EM (X_{raw})

Fig. 5. Direct approach.

4.2. A traditional multi-block approach using auxiliary data

In this section we propose an approach which exploits the X_{raw} to allow three matrices to be combined to form a single multi-block matrix. PCA with EM can then be performed on the augmented multi-block matrix. The schematic procedure is shown in Fig. 6. The new rubber property matrix X^{*}_{raw} consists of repeated rows of the original rubber property matrix X_{raw} corresponding to the rubber material used in each blending experiment corresponding to the rows of X_{ratio}



Estimation for missing data = PCA with EM (XY_{tra})

Fig. 6. A traditional multi-block approach.

and Y_{property}. The row data in X^*_{raw} is repeated as long as the same rubber is used, even if the mixture ratio is changed. PCA with EM is applied to new combined matrix XY_{tra}. Note that a limitation of this approach is that the number of the rubber materials used in each blending experiment of X_{ratio} necessarily has to be only ONE. If there are mixture data using two or more rubbers, they must be discarded. In industrial situations, several rubbers with different properties are often used in the same blend. For examples, in the first industrial study of Section 5 approximately 30% of the blends contained more than one rubber. This is therefore a severe constraint on this approach.

A second problem with this approach is that each row of X_{raw} may appear multiple times in X^*_{raw}, and so multiple estimates of the same missing values will be obtained by applying EM to the PCA on XY_{tra}. The mean values of these repeated estimated missing data points of X_{raw} in X^*_{raw} can be used as the estimated values. Using these mean values at each iteration of the EM algorithm could potentially pose a convergence problem since the estimated values at the end of the M-step are not being used for the following estimation step. A better approach could be to build into the EM algorithm this constraint that unknown values in the repeated rows are the same. Since in this study, convergence using the mean values outside the EM algorithm was obtained, this simpler approach was used.

4.3. New projection approach for using auxiliary data

In this section, we propose a new multi-step approach that effectively utilizes X_{ratio} and Y_{property} as well as X_{raw}, without the constraints that arise in the proceeding approach. A schematic of the procedure is shown in Fig. 7.

The procedure of this approach is as follows:

(1) Construct a PLS model between X_{ratio} and Y_{property}
(2) Substitute 100% ratios for the K rubber materials in X_{ratio} and use the PLS model to predict the $(K \times L)$ 100% property matrix $Y_{100\%\ \text{property}}$ (Projection of PLS model to 100% rubber properties)
(3) Combine the corresponding rows of X_{raw} and $Y_{100\%\text{property}}$ into the new matrix $XY_{\text{new}}=[X_{\text{raw}}\ Y_{100\%\ \text{property}}]$
(4) Implement PCA with EM for XY_{new}.

Some of the features of this approach are as follows: (1) It does not have the constraints that arise in the traditional multi-block approach. Namely, the number of the rubber materials used at each blending experiment at X_{ratio} can be greater than one, and it does not face the problem of converging to different estimates of the same missing value. (2) The new projection approach is flexible and can be applied, even for the analysis of only part of the group of raw materials, like rubber. (3) The new approach provides some insights into the physical relationships between

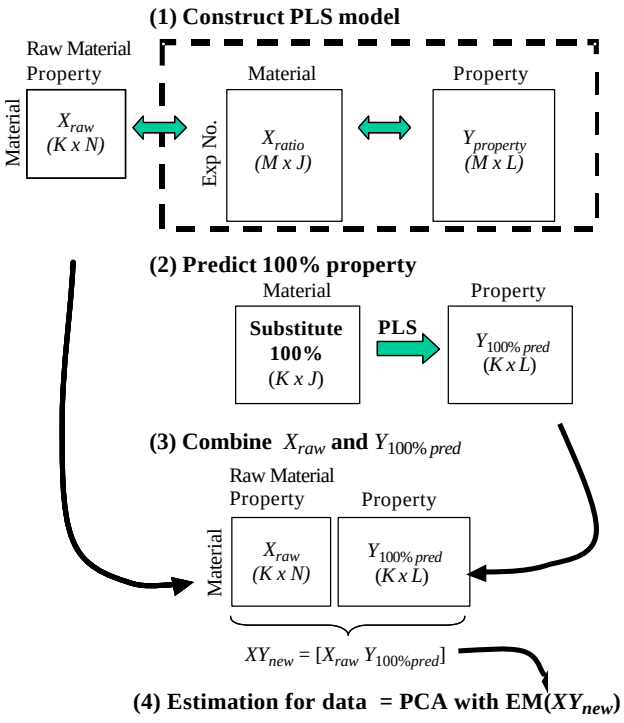


Fig. 7. New multi-step projection approach.

mixture properties and rubber properties. (4) The analytical procedure of the new method is very similar to the thinking process of polymer researchers performing the mixture design. Indeed, the new projection approach has been devised from that idea.

In this projection approach we make no claim that the 100% predictions are in any way an accurate prediction of the true blend properties if 100% rubbers were used. In fact we would expect this large extrapolation to be very poor. However, the prediction of the 100% properties is only used as a means of projecting the information in the auxiliary high-dimensional matrices (X_{ratio} and Y_{property}) to a lower dimensional space ($Y_{100\%\ \text{property}}$) that is of a dimension compatible with the X_{raw} matrix and which should reflect the information in the auxiliary matrices that is related to X_{raw}. In other words, $Y_{100\%\ \text{property}}$ should be viewed as a projection of the information in the auxiliary data to the low dimensional space of X_{raw}.

5. Industrial examples

This section presents two industrial examples of polymer blending with missing data in the raw material property matrix. Both examples feature the manufacture of thermo-plastic materials which blend some types of polypropylene, oil, and rubber. It can be assumed that the mixture property matrix Y_{property} is influenced by only the mixture ratio and the property of the raw materials, since the same manufacturing equipment was used and all

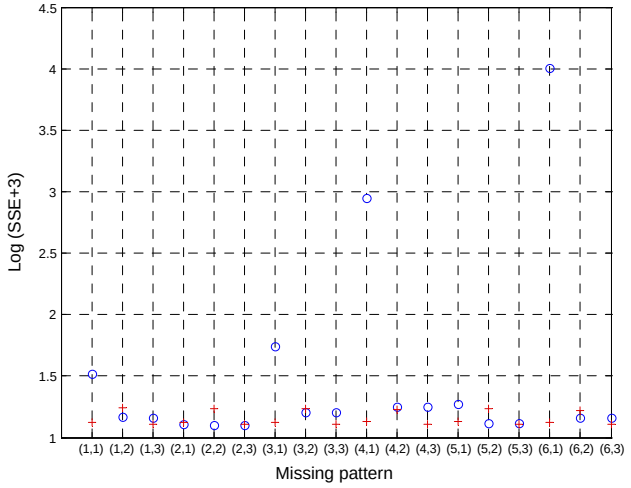


Fig. 8. Comparative estimation error results of two approaches for all possible cases of a single missing measurement (○: direct approach; +: new projection approach using auxiliary data).

processing conditions such as extruder and mold temperatures were kept constant. However, any variations in process conditions could be accounted for with an additional auxiliary matrix $X_{process}$ [5].

In both industrial examples a data set with complete matrices (no missing data) has been assembled. We then delete selected sets of measurements in the rubber material property matrix X_{raw} to simulate missing data. The estimates of the "missing data" are then compared to the known measured values in X_{raw} to verify the effectiveness of each approach. We assume there are one or two missing points in the raw material data matrix X_{raw} in each simulation. In selecting the missing points, all the combination patterns of one and two at time are investigated. As a criterion to evaluate the approaches, the sum of the squared estimation errors (SSEs) between observed and predicted values is calculated. The scaled values after unit variance scaling and mean centering are used in calculating the SSE. Constant means and scaling factors for mean centering and unit variance scaling calculated from the full data matrix are used for all the simulations, in order to avoid the effect of different means and scaling factors on the estimated results. The means and scaling factors calculated on the data matrices with missing points are highly changeable since the size of raw material property data matrix X_{raw} handled in this analysis is very small. For all approaches, the number of the latent variables in PCA and PLS are selected so that cumulative cross-validated coefficient (Q2) of the models are maximized. For part of the mixture ratio matrix X_{ratio}, a *D-optimal design* has been applied and so the conditioning of the auxiliary data is relatively good.

5.1. Example 1

In this first industrial blending example the (6×3) rubber material property matrix X_{raw} consists of 6 rubbers and 3 properties, the (37×11) mixture ratio matrix X_{ratio} consists of 37 blends and the ratios of 11 materials used in each blend and the (37×8) mixture property matrix $Y_{property}$ consists of 8 polymer blend properties measured on each of the 37 blends. The 11 raw materials include 6 rubbers, 3 polypropylenes, and 2 oils. A feature of X_{ratio} is that approximately 30% of the blends use two or more rubbers. If the traditional multi-block approach is used, it will entail discarding 30% of the data. Therefore, in this example, the comparative study is performed only between the direct approach and the new projection approach using all of the auxiliary data. From past experience with these blends, the blend properties in $Y_{property}$ are highly correlated with the rubber properties in X_{raw}. This kind of the expectation is necessary for the auxiliary data to be useful.

In the first study, only one missing measurement at a time is assumed present in X_{raw}. To cover all possible cases, the total number of simulations is 18 (=6×3) patterns. The comparative results of the two approaches are shown at Fig. 8. The abscissa designates the location (i,j) in X_{raw} of each missing data pattern corresponding to the 18 cases. The ordinate designates the log of the SSEs between the observed and estimated values at each simulation. The circle mark "○" indicates the result by the direct approach, and the plus mark "+" indicates the result by the new projection approach.

The estimation by the projection approach is more accurate than the direct approach for all the missing patterns. The direct approach is very inaccurate for some missing patterns. The auxiliary information of the new approach contributes highly to a better estimation in these latter instances. In Section 6, reasons for the occurrence of errors in the direct approach and for their disappearance when using the auxiliary data are discussed.

The comparative results for all combinations of two missing measurements are presented in Fig. 9. In all, 153 combination patterns of 18 (=6×3) elements, taken 2 at a time, are investigated ($_{18}C_2$=18!/18!(18−2)!=153).

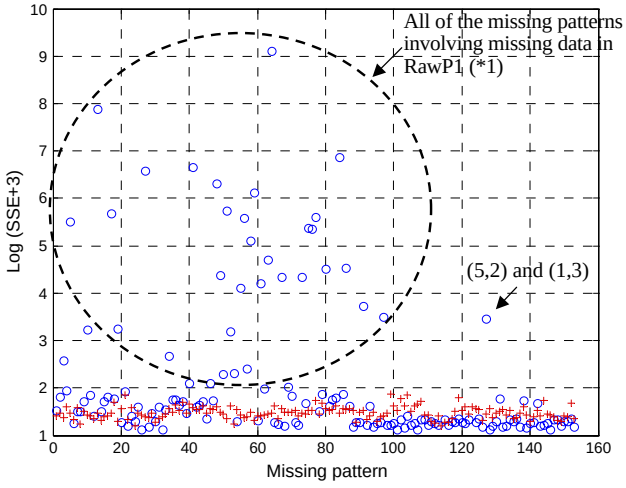


Fig. 9. Comparative estimation error results for the case of all possible cases of two simultaneous missing measurements (○: direct approach, +: new approach) ((*1): first raw material property in X_{raw}).

Table 1
Sum of squared error (SSE) results for example 1

	Direct approach	New approach
One missing data per simulation	75.4	3.2
Two missing data per simulation	18,158.0	219.4

As can be seen in Fig. 9, in the case of two missing data, the estimation by the new projection approach again is much better than that by the direct approach. Quantitative results for the sum of the squared errors by the two approaches are shown in Table 1. The new projection approach drastically improves the accuracy for the estimation of certain combinations of missing data. Explanations for the large SSE values in Fig. 9 are discussed in the Section 6.

5.2. Example 2

This section illustrates a second industrial example of the blending of rubber, polypropylene, and oil, but the data are from a different study and for very different products. The (5×3) raw material property matrix X_{raw} consists of 3 properties measured on 5 rubbers; the (42×8) mixture ratio matrix X_{ratio} consists of 42 blends of 8 materials; and the (42×7) mixture property matrix $Y_{property}$ consists of 7 property measurements on the 42 blends. The rubber properties and mixture properties are almost the same as those of the first example. The different feature of this data is that only one rubber is used for each of the blends in X_{ratio}. Therefore, the traditional multi-block approach can also be used. The main objective is to conduct a comparative study among three approaches for estimating the missing data.

The results are shown in Table 2. The results in the first row are for all possible cases where there is only one missing observation in X_{raw}, the total number of the simulations being 15 (=5×3). The second row presents the results for all possible combinations of two missing measurements (105 combination patterns of 15 (=5×3) measurements taken 2 at a time). The estimations of the missing data by the two approaches using the auxiliary data on the mixture information are much better than the direct approach that uses only X_{raw}. The difference between the new projection approach and the traditional multi-block approach is small when there is only a single missing measurement, but is slightly better when there are two missing measurements. Both approaches using auxiliary data provide good estimation, but the results of the direct approach are unacceptable.

Table 2
Sum of squared error (SSE) results for example 2

	Direct approach	New approach	Traditional approach
One missing data per simulation	143.41	2.46	2.47
Two missing data per simulation	928.95	49.06	66.83

Table 3
Sum of squared error results for example 1 (after discarding blends using more than two rubbers at each blends)

	Direct approach	New approach	Traditional approach
One missing data per simulation	1080.90	3.26	3.31
Two missing data per simulation	4745.90	48.93	58.12

5.3. Example 1 revisited

As can be seen in Table 2 for example 2, the results of the projection approach are slightly better than those of the traditional multi-block approach in terms of the estimation error. In particular, the difference between the two approaches is larger in the case of two simultaneous missing data. In order to further verify the validity of the result, another comparative study using the example 1 data is performed. Since approximately 30% the blends in X_{ratio} of the example 1 consists of blends using more than two rubbers, these blends must be discarded in order to use the multi-block approach. As the result, the size of three matrices is as follows; the rubber material property matrix X_{ratio} in (4×3), the mixture ratio matrix X_{ratio} in (26×9), the mixture property matrix $Y_{property}$ in (26×8). The results shown in Table 3 are very similar to the results in Table 2 for example 2.

The use of the auxiliary data has enabled both approaches to give greatly improved results, compared to direct approach. The two auxiliary data methods also give very comparable results, but again the new projection approach gives slightly better results when there are two missing measurements. We offer the following conjecture for this slight improvement. In the traditional multi-block approach, the PCA model of XY_{tra} must simultaneously explain three blocks; the raw material property matrix X^*_{raw}, the larger mixture ratio matrix X_{ratio} mixture material properties matrix $Y_{property}$. The X_{ratio} and $Y_{property}$ matrices, resulting in part from an experimental design, require a large number of PCs to model. On the other hand, in the projection approach PLS is used to extract only the information related to the rubber materials from these matrices X_{ratio} and $Y_{property}$ and project it onto a low dimensional space $Y_{100\%\ pred}$. By eliminating the extra dimensions, unrelated to the rubber properties, the lower dimensional PCA model on XY_{new} gives a smaller variance of prediction.

6. Discussion

The above two examples have illustrated the value and ability of each of three approaches. This section will further investigate the importance of using auxiliary data by: (1) examining why certain missing measurements or combina-

tions of missing measurements lead to large errors; (2) demonstrating why and when the auxiliary data approaches work well; (3) showing how the new projection approach can provide additional insights into the relationship between raw material properties and mixture material properties.

6.1. Diagnostic analysis

The results from example 1 are used to diagnose why certain combinations of missing data lead to large errors in the direct approach (using X_{raw} only). All the cases for which large errors arise in Fig. 8 (i.e. for one missing measurement only) involve missing data on property RawP1. This is because RawP2 and RawP3 are highly correlated to each other and RawP1 is almost orthogonal to these two variables, as can be seen the loading plot in Fig. 10. Therefore, missing a measurement on RawP1 is equivalent to losing information on one of the latent variable dimensions. The reason for the very large error when the measurement X_{raw} (6,1) is missing is due not only to RawP1 being missing, but also to the large leverage of the Rub6 sample in that same direction (see score plot in Fig. 11). The reason for the large error when X_{raw} (4,1) is missing in spite of the small leverage of the Rub4 (Fig. 11) is due to high SPE for Rub4 in the model (Fig. 12).

All the missing patterns for which large errors arises in Fig. 9 (i.e. two simultaneous missing measurements) also include situations where RawP1 is one of the missing measurements, except for the missing data pattern (X_{raw} (5,2) and X_{raw} (1,3)). The sensitivity of this missing combination is due simultaneously missing properties RawP2 and RawP3 which define one latent variable direction (Fig. 10), and to this occurring simultaneously in rubbers 1 and 5 both of which have high leverage in this direction (see Fig. 11).

Clearly, for various reasons there exist sensitive combinations of missing measurements that have a large impact on the direct EM method using the X_{raw} data only.

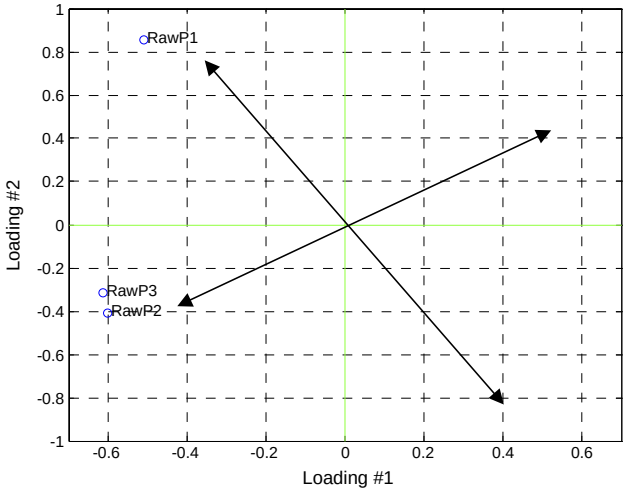


Fig. 10. Loading plot of PCA in X_{raw} (3 variables: RawP1–RawP3).

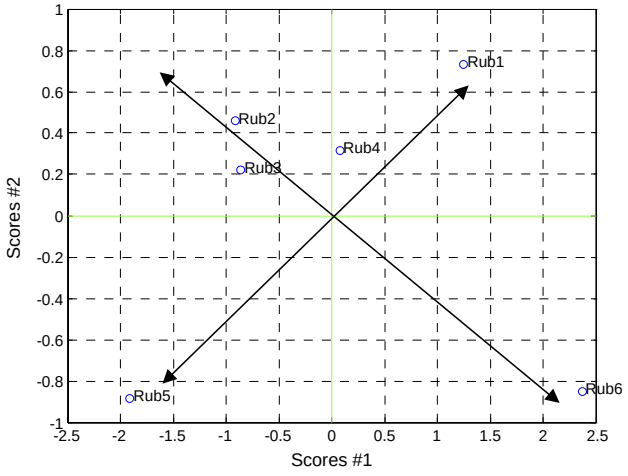


Fig. 11. Scores plot of PCA in X_{raw} (6 rubber samples: Rub1–Rub6).

The next section reveals how use of the auxiliary data largely eliminates the sensitivity of these missing data combinations.

6.2. Effectiveness of auxiliary information

The addition of auxiliary information on the rubber properties obtained from the projection of the (X_{ratio} and $Y_{property}$) experimental data to a $Y_{100\%\ pred}$ matrix provides additional variables in the augmented matrix XY_{new} that are correlated with the missing data. To investigate the strength of the correlation of each variable in X_{raw} with the other variables in X_{raw} alone and with the other variables in the augmented matrix XY_{new}, several PLS models are built. Each property in X_{raw} is set as the Y variable, and all the other variables are set as X variables. In the direct approach, the number of X variables is always 2. In the new projection approach, the number of X variables is always 10 (=2+8). The R2 (cumulative sum of squares of Y explained) and the Q2 (cumulative cross-validated R2) from the PLS models are shown in Table 4.

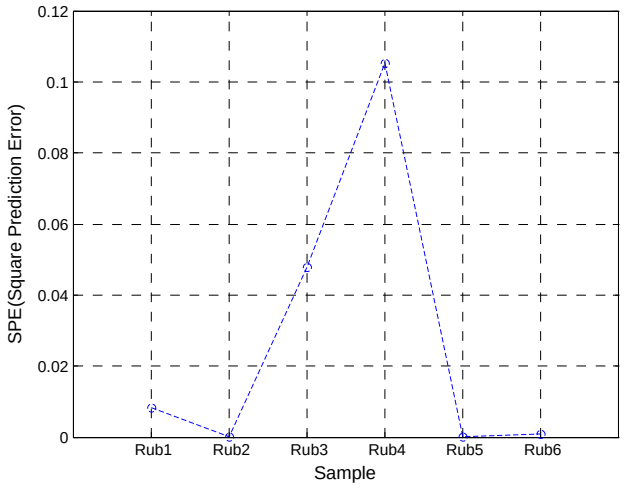


Fig. 12. SPE Plot of PCA model for X_{raw}.

For all the properties studied, the R2 and Q2 values are much better using the auxiliary information. In particular, the improvement in the prediction of the raw material property RawP1 is remarkable. Clearly, using the direct approach to estimate the missing data property, RawP1 is highly questionable because of its lack of correlation with the remaining variables. The fact that RawP1 is poorly correlated with the RawP2 and RawP3 was also seen through the loading plot of X_{raw} in Fig. 10. On the other hand, RawP2 and RawP3 are highly correlated with each other, and so even by using the direct approach, they can be estimated reasonably accurately if only one measurement is missing, as shown in Table 4. However, even for these cases, the estimation is enhanced by the addition of the auxiliary information. In order to further illustrate the useful information coming from the auxiliary data, the PCA loading plot of XY_{new} is shown in Fig. 13. It can be seen that the variables in the projection matrix ($Y_{100\%\ pred}$) coming from the auxiliary data, namely MixP1 to MixP8 are very highly correlated with the original raw material properties in X_{raw} (RawP1 to RawP3). In this situation missing any one measurement or any combination of two measurements in X_{raw} will have little impact on the ability to estimate their missing values using EM on XY_{new}. None of these missing combinations will result in a significant loss of information, in one of the latent variable directions, as was the case in Fig. 10.

In summary, the auxiliary data methods will make a significant improvement if the auxiliary data is highly correlated with the missing measurements. As a counter example, if all the blend properties were highly influenced by only polypropylene, then the use of the auxiliary blend data would not provide much improvement for estimating the missing rubber properties (However, in this case, given that the rubber materials have little influence on the blended product, one would have little interest in estimating their missing properties).

6.3. Insight on relationships between raw material properties and mixture material properties

It is very important for polymer researchers to have a feel for variable relationships between raw material properties and mixture material properties in order to help them in their efforts at designing new blended products. A side of benefit of the new projection approach is that it can be used to provide qualitative insight into these relationships through the loading plot in Fig. 13 of the PCA model on the augmented matrix XY_{new}. Fig. 13 clearly shows how the rubber raw material properties (RawP1 to RawP3) are correlated with the projected mixture properties (MixP1–MixP8). Most of these correlations in this figure are quite consistent with the knowledge of the experienced polymer chemists.

Table 4
R2 and Q2 for each property using PLS (direct approach and new approach)

		RawP1	RawP2	RawP3
Direct approach	R2(%)	39.5	81.7	86.9
	Q2(%)	8.0	68.5	77.2
New approach	R2(%)	99.4	98.7	99.6
	Q2(%)	91.2	76.6	93.8

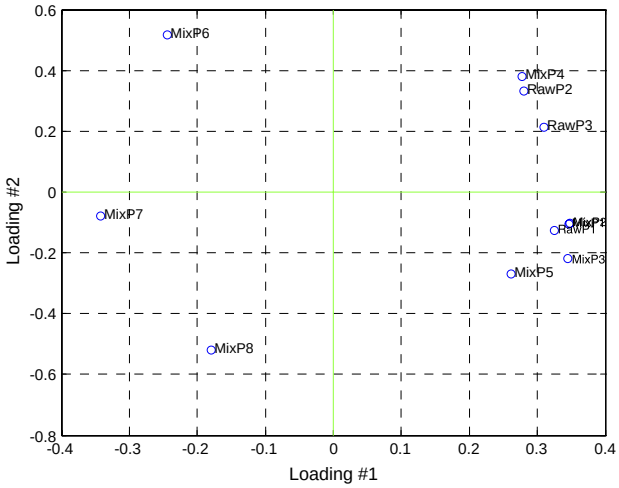


Fig. 13. Loading plot of PCA model for XY_{new} (3 raw material property variables: RawP1–RawP3, 8 projected mixture material property variables: MixP1–MixP8).

7. Conclusion

This paper presents latent variable approaches that exploit auxiliary data, for estimating missing data in a matrix. Two approaches to incorporating the auxiliary information are presented; a traditional multi-block approach and a new projection approach. Both approaches greatly improve the estimates of the missing raw material data in industrial polymer blending problems, when compared to the direct approach that uses only the original data matrix itself. The new projection approach has several advantages over the traditional multi-block approach. It is not as restrictive in the type of auxiliary data that can be used, it provides additional insight into the relationships among the raw material properties and the auxiliary blend properties, and it gives slightly better estimates for the missing data. The approaches can be easily used in other blending problems such as encountered in pharmaceutical, food and biotechnology industries, and the concept can be extended to many other types of missing data problems where auxiliary data is available.

References

[1] Francisco Arteaga, Alberto Ferrer, Dealing with missing data in MSPC: several methods, different interpretations, some examples, Journal of Chemometrics 16 (2002) 408–418.

[2] Philip R.C. Nelson, "The Treatment of Missing Measurements in PCA and PLS Models". PhD thesis in McMaster University, 2002.
[3] Philip R.C. Nelson, Paul A. Taylor, John F. MacGregor, Missing data methods in PCA and PLS: score calculations with incomplete observations, Chemometrics and Intelligent Laboratory Systems 42 (1996) 121039.
[4] Philip R.C. Nelson, Paul A. Taylor, John F. MacGregor, Impact of missing data, Chemometrics and Intelligent Laboratory Systems (2004) (in press).
[5] Nouna Kettaneh-Wold, Analysis of mixture data with partial least squares, Chemometrics and Intelligent Laboratory Systems 14 (1992) 57–69.
[6] B. Walczak, D.L. Massart, Dealing with missing data: Part I, Chemometrics and Intelligent Laboratory Systems 58 (2001) 1117.
[7] B. Walczak, D.L. Massart, Dealing with missing data: Part II, Chemometrics and Intelligent Laboratory Systems 58 (2001) 29–42.
[8] S. Wold, Nonlinear estimation by iterative least squares procedure, in: F. David (Ed.), Research papers in statistics, Wiley, New York, 1966, pp. 411–444.
[9] Donald B. Rubin, Roderick J.A. Little, Statistical Analysis with Missing Data, Second edition, Wiley Inter-Science, 2002.