Ensemble Modeling for Data Fusion in Manufacturing Process Scale-up

Ran Jin[1]and Xinwei Deng[2]

[1]*Grado Department of Industrial and Systems Engineering, Virginia Tech*

[2]*Department of Statistics, Virginia Tech*

Abstract: In modern manufacturing scale-up, design of experiments is widely used to identify optimal process settings, followed by production runs to validate the process settings. Both experimental data and observational data are collected in the manufacturing process. However, current methodologies often use a single type of data to model the process. This work presents an innovative method to efficiently model a manufacturing process by integrating the two types of data. We propose an ensemble modeling strategy through the constrained likelihood approach, where the constraints incorporate the sequential nature and inherent features of the two types of data. It therefore achieves better estimation and prediction than the conventional methods. Simulations and a case study of wafer manufacturing are provided to illustrate the merits of the proposed method.

Keywords: Data fusion; Ensemble model; Manufacturing scale-up; Model selection; Nonnegative garrotte; Variation reduction.

1 Introduction

In a product realization cycle, it contains several important steps, including (1) product design, (2) manufacturing process design, (3) manufacturing operation planning, and (4) quality inspection and control. Finally, products are delivered to customers through supply chain systems. The highly dynamic market needs require modern manufacturing to produce customized products with high quality in a timely manner. Therefore, it is crucial to shorten the lead time of the product realization cycle for effectively improving the manufacturing system performance. In order to do so, it is important to shorten the time period in manufacturing scale-up.

Manufacturing scale-up is an important step in product realization. It transfers a pilot operation at experimental scale to manufacturing production at large scale (Parker, 2002). It is generally time consuming to fulfill such a scale-up effort because it requires multiple rounds of adjusting technology, equipment, process settings and so on. In manufacturing scale-up efforts, a typical problem is to optimize the process settings in the large manufacturing scale, given that existing manufacturing equipment are running normally. For example, in a wafer manufacturing scale-up process, design of experiments are conducted to identify optimal settings of the process to improve quality in a lapping process, as shown in Figure 1 (Ning et al., 2012, with authors' permission). Production runs are then conducted after the experiments to validate the optimized settings, i.e., the optimal settings or the settings in the neighborhood of the optimal settings obtained from the experiments will be used to evaluate the quality performance. Such experiment-validation process may take several rounds until the quality requirements of wafers are satisfied. The whole process can take several weeks to finish and cost a large amount of materials and energy. This calls for a pressing need to accelerate the scale-up efforts for effectively reducing the cost and time, while improving the performance of manufacturing process.

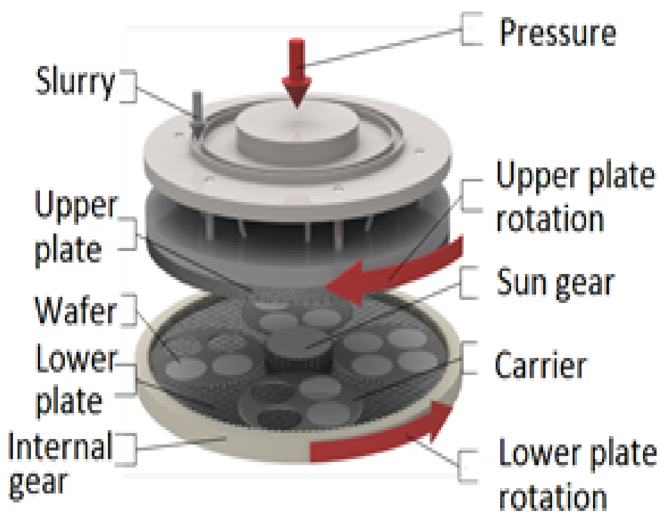


Figure 1: A diagram of the lapping process. Wafers are lapped between the upper and lower plates. Details described in Section 4 (Ning et al., 2012, with authors' permission).

Motivated by the wafer manufacturing scale-up example, one research objective is to quickly obtain an adequate quality-process model for quantifying the relationship between product quality and process variables. Then the model can be used to optimize process settings, thus improving the product quality in manufacturing scale-up. In general, constructing such a quality-process model is expensive in terms of two aspects. First, the process needs both designed experiments and validation production runs, which are time consuming. Design of experiments (DOE) are conducted to identify important process variables and determine the optimized initial recipe. After DOE, validation production runs are performed to get observational (OBS) data for validating the manufacturing process settings. In most cases, the initial recipe from the DOE is used in the validation runs. The actual settings of the process variables in the validation may vary in the neighborhood of the initial recipe due to various reasons. For example, depending on the control precision of manufacturing equipment, the real values of process variables may not always be identical to the targeted settings. Instead, the real values can fluctuate around the targeted settings of the initial recipe. Second, it often needs several rounds of DOE and validation production runs to obtain an adequate model for process optimization. Nominal-the-best or smaller-the-better objectives are usually adopted to determine the optimal process settings. Although both DOE and OBS data are collected, current research mainly targets on analyzing a single type of data. In particular, the DOE data are used for modeling and optimization, while the OBS data are used for validation. If models are obtained based on the two types of data separately, the resulting models may fail to consistently describe the quality-process relationship. The model from the DOE data may have different significant variables and parameter estimations with the model from the OBS data, even if both types of data come from the same manufacturing process. The model based on one type of data can have poor prediction performance on the other type of data. This phenomena has been observed in the latter simulations and case studies. Consequently, it requires additional trial-and-error to conduct more DOE and production runs to optimize manufacturing process settings, which largely increases the lead time and cost for product realization.

In the literature, the two types of data are commonly used to model the manufacturing process, respectively. Regression models based on the DOE data have been developed in different perspectives. Various process optimizations and controls are performed to reduce the variation of quality variables and to improve the yield, such as Robust Parameter Design (RPD) (Wu and Hamada, 2009), RPD based feedforward or feedback controls (Joseph, 2003; Dasgupta and Wu, 2006), and

DOE-based automatic process control (Jin and Ding, 2004; Zhong et al., 2010). These methods have been widely used in discrete part manufacturing, nanostructured material fabrications (Basumallick et al., 2003; Dasgupta et al., 2008) and other applications. Although DOE has been successfully used for manufacturing processes, the high cost of physical experiments prohibits a large number of runs for the modeling and optimization purpose (Shi, 2006). On the other hand, OBS data from production runs are also widely used to model manufacturing systems. For example, in quality engineering, regression-based variation analysis (Fong and Lawless, 1998) is proposed to model the quality-process relationship from OBS data. In Stream-of-Variation theory (Shi, 2006), state space models are constructed to link the quality variables with the process and upstream variables. OBS data are further used to estimate and calibrate the model parameters. Recently, data mining approaches are also introduced to model and improve manufacturing processes (Jin and Shi, 2012). Although the models based on the OBS data have demonstrated success in various applications, they may not be directly applicable to the unstable testing production, where the data contains high uncertainties.

Table 1 summarizes the characteristics of the two types of data. For the DOE data, they are usually collected in well-designed settings and well-controlled production environment, which reduces the collinearity of the factors as well as the impact of the noise factors. The ranges of factors are usually properly selected to explore more possible combinations of the settings. However, the sample is often limited, which could result in inaccurate estimation of parameters. For the OBS data, they have a large sample size but it may contain high uncertainty due to the uncontrolled covariate factors. The covariates can be intermediate quality variables or environmental variables which cannot be controlled, but still play important roles for the final process performance. The process variables are usually constrained in a small neighborhood of the manufacturing process settings. The corresponding model may not work well in the extrapolated region. Thus, the estimated optimal settings based on the OBS data could be a local optimal.

As both types of data are readily available for the scale-up efforts, it is natural to integrate both types of data in a proper manner. Using DOE and OBS data, we propose an ensemble modeling strategy to model manufacturing processes of experiment-validation in the scale-up. It can outperform models estimated from a single type of data, with the following attractive features.

Table 1: Characteristics of DOE data and OBS data.

Data Type	Sample Size	Uncertainty	Range
DOE	Small	Low	Large
OBS	Large	High	Small

First, the proposed method enables the use of DOE data to better identify significant factors, while integrates the OBS data to enhance the model estimation and prediction. Second, a meaningful variable selection is achieved by incorporating the sequential nature and inherent features of two types of data. The sequential nature refers to the fact that the two types of data are usually collected sequentially. Following the DOE, the OBS data are obtained by conducting validations runs with setting based on the optimal recipe from DOE data. The inherent features refer to that significant predictors in modeling the DOE data are expected keeping their significance in modeling the OBS data. The proposed method adopts the constrained likelihood approach, where the constraints address the sequential nature and inherent features of two types of data sets in variable selection. Therefore, when obtaining a more appropriate model with better prediction and variable selection, we can reduce the number of rounds for experiments and validation production runs in the scale-up, leading to significant saving of time and cost.

The remainder of the paper is organized as follows. Section 2 describes in detail the proposed ensemble modeling method. The statistical property of the estimation will also be discussed. The simulation study is reported in Section 3 to show the effectiveness of the proposed method. A real case study of wafer manufacturing process is used to elaborate the proposed method in Section 4. Finally, Section 5 concludes this work with discussions.

2 Ensemble Modeling

In this section, we consider jointly estimating two models, one for the DOE data and the other for the OBS data. We use DOE model to refer the model based on the DOE data, and OBS model to refer the model based on the OBS data. Several assumptions are made for developing the proposed method: (1) the types of data come from the same manufacturing process with the same process input variables and quality response variable. The first data set is collected from the DOE and the second data set is collected from the validation production runs after the DOE. (2) The manufacturing process is static in the modeling effort, which indicates that the underlying model will remain unchanged for the significant variables and coefficients. Here assumption (2) implies that additional uncertainty in the OBS data is introduced by uncontrolled noise factors. (3) The significant variables identified from the DOE model are suggested to be significant in the final model. Here assumption (3) implies that the DOE model usually have better capability than the OBS model in identifying significant variables. We will incorporate this assumption as constraints in the proposed method.

Let us denote the DOE data are $(\boldsymbol{z}_i^{(1)}, y_i^{(1)}), i = 1, \ldots, n_1$ where $\boldsymbol{z}_i^{(1)} = (z_{i1}^{(1)}, \ldots, z_{ip}^{(1)})$, and the OBS data are $(\boldsymbol{z}_j^{(2)}, y_j^{(2)}), j = 1, \ldots, n_2$ where $\boldsymbol{z}_j^{(2)} = (z_{j1}^{(2)}, \ldots, z_{jp}^{(2)})$. Here $y^{(k)}, k = 1, 2$ is univariate response. To model the quality-process relationship, we consider linear models with the main effects and two-factor interaction effects as predictors. Here we only consider the two-factor interaction effects as most optimization problems, such as RPD, mainly emphasize on control-noise (control-covariates) interactions. Other interaction terms can be easily adopted in our framework. Specifically, we model DOE data and OBS data respectively as follows

$$y_i^{(1)} = \boldsymbol{x}_i^{(1)\prime} \boldsymbol{\beta}^{(1)} + \epsilon_i^{(1)}, \ \epsilon_i^{(1)} \sim N(0, \sigma_1^2), \tag{1}$$

$$y_j^{(2)} = \boldsymbol{x}_j^{(2)\prime} \boldsymbol{\beta}^{(2)} + \epsilon_j^{(2)}, \ \epsilon_j^{(2)} \sim N(0, \sigma_2^2), \tag{2}$$

where $\epsilon_i^{(1)}$ and $\epsilon_j^{(2)}$ are independent and identically distributed random errors. The predictor vector $\boldsymbol{x}_i^{(m)}, m = 1, 2$ is written as $\boldsymbol{x}_i^{(m)} = (x_{i1}^{(m)}, \ldots, x_{ip}^{(m)}, x_{i1}^{(m)} x_{i2}^{(m)}, \ldots, x_{i,p-1}^{(m)} x_{ip}^{(m)})'$. The $\boldsymbol{\beta}^{(m)} = (\beta_1^{(m)}, \beta_p^{(m)}, \beta_{12}^{(m)}, \ldots, \beta_{p-1,p}^{(m)})'$ is the corresponding parameter coefficients. It means that the predictor variables in the model includes the main effect $x_k, k = 1 \ldots, p$ and their 2-factor interactions $x_k x_l$. In this model formulation, we assume that the DOE model and the OBS model can have different structures and parameters. This does not imply that the underlying true model would vary for the generation of the DOE and OBS data. We assume that the underlying model is static and remains unchanged for the whole process. The model structures in the DOE and OBS models reflect different information from the two types of data. The DOE model intends to capture the significant predictors from the DOE data. The OBS model attempts to enhance the parameter estimation using the OBS data, while preserving the significant variables from the first model. Recall that the OBS data are collected from the validation production runs after conducting the experimental designs. For the significance of predictors in both models, the assumption (3) indicates that when the kth predictor variable is significant in the DOE model, we expect that it should be also significant in the OBS model. It means that if the kth predictor variable is not significant in the OBS model, then we expect that it is also not significant in the DOE model. However, if the kth predictor variable is not significant in the DOE model, it is possible that it becomes significant in the OBS model. The significance relationship of the predictors will be reflected through the constraints in maximizing the likelihood function. For the proposed method, the OBS model structure will be used as the final model structure for the manufacturing process, which leads to the final manufacturing process settings.

2.1 The Proposed Method

To incorporate the sequential nature and inherent features of the two types of data sets, we propose a novel regularized approach to estimate the model parameters. Specifically, we adopt the nonnegative garrotte to achieve the joint variable selection and estimation. The original nonnegative garrote estimator is introduced by Breiman (1995), which can be viewed as a scaled version of the least squares estimation. Theoretical properties of nonnegative garrotes can be found in Yuan and Lin (2007). The key idea of nonnegative garrote is to re-parameterize the coefficients in (1) and (2) by

$$\beta_k^{(m)} = \theta_k^{(m)} \tilde{\beta}_k^{(m)}, \beta_{kl}^{(m)} = \theta_{kl}^{(m)} \tilde{\beta}_{kl}^{(m)},\ m = 1, 2,$$

where $\tilde{\beta}_k^{(m)}$ and $\tilde{\beta}_{kl}^{(m)}$ are least squares estimates. The $\theta_k^{(m)} \geq 0$ and $\theta_{kl}^{(m)} \geq 0$ are shrinkage coefficients, which will be estimated from the data. Note that when $\theta_k^{(m)} = 1$ and $\theta_{kl}^{(m)} = 1$, the nonnegative garrotte method becomes exactly the least squares estimation.

Now we can define the transformed data points $\tilde{\boldsymbol{x}}_i^{(m)} = \boldsymbol{B}\boldsymbol{x}_i^{(m)}, m = 1, 2$ for DOE and OBS data, where $\boldsymbol{B} = diag(\tilde{\beta}_1^{(m)}, \ldots, \tilde{\beta}_p^{(m)}, \tilde{\beta}_{12}^{(m)}, \ldots, \tilde{\beta}_{p-1,p}^{(m)})$. Defining $\boldsymbol{\theta}^{(m)} = (\theta_1^{(m)}, \ldots, \theta_p^{(m)}, \theta_{12}^{(m)}, \ldots, \theta_{p-1,p}^{(m)})'$, then the DOE and OBS models in (1) and (2) can be rewritten as

$$y_i^{(1)} = \tilde{\boldsymbol{x}}_i^{(1)'}\boldsymbol{\theta}^{(1)} + \epsilon_i^{(1)},\ \epsilon_i^{(1)} \sim N(0, \sigma_1^2), \tag{3}$$

$$y_j^{(2)} = \tilde{\boldsymbol{x}}_j^{(2)'}\boldsymbol{\theta}^{(2)} + \epsilon_j^{(2)},\ \epsilon_j^{(2)} \sim N(0, \sigma_2^2). \tag{4}$$

Such parametrization provides us the flexibility to impose various constraints for estimating parameters. The negative log-likelihood function based on the above models can be written as

$$n_1\big[\log\sigma_1^2 + \frac{1}{n_1}\sum_{i=1}^{n_1}\frac{(y_i^{(1)} - \tilde{\boldsymbol{x}}_i^{(1)'}\boldsymbol{\theta}^{(1)})^2}{\sigma_1^2}\big] + n_2\big[\log\sigma_2^2 + \frac{1}{n_2}\sum_{j=1}^{n_2}\frac{(y_j^{(2)} - \tilde{\boldsymbol{x}}_j^{(2)'}\boldsymbol{\theta}^{(2)})^2}{\sigma_2^2}\big], \tag{5}$$

up to some constant. Note that both DOE and OBS models contain main effects and 2-factor interactions. In engineering practice, the significance relationships for main effects and 2-factor interactions commonly follows the heredity principle (Wu and Hamada, 2009). The weak heredity principle states that if a 2-factor interaction $x_k x_l$ is significant only if at least one of its parents $\{x_k, x_l\}$ is significant, while the strong heredity principle requires both parents to be significant to allow a significant 2-factor interaction.

To accommodate the heredity principle, we impose proper linear constraints of shrinkage coefficients onto minimizing the negative log-likelihood function. Incorporating the heredity structures for variable selection through nonnegative garrotes was originally developed in Yuan et al. (2009). In this paper, we focus on the weak heredity for the proposed method. The constraint for the weak heredity is $\theta_{kl}^{(m)} \leq max\{\theta_k^{(m)}, \theta_l^{(m)}\}$, $m = 1, 2$. However, such a constraint for the weak heredity is not convex. To circumvent this difficulty, we consider a relaxed version of the linear constraint

$$\theta_{kl}^{(m)} \leq \theta_k^{(m)} + \theta_l^{(m)}.$$

For the strong heredity, one can formulate the constraints as $\theta_{kl}^{(m)} \leq \theta_k^{(m)}, \theta_{kl}^{(m)} \leq \theta_l^{(m)}$. Heredity structures for variable selection are also used in support vector machines (Wu et al., 2008) and the hierarchical modeling (Choi et al., 2010).

Moreover, the assumption (3) implies that if one significant variable is identified from the DOE model, it is very likely to be significant in the OBS models as well. We formulate such information into the following constraints:

$$\theta_k^{(1)} \le \theta_k^{(2)},\ \forall k = 1, \ldots, p,$$

$$\theta_{kl}^{(1)} \le \theta_{kl}^{(2)},\ \forall k \ne l.$$

Therefore, we propose to estimate the shrinkage coefficients by using constrained likelihood estimation. Specifically, the estimation problem can be formulated as

$$\begin{aligned}
\min & \left\{ n_1 \left[\log \sigma_1^2 + \frac{1}{n_1} \sum_{i=1}^{n_1} \frac{(y_i^{(1)} - \tilde{\boldsymbol{x}}_i^{(1)'} \boldsymbol{\theta}^{(1)})^2}{\sigma_1^2} \right] + n_2 \left[\log \sigma_2^2 + \frac{1}{n_2} \sum_{j=1}^{n_2} \frac{(y_j^{(2)} - \tilde{\boldsymbol{x}}_j^{(2)'} \boldsymbol{\theta}^{(2)})^2}{\sigma_2^2} \right] \right\} \\
s.t. & \quad \sum_{k=1}^{p} \theta_k^{(1)} + \sum_{k=1}^{p} \theta_k^{(2)} \le M, \\
& \theta_k^{(1)} \ge 0, \forall k,\ \theta_k^{(2)} \ge 0, \forall k, \\
& \theta_k^{(1)} \le \theta_k^{(2)},\ k = 1, \ldots, p, \\
& \theta_{kl}^{(1)} \le \theta_{kl}^{(2)},\ \forall k \ne l, k, l = 1, \ldots, p, \\
& \theta_{kl}^{(1)} \le \theta_k^{(1)} + \theta_l^{(1)},\ \forall k \ne l, k, l = 1, \ldots, p, \\
& \theta_{kl}^{(2)} \le \theta_k^{(2)} + \theta_l^{(2)},\ \forall k \ne l, k, l = 1, \ldots, p,
\end{aligned} \tag{6}$$

where $M \ge 0$ is a tuning parameter. The first two constraints here are used to encourage a general variable selection for both models, while the remaining constraints accommodate the sequential nature and the weak heredity principle of the DOE and OBS data. Note that the optimization in (6) is a constraint convex programming. It can be solved efficiently with achieving a global optimal solution (Boyd and Vandenberghe, 2004).

2.2 Computational Algorithm

The decision variables in (6) are σ_1^2, σ_2^2, and $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}$. Although the optimization may not be solved straightforwardly in terms of the whole parameter set $\{\sigma_1^2, \sigma_2^2, \boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}\}$, they can be solved in an efficient fashion by iteratively estimating σ_1^2, σ_2^2 and $\boldsymbol{\theta}^{(1)}, \boldsymbol{\theta}^{(2)}$. The procedure is to firstly optimize $\hat{\boldsymbol{\theta}}^{(1)}, \hat{\boldsymbol{\theta}}^{(2)}$ by fixing $\hat{\sigma}_1^2, \hat{\sigma}_2^2$, and then estimate $\hat{\sigma}_1^2, \hat{\sigma}_2^2$ by given $\hat{\boldsymbol{\theta}}^{(1)}, \hat{\boldsymbol{\theta}}^{(2)}$, which have the close form solutions.

Given $\hat{\boldsymbol{\theta}}^{(1)}, \hat{\boldsymbol{\theta}}^{(2)}$, the solution to σ_1^2, σ_2^2 can be obtained explicitly. That is

$$\hat{\sigma}_1^2 = \frac{1}{n_1}\sum_{i=1}^{n_1}(y_i^{(1)} - \tilde{\boldsymbol{x}}_i^{(1)'}\hat{\boldsymbol{\theta}}^{(1)})^2, \tag{7}$$

$$\hat{\sigma}_2^2 = \frac{1}{n_2}\sum_{i=1}^{n_2}(y_i^{(2)} - \tilde{\boldsymbol{x}}_i^{(2)'}\hat{\boldsymbol{\theta}}^{(2)})^2. \tag{8}$$

Given $\hat{\sigma}_1^2, \hat{\sigma}_2^2$, the solution of $\boldsymbol{\theta}^{(1)}$ and $\boldsymbol{\theta}^{(2)}$ can be solved through a quadratic programming with linear constraints. That is

$$\begin{aligned}
&\min \left\{ [\sum_{i=1}^{n_1} \frac{(y_i^{(1)} - \tilde{\boldsymbol{x}}_i^{(1)'}\boldsymbol{\theta}^{(1)})^2}{\hat{\sigma}_1^2}] + [\sum_{j=1}^{n_2} \frac{(y_j^{(2)} - \tilde{\boldsymbol{x}}_j^{(2)'}\boldsymbol{\theta}^{(2)})^2}{\hat{\sigma}_2^2}] \right\} \\
&s.t. \quad \sum_{k=1}^{p}\theta_k^{(1)} + \sum_{k=1}^{p}\theta_k^{(2)} \leq M, \\
&\qquad \theta_k^{(1)} \geq 0, \forall k,\ \theta_k^{(2)} \geq 0, \forall k, \\
&\qquad \theta_k^{(1)} \leq \theta_k^{(2)},\ k = 1, \ldots, p, \\
&\qquad \theta_{kl}^{(1)} \leq \theta_{kl}^{(2)},\ \forall k \neq l, k, l = 1, \ldots, p, \\
&\qquad \theta_{kl}^{(1)} \leq \theta_k^{(1)} + \theta_l^{(1)},\ \forall k \neq l, k, l = 1, \ldots, p, \\
&\qquad \theta_{kl}^{(2)} \leq \theta_k^{(2)} + \theta_l^{(2)},\ \forall k \neq l, k, l = 1, \ldots, p.
\end{aligned} \tag{9}$$

Because of quadratic programming, the solution can be efficiently obtained with global optimal convergence. Specifically, the iterative algorithm is described as follows:

Algorithm 1.

Step 1: *Set initial estimates* $\sigma_1^2 > 0, \sigma_2^2 > 0$.

Step 2: *Obtain the estimates* $\hat{\boldsymbol{\theta}}^{(1)}, \hat{\boldsymbol{\theta}}^{(2)}$ *by solving the optimization in* (9).

Step 3: *Obtain the estimates* $\hat{\sigma}_1^2, \hat{\sigma}_2^2$ *by plugging* $\hat{\boldsymbol{\theta}}^{(1)}, \hat{\boldsymbol{\theta}}^{(2)}$ *obtained in Step 2 into* (7) *and* (8).

Step 4: *Check if both* $\|\hat{\sigma}_1^2 - \sigma_1^2\|_2^2$ *and* $\|\hat{\sigma}_2^2 - \sigma_2^2\|_2^2$ *are less than a pre-specified positive tolerance value. Otherwise, set* $\sigma_1^2 = \hat{\sigma}_1^2$, $\sigma_2^2 = \hat{\sigma}_2^2$, *and go back to* **Step 2**.

2.3 Tuning Parameters Selection

Note that M in (6) is a tuning parameter, which needs to be specified based on the data. The common methods to select tuning parameters include cross-validation and information criterion approaches such as Akaike information criterion (AIC), Bayesian information criterion (BIC), and C_p criteria (Burnham and Anderson, 2002). In this work, we use the BIC for finding an optimal value of the tuning parameter M. The BIC for the proposed model can be written as

$$BIC(M) = n_1 \log \hat{\sigma}_1^2 + n_2 \log \hat{\sigma}_2^2 + q \log(n_1 + n_2), \tag{10}$$

where q is the number of nonzero estimates of parameters, i.e.,

$$q = \sum_{m=1}^{2} \big[\sum_{k=1}^{p} I(\hat{\theta}_k^{(m)} \neq 0) + \sum_{k<l} I(\hat{\theta}_{kl}^{(m)} \neq 0) \big].$$

Here $\hat{\theta}_k^{(m)}$, $\hat{\theta}_{kl}^{(m)}$, $\hat{\sigma}_1^2$, and $\hat{\sigma}_2^2$ are parameter estimates in (6) given the value of M. Specifically, we can generate a grid for M such that the value of $M \in \mathcal{C} = \{m_1, \ldots, m_t\}$. For each grid point m_j in $\mathcal{C}$, we evaluate the corresponding BIC score, and find the optimal choice of M which has the minimal value of BIC among all grid points in $\mathcal{C}$.

2.4 Statistical Properties

To obtain more insight for the proposed method, we study the statistical properties for parameter estimation of $\boldsymbol{\beta}$ in (1) and (2). Assume that the mechanism of the true data satisfies the weak heredity principle as well as the assumption (3). Then we can show that the proposed method can have the root-n consistency for the nonzero components of $\boldsymbol{\beta}$, and the zero components of $\boldsymbol{\beta}$ can be estimated by zeros with probability one as the sample size goes to infinity. Let us denote $\mathcal{I}^{(1)} = \{j : \beta_j^{(1)} \neq 0\}$ as the indices of nonzeros components in $\boldsymbol{\beta}^{(1)}$ for the DOE model in (1), and $\mathcal{I}^{(2)} = \{j : \beta_j^{(1)} \neq 0\}$ as the indices of nonzeros components in $\boldsymbol{\beta}^{(2)}$ for the OBS model in (2). Define $\hat{\boldsymbol{\beta}}^{(1)}$ and $\hat{\boldsymbol{\beta}}^{(2)}$ to be the coefficient estimates from the proposed method. Note that the corresponding shrinkage coefficients $\hat{\boldsymbol{\theta}}^{(1)}$ and $\hat{\boldsymbol{\theta}}^{(2)}$ from (6) can be obtained from an equivalent formulation by minimizing

$$n_1 \big[\log \sigma_1^2 + \frac{1}{n_1} \sum_{i=1}^{n_1} \frac{(y_i^{(1)} - \tilde{\boldsymbol{x}}_i^{(1)'} \boldsymbol{\theta}^{(1)})^2}{\sigma_1^2} \big] + n_2 \big[\log \sigma_2^2 + \frac{1}{n_2} \sum_{j=1}^{n_2} \frac{(y_j^{(2)} - \tilde{\boldsymbol{x}}_j^{(2)'} \boldsymbol{\theta}^{(2)})^2}{\sigma_2^2} \big] + \lambda_n \left(\sum_{k=1}^{p} \theta_k^{(1)} + \sum_{k=1}^{p} \theta_k^{(2)} \right)$$

subject to $\theta_k^{(1)} \geq 0, \theta_k^{(2)} \geq 0$, $\theta_k^{(1)} \leq \theta_k^{(2)}, \theta_{kl}^{(1)} \leq \theta_{kl}^{(2)}$, and $\theta_{kl}^{(1)} \leq \theta_k^{(1)} + \theta_l^{(1)}, \theta_{kl}^{(2)} \leq \theta_k^{(2)} + \theta_l^{(2)}$ for some $\lambda_n \geq 0$.

Proposition 1. *Suppose* $\frac{1}{n_1}\sum_{i=1}^{n_1}\boldsymbol{x}_i^{(1)}\boldsymbol{x}_i^{(1)'} \to \boldsymbol{\Sigma}_1$ *as* $n_1 \to \infty$ *and* $\frac{1}{n_2}\sum_{i=1}^{n_2}\boldsymbol{x}_i^{(2)}\boldsymbol{x}_i^{(2)'} \to \boldsymbol{\Sigma}_2$ *as* $n_1 \to \infty$. *Both* $\boldsymbol{\Sigma}_1$ *and* $\boldsymbol{\Sigma}_2$ *are positive definite. Assume that the true model satisfies the weak heredity principle as well as the engineering knowledge described in assumption (3). Let* $n = min(n_1, n_2)$. *When* $\lambda_n \to \infty$ *with rate* $\lambda_n = o(\sqrt{n})$ *as* $n \to \infty$, *we have*
(1) $\forall j \notin \mathcal{I}^{(1)}, \hat{\beta}_j^{(1)} = 0$ *with probability* 1, *and* $\forall j \in \mathcal{I}^{(1)}, \hat{\beta}_j^{(1)} - \beta_j^{(1)} = O_p(1/\sqrt{n})$.
(2) $\forall k \notin \mathcal{I}^{(2)}, \hat{\beta}_k^{(2)} = 0$ *with probability* 1, *and* $\forall k \in \mathcal{I}^{(2)}, \hat{\beta}_k^{(2)} - \beta_k^{(2)} = O_p(1/\sqrt{n})$.

The proof of Proposition 1 closely follows the proof of Theorem 1 in Yuan et al. (2009), thus it is omitted here.

3 Simulation

To demonstrate the effectiveness of the proposed method, we evaluate the performance of the prediction and variable selection through several simulated data sets. The following three examples are considered for generating the data in each simulation run. For each example, we consider p main factors and $p(p-1)/2$ two-factor interactions in the full model with the underlying true model as follows

- Example 1: Let $p = 5$. The model follows the weak heredity,

$$y = 2.88x_1 + 2.32x_2 + 3.22x_3 + 1.30x_1x_2 + 1.85x_1x_3 + 2.63x_1x_4 + 2.84x_1x_5 + 2.23x_4x_5 + \epsilon. \tag{11}$$

- Example 2: Let $p = 10$. This model follows the strong heredity,

$$\begin{aligned} y = {} & 2.44x_1 + 2.82x_2 + 2.20x_3 + 3.67x_4 + 4.37x_7 + 2.34x_8 + 3.80x_9 + 0.60x_1x_2 \\ & + 2.22x_1x_3 + 3.29x_1x_4 + 3.71x_1x_7 + 1.95x_1x_8 + 3.68x_1x_9 + 3.59x_2x_3 + 3.77x_2x_4 \\ & + 1.67x_2x_7 + 2.49x_2x_8 + 4.17x_2x_9 + 2.30x_3x_4 + 3.67x_7x_8 + 4.23x_7x_9 + 2.87x_8x_9 + \epsilon \end{aligned} \tag{12}$$

- Example 3: Similar as Example 2, but the model follows a weak heredity.

$$y = 1.60x_1 + 4.01x_2 + 3.51x_3 + 2.36x_4 + 1.40x_7 + 1.93x_8 + 2.48x_9 + 4.66x_1x_2 + 3.78x_1x_3$$
$$+ 2.34x_1x_4 + 3.33x_1x_7 + 4.85x_1x_8 + 2.87x_1x_9 + 1.45x_2x_3 + 3.40x_2x_4 + 3.34x_2x_7$$
$$+ 5.20x_2x_8 + 1.89x_2x_9 + 2.33x_3x_4 + 1.97x_7x_8 + 4.91x_8x_9 + 2.44x_8x_{10} + \epsilon \quad (13)$$

The detailed settings of these three examples are summarized in Table 2, with the parenthesis as the number of significant variables. Take Example 1 for illustration. There are four controllable variables $x_1 - x_4$, one covariate x_5 and 10 two-factor interactions of the controllable variables and the covariate. The controllable variables can be changed during the DOE, while the covariate x_5 is uncontrollable but measurable. For generating the data, we consider 27 different scenarios by varying the settings of uncertainty (i.e., standard deviation of the errors), sample size, and range. Specifically, the standard deviation σ_1 is set to be 2 in the DOE model, and vary σ_2 to be 2, 10 and 20 respectively in the OBS model (corresponding to $\sigma_2/\sigma_1 = 1, 5, 10$). A 2^{4-1} fractional factorial design with levels -1 and 1 is constructed for the four controllable variables, and the DOE data set has the sample size $n_1 = 24$ (3 replications for each DOE setting). While the sample size of the OBS data set is varied as $n_2 = 24$, 72 and 120, respectively (corresponding to $n_2/n_1 = 1, 3, 5$). The range of predictors for the DOE data is $R_{DOE} = [-1, 1]$, and the range for the OBS data R_{OBS} varies from [-1,1], [-0.5, 0.5], to [-0.3, 0.3], respectively (corresponding to the range shrinkage $R_{DOE}/R_{OBS} = 1, 0.5, 0.3$). In Examples 2 and 3, the predictor variables include 10 factors with six

Table 2: A Summary of models and 27 simulation scenarios in three examples.

	Control	Covariate	Interaction	σ_1	σ_2/σ_1	n_1	n_2/n_1	R_{DOE}	R_{OBS}/R_{DOE}
Example 1	4(3)	1(0)	10(5)	2	1,5,10	24	1,3,5	[-1, 1]	1,0.5,0.3
Example 2	6(4)	4(3)	45(15)	2	1,5,10	64	1,3,5	[-1, 1]	1,0.5,0.3
Example 3	6(4)	4(3)	45(15)	2	1,5,10	64	1,3,5	[-1, 1]	1,0.5,0.3

controllable variables $x_1 - x_6$ and four covariates $x_7 - x_{10}$, and their 45 two-factor interactions. A

2^{6-2} factional factorial designs with levels -1 and 1 are used as the design matrix for control factors in both Examples 2 and 3, where the DOE data set has the sample size $n_1 = 64$ (4 replications for each DOE setting). For all models in Examples 1-3, the range of covariates in DOE data is in [-1, 1]. The coefficient values of the significant predictors are generated randomly from a normal distribution $N(3, 1)$.

We generate 50 simulation replicates for each scenario of the simulation. In the simulation, the underlying true models will remain unchanged in each scenario. They are used to construct the DOE data and the OBS data. Specifically, in each replicate of every example, we generate a training set for DOE model, and a training set for the OBS model, respectively. When merging the two sets, we denote it as a combined training data (CBD) set. For the test set, we generate a data set with the value of predictor variables uniformly distributed to the same range of the variables for the DOE data ([-1, 1]). We compare the proposed ensemble model (denoted as EM) with three benchmark regression models for the prediction based on the testing set: (1) the regression model from the training set of DOE model with variable selected using BIC (denoted as BM_{DOE}), (2) the regression model based on the training set of OBS model with variable selected using BIC (denote as BM_{OBS}), and (3) the regression model based on the CBD with variable selected using BIC (denoted as BM_{CBD}). Then all models will be evaluated based on the test data. Tables 3-5 report the average of root mean squared prediction errors (RMSPE) and standard errors in parenthesis based on 50 simulation replicates of the test data. Each table contains the result for 27 scenarios under different ratios of sample size n_2/n_1, uncertainty σ_2/σ_1, and range R_{OBS}/R_{DOE}. We further evaluate the variable selection performances based on the training data set, which are shown in Tables 6-8.

From the results in Tables 3-5, the proposed EM method has the best prediction performance in most scenarios. For the situation of $\sigma_1/\sigma_2 = 1$ and $R_{OBS}/R_{DOE} = 1$, it implies the OBS data have similar information as the DOE data. In this case, the results in EM approach generally have comparable prediction performance to BM_{CBD}. When the ratio σ_1/σ_2 becomes larger, and the range of OBS data R_{OBS} shrinks, the proposed EM significantly outperforms BM_{CBD} and other methods. Note that BM_{CBD} is obtained from simply combining two training data, without addressing the sequential nature and inherent feature of the two types of data. The proposed EM

Table 3: Averages and standard errors of testing RMSPE from 50 simulation runs for Example 1.

		$R_{OBS}/R_{DOE}=1$			$R_{OBS}/R_{DOE}=0.5$			$R_{OBS}/R_{DOE}=0.3$		
n_2/n_1	Method	$\frac{\sigma_2}{\sigma_1}=1$	$\frac{\sigma_2}{\sigma_1}=5$	$\frac{\sigma_2}{\sigma_1}=10$	$\frac{\sigma_2}{\sigma_1}=1$	$\frac{\sigma_2}{\sigma_1}=5$	$\frac{\sigma_2}{\sigma_1}=10$	$\frac{\sigma_2}{\sigma_1}=1$	$\frac{\sigma_2}{\sigma_1}=5$	$\frac{\sigma_2}{\sigma_1}=10$
	BM_{DOE}	3.57	2.67	2.99	2.98	3.10	3.69	2.66	3.64	3.15
		(0.37)	(0.24)	(0.27)	(0.22)	(0.27)	(0.41)	(0.22)	(0.34)	(0.19)
	BM_{OBS}	7.67	28.29	54.75	26.95	124.11	198.70	56.34	315.42	557.63
1		(0.48)	(2.80)	(5.31)	(2.27)	(11.15)	(22.99)	(7.28)	(34.52)	(74.77)
	BM_{CBD}	**1.61**	3.58	5.90	2.51	2.51	4.53	2.10	3.13	5.26
		(0.07)	(0.33)	(0.79)	(0.20)	(0.40)	(0.41)	(0.16)	(0.12)	(0.34)
	EM	1.69	**1.79**	**2.93**	**2.04**	**1.78**	**2.98**	**1.79**	**2.59**	**2.68**
		(0.05)	(0.14)	(0.20)	(0.05)	(0.17)	(0.27)	(0.11)	(0.13)	(0.15)
	BM_{DOE}	4.00	3.71	3.09	3.56	3.02	2.99	3.30	3.11	3.22
		(0.33)	(0.51)	(0.29)	(0.35)	(0.31)	(0.28)	(0.29)	(0.24)	(0.26)
	BM_{OBS}	3.45	11.90	16.96	9.53	34.49	65.28	22.30	110.37	189.93
3		(0.12)	(1.01)	(2.11)	(0.76)	(3.81)	(8.50)	(2.56)	(14.61)	(26.18)
	BM_{CBD}	**1.70**	3.45	4.34	2.42	3.34	4.00	2.17	3.23	4.11
		(0.08)	(0.19)	(0.58)	(0.10)	(0.17)	(0.19)	(0.18)	(0.06)	(0.03)
	EM	2.04	**2.02**	**2.60**	**2.23**	**1.85**	**2.48**	**1.95**	**1.78**	**2.48**
		(0.06)	(0.05)	(0.21)	(0.06)	(0.05)	(0.29)	(0.05)	(0.09)	(0.25)
	BM_{DOE}	3.63	2.94	3.51	3.58	3.33	2.66	3.81	3.45	2.73
		(0.29)	(0.23)	(0.33)	(0.35)	(0.32)	(0.24)	(0.44)	(0.32)	(0.21)
	BM_{OBS}	2.85	7.38	12.10	5.66	21.26	50.93	11.52	43.95	60.67
5		(0.10)	(0.60)	(1.23)	(0.45)	(2.80)	(6.11)	(1.75)	(7.96)	(13.98)
	BM_{CBD}	**1.37**	3.34	4.73	**2.27**	3.81	3.61	2.33	3.66	3.63
		(0.07)	(0.12)	(0.52)	(0.12)	(0.60)	(0.18)	(0.19)	(0.15)	(0.02)
	EM	1.79	**2.45**	**3.78**	2.34	**2.00**	**2.71**	**2.17**	**2.09**	**2.07**
		(0.05)	(0.12)	(0.39)	(0.06)	(0.11)	(0.24)	(0.14)	(0.05)	(0.27)

Table 4: Averages and standard errors of testing RMSPE from 50 simulation runs for Example 2.

		$R_{OBS}/R_{DOE}=1$			$R_{OBS}/R_{DOE}=0.5$			$R_{OBS}/R_{DOE}=0.3$		
n_2/n_1	Method	$\frac{\sigma_2}{\sigma_1}=1$	$\frac{\sigma_2}{\sigma_1}=5$	$\frac{\sigma_2}{\sigma_1}=10$	$\frac{\sigma_2}{\sigma_1}=1$	$\frac{\sigma_2}{\sigma_1}=5$	$\frac{\sigma_2}{\sigma_1}=10$	$\frac{\sigma_2}{\sigma_1}=1$	$\frac{\sigma_2}{\sigma_1}=5$	$\frac{\sigma_2}{\sigma_1}=10$
	BM_{DOE}	12.53	11.38	11.84	11.72	12.87	11.23	12.04	10.39	11.19
		(0.77)	(0.64)	(0.66)	(0.60)	(0.81)	(0.55)	(0.75)	(0.62)	(0.71)
	BM_{OBS}	18.43	74.43	125.71	70.39	255.12	509.31	192.33	713.10	1715.36
1		(0.99)	(5.63)	(9.10)	(5.45)	(18.38)	(41.39)	(14.61)	(55.29)	(99.95)
	BM_{CBD}	**2.89**	7.38	10.41	**3.98**	5.14	6.26	5.14	**4.49**	**5.34**
		(0.12)	(0.38)	(0.64)	(0.24)	(0.19)	(0.28)	(0.31)	(0.06)	(0.04)
	EM	3.47	**5.67**	**7.09**	4.30	**5.02**	**6.21**	**4.66**	4.64	5.39
		(0.10)	(0.24)	(0.48)	(0.14)	(0.16)	(0.36)	(0.13)	(0.13)	(0.28)
	BM_{DOE}	11.27	12.08	11.68	13.03	12.61	12.24	10.70	12.77	12.43
		(0.88)	(0.68)	(0.56)	(0.97)	(0.76)	(0.66)	(0.57)	(0.78)	(0.72)
	BM_{OBS}	4.36	13.18	21.17	10.60	36.71	72.78	23.37	93.38	177.44
3		(0.11)	(0.63)	(1.49)	(0.57)	(3.16)	(6.34)	(1.95)	(8.15)	(17.80)
	BM_{CBD}	**2.08**	6.71	9.67	**2.79**	5.52	6.89	4.36	5.01	6.71
		(0.05)	(0.33)	(0.49)	(0.16)	(0.24)	(0.39)	(0.24)	(0.15)	(0.09)
	EM	2.47	**4.43**	**6.14**	3.48	**5.00**	**4.51**	**3.94**	**4.54**	**5.44**
		(0.05)	(0.15)	(0.23)	(0.15)	(0.16)	(0.22)	(0.09)	(0.15)	(0.13)
	BM_{DOE}	11.50	11.10	11.46	11.41	11.21	11.61	12.52	11.88	11.67
		(0.57)	(0.55)	(0.62)	(0.55)	(0.68)	(0.63)	(0.66)	(0.58)	(0.81)
	BM_{OBS}	3.39	9.99	14.62	7.59	23.05	53.25	15.29	52.72	104.10
5		(0.06)	(0.33)	(0.92)	(0.27)	(2.00)	(4.52)	(1.07)	(5.91)	(12.20)
	BM_{CBD}	**1.68**	5.84	8.33	**3.40**	5.19	6.49	4.46	5.26	6.28
		(0.05)	(0.19)	(0.37)	(0.20)	(0.11)	(0.33)	(0.23)	(0.20)	(0.08)
	EM	2.20	**4.83**	**6.09**	3.69	**4.80**	**4.05**	**3.91**	**4.75**	**4.40**
		(0.06)	(0.08)	(0.09)	(0.10)	(0.06)	(0.07)	(0.06)	(0.13)	(0.14)

Table 5: Averages and standard errors of testing RMSPE from 50 simulation runs for Example 3.

		$R_{OBS}/R_{DOE}=1$			$R_{OBS}/R_{DOE}=0.5$			$R_{OBS}/R_{DOE}=0.3$		
n_2/n_1	Method	$\frac{\sigma_2}{\sigma_1}=1$	$\frac{\sigma_2}{\sigma_1}=5$	$\frac{\sigma_2}{\sigma_1}=10$	$\frac{\sigma_2}{\sigma_1}=1$	$\frac{\sigma_2}{\sigma_1}=5$	$\frac{\sigma_2}{\sigma_1}=10$	$\frac{\sigma_2}{\sigma_1}=1$	$\frac{\sigma_2}{\sigma_1}=5$	$\frac{\sigma_2}{\sigma_1}=10$
	BM_{DOE}	11.96	11.99	10.58	11.95	11.82	11.49	10.62	11.40	12.49
		(0.93)	(0.61)	(0.54)	(0.60)	(0.70)	(0.63)	(0.53)	(0.66)	(0.73)
	BM_{OBS}	19.54	66.92	142.64	72.10	316.06	541.61	180.67	819.63	1569.76
1		(1.02)	(5.00)	(11.88)	(5.60)	(21.37)	(39.24)	(11.47)	(54.90)	(131.05)
	BM_{CBD}	**2.81**	7.09	12.23	4.67	5.49	6.86	5.17	5.84	5.83
		(0.09)	(0.35)	(0.79)	(0.32)	(0.26)	(0.42)	(0.30)	(0.48)	(0.04)
	EM	3.50	**6.24**	**8.36**	**4.47**	**5.18**	**5.70**	**4.70**	**5.75**	**5.70**
		(0.10)	(0.25)	(0.43)	(0.19)	(0.25)	(0.32)	(0.15)	(0.18)	(0.21)
	BM_{DOE}	11.37	10.83	12.99	11.38	11.63	11.46	12.01	12.05	12.42
		(0.68)	(0.55)	(0.58)	(0.63)	(0.67)	(0.61)	(0.61)	(0.59)	(0.72)
	BM_{OBS}	4.39	10.90	18.47	10.78	38.23	72.78	27.26	73.52	189.51
3		(0.11)	(0.58)	(1.38)	(0.59)	(3.59)	(6.65)	(1.81)	(8.64)	(18.98)
	BM_{CBD}	**1.97**	5.25	8.33	**3.49**	5.51	6.45	4.97	5.03	5.97
		(0.07)	(0.23)	(0.42)	(0.21)	(0.31)	(0.17)	(0.29)	(0.05)	(0.08)
	EM	2.56	**4.05**	**5.09**	3.65	**4.27**	**4.85**	**4.69**	**4.52**	**4.82**
		(0.06)	(0.13)	(0.09)	(0.12)	(0.06)	(0.15)	(0.13)	(0.05)	(0.08)
	BM_{DOE}	12.69	11.25	13.06	12.59	10.67	11.55	11.63	11.98	11.73
		(0.90)	(0.68)	(0.76)	(0.75)	(0.49)	(0.72)	(0.57)	(0.64)	(0.63)
	BM_{OBS}	3.48	8.41	14.56	7.89	22.92	47.88	13.48	60.87	100.21
5		(0.07)	(0.36)	(1.06)	(0.31)	(1.81)	(5.03)	(1.13)	(6.41)	(12.34)
	BM_{CBD}	**1.62**	4.98	8.97	**3.61**	4.66	7.21	3.97	5.09	6.85
		(0.06)	(0.16)	(0.40)	(0.15)	(0.22)	(0.41)	(0.28)	(0.09)	(0.11)
	EM	2.02	**4.16**	**5.99**	3.72	**3.80**	**5.28**	**3.39**	**4.44**	**4.93**
		(0.07)	(0.07)	(0.20)	(0.10)	(0.06)	(0.20)	(0.06)	(0.07)	(0.08)

Table 6: The number of false selection for Example 1, average of 50 simulation replicates.

		$R_{OBS}/R_{DOE}=1$			$R_{OBS}/R_{DOE}=0.5$			$R_{OBS}/R_{DOE}=0.3$		
n_2/n_1	Method	$\frac{\sigma_2}{\sigma_1}=1$	$\frac{\sigma_2}{\sigma_1}=5$	$\frac{\sigma_2}{\sigma_1}=10$	$\frac{\sigma_2}{\sigma_1}=1$	$\frac{\sigma_2}{\sigma_1}=5$	$\frac{\sigma_2}{\sigma_1}=10$	$\frac{\sigma_2}{\sigma_1}=1$	$\frac{\sigma_2}{\sigma_1}=5$	$\frac{\sigma_2}{\sigma_1}=10$
	BM_{DOE}	3.72	3.58	3.48	3.72	3.54	3.66	3.72	3.16	3.66
1	BM_{OBS}	6.96	7.52	7.40	7.38	7.64	7.78	7.74	7.76	7.58
	BM_{CBD}	2.76	5.34	7.84	2.80	5.86	8.00	3.06	5.00	7.74
	EM	4.02	3.32	4.14	4.00	4.02	3.40	3.52	3.14	3.28
	BM_{DOE}	3.98	3.62	3.70	3.58	3.42	3.68	3.58	3.72	3.80
2	BM_{OBS}	5.22	7.52	7.82	6.50	7.66	7.80	7.24	7.86	8.26
	BM_{CBD}	1.86	5.14	7.86	2.26	5.70	7.96	2.24	7.12	8.00
	EM	3.14	2.96	4.56	2.62	2.88	4.28	2.94	2.98	3.96
	BM_{DOE}	4.12	3.46	3.48	3.50	3.96	3.48	3.46	3.60	3.96
3	BM_{OBS}	4.22	7.46	7.98	6.54	8.08	8.14	7.64	7.92	7.90
	BM_{CBD}	1.50	5.68	8.10	2.34	5.84	8.02	2.46	6.38	8.00
	EM	2.72	3.20	4.90	4.04	3.70	3.74	2.74	2.78	3.46

method considers the precedence structure of two data sets, hence leading to better prediction performance. In the real manufacturing scale-up environment as described in Table 1, the differences of the sample size, uncertainty, and range often become large. In these cases, these reported simulation shows that the proposed EM method achieves a better prediction performance compared with other methods. For some scenarios in Example 2 such as $n_2/n_1 = 1, R_{OBS}/R_{DOE} = 0.3, \sigma_2/\sigma_1 = 5$, the proposed EM may have slightly larger prediction error than BM_{CBD}. This is probably because Example 2 follows the strong heredity principle, which violates the weak heredity assumption used in the proposed method. It is also worth to note that the standard errors of RMSPE in parenthesis for EM are overall smaller than those in the other methods. It implies that the proposed EM provides a reliable and stable prediction performance.

Moreover, Tables 6-8 examine the performance of variable selection in the three examples. Here we compare the number of false selection, i.e., the summation of the number of variables

Table 7: The number of false selection for Example 2, average of 50 simulation replicates.

		$R_{OBS}/R_{DOE}=1$			$R_{OBS}/R_{DOE}=0.5$			$R_{OBS}/R_{DOE}=0.3$		
n_2/n_1	Method	$\frac{\sigma_2}{\sigma_1}=1$	$\frac{\sigma_2}{\sigma_1}=5$	$\frac{\sigma_2}{\sigma_1}=10$	$\frac{\sigma_2}{\sigma_1}=1$	$\frac{\sigma_2}{\sigma_1}=5$	$\frac{\sigma_2}{\sigma_1}=10$	$\frac{\sigma_2}{\sigma_1}=1$	$\frac{\sigma_2}{\sigma_1}=5$	$\frac{\sigma_2}{\sigma_1}=10$
1	BM_{DOE}	22.70	21.68	22.68	22.76	22.74	21.74	21.52	20.58	21.90
	BM_{OBS}	26.32	27.66	26.26	27.00	26.74	27.58	27.58	26.84	28.48
	BM_{CBD}	6.74	14.26	20.10	7.64	14.54	20.00	9.24	13.20	21.72
	EM	11.04	15.88	15.76	13.74	13.40	16.10	14.74	13.92	15.16
2	BM_{DOE}	20.98	22.02	21.82	22.38	21.66	23.68	20.78	22.06	22.04
	BM_{OBS}	13.10	21.74	22.30	20.06	22.28	22.62	20.38	22.06	22.36
	BM_{CBD}	5.16	13.94	20.70	6.00	14.80	21.00	7.56	14.64	20.48
	EM	8.08	11.56	14.76	9.86	12.56	11.80	10.48	12.16	13.08
3	BM_{DOE}	21.70	21.32	21.92	22.54	21.68	22.70	22.92	22.72	21.74
	BM_{OBS}	9.30	20.50	21.86	15.90	21.88	22.52	18.92	22.10	22.10
	BM_{CBD}	2.94	12.64	21.30	5.90	14.76	21.52	7.72	14.10	21.24
	EM	5.06	10.68	16.66	9.06	12.78	12.82	10.08	11.36	12.38

which are false positive and false negative. A smaller number of false selection indicates more accurate selection. Note that Examples 1-3 have 15, 55 and 55 predictors, respectively. The results show that the proposed EM generally have better variable selection accuracy than the other three methods. When n_2/n_1 becomes larger, the variable selection accuracy is improved for all four models. However, the change of R_{OBS}/R_{DOE} gives comparable variable selection performances for all four models. When σ_2/σ_1 is small, BM_{CBD} has the best variable selection performance. But when σ_2/σ_1 becomes larger, the proposed EM method provides more accurate variable selection than BM_{DOE} in Examples 2-3, and has comparable variable selection performance as BM_{DOE} in Example 1. In the all scenarios, the proposed EM method has better variable selection performance than BM_{OBS}. This finding indicates that the EM can borrow the strength of variable selection from the DOE data set.

Table 8: The number of false selection for Example 3, average of 50 simulation replicates.

		$R_{OBS}/R_{DOE} = 1$			$R_{OBS}/R_{DOE} = 0.5$			$R_{OBS}/R_{DOE} = 0.3$		
n_2/n_1	Method	$\frac{\sigma_2}{\sigma_1} = 1$	$\frac{\sigma_2}{\sigma_1} = 5$	$\frac{\sigma_2}{\sigma_1} = 10$	$\frac{\sigma_2}{\sigma_1} = 1$	$\frac{\sigma_2}{\sigma_1} = 5$	$\frac{\sigma_2}{\sigma_1} = 10$	$\frac{\sigma_2}{\sigma_1} = 1$	$\frac{\sigma_2}{\sigma_1} = 5$	$\frac{\sigma_2}{\sigma_1} = 10$
	BM_{DOE}	22.68	23.00	20.32	22.84	22.64	22.22	21.24	20.40	23.04
1	BM_{OBS}	26.90	26.76	27.86	27.26	28.88	26.90	27.28	27.12	27.18
	BM_{CBD}	5.78	13.00	19.68	8.06	14.78	20.16	9.42	13.56	21.32
	EM	11.24	13.76	16.08	13.76	14.88	14.76	14.68	15.84	14.96
	BM_{DOE}	21.86	21.74	24.18	20.72	22.24	21.60	22.24	23.02	22.30
2	BM_{OBS}	13.30	21.74	22.18	17.62	22.46	22.28	20.74	22.12	22.62
	BM_{CBD}	4.76	15.42	21.08	6.08	16.08	20.24	8.20	17.34	20.98
	EM	7.90	13.24	13.12	8.36	13.78	13.26	11.68	14.64	12.04
	BM_{DOE}	22.50	21.96	23.12	21.52	21.82	21.56	21.84	21.98	20.52
3	BM_{OBS}	11.18	21.36	21.94	16.24	21.66	22.32	19.36	22.16	22.16
	BM_{CBD}	4.10	16.52	21.10	5.50	15.50	21.90	8.32	15.42	21.60
	EM	7.06	13.44	13.34	7.42	12.38	12.88	10.68	11.52	12.30

4 Case Study: Wafer Manufacturing

For further demonstrating the effectiveness of the proposed method, a real wafer manufacturing case is studied and discussed here (Ning et al., 2012). Recall the lapping process described in Section 1. In the wafer manufacturing scale-up, the lapping process is an important step to reduce the thickness variation of wafers. As shown in Figure 1, the wafers are placed on the lower plate, while the upper plate will press against the lower plate with rotations in opposite directions. At the same time, abrasive slurry will be fed to remove the silicon material. A lapping process is a key operation to reduce the variation of the geometric variables of wafers, which are treated as major quality measures in wafer manufacturing. Therefore, it is important to identify proper process settings, such that the variation can be reduced. In this case study, the thickness of the wafers after the lapping process (CTHK1) is considered as the quality response of the model, which is to be predicted based on the 10 factors. Detail of these factors are summarized in Table 9. Among the 10 factors, four process variables are controllable variables to affect CTHK1. These four controllable process variables can be adjusted during the DOE and validation production runs. There are also six covariates, which are the quality variables of wafers from the upstream production. These covariates are automatic measured before the lapping process, but cannot be adjusted during the manufacturing process.

Table 9: Measured variables in the lapping process.

Variable Type	**Variable Name**	**Physical Meaning**
Controllable Process Variable	Pressure (N/m^2)	The high pressure of the upper to lower plate
	Rotation (Rpm)	The rotation speed
	LowPTime (Sec.)	The time for low pressure
	HighPTime (Sec.)	The time for high pressure
Covariate	CTHK0 (μm)	Central thickness of wafers
	TTV0 (μm)	Total thickness variation of wafers
	TIR0 (μm)	Total indicator reading of wafers
	STIR0 (μm)	Site total indicator reading of wafers
	BOW0 (μm)	Deviation of local warp at the center of wafers
	WARP0 (μm)	Maximum of local warp of wafers
Quality Response	CTHK1 (μm)	Central thickness of wafers after lapping

In this scale-up effort, an experiment of 2^{4-1} fractional factorial design at level -1 and 1 with two center points at 0 is firstly planned for the controllable process variables. For each run, there are 10 replicates, resulting in 100 samples for the DOE data. After the DOE, further validation production runs are carried out, where 231 samples of the OBS data are used to validate the initial process setting. The initial process setting is optimized to change the values of the four controllable variables based on the covariates (quality measurements from the upstream stages). Then in the validation production runs, the values of the process variables are in the neighborhood of the initial process setting from DOE. For the controllable variables in OBS data, the ranges for Pressure, Rotation, and LowPTime are in [-0.2, 0.2], and the range for HighPTime is in [-0.9, 0.3]. For the covariates, their ranges are in [-3, 3] for both DOE and OBS data.

The proposed EM method is compared with the three benchmark regression models, BM_{DOE}, BM_{OBS} and BM_{CBD}, which follow the same definitions in Section 3. The three benchmark models are estimated using BIC variable selection. The data are randomly partitioned into a training set and a test set with equal sample sizes. The training set is used for variable selection and parameter estimation, and the test set is used to evaluate the model performance. All models in comparison are evaluated on six different data sets: the training and test sets from DOE data ($\boldsymbol{D}_{tr}$ and $\boldsymbol{D}_{ts}$), the training and test set from the OBS data ($\boldsymbol{O}_{tr}$ and $\boldsymbol{O}_{ts}$), and the training and test set from the combined data ($\boldsymbol{C}_{tr}$ and $\boldsymbol{C}_{ts}$). The performance comparison of RMSPE is summarized in Table 10.

From the comparison results, the proposed EM has comparable prediction performance with BM_{CBD} for the DOE test set $\boldsymbol{D}_{ts}$ (2.677 vs. 2.656). In contrast, the proposed EM provides a better prediction than BM_{OBS} for the OBS test set $\boldsymbol{O}_{ts}$ (3.764 vs. 6.270). Under the combined test data $\boldsymbol{C}_{ts}$, the proposed EM method gains the best prediction performance among all four models. It shows that the proposed EM method, obtained through the effective fusion of DOE and OBS data, achieves the best prediction performance compared with the other three benchmark models. Figure 2 demonstrates the variable selection results of the four models. In the figure, each row and each column represents one variable, respectively. The orders of the variables (from left to right, and from top to bottom) follows the order of predictors listed in Table 9. The diagonal blocks represent the main effects of the variables, and the off diagonal blocks represent their two-factor interactions. The dark color indicates that the corresponding predictor is significant. Comparing the patterns from Figure 2(a) and Figure 2(b), we note that the BM_{DOE} and BM_{OBS} are not consistent in terms of significant predictor variables. In contrast, BM_{CBD} has a very similar variable selection performance as BM_{OBS} shown in Figure 2(b) and Figure 2(c). This can be due to the fact that the sample size of the OBS data is more than twice as the DOE data. The variable selection could

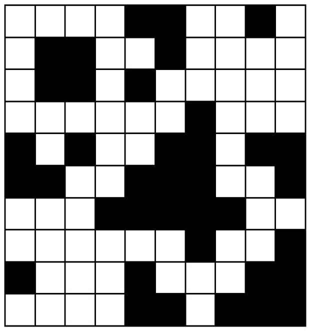
Table 10: Comparison of the testing RMSPE for the lapping process case.

	$\boldsymbol{D}_{tr}$	$\boldsymbol{D}_{ts}$	$\boldsymbol{O}_{tr}$	$\boldsymbol{O}_{ts}$	$\boldsymbol{C}_{tr}$	$\boldsymbol{C}_{ts}$
BM_{DOE}	1.167	3.847	13.701	20.495	11.467	17.241
BM_{OBS}	3.950	3.466	3.892	7.475	3.910	6.525
BM_{CBD}	2.461	2.656	4.102	6.270	3.684	5.435
EM	2.461	2.677	4.348	3.764	3.877	3.471

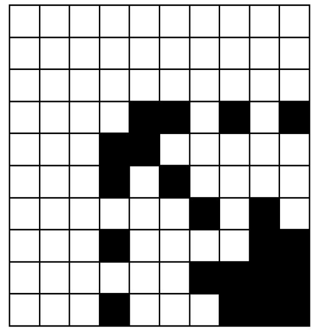
be influenced more by the OBS data set. In this case, none of the first four controllable process variables for the DOE factors are identified as significant variables. A possible explanation is that BM_{CBD} model overlooks the sequential nature and inherent features of the two types of data. As shown in Figure 2(d), the proposed EM successfully identifies some significant variables from DOE in ensemble modeling, illustrating the effectiveness of variable selection in our proposed ensemble modeling strategy.

5 Conclusion and Discussion

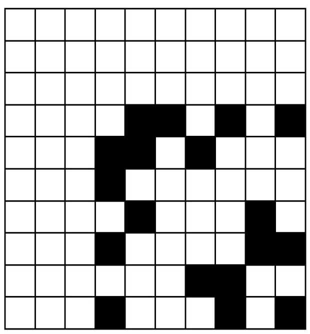
Manufacturing scale-up is an important, yet a time-consuming and expensive process in product realization. It involves both experiments and validation production runs of a manufacturing process for obtaining an adequate model. In this paper, we propose an ensemble modeling strategy to integrate both DOE and OBS data for the manufacturing scale-up. The proposed method can provide an accurate model, in which the selected significant variables reflect the sequential nature and inherent feature of the two types of data. Thus the variable selection from the propose method is more meaningful to reflect the manufacturing process. This helps us identify an adequate model more quickly in the manufacturing scale-up. As a result, fewer rounds of data collection and modeling are expected. It is worth to point out that the proposed ensemble modeling method will not only be suitable for quality-process modeling, but also applicable for improving yield and reducing cost, where the regression analysis can be generally used. The proposed method therefore



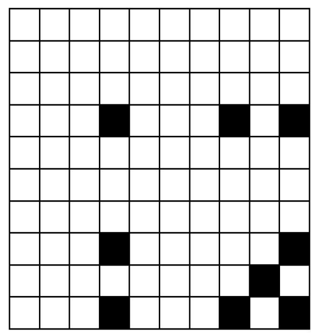
(a)



(b)



(c)



(d)

Figure 2: Variable selection on the wafer data for (a) BM_{DOE} (b) BM_{OBS} (c) BM_{CBD} (d) EM. The order of predictors are Pressure, Rotation, LowPTime, HighPTime, CTHK0, TTV0, TIR0, STIR0, BOW0, WARP0.

can significantly reduce lead time in the manufacturing scale-up.

In the proposed method, we consider the frequentists' likelihood estimation approach, with constraints to encourage the significant predictors in the DOE model also to be significant in the OBS model. It relies on the correctness and completeness of the DOE data to identify significant predictors. In addition to the likelihood approach, one can also consider the Bayesian analysis (Reese et al., 2004) to integrate two types of data, where the findings from DOE serve as the prior information for modeling the OBS data. On the other hand, if additional engineering knowledge is useful to identify some significant predictor variables and/or their interactions, we can extend the proposed method by adding more constraints into the optimization problem (9), enabling an engineering-driven data fusion framework.

In this work, we treat the quality response as a continuous variable, where a linear model is used to link the quality response and predictor variables. When the response is binary or categorical, the proposed method can be generalized by using more flexible models such as generalized linear models (McCullagh and Nelder, 1989). Beside using the nonnegative garrotes for variable selection, a future research direction is to investigate other variable selection methods (Hastie et al., 2009) for the efficient fusion of different data sets.

Another future research direction is to advance the improvement on modeling DOE data and OBS data for maximizing the overall prediction accuracy. When the DOE is poorly designed, the DOE data cannot provide an adequate model for significant predictors. In addition, if the OBS data contain very high uncertainty, it may not improve the overall modeling accuracy. Efforts are needed to make the overall ensemble modeling accuracy satisfy the manufacturing scale-up requirements.

References

Basumallick, A., Das, G. C., and Mukherjee, S. (2003) Design of experiments for synthesizing in situ Ni-SiO2 and CO-SiO2 nanocomposites by non-isothermal reduction treatment, *Nanotechnology.* **14**(8), 903–906.

Breiman, L. (1995) Better subset regression using the nonnegative garrote. *Technometrics*, **37**, 373–384.

Boyd, S. and Vandenberghe., L. (2004) *Convex Optimization*, Cambridge University Press, New York, NY.

Burnham, K. P. and Anderson, D. R. (2002) *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, 2nd Edition, Springer-Verlag, New York, NY.

Choi, N. H., Li, W., and Zhu, J. (2010) Variable selection with the strong heredity constraint and its oracle property. *Journal of the American Statistical Association*, **105**(489), 354–364.

Dasgupta, T., Ma, C., Joseph, V. R., Wang, Z. L., and Wu, C. F. J. (2008) Statistical modeling and analysis for robust synthesis of nanostructures. *Journal of the American Statistical Association*, **103**, 594–603.

Dasgupta, T. and Wu, C. F. J. (2006) Robust parameter design with feedback control. *Technometrics*, **48**, 349–360.

Fong, D. and Lawless, J. F. (1998) The analysis of process variation transmission with multivariate measurements. *Statistica Sinica*, **8**, 151–164.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2009) *The Element of Statistical Learning*, 2nd Edition, Springer-Verlag, New York, NY.

Jin, J. and Ding, Y. (2004) Online automatic process control using observable noise factors for discrete part manufacturing. *IIE Transactions*, **36**, 899–911.

Jin, R. and Shi, J. (2012) Reconfigured piecewise linear regression tree for multistage manufacturing process control. *IIE Transactions*, **44**, 249–261.

Joseph, R. (2003) Robust parameter design with feed-forward control. *Technometrics*, **45**, 284–292.

McCullagh, P. and Nelder, J. (1989) *Generalized Linear Models*, 2nd Edition, Chapman and Hall/CRC, Boca Raton, FL.

Ning, Y., Bian, Y., and Liu, B (2012) Improving a lapping process using robust parameter design and run-to-run control. *Journal of the Chinese Institute of Industrial Engineers*, **29**(2), 111–124.

Parker, S. (2002) *McGraw-Hill Dictionary of Scientific and Technical Terms*, 6th Edition, McGraw-Hill, Inc., New York, NY.

Reese, C., Wilson, A., Hamada, M., Martz, H. F., and Ryan, K. J. (2004) Integrated analysis of computer and physical experiments. *Technometrics*, **46**, 153–164.

Shi, J. (2006) *Stream of Variation Modeling and Analysis for Multistage Manufacturing Processes*, CRC Press, Taylor & Francis Group, Boca Raton, FL.

Wu, C. F. J. and Hamada, M. (2009) *Experiments: Planning, Analysis, and Optimization*, 2nd Edition, Wiley, New York, NY.

Wu, S., Zou, H., and Yuan, M (2008) Structured variable selection in support vector machines. *Electronic Journal of Statistics*, **2**, 103–117.

Yuan, M., Joseph, R., and Zou, H. (2009) Structured variable selection and estimation. *Annals of Applied Statistics*, **3(4)**, 1738–1757.

Yuan, M. and Lin, Y. (2007) On the nonnegative garrote estimator. *Journal of the Royal Statistical Society, Series B*, **69**(2), 143–161.

Zhong, J., Shi, J., and Wu, C. F. J. (2010) Design of DOE-based automatic process controller with consideration of model and observation. *IEEE Transactions on Automation Science and Engineering*, **7**, 266–273.