AIChE

Predictive Modeling of Pharmaceutical Processes with Missing and Noisy Data

Fani Boukouvala, Fernando J. Muzzio, and Marianthi G. Ierapetritou
Dept. of Chemical and Biochemical Engineering, Rutgers University, Piscataway, NJ 08854

DOI 10.1002/aic.12203
Published online March 26, 2010 in Wiley Online Library (wileyonlinelibrary.com).

Lack of knowledge of the first principles, that describe the behavior of processed particulate mixtures, has created significant attention to data-driven models for characterizing the performance of pharmaceutical processes—which are often treated as black—box operations. Uncertainty contained in the experimental data sets, however, can decrease the quality of the produced predictive models. In this work, the effect of missing and noisy data on the predictive capability of surrogate modeling methodologies such as Kriging and Response Surface Method (RSM) is evaluated. The key areas that affect the final error of prediction and the computational efficiency of the algorithm were found to be: (a) the method used to assign initial estimate values to the missing elements and (b) the iterative procedure used to further improve these initial estimates. The proposed approach includes the combination of the most appropriate initialization technique and the Expectation Maximization Principal Component Analysis algorithm to impute missing elements and minimize noise. Comparative analysis of the use of different initial imputation techniques such as mean, matching procedure, and a Kriging-based approach proves that the two former used approaches give more accurate, "warm-start" estimates of the missing data points that can significantly reduce computational time requirements. Experimental data from two case studies of different unit operations of the pharmaceutical powder tablet production process (feeding and mixing) are used as examples to illustrate the performance of the proposed methodology. Results show that by introducing an extra imputation step, the pseudo complete data sets created, produce very accurate predictive responses, whereas discarding incomplete observations leads to loss of valuable information and distortion of the predictive response. Results are also given for different percentages of missing data and different missing patterns.
AIChE J, 56: 2860–2872, 2010
Keywords: missing and noisy data, Kriging, response surface, pharmaceutical, data-driven models, imputation, EM-PCA

Introduction

Current trends in pharmaceutical product development focus on the fundamental understanding and characterization of all process unit operations, which will lead to the development of predictive models to minimize the variability in product performance. The pharmaceutical industry is one of the most characteristic examples where high quality products—with strict predefined performance specifications—must be produced. The successful completion of this task is obstructed by the fact that mechanical and physiochemical properties of Active Pharmaceutical Ingredients (APIs) and excipients used are not completely understood.[1] Other industrial fields (food, ceramics, catalyst manufacturing, etc.)—

Correspondence concerning this article should be addressed to M. G. Ierapetritou at marianth@soemail.rutgers.edu.

that deal with processing of materials in powder form—are faced with the same challenging issue simply because powders share both characteristics of fluids and solids. In these types of processes it is very difficult to model the behavior of granular materials using first principles without making a series of assumptions and introducing a relative amount of uncertainty. This is mainly the reason why data driven modeling methods have attracted a lot of attention in these types of studies. Through them, the system input and output variables are connected by using the available experimental data and only a limited knowledge of the physical behavior of the system. In this case, large multivariate data matrices are the resources for creating the predictive models which will capture the effects of multiple variables, to the final desired output—as well as interactions and correlations between variables. Specifically, the data sets contain rows—which represent observations or cases—and columns—which represent the variables measured at different levels.

In literature, several other approaches have been proposed to produce models for individual pharmaceutical process operations. Methods such as Monte Carlo simulations, particle-dynamic simulations, heuristic models, and kinetic theory models have been proposed to characterize powder flow during the process of mixing using continuous blenders.[2–4] A number of articles have used the discrete-element method (DEM) to model the behavior of granular flow in continuous mixers.[5] The drawback to these simulations, however, is their computational efficiency.[6] Work has also been done to improve the design of feeders,[7,8] however, limited attention has been directed at the analytical study of feeder performance in terms of the input process variables and operating conditions. Unlike all the above examples of first principle-based models, data driven methods were used by Jia et al.[9] to develop a less computationally expensive model based on experimental data alone. Specifically, Kriging and Response Surface methodologies were used to predict the output flow variability of a loss-in-weight feeder and the residence time distribution of a continuous mixer.

Response Surface Methodology (RSM) and Kriging are both types of data-based methods which can be differentiated since RSM produces noninterpolating surfaces (sum of squares error from a predefined function is minimized), while Kriging produces interpolating surfaces (passing through all the experimental points). A review and comparison of non and interpolating response surface methodologies is carried out in[10] concluding that interpolating methods can more often capture the true underlying function of the data, while noninterpolating (usually fitted quadratic surfaces) can sometimes fail and can result to an unreliable optimum value. Both methods have captured a great amount of attention lately since they are fairly simple and computationally efficient optimization methods.[11–14] In this work, both methods are used for producing predictive models for each pharmaceutical case study so as to compare the results and accuracy of each method. The theory of these two methods is briefly discussed in the next section.

The quality of a response surface produced using surrogate-based methods, however, strongly depends on the quality of the data sets. In addition, most of the statistical techniques for developing data-driven models require complete data sets. Unfortunately, however, experimental data sets are strongly susceptible to many forms of uncertainty such as missing information and noise. Thus, it is very important to understand the effect of the aforementioned types of uncertainty in the predictive ability of the data-driven models.

The rest of the article is organized as follows. In section "Data-Based Models," the theory behind the data-driven modeling techniques used is described. "Modeling with Missing and Noisy Data," is an introduction to modeling with missing and noisy data along with a literature review of proposed methods for treating these problems. A detailed explanation of the proposed algorithm used to handle the problematic data sets along with the reasons for choosing this approach for the specific cases of pharmaceutical experimental data is described in "Proposed (EM-PCA) Algorithm." Different methodologies which are used for calculating the initial approximations of the hidden states are also described in this section. Two case studies from pharmaceutical manufacturing are presented in Section "Results" together with the analysis of the results obtained, and the article concludes with the next Section.

Data-Based Models

Response surface methodology

RSM was first introduced by Box and Wilson in 1951[15] and is a tool that has been widely used for the optimization of noisy processes. RSM is a local optimization technique whereby an optimum is found after sequential optimization of localized sampling-based models.[16] There are three basic steps to the algorithm: (1) specification of a sampling set within the local region, usually accomplished with design of experiment tools (DOE), (2) construction of a local model centered at a nominal sampling point, and (3) model optimization with respect to the local region to determine the location at which process improvement is maximized. There are two major questions associated with model construction, namely (1) spatial location of sampling points and (2) quantification of model uncertainty. To address the first issue, DOE tools are applied. The experimental design is defined as the specification of a number of treatment levels for each input variable, the experimental units by which responses are measured, and the mechanism by which treatments are assigned to units.

For a problem containing n continuous input variables, an n-dimensional quadratic polynomial is used as the local model since quadratic behavior describes the mathematical geometry in the neighborhood of an optimum. Model accuracy can be improved if bilinear terms capturing the interaction effects between two inputs may are also incorporated into the local model. A general second-order response surface model has the following form:

$$z = \beta_0 + \sum_j \beta_j x_j + \sum_{i<j} \beta_{ij} x_i x_j + \sum_j \beta_{jj} x_j^2 \qquad (1)$$

where x_j are input variables, β_0, β_j, β_{ij}, and β_{jj} are model coefficients, and z represents the response that describes the predicted output behavior.

Kriging methodology

Kriging was first developed as an inverse distance weighting method to describe the spatial distribution of mineral

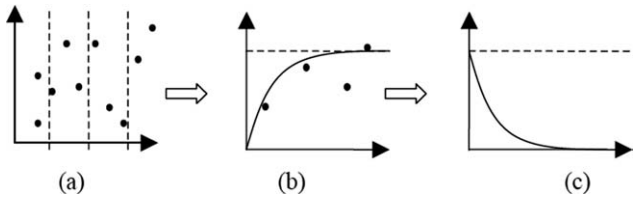

Figure 1. (a) Variogram model, (b) semivariance, and (c) covariance plot.

deposits.[17,18] This method has attracted a lot of attention recently due to it's capability of modeling complex functions and providing error estimates.[9] The prediction f_{pred} is expressed as a weighted sum of the observed function values at sampling points that fall within a set interval around the point that is predicted. The basic idea of Kriging is that a function value for a sampling point located close to the test point is generally weighted more heavily in contrast to the function value corresponding to a sampling point located farther away. Lower weight is placed on function values whose sampling points are clustered together to minimize the possibility of generating biased estimates. Since a variance for each test point is also calculated, regions where subsequent sampling is required can be linked to a high variance at the regional points.

The first step of the Kriging methodology is the determination of variogram coefficients from an experimental sampling set consisting of N sampling points. The variogram is a quantitative descriptive statistic that graphically characterizes data set roughness (and the information obtained complements that which is obtained using histograms and other common descriptive statistics). The variogram coefficients are determined as follows. First, $(N)(N-1)/2$ squared function differences are obtained for each sampling pair. The squared function differences are then plotted with respect to the L^2-norm sampling pair distance as illustrated in Figure 1a.

As shown in Figure 1a, the best variogram model to use may not immediately be apparent. Data smoothing is used to improve the fit by replacing clustered scatterpoints falling within an interval $[h_i-\text{tol}, h_i+\text{tol}]$, with average values defined as semivariances. The semivariance $[\gamma(h)]$ is determined according to the equation:

$$\gamma(h) = \frac{1}{2N(h)} \sum_{N(h)} [f(x_i) - f(x_j)]^2 \tag{2}$$

where $N(h)$ is the number of sampling pairs whose Euclidean distance falls within the range $[h_i-\text{tol}, h_i+\text{tol}]$. Variogram model coefficients are then obtained from regression of the semivariance scatterpoints, to one of the five elementary types: spherical, Gaussian, exponential, power, or linear. It is considered that the one that captures the trend of the semivariance best is the one whose least square error is the lowest. The covariance function which is a complementary function of the semivariance is used to calculate the Kriging weights (w_i) by solving the system:

$$\begin{bmatrix} \text{Cov}(d_{1,1}) & \cdots & \text{Cov}(d_{1,M}) & 1 \\ \vdots & \ddots & \vdots & \vdots \\ \text{Cov}(d_{M,1}) & \cdots & \text{Cov}(d_{M,M}) & 1 \\ 1 & \cdots & 1 & 0 \end{bmatrix} \times \begin{bmatrix} w_1 \\ \vdots \\ w_M \\ \lambda \end{bmatrix} = \begin{bmatrix} \text{Cov}(d_{1,k}) \\ \vdots \\ \text{Cov}(d_{M,k}) \\ 1 \end{bmatrix} \tag{3}$$

where d_{ij} is the distance between sampling point x_i and sampling point x_j, and d_{ik} is the distance between sampling point x_i and test point x_k. Similarly, $\text{Cov}(d_{ij})$ and $\text{Cov}(d_{ik})$ represent the modeled covariances between sampled function data whose corresponding input vectors are a distance x_i-x_j or x_i-x_k apart, respectively. The Kriging prediction f_k is then evaluated by the following form:

$$f_k = \sum_{i=1}^{M} w_i f_i \tag{4}$$

where w_i and f_i represent the weight and observed value at sampling point i, respectively. For each test point x_k, a variance σ_k is also obtained as follows:

$$\sigma_k^2 = \sigma_{\max}^2 - \sum_{i=1}^{M} w_i \text{Cov}(d_{ik}) - \lambda \tag{5}$$

where $\text{Cov}(d_{ik})$ corresponds to the right hand side of Eq. 3.

Additional experiments are conducted at test points which have any of the following characteristics: (1) high variance, (2) high difference between Kriging predictions for consecutive iterations, and (3) minimum prediction values, or the set of points in which the model estimate is lower than the corresponding estimates obtained for model predictions at a set of nearest-neighbor sampling vectors. The Kriging procedure stops when the average Kriging prediction values for the current iteration do not change significantly relative to the average Kriging value for the previous iteration.[9]

Modeling with Missing and Noisy Data

Huh et al.[19] identify that the most important features of data quality are: accuracy, consistency, completeness, and currency. In this work, the effect of the first three aspects is considered, since all of the data sets used are up-to-date. The problem of incomplete observations may often arise due to insufficient sampling, sensor failures, or even when data fusion techniques are required to combine information from different sources. Missingness is addressed very often in the pharmaceutical engineering research not only due to the aforementioned reasons but also because there are so many different possible combinations of input and output variables to be measured that it is sometimes impossible to gather all the required information.[20]

By simply discarding the incomplete observations, valuable information for subsequent analysis of the data can be lost. Additional uncertainty can exist in any experimental data set in the form of noise. The main causes of noisy sampling are data acquisition errors and the effects of uncontrolled external factors. In fact, it is very common that a data point can contain so much noise that it is better to discard it and treat is as a missing observation.

In literature, many different methods exist for handling incomplete data sets, while the appropriateness and efficiency of each method depends on (1) the existence and type of a missing data mechanism, (2) amount of missing data, and (3) the nature of the matrix. Little and Rubin[21] have classified the mechanisms by which missing information occurs in data. When a mechanism that causes the

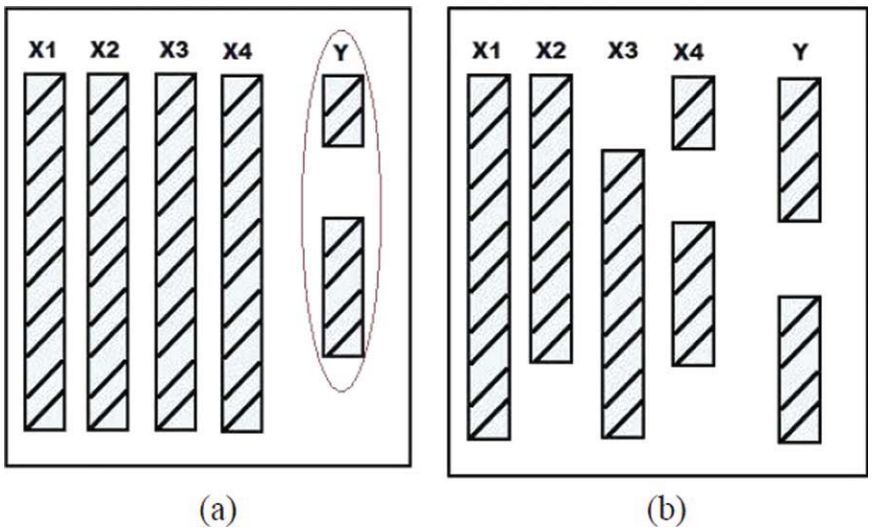

Figure 2. (a) Univariate missing pattern and (b) general multivariate missing pattern.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

missing data is known, then it should be assessed whether or not it could be ignored without affecting the results of the analysis. In most cases, it is assumed that the missingness mechanism is ignorable, which means that missing data is completely random. This is the case for all the case studies addressed in this work. Missing data structures can be univariate when they refer to a single outcome variable or multivariate. Figure 2a shows the simplest univariate case, where data is only missing from the measured response and it can be handled by simple imputation techniques. As the missing data patterns get more complicated, (Figure 2b) however, more sophisticated and complex methods are required to be used.

The methods that appear in the literature that deal with cases of missing data can be grouped into the following categories: (1) procedures based on completely recorded units, (2) single imputation procedures, (3) multiple imputation procedures, and (4) model-based imputation procedures.[21] The method in the first class discards any incomplete observation. This is a simple method, which is efficient only when the amount of missing data is small, otherwise, it can lead to biased and insufficient results.[22] A variety of single imputation methods exist in the literature.[23] The most popular imputation procedures include: (i) hot deck imputation where the most popular recorded units in the sample are substituted, (ii) mean imputation, where means from sets of recorded values are substituted, and (iii) regression imputation, where predicted values are calculated from regression of the known variables of the set. Multiple imputation is proposed by Rubin[24] and it is based on the main idea that for each missing element, several values representing a distribution of possibilities are imputed. This allows calculation of the variance of parameter estimate through the variability of estimates from within each imputed data set. Model-based imputation includes sophisticated models that evaluate parameters which will lead to the estimation of the missing values, e.g., Maximum Likelihood, Gibbs Sampler. These models are based on statistics and involve iterative procedures from which different methods for estimating conditional probability distributions are used. Usually, as the amount of missing data increases, the number of iterations increases exponentially. The most popular method for estimating conditional probabilities from incomplete data sets is the expectation maximization (EM) algorithm.

The EM algorithm[25,26] is an iterative procedure that is used as a single model-based imputation technique for missing data cases. Even if the exact values of the imputed points are not calculated with EM algorithm, parameters describing the data set—such as mean or covariance (μ, σ) are approximated—in the sense of maximizing the likelihood of the imputed values of the missing observations. Each iteration consists of an expectation step (E-step) where the conditional expectation of the sufficient statistics given the observed data and the current estimated parameters is calculated, and a maximization step (M-step) where Maximum Likelihood (ML) estimation is performed as if there were no missing data.

The EM algorithm is closely related to ML estimation where it is assumed that the distribution of a data set (X) can be described by a probability density function $p(X|\Theta)$ where if p is a Gaussian, $\Theta = (\mu, \sigma)$ corresponds to the mean and covariance of the data set. Any function proportional to p, is called a likelihood function and it is always a function of parameters for given X. The goal of a ML problem is to find the right Θ that will maximize the likelihood function. Solving ML problems can be easy. Specifically, if the likelihood function is differentiable, the maximization problem is solved by calculating the derivative and finding the parameter Θ, that sets it to zero. In general, however, explicit ML estimates cannot be calculated—due to the complexity of the statistics describing the data set—and thus iterative approaches must be applied (EM algorithm). The advantage of this algorithm is that it converges reliably since the maximum likelihood estimate increases in each iteration, but it converges quite slowly when the amount of missing data exceeds a certain threshold.

Many extensions and hybrid forms of the EM algorithm have been applied to analyze incomplete data sets in

different fields. Donnet and Samson[27] used a stochastic approximation EM algorithm (SAEM) to estimate parameters of incomplete data models defined by ordinary differential equations that describe biological processes. An interesting approach is reported in,[26,28–32] where the EM algorithm is combined with Principal Component Analysis (PCA) to produce a hybrid iterative algorithm that deals with three main data analysis issues: (1) missing values imputation, (2) dimensionality reduction, and (3) noise minimization. This methodology is thus proposed to be used to handle data sets of the pharmaceutical manufacturing processes. The properties of the EM-PCA algorithm will be discussed in detail in Section "Proposed (EM-PCA) algorithm."

Several methods, such as Partial Least Squares,[31] Fuzzy B-Splines,[33] and Principal Component Regression[34] are used in literature to handle noisy data and eliminate the effects of imperfect observations. The former, is based on PCA methodology and has a second interesting feature—that of dimensionality reduction—which is very useful when dealing with large data sets.

Proposed (EM-PCA) Algorithm

The hybrid EM-PCA algorithm was first introduced by Roweis,[35] for approximating the Principal Components of extremely large or incomplete data sets to decrease computational cost. In this methodology, the covariance matrix used for calculating principal components is not explicitly calculated, but it is approximated through the EM algorithm. It is an iterative procedure, during which each time the incomplete data set is completed with estimates (imputations), PCA is performed, and a new matrix of the significant principal components is constructed. From this matrix, new predicted values of the missing values are obtained and the procedure is repeated until the imputed values do not change significantly in the preceding iteration.

PCA is a very popular method for handling and analyzing large data sets and it is based on the idea of finding a lower dimensional set of vectors which are able to describe the data matrix.[36] It is used to develop models from data sets composed of observations on large numbers of highly correlated variables. PCA is used for dimensionality reduction since it helps to identify the most important factors when a problem is multidimensional. PCA is especially useful in pharmaceutical processes data analysis where there is always a large number of input variables, but by transforming into PC space, the important components are identified and kept, while the rest are eliminated. Following this approach, noise minimization can be achieved since the Principal Components that do not contain significant information (variability) are eliminated to smooth out the reconstructed data and reveal the significant relations between the variables. PCA can also be viewed as a minimization problem where the optimum matrix C (principal components or eigenvectors) must be found to transform matrix y into matrix of lower dimension x minimizing noise ε (Eq. 6).

$$y = Cx + \varepsilon \tag{6}$$

PCA does not substitute statistical analysis such as analysis of variance techniques (ANOVA), which is used for preliminary analysis of any experimental data[37] presents a combined (ANOVA-PCA) methodology to improve the detection of significant factors.

Despite all the positive features of PCA, there are a few limitations. Specifically, regular PCA methodology does not allow the calculation of the optimum principal component space when some elements are missing. This shortcoming, however, can be eliminated through the use of EM-PCA through which the covariance structure of the data is approximated through the following iterative procedure:

Step 1: Initialize the Principal Component Space

Step 2: Project the original data set using desired amount of Principal Components. Replace missing values with their approximations while observed elements remain unchanged. (Expectation Step)

Step 3: Fix values of missing elements and calculate new PC space based on the pseudo-completed data set. (Maximization Step)

Step 4: Check if missing element values have changed significantly since last iteration. Go to Step 2 if this convergence criterion is not satisfied.

Effect of Initial Guess

The most common method for imputing an initial value of all the missing points is the mean imputation methodology. In this simple procedure, the average of all observed points of each column is substituted as an initial guess for all missing points. The nature of the data sets produced from designed experiments, however, has some general characteristics that make the columns of independent variables (inputs) take values from a number of set/discrete levels, for which each factor is investigated. Thus, substituting any missing observations as the mean of each column, might lead to infeasible imputation values for the input variables. In an effort to make this initial guess better and feasible, the following approaches are proposed.

Matching procedure

For each incomplete observation, the most similar complete observation is found, and any missing points of the incomplete observation are substituted with the observation of the most similar complete point. To identify the measure of similarity, the sum of absolute differences between the observed elements of the same column is calculated. To eliminate the large effects of variables that have large ranges, it is made certain that all data sets used are always normalized.

Kriging-based imputation

This method uses the Kriging algorithm as an imputation technique. To apply this method, however, an extra requirement must be satisfied: each incomplete observation must contain a maximum number of one missing elements. This restriction does not allow the imputation of more than one missing value in a single row of the data matrix. It is an online imputation technique, where for each incomplete observation the original data set is reconstructed to have as a response the column with the missing element. Following the procedure described in Section "Kriging Methodology,"

the sampling set is defined as the set of complete observations, while the test set is the current incomplete observation. The data set is reconstructed back to its original form containing the new pseudo-completed observation and this iterative procedure is repeated for all the incomplete observations.

As shown in the Results Section, the two above described methods not only give better approximations of the initial guesses of the missing elements but also reduce the number of iterations that the EM-PCA algorithm requires to converge.

Cross Validation

Evaluating the predictive ability of any model is the key to any predictive modeling procedure. Cross validation is one of the most commonly used model evaluation methods, as it performs better than the full-sample validation method where the performance of the model is tested using a different data set. The general idea is to split the data into subsets such that the analysis is initially performed on the first subset, while the other subset is used for validating the model. Many different methods for performing cross validation exist in literature such as *k*-nearest neighbor or leave-one-out techniques.[38] In this study, *k*-fold cross validation is performed in two of the stages of our approach: (1) During the process of random deletion of data points to examine whether the location of the deleted data points affects the imputation accuracy and (2) During the development of the predictive models with Kriging and RSM algorithms, to test the robustness of the final models for each case study. The main idea of *k*-fold cross validation is that in each fold, a different training set is used in order for the rest of the data to be either imputed or predicted.

Results

To test the results and behavior of the available missing-value-treatment methods, pseudo incomplete data sets are created, by random deletion of observations from originally complete matrices. The next step is simply the "treatment" of the incomplete data sets to either ignore the incomplete cases or impute the missing points and then the resulting matrix is used as the input of the data-driven predictive model. The error of final prediction when: (1) using complete data sets, (2) discarding any incomplete observations, and (3) imputing missing data points using the proposed approaches are compared and the results are reported.

Case Study 1: Powder loss-in-weight feeder (K35- by Ktron)

For this case study, experimental data were generated using loss-in-weight feeders provided by Ktron. The input variables are comprised of the set feed rate of the powder mixture and the powder flow index. FMC Avicel PH102 and ceolus KG802 were used for producing the powder mixtures used in this experiment. The difference in these materials is quantified by their flow index (38 and 49.2, respectively). This variable is connected to the flowability of the powder mixtures, where a high value can be translated to a poorly flowing powder.[39] Samples were taken at every 1-s time intervals for a total duration of 30 min to get the second variable—which is the set feed rate of the feeder. Specifically, the actual flow rate can be simply derived dividing the mass by the corresponding time interval. The output of this model is the standard deviation from the mean flow rate which is considered to be a measure of the flow variability. The purpose of producing such a model is to find the optimum range of input variables to achieve the minimum output flow variability.

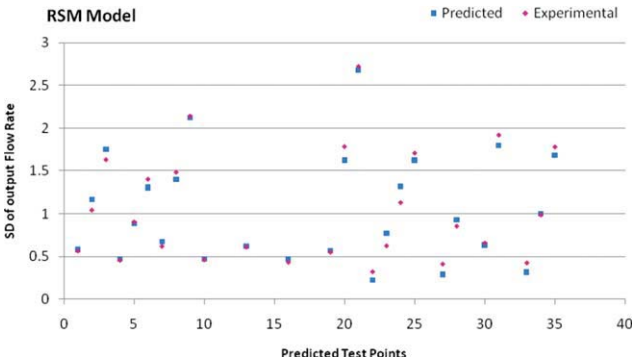

Figure 3. RSM predicted points vs. experimental points for Case Study 1.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Complete data scenario (no missing elements)

The complete data set consists of three columns (two input variables and one output) and a total of 35 rows (observations). The final errors of prediction were calculated as the average of the residuals of each predicted point compared to the experimental value, calculated using the following formula:

$$\text{Error} = \frac{1}{N}\sum_{i=1}^{N}\frac{\text{abs}(y^{\text{i}}_{\text{pred}} - y^{i}_{\text{exp}})}{y^{i}_{\text{exp}}} \tag{7}$$

where N is the number of sampling points.

For RSM, the average error of prediction is 15.31%, while with Kriging the error is 10.71%. Results for the complete cases are shown in Figures 3–5 that show the predicted points using RSM and Kriging vs. the actual experimental points.

Figure 4 represents the calculated response surface obtained by RSM using the calculated coefficients that provide the best possible fit for the experimental data points. The advantage of RSM is that it provides an analytical surface—compared to Kriging which approximates the value and a variance of the output at a given set of operating conditions. To assess the accuracy of the Kriging predictions, a fraction of the available experimental points are used to predict the remaining points that are plotted in Figure 5. The explicit form of the response surface when using RSM to fit a full quadratic model (Eq. 1) is the following:

$$z = 0.0316x_1 - 0.0363x_2 + 0.0012x_1x_2 - 0.0007x_1^2 \tag{8}$$

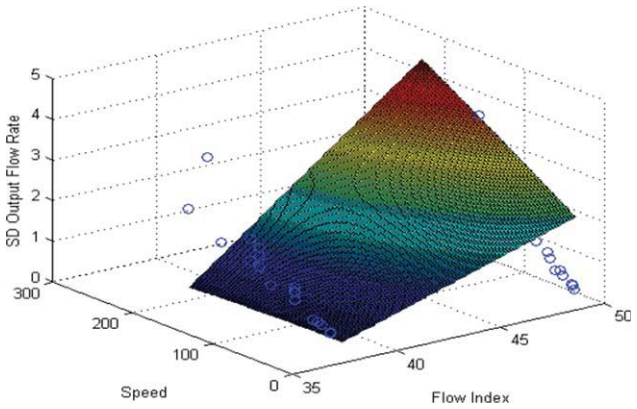

Figure 4. Response surface vs. experimental points for Case Study 1.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

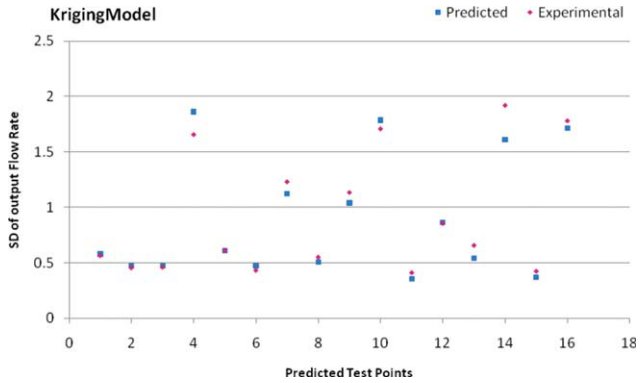

Figure 5. Kriging predicted points vs. experimental points for Case Study 1.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

where z is the output, x_1 represents the flow index, and x_2 represents the speed.

Incomplete data scenarios

To compare the two approaches: (1) when missing elements are imputed and (2) when observations with missing elements are discarded, their effects on the accuracy of the predictive models are compared through the final average prediction errors (Eq. 7) of RSM (Table 1). The response surfaces are also compared in Figure 6, where it is obvious that the treated data set is almost identical to the original response surface. Specifically, as the amount of missing data increases, the surfaces that are produced by discarding the incomplete observations grow farther apart and follow different trends. This is very well illustrated in Figure 6c, where the case in which missing elements are discarded has no similarity with the other two responses.

The goal of this case study is to identify the optimum operating conditions (Speed) and material properties (Flow Index) that will minimize the variability in the output flow rate of a feeder. Table 2 contains the objective values calculated from solving the NLP problem of minimizing the SD of the output flow rate as a function of the two input variables. The optimization problem is solved using the software GAMS. It is obvious that the objective values are closer to the real value when using the proposed approaches to complete the missing values.

In this section, it is shown that the quality of the produced response surfaces is affected in the presence of missing data. The resulting final response surface obtained by the pseudo-completed data set, when using the proposed approach, is compared to the surface obtained by the data set from which all incomplete rows are deleted. As expected, the amount of missing data does affect the final error of prediction of the derived model. The lowest errors were found for using the matching procedure as an initial guess methodology and EM-PCA as the iterative procedure for the random missing data pattern. On the contrary, when the missing data pattern is univariate, the Kriging-based approach is found to be the most efficient imputation method for minimizing the final prediction error. The results reported in the tables and figures of this section are for the aforementioned approaches, based on their suitability for each case.

Effect of initial guess

The most common method for estimating an initial guess for the iterative procedure in literature has been the mean value estimation. However, the main disadvantage of this approach is that the estimates are not only very different to the actual values but are also infeasible since the Flow Index can take values from a set of three possible levels (See Appendix and Figure 7). This limitation can be overcome by using the matching procedure—since the values that are used as initial guesses come from the set of feasible levels. In Figure 7, the experimental values of an operating variable (Flow Index) are plotted against their initial imputed

Table 1. Average Prediction Errors (%)

		Univariate Missing Data Pattern (Kriging Imputation)		Random Missing Data Pattern (EM-PCA Imputation)	
		Average Error (%) when Incomplete Observations Discarded	Average Error (%) when Missing Elements are Imputed	Average Error (%) when Incomplete Observations Discarded	Average Error (%) when Missing Elements are Imputed
Amount missing (%)	0	15.3	–	15.3	–
	10	27.8	17.7	18.4	16.2
	20	30.2	18.9	26.3	18.3
	30	35.9	20.8	37.6	20.4

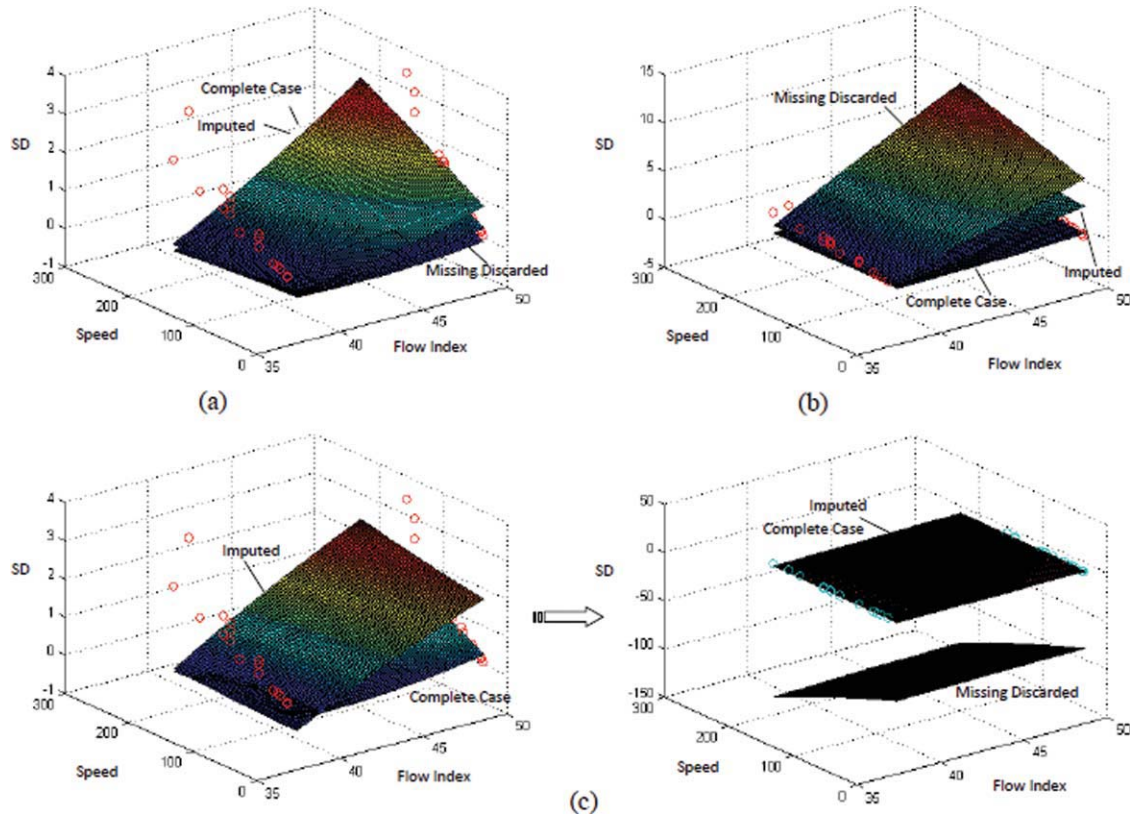

Figure 6. Response surfaces of (a) 10% missing data, (b) 20% missing data, and (c) 30% missing data in a multivariate random pattern.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

estimates using the two methods (matching and mean). As it can be seen, the matching procedure can be very accurate in approximating the real values of the missing elements, and even if it sometimes fails, the iterative procedure that follows does improve the accuracy of the imputation. Another advantage of providing an accurate warm start to the EM-PCA algorithm is the reduction of computational time. It is observed that not only the initial error of prediction but also the iterations necessary for EM-PCA to converge can be significantly reduced when the matching procedure is used (Table 3).

Case Study 2: Effect of powder mixing shear rate and composition on LIBS intensity measurements

In this case study, Laser Induced Breakdown Spectroscopy (LIBS) measurements are used to quantify the amount of lubricant in the tablet. The two input variables in this case are the concentration of MgSt in the powder mixture and the shear intensity in the Mixer. MgSt is a commonly used lubricant because of its advantages of facilitating tablet ejection and providing internal failure points during compression. Distribution of MgSt in the tablet affects the compression process and properties of tablets, such as tablet hardness, disintegration, and drug release rate. LIBS is a destructive procedure that provides, however, measurements from many sites (positions) of the tablet. If it can be shown that LIBS intensity is a function of the amount of MgSt in the tablet, the variance of the measurements at different sites can be used as a measure of tablet uniformity.

However, there is a lack of understanding of the effect of process parameters on the output measurement of LIBS. The main target of this study is to model the effect of shear intensity in the mixer and the concentration of MgSt to a

Table 2. Objective Values for Case Study 1

		Univariate Missing Data Pattern (Kriging Imputation)		Random Missing Data Pattern (EM-PCA Imputation)	
		Objective Value when Incomplete Observations Discarded	Objective Value when Missing Elements are Imputed	Objective Value when Incomplete Observations Discarded	Objective Value when Missing Elements are Imputed
Amount missing (%)	0	0.155	–	0.155	–
	10	0.109	0.150	0.119	0.148
	20	0.310	0.145	0.305	0.133
	30	0.324	0.163	0.301	0.183

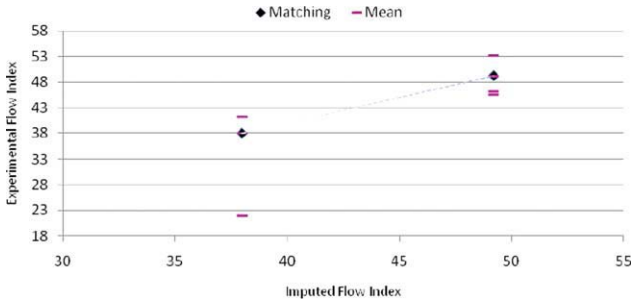

Figure 7. Mean imputation vs. matching procedure initial guesses of missing flow index.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

measurement that gives valuable information about a property of a produced tablet.

Complete data scenario (no missing elements)

The complete data set consists of three columns (two input variables and one output) and a total of 179 rows (observations). The final errors of prediction were calculated using Eq. 7. This data set is a product of a designed experiment where 10 replicates were performed for each measurement.

The two models produced using Kriging and RSM illustrate that the output can be predicted with great accuracy. The average errors of prediction are very similar, 3.42% when using RSM and 4.03% when using Kriging. Figures 8–10 show the results of this case study, whereas Eq. 9 represents the analytical form of the response surface of Figure 9 produced by fitting Eq. 1 to the experimental data.

$$z = 9012.903 + 26.842x_1 + 14675.15x_2 - 0.828x_1x_2 - 0.029x_1^2 - 664.282x_2^2 \quad (9)$$

where z represents the LIBS intensity, x_1 the shear intensity, and x_2 is the MgSt concentration.

Incomplete data scenarios

To compare the two approaches: (1) when missing elements are imputed and (2) when observations with missing elements are discarded, their effects on the accuracy of the predictive models are compared. Table 4 contains the final average prediction errors (Eq. 7) of RSM, whereas in Figure 11, the calculated response surfaces can be compared for each of the different approaches.

Table 3. Effect of Initial Guess

		Mean Imputation		Matching Procedure Imputation	
		Initial Error (%)	Iterations	Initial Error	Iterations
Amount missing (%)	10	30.45	36,143	11.14	31,948
	20	43.18	58,560	24.99	48,791
	30	43.59	73,914	24.63	16,723

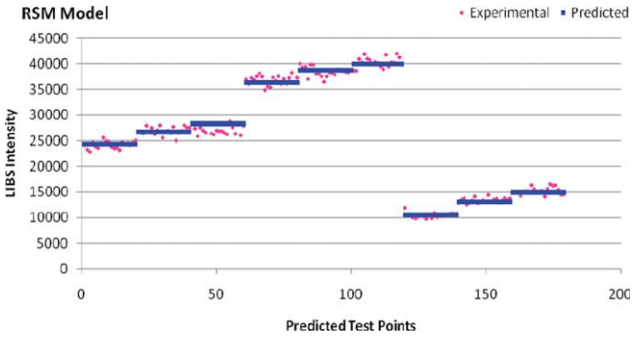

Figure 8. RSM predicted points vs. experimental points for Case Study 2.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

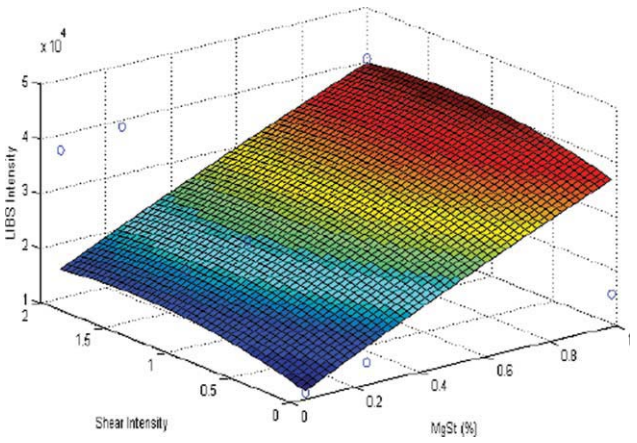

Figure 9. Response surface vs. experimental points for Case Study 2.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

From Figure 11, it can be observed that the response surfaces calculated in all approaches do not differ as much as in Case Study 1 and this can be explained by the fact that the data set in this case study is larger and contains many replicates for each combination of input variables. As a

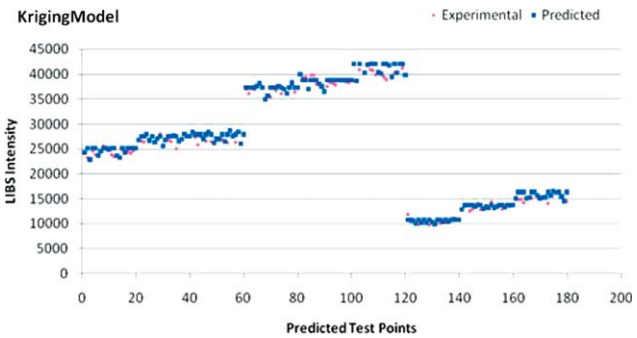

Figure 10. Kriging predicted points vs. experimental points for Case Study 2.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Table 4. Average Prediction Errors (%) for Case Study 2

		Univariate Missing Data Pattern (Kriging Imputation)		Random Missing Data Pattern (EM-PCA Imputation)	
		Average Error (%) when Incomplete Observations Discarded	Average Error (%) when Missing Elements are Imputed	Average Error (%) when Incomplete Observations Discarded	Average Error (%) when Missing Elements are Imputed
Amount missing (%)	0	3.4	–	3.4	–
	10	5.3	4.2	4.3	3.8
	20	7.7	6.8	6.5	5.4
	30	18.3	15.9	3.4	12.2

result, the effect of the missing values becomes less important in general, but increases as the amount of missing data increases. Specifically, for only 10% of incomplete observations, one cannot visually see the differences between the three calculated surfaces (Figure 11a).

It is shown here that replication is a method for minimizing the effect of missing and noisy data. However, the quality of the produced response surfaces is affected in the presence of missing data, as the amount of missingness increases. The resulting final response surface obtained by the pseudo-completed data set, when using the proposed approach, is compared to the surface obtained by the data set from which all incomplete rows are deleted. The lowest errors were found for using the matching procedure as an initial guess methodology and EM-PCA as the iterative procedure for the random missing data pattern, whereas the

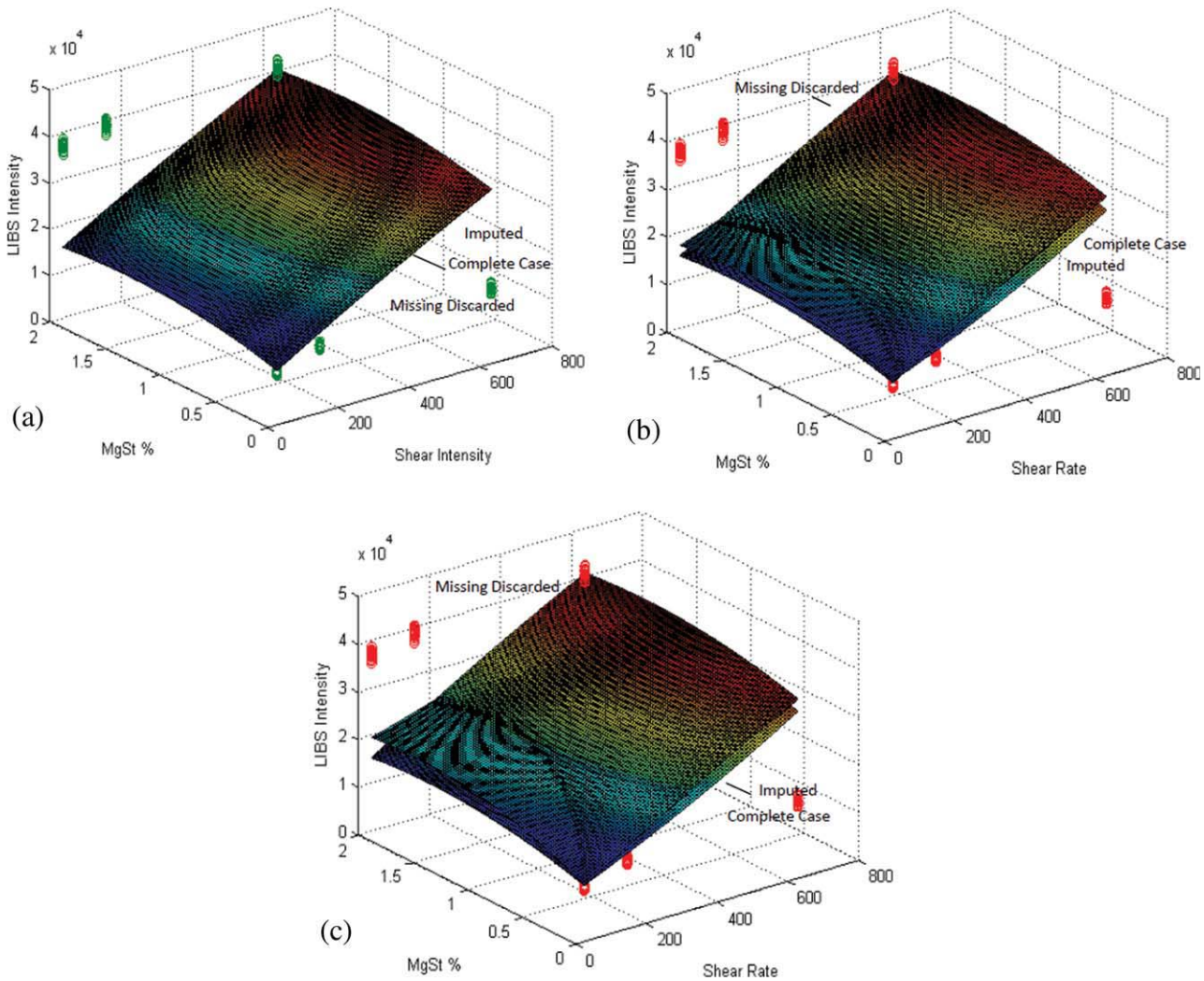

Figure 11. Response surfaces for (a) 10% missing data, (b) 20% missing data, and (c) 30% missing data in multivariate random pattern.

[Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

Table 5. Effect of Initial Guess

		Mean Imputation		Matching Procedure Imputation	
		Initial Error (%)	Iterations	Initial Error	Iterations
Amount missing (%)	10	337.27	80,045	24.63	16,723
	20	369.49	92,301	56.63	34,980
	30	415.10	103,980	82.13	38,124

Kriging-based approach is found to be the most efficient imputation method for minimizing the final prediction error for the univariate pattern. Again, due to their suitability, these results are only reported in Table 4 and Figures 8–11.

Effect of initial guess

In this case study, the effect of the initial guess has a very large influence on the performance of the algorithm (Table 5). In Table 5, it is shown clearly that not only the initial guess error is decreased dramatically when the matching procedure is used but also the number of iterations for convergence of the EM-PCA algorithm. This result supports out initial assumption that the nature of the data set has an effect on the appropriateness of the imputation technique. This data set is very large and has a total number of 10 repetitions for each measurement. This characteristic, makes the matching procedure very efficient since even if some observations are missing, there is a greater chance that a very similar observation exists, from which the missing points can be estimated and imputed. In fact, it is actually observed that this procedure results into a very accurate initial pseudo-completed data set—that is not significantly further improved by the iterative procedure of EM-PCA.

Conclusions

In this work, EM-PCA algorithm is used to handle noisy and incomplete data sets, describing unit process operations of a pharmaceutical tablet production processes, whereas Kriging and RSM are used to produce predictive models for the processes studied in the two case studies and the results of the completely observed cases were compared to the cases of: (1) smaller data sets that are formed by discarding any incomplete observation and (2) data sets of the same size of the original data, with imputed missing values using the proposed approach. Even though it is impossible to predict the exact values of missing information, the pseudo-completed data sets are accurate approximations that produce very similar responses to the original raw data cases for describing the behavior of the processes.

In this work, we have shown that accurate predictive models for unit operations of the tablet manufacturing process can be produced, even when up to 30% of the data is missing—without distortion of the trends and major effects of the data set. Special attention must be given to the choice of the initial guess of the proposed iterative procedure—since the appropriate method can minimize computational time and improve final prediction error.

Future work includes the testing of the proposed approach on larger data sets from pharmaceutical manufacturing to verify its ability as a dimensionality reduction technique. Another issue that has to be addressed during this study is the amount of data that are required to generate an accurate predictive model. Work is devoted to use the proposed idea to address this question which will have a huge impact in reducing the number of experimental conditions that need to be considered and, thus, lead to significant cost reduction.

Acknowledgments

The authors acknowledge the support provided by the ERC (NSF-0504497, NSF-ECC 0540855). They also thank their colleagues Bill Englisch and Aditya Vanarase for providing the experimental data.

Literature Cited

1. Portillo P, Ierapetritou M, Tomassone S, Mc Dade C, Clancy D, Avontuur P, Muzzio F. Quality by design methodology for development and scale-up of batch mixing processes. *J Pharm Innovation*. 2008;3:258–270.
2. Berthiaux H, Marikh K, Mizonov V, Ponomarev D, Barantzeva. Modeling continuous powder mixing by means of the theory of Markov chains. *Part Sci Technol*. 2004;22:379–389.
3. Ghaderi A. On characterization of continuous mixing of particulate materials. *Part Sci Technol*. 2003;21:271–282.
4. Sudah OS, Chester AW, Kowalski JA, Beeckman JW, Muzzio FJ. Quantitative characterization of mixing processes in rotary calciners. *Powder Technol*. 2002;126:166–173.
5. Portillo PM, Muzzio FJ, Ierapetritou MG. Characterizing powder mixing processes utilizing compartment models. *Int J Pharm*. 2006; 320:14–22.
6. Patricia MP, Marianthi FJM, Ierapetritou G. Hybrid DEM-compartment modeling approach for granular mixing. *AIChE J*. 2007;53: 119–128.
7. Gundogdu MY. Design improvements on rotary valve particle feeders used for obtaining suspended airflows. *Powder Technol*. 2004;139:76–80.
8. Reist PC, Taylor L. Development and operation of an improved turntable dust feeder. *Powder Technol*. 2000;107:36–42.
9. Jia Z, Davis E, Muzzio F, Ierapetritou M. Predictive Modeling for Pharmaceutical Processes Using Kriging and Response Surface. *Journal of Pharmaceutical Innovation* 2009;4:174–186.
10. Jones D. A taxonomy of global optimization methods based on response surfaces. *J Global Optim*. 2001;21:345–383.
11. Davis E, Ierapetritou M. A Kriging-based approach to MINLP containing black-box models and noise. *Ind Eng Chem Res*. 2008;47: 6101–6125.
12. Davis E, Ierapetritou M. A Kriging based method for the solution of mixed-integer nonlinear programs containing black-box functions. *J Global Optim*. 2009;43:191–205.
13. Gatti MN, Milocco RH, Giaveno A. Modeling the bacterial oxidation of ferrous iron with Acidithiobacillus ferrooxidans using kriging interpolation. *Hydrometallurgy*. 2003;71:89–96.
14. Hernandez AF, Grover MA, Rita Maria de Brito Alves CAOdNaECB Jr. Stochastic dynamic predictions using Kriging for nanoparticle synthesis. *Comput Aid Chem Eng*. 2009;357–362.
15. Box GEP, Wilson KB. On the experimental attainment of optimum conditions. *J Roy Stat Soc Ser B (Methodological)*. 1951;13:1–45.
16. Raymond HM, Douglas CM. *Response Surface Methodology: Process and Product in Optimization Using Designed Experiments*. New York: John Wiley and Sons, Inc., 1995: 728.
17. Cressie N. *Statistics for Spatial Data (Wiley Series in Probability and Statistics)*. New York: Wiley-Interscience, 1993.
18. Isaaks E. *Applied Geostatistics*. New York: Oxford University Press, 1989.
19. Huh YU, Keller FR, Redman TC, Watkins AR. *Data Quality*. 1990; 32:559–565. DOI: http://dx.doi.org/10.1016/0950-5849(90)90146-I
20. Adams E, Walczak B, Vervaet C, Risha PG, Massart DL. Principal component analysis of dissolution data with missing elements. *Int J Pharm*. 2002;234:169–178.
21. Little RJA, Rubin DB. *Statistical Analysis with Missing Data*, 2nd ed. New York: Wiley, 2002.

22. Dodge Y. *Analysis of Experiments with Missing Data. Probability and Mathematical Statistics*. New York: John Wiley & Sons, 1985.
23. Schafer JL. *Analysis of Incomplete Multivariate Data*, 1st ed. London: Chapman & Hall/CRC, 1997.
24. Rubin DB. Inference and missing data. *Biometrika*. 1976;63:581–592.
25. Moon TK. The expectation-maximization algorithm. *Signal Proc Mag IEEE*. 1996;13:47–60.
26. Walczak B, Massart DL. Dealing with missing data: Part II. *Chem Int Lab Syst*. 2001;58:29–42.
27. Donnet S, Samson A. Estimation of parameters in incomplete data models defined by dynamical systems. *J Stat Plan Infer*. 2007;137: 2815–2831.
28. Andrews DT, Wentzell PD. Applications of maximum likelihood principal component analysis: incomplete data sets and calibration transfer. *Anal Chim Acta*. 1997;350:341–352.
29. Ho P, Silva MCM, Hogg TA. Multiple imputation and maximum likelihood principal component analysis of incomplete multivariate data from a study of the ageing of port. *Chem Int Lab Syst*. 2001;55:1–11.
30. Stanimirova I, Serneels S, Van Espen PJ, Walczak B. How to construct a multiple regression model for data with missing elements and outlying objects. *Anal Chim Acta*. 2007;581:324–332.
31. Walczak B, Massart DL. Dealing with missing data: Part I. *Chem Int Lab Syst*. 2001;58:15–27.
32. Wentzell DP, Andrews TD, Hamilton CD, Faber K, Kowalski RB. Maximum likelihood principal component analysis. *Journal of Chemometrics* 1997;11:339–366.
33. Anile AM, Falcidieno B, Gallo G, Spagnuolo M, Spinello S. Modeling uncertain data with fuzzy B-splines. *Fuzzy Sets Syst*. 2000;113:397–410.
34. Walczak B, Massart DL. Robust principal components regression as a detection tool for outliers. *Chem Int Lab Syst*. 1995;27:41–54.
35. Sam R. EM Algorithms for PCA and SPCA. In Proceedings of the 1997 Conference on Advances in Neural Information Processing Systems 10. Denver, CO: MIT Press, 1998.
36. Jolliffe IT. *Principal Component Analysis*. New York: Springer, 2002.
37. Climaco-Pinto R, Barros AS, Locquet N, Schmidtke L, Rutledge DN. Improving the detection of significant factors using ANOVA-PCA by selective reduction of residual variability. *Anal Chim Acta*. 2009;653:131–142.
38. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. In International Joint Conference On Artificial Intelligence. Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2:1137–1145. Montreal, Canada: 1995. Morgan Kaufmann Publishers.
39. Faqih A, Chaudhuri B, Alexander AW, Davies C, Muzzio FJ, Silvina TM. An experimental/computational approach for examining unconfined cohesive powder flow. *Int J Pharm*. 2006;324:116–127.

Appendix

The experimental data sets used for the two case studies are given below.

Table A1. Case Study 1

	Inputs		Outputs
Point	Input 1: Flow Index	Input 2: Speed	Output: SD Flowrate
1	38	42	0.563042
2	38	105	1.043764
3	38	168	1.630974
4	38	30	0.453161
5	38	75	0.904712
6	38	120	1.402104
7	38	52	0.617771
8	38	130	1.483300
9	38	208	2.141842
10	38	30	0.458309
11	38	75	1.119623
12	38	120	1.655917
13	38	46	0.609757
14	38	130	1.992328
15	38	184	3.612980
16	38	30	0.430068
17	38	75	1.228206
18	38	120	1.861055
19	38	62	0.823138
20	38	155	1.928481
21	38	248	3.024431
22	38	55	0.617719
23	38	137	1.233741
24	38	220	1.947367
25	49.2	31	0.548715
26	49.2	77.5	1.784458
27	49.2	124	2.726952
28	49.2	16	0.318950
29	49.2	40	0.623242
30	49.2	64	1.131594
31	49.2	77.5	1.707953
32	49.2	124	3.265501
33	49.2	18.8	0.408787
34	49.2	47	0.854719
35	49.2	75.2	1.210570

Table A2. Case Study 2

Point	Shear Rate (revs)	MgSt Conc (%)	LIBS intensity1
1	40	1	24207.4
2	40	1	23264.96
3	40	1	22860.71
4	40	1	24610.76
5	40	1	23837.15
6	40	1	23612.67
7	40	1	24517.5
8	40	1	25606.89
9	40	1	24937.81
10	40	1	24861.93
11	40	1	23848.72
12	40	1	23515.67
13	40	1	23627.62
14	40	1	23196.57
15	40	1	24782.41
16	40	1	24273.53
17	40	1	24413.5
18	40	1	24163.63
19	40	1	24635.84
20	40	1	25096.57
21	160	1	26730.54
22	160	1	26554.91
23	160	1	26449.51
24	160	1	27926.23
25	160	1	26695.15
26	160	1	27482.5
27	160	1	26289.89
28	160	1	27101.16
29	160	1	27997.91
30	160	1	25566.45
31	160	1	26753.42
32	160	1	26912.83
33	160	1	26489.06
34	160	1	27710.69
35	160	1	25022.77
36	160	1	26547.02
37	160	1	26888.86
38	160	1	28009.09
39	160	1	27557.38
40	160	1	27522.27
41	640	1	28409.68
42	640	1	27362.09
43	640	1	25850.23
44	640	1	27539.89
45	640	1	26997.02

Table A2. (Continued)

Point	Shear Rate (revs)	MgSt Conc (%)	LIBS intensity1
46	640	1	26596.1
47	640	1	28329.32
48	640	1	26421.45
49	640	1	26219.28
50	640	1	26996.36
51	640	1	26893.7
52	640	1	26878.21
53	640	1	26525.67
54	640	1	26263.56
55	640	1	28741.13
56	640	1	27576.29
57	640	1	26347.94
58	640	1	28392.56
59	640	1	26066.5
60	640	1	27900.12
61	40	2	36984.63
62	40	2	36125.68
63	40	2	37217.58
64	40	2	36898.07
65	40	2	37565.06
66	40	2	38169.62
67	40	2	37497.26
68	40	2	34902.47
69	40	2	35647.56
70	40	2	35435.29
71	40	2	37314.64
72	40	2	36678.76
73	40	2	37599.42
74	40	2	36102.49
75	40	2	37017.96
76	40	2	36101.97
77	40	2	37211.02
78	40	2	38277.73
79	40	2	36412.36
80	40	2	37270.53
81	160	2	40007.59
82	160	2	39265.19
83	160	2	39403.09
84	160	2	37026.29
85	160	2	39839.31
86	160	2	39805.9
87	160	2	38007.71
88	160	2	38065.73
89	160	2	37533.71
90	160	2	36491.58
91	160	2	37533.46
92	160	2	38391.35
93	160	2	38218.28
94	160	2	37861
95	160	2	38890.77
96	160	2	38715.15
97	160	2	38629.31
98	160	2	38463.13
99	160	2	38302.91
100	160	2	38800.14
101	640	2	38689.53
102	640	2	38639.64
103	640	2	40910.95
104	640	2	11952.6
105	640	2	40301.59
106	640	2	41932.06
107	640	2	40974.37
108	640	2	40747.56
109	640	2	39865.14
110	640	2	40356
111	640	2	40130.88
112	640	2	39462.34
113	640	2	38943.04
114	640	2	41797.13
115	640	2	39460.95
116	640	2	40247.03

Table A2. (Continued)

Point	Shear Rate (revs)	MgSt Conc (%)	LIBS intensity1
117	640	2	40230.33
118	640	2	42037.25
119	640	2	41254.84
120	640	2	39838.19
121	40	0.05	11912.22
122	40	0.05	10467.56
123	40	0.05	10581.09
124	40	0.05	9999.59
125	40	0.05	9855.93
126	40	0.05	10187.8
127	40	0.05	10563.07
128	40	0.05	10195.2
129	40	0.05	9751.14
130	40	0.05	10343.94
131	40	0.05	9880.78
132	40	0.05	10882.53
133	40	0.05	10065.94
134	40	0.05	10352.83
135	40	0.05	10254.15
136	40	0.05	10491.59
137	40	0.05	10627.67
138	40	0.05	10808.99
139	40	0.05	10787.01
140	40	0.05	10793.91
141	160	0.05	12849.73
142	160	0.05	13433.1
143	160	0.05	13714.36
144	160	0.05	12551.3
145	160	0.05	12969.43
146	160	0.05	13382.02
147	160	0.05	14091.38
148	160	0.05	12843.29
149	160	0.05	12956.12
150	160	0.05	13337.62
151	160	0.05	13135.12
152	160	0.05	14402.31
153	160	0.05	13124.56
154	160	0.05	13456.24
155	160	0.05	13658.77
156	160	0.05	12938.29
157	160	0.05	13239.78
158	160	0.05	13761.89
159	160	0.05	13407.54
160	160	0.05	13712.8
161	640	0.05	15102.93
162	640	0.05	14857.98
163	640	0.05	14906.57
164	640	0.05	14209.05
165	640	0.05	15010.83
166	640	0.05	15204.89
167	640	0.05	15118.87
168	640	0.05	16332.68
169	640	0.05	15621.95
170	640	0.05	15119.14
171	640	0.05	15181.68
172	640	0.05	15271.55
173	640	0.05	14091.06
174	640	0.05	15675.45
175	640	0.05	16537.62
176	640	0.05	16218.54
177	640	0.05	16326.35
178	640	0.05	15435.5
179	640	0.05	14410.53
180	640	0.05	14474.1

Manuscript received Nov. 26, 2009, and revision received Jan. 25, 2010.