IIE Transactions

ISSN: 0740-817X (Print) 1545-8830 (Online) Journal homepage: http://www.tandfonline.com/loi/uiie20

Logistic regression for crystal growth process modeling through hierarchical nonnegative garrote-based variable selection

Hongyue Sun, Xinwei Deng, Kaibo Wang & Ran Jin

To cite this article: Hongyue Sun, Xinwei Deng, Kaibo Wang & Ran Jin (2016) Logistic regression for crystal growth process modeling through hierarchical nonnegative garrote-based variable selection, IIE Transactions, 48:8, 787-796, DOI: 10.1080/0740817X.2016.1167286

To link to this article: http://dx.doi.org/10.1080/0740817X.2016.1167286

IIE TRANSACTIONS
2016, VOL. 48, NO. 8, 787–796
http://dx.doi.org/10.1080/0740817X.2016.1167286

Taylor & Francis
Taylor & Francis Group

Logistic regression for crystal growth process modeling through hierarchical nonnegative garrote-based variable selection

Hongyue Sun[a], Xinwei Deng[b], Kaibo Wang[c] and Ran Jin[a]

[a]Grado Department of Industrial and Systems Engineering, Virginia Tech, Blacksburg, VA, USA; [b]Department of Statistics, Virginia Tech, Blacksburg, VA, USA; [c]Department of Industrial Engineering, Tsinghua University, Beijing, People's Republic of China

ABSTRACT

Single-crystal silicon ingots are produced from a complex crystal growth process. Such a process is sensitive to subtle process condition changes, which may easily become failed and lead to the growth of a polycrystalline ingot instead of the desired monocrystalline ingot. Therefore, it is important to model this polycrystalline defect in the crystal growth process and identify key process variables and their features. However, to model the crystal growth process poses great challenges due to complicated engineering mechanisms and a large amount of functional process variables. In this article, we focus on modeling the relationship between a binary quality indicator for polycrystalline defect and functional process variables. We propose a logistic regression model with hierarchical nonnegative garrote-based variable selection method that can accurately estimate the model, identify key process variables, and capture important features. Simulations and a case study are conducted to illustrate the merits of the proposed method in prediction and variable selection.

ARTICLE HISTORY
Received 12 April 2014
Accepted 29 February 2016

KEYWORDS
Crystal growth; logistic regression; polycrystalline; process modeling; variable selection

1. Introduction

Wafer manufacturing is an important upstream process for many high-tech products, such as computer electronics, automatic control devices, solar cells, etc. Such a manufacturing process consists of many stages, including crystal growth, wire slicing, etching, lapping, polishing, etc. The crystal growth process is the first step to produce a silicon ingot, which determines the initial quality for downstream products. Therefore, it is extremely important to control the quality at this stage.

The majority of crystal ingots used in industry are grown by the Czochralski crystal growth process (CZ process); see Fisher *et al.* (2012) for details. A successful CZ process is maintained at an extremely high temperature for more than 60 hours. The process can be divided into the following phases (Zulehner, 1983; Dhanaraj *et al.*, 2010). First, the polycrystalline silicon is melted in a silica crucible. Then, a precisely oriented seed crystal is dipped into the melt. Then, by jointly controlling the thermal gradient and pulling speed, the ingot grows to the desired diameter. Afterwards, the ingot is slowly pulled upwards and simultaneously rotated. This pulling and rotation process lasts for more than 20 hours, which is called the "body growth phase." This body growth phase is the most important phase during a CZ process, since the majority of an ingot is grown in this phase. Finally, the ingot finishes its growth after a tailing phase. The above ingot growth process takes place in industrial CZ furnaces, as shown in Fig. 1(a) (Zhu *et al.*, 2014). Inside the furnace, the structure and operation conditions in the hot zone are critical for the ingot growth (Fig. 1(b), Zhang *et al.*, 2014).

Due to the high energy consumption and long cycle time in the CZ process, any quality defect of the ingot results in large levels of waste in terms of energy, time, and cost. The quality defects include microscopic defects and macroscopic defects (Dhanaraj *et al.*, 2010). Examples of microscopic defects are voids, interstitials, dislocations, etc., which affect the electronic and mechanical properties of downstream products (Mahajan, 2000). The macroscopic defects are more severe and may cause failure of the entire growth process. In such a situation, the manufacturer has to discard the nonconforming segments of the ingot or remelt the material and repeat the growth process, which leads to further waste. Among these macroscopic defects, polycrystalline defects are the most frequently observed type. Polycrystalline defects refer to the phenomenon that the desired monocrystalline ingot becomes polycrystalline. Once a segment of the ingot becomes polycrystalline, the entire segment has to be discarded (Zhang *et al.*, 2014). Thus, it is critical to reduce this type of quality defect during the manufacturing process. In the literature, defect analysis in crystal growth is mainly focused on microscopic defects (Voronkov, 1982; Sinno *et al.*, 2000; Brown *et al.*, 2001; Dhanaraj *et al.*, 2010). In this article, we focus on modeling polycrystalline defects during the body growth phase, since the majority of polycrystalline defects appear in this phase.

To model the polycrystalline defect, we use a binary variable as the indicator for the formation of polycrystalline defects and propose a logistic regression model to model the binary quality variable (response) with the functional process variables (predictors). Engineering knowledge suggests that the features of

CONTACT Ran Jin jran5@vt.edu
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/uiie.
Supplemental data for this article can be accessed at www.tandfonline.com/uiie

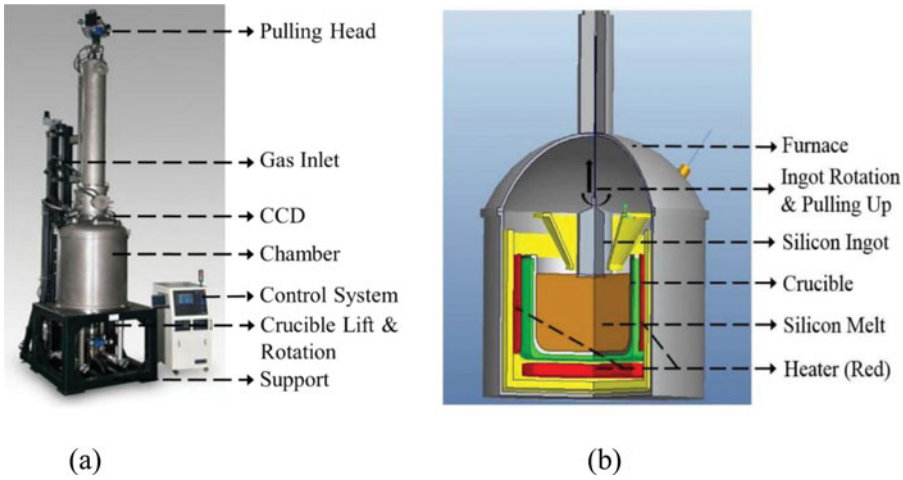

Figure 1. A schematic of a crystal growth furnace: (a) the furnace and (b) the internal structure of the chamber (hot zone). Reproduced from Zhu *et al.* (2014) and Zhang *et al.* (2014).

the process variables should be captured, as sudden changes in the process variables are potential root causes for polycrystalline defects. Therefore, we adopt wavelet analysis for each functional process variable. Wavelet analysis is selected due to its excellent performance in extracting features from local time and frequency (Mallat, 1989). Thus, all of the wavelet coefficients of a functional process variable form a group of features. In this article, the wavelet coefficients of a process variable are called "features" or "local features" and each process variable has a "group" of corresponding features. The objective is to identify both key process variables and significant features. Therefore, a logistic regression with Hierarchical Non-Negative Garrote (HNNG)-based variable selection is used.

The Non-Negative Garrote (NNG) proposed by Breiman (1995) is a shrinkage method for estimating a parsimonious model. The NNG was first proposed for variable selection in linear models (Breiman, 1995; Jin and Deng, 2015). Makalic and Schmidt (2011) developed an NNG for logistic regression models. Consistency in prediction and variable selection of the NNG was studied in Yuan and Lin (2007). However, none of the existing NNG-based variable selection methods can address the aforementioned two-level variable selection problem in a logistic regression model. In this article, the newly proposed HNNG method can identify significant groups (representing functional process variables) as well as local features (representing wavelet coefficients from the functional process variables) to predict the binary response. The advantages of the HNNG method lie in several aspects. First, the proposed HNNG method performs simultaneous variable selection for both significant groups and features. Second, the computation issues are addressed by quadratic approximation of the objective function. Third, the polycrystalline defect can be predicted in a timely manner based on the measurements. Specifically, we divide the measurements into windows with binary quality labels given by the domain expert. In each time window, wavelet analysis is adopted for the measurements and the corresponding wavelet coefficients are treated as predictors in the logistic regression. Therefore, the model can predict whether the ingot becomes polycrystalline for each window.

The rest of this article is organized as follows. In Section 2, the state-of-the-art for CZ process modeling, variable selection, and wavelet analysis are reviewed. Section 3 illustrates the proposed method and the computation algorithm. We demonstrate the effectiveness of the proposed method in prediction and variable selection by using simulations and a case study in Sections 4 and 5, respectively. Finally, conclusions and future research are discussed in Section 6.

2. State-of-the-art

Engineering models are available for simulation and defect analysis of CZ processes. Simulation models mainly focused on predicting the thermal field distribution of the system for equipment design. Such models are typically based on Partial Differential Equations (PDEs) that are used to describe the growth dynamics (Derby and Brown, 1986; Fischer *et al.*, 2005). Müller (2002) proposed the concept of reverse simulation, which aimed at controlling a certain kind of defect given the defect–growth process relationships. In most cases, these simulation models were solved offline using finite element methods. The performance of simulation models depends on the engineering assumptions, boundary conditions, and accuracy of the material property characterizations. These models cannot be used to model the creation of polycrystalline defects with potential online prediction requirements. Another category of models focus on microscopic defects; they are typically used to model the distribution of microscopic defects as a function of process variables. Voronkov (1982) concluded that the ratio of the crystal pulling speed to the magnitude of temperature gradient above the solid–liquid interface determined the formation of point defects. The formation of larger-scale defects, such as oxidation-induced stacking faults, was also modeled. Comprehensive reviews of defect modeling have been presented by Sinno *et al.* (2000) and Brown *et al.* (2001). However, these models focused on microscopic defects, and there were limited engineering-driven models that could be used to quantitatively predict the polycrystalline defects.

Researchers have attempted to model the CZ processes by using statistics, optimization, and data mining methods. For instance, time series analysis for the dynamic properties of striations in the ingot has been explored (Miyano and Shintani, 1993;

Shintani *et al.*, 1995). Back-propagation, regularization, and perceptron neural networks have been used to analyze the creation of ingot striation patterns. In addition, a genetic algorithm, coupled with a PDE to describe thermal effects, was used to optimize the configuration of the heat shield on a CZ furnace (Fühner and Jung, 2004). As another example, Avci and Yamacli (2010) used an artificial neural network to modify a PDE that was used to describe the defect concentration. This method yielded a highly accurate prediction for the defect concentration.

To model a binary quality variable using functional process variables, one can formulate this problem as a classification problem. Data mining methods—for instance, linear discriminant analysis, support vector machines, classification and regression tree, and random forests—can be applied. See Hastie *et al.* (2009) for details. A functional logistic regression model can also be used to link the binary response and functional predictors (Ratcliffe *et al.*, 2002). In this article, we adopt the latter approach. To improve the performance of the model as well as its interpretability, different kinds of variable selection methods have been proposed in the literature. These methods include subset and stepwise regression (Miller, 2002), Akaike information criterion (Akaike, 1974), Bayesian information criterion (BIC; Schwarz (1978)), Lasso (Tibshirani, 1996), NNG (Breiman, 1995), smoothly clipped absolute deviation (Fan and Li, 2001), and elastic net (Zou and Hastie, 2005). However, the penalization methods introduced above may not perform well for variable selection with a group structure. To address this problem, Yuan and Lin (2006) proposed the Group Lasso approach. Zhao *et al.* (2009) proposed the use of flexible composite absolute penalties. Meier *et al.* (2008) studied the group variable selection for logistic regression via Group Lasso (GrpLasso). Although these methods usually have a better performance than traditional methods, they can only select the group as a whole and cannot select features within the group, as stated by Huang *et al.* (2009), Zhou and Zhu (2010), and Paynabar *et al.* (2015).

To deal with the hierarchical variable selection problem, Huang *et al.* (2009) proposed the Group Bridge (GrpBridge) approach. However, the GrpBridge penalty is not always differentiable and tends to be inconsistent for feature selection (Huang *et al.*, 2012). Zhou and Zhu (2010) proposed the Hierarchical Lasso (HLasso) approach, which penalizes the coefficients using two levels of L_1 penalty. Paynabar *et al.* (2015) claimed that this method may fall into a local optimum. They proposed a hierarchical NNG for group variable selection in linear regression: first identify the important groups and then the important features within the selected groups in two separate steps. They demonstrated that their hierarchical NNG performed well in prediction and variable selection for linear regression models. In this article, we explore hierarchical variable selection for a logistic regression via HNNG. The advantage of HNNG is that it can simultaneously select important groups and features in one step. It should be noted that the hierarchical NNG method proposed by Paynabar *et al.* (2015) focused on linear regression models, whereas we focus on logistic regression models.

In this study, wavelet analysis is used to transform a functional variable into a group of wavelet features. Wavelet analysis is a multi-resolution analysis tool that can provide both localized time and frequency information (Mallat, 1989). We use wavelet analysis so that the features from local time and frequency can represent the subtle changes in process variables, which might lead to polycrystalline defects. Wavelet analysis has been widely adopted in engineering applications for quality improvement. For instance, Jin and Shi (1999) applied wavelet analysis for data compression of the force signal in a stamping process. Subsequently, Jin and Shi (2001) used the wavelet analysis approach to diagnose faults in the stamping process. Other applications include nano-machining (Ganesan *et al.*, 2004), a forging process (Zhou and Jin, 2005), structural health monitoring (Bukkapatnam *et al.*, 2005), antenna (Jeong *et al.*, 2006), a rolling process (Li *et al.*, 2007), and an engine assembly process (Paynabar and Jin, 2011).

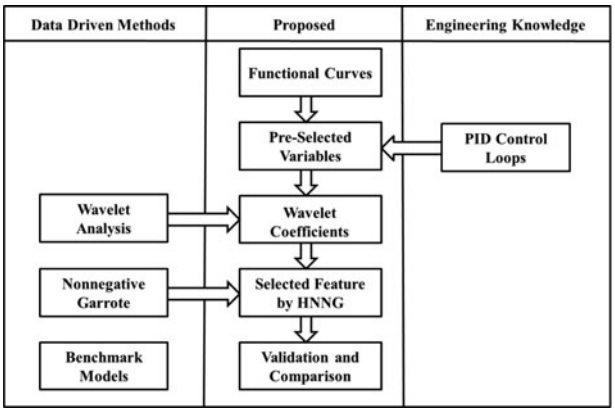

Figure 2. Overview of the proposed method.

3. The proposed method

3.1. Overview of the proposed method

An overview of the proposed method is shown in Fig. 2. The potentially important process variables are selected for the modeling study based on the Proportional-Integral-Derivative (PID) control loops of the CZ process. Wavelet analysis is then adopted for each process variable. Then we use HNNG-based logistic regression to predict the binary response based on groups of wavelet coefficients. Finally, our proposed method is compared with other benchmark methods.

3.2. Data structure

Assuming that we have p functional process variables to be modeled, the number of dilations in the wavelet analysis is set to be m. After wavelet decomposition, we have m levels of detailed coefficients and one level of coarse coefficients. The original process variable is formulated in the structure shown in Table 1, where $p_1, p_2, \ldots, p_m$ and p_c are the number of wavelet coefficients in each level. We denote $P_j = \sum_{i=1}^{m} p_i + p_c$ to be the number of features in the jth process variable and $P = \sum_{j=1}^{p} P_j$ to be the total number of features for p process variables. For each sample, there are P predictors with the structure shown in Table 1 and one binary response y_i. In total, there are n samples for modeling.

Table 1. Data structure after wavelet decomposition.

Detail level 1	*Detail level 2*	…	*Detail level m*	*Coarse level*
$x_{1,1}\, x_{2,1}\, \cdots\, x_{p_1,1}$	$x_{1,2}\, x_{2,2}\, \cdots\, x_{p_2,2}$	…	$x_{1,m}\, x_{2,m}\, \cdots\, x_{p_m,m}$	$x_{1,c}\, x_{2,c}\, \cdots\, x_{p_c,c}$

3.3. HNNG-based logistic regression model

The logistic regression model has the form illustrated in Equation (1):

$$\text{logit}\left(E\left[y_i|\boldsymbol{x}_i\right]\right) = \log\frac{p\left(\boldsymbol{x}_i\right)}{1-p\left(\boldsymbol{x}_i\right)} = \boldsymbol{x}_i{}^T\boldsymbol{\beta},\ i=1,\ \ldots,\ n, \tag{1}$$

where y_i is the binary response for the ith sample, with $y_i = 0$ indicating a conforming growth sample and $y_i = 1$ indicating a polycrystalline growth sample; $p(\boldsymbol{x}_i)$ is the probability that the ith sample is polycrystalline (i.e., $y_i = 1$); $\boldsymbol{x}_i = (\boldsymbol{x}_{1,i}^T, \boldsymbol{x}_{2,i}^T, \ldots, \boldsymbol{x}_{p,i}^T)^T = (x_{1,1,i}, x_{2,1,i}, \ldots, x_{P_1,1,i}, x_{1,2,i}, x_{2,2,i}, \ldots, x_{P_2,2,i}, \ldots, x_{1,p,i}, x_{2,p,i}, \ldots, x_{P_p,p,i})^T$ is the predictor vector for the ith sample, where $x_{k,j,i}$ is the kth feature in the jth group for the ith sample. In the above notations, there are p groups of process variables and P_j features in each process variable. $\boldsymbol{\beta} = (\beta_1^{(1)},\ \beta_2^{(1)}, \ldots, \beta_{P_1}^{(1)}, \beta_1{}^{(2)}, \beta_2^{(2)}, \ldots, \beta_{P_2}^{(2)}, \ldots,\ \beta_1^{(p)},\ \beta_2^{(p)}, \ldots, \beta_{P_p}^{(p)})^T$ is model coefficient vector with $\beta_k^{(j)}$ the coefficient for the kth feature in the jth group.

As discussed above, the NNG can be used to enforce a parsimonious model. It reparameterizes the model coefficient vector $\boldsymbol{\beta} = \boldsymbol{\theta}\cdot\tilde{\boldsymbol{\beta}}$, where $\boldsymbol{\theta} = (\theta_1^{(1)},\ \theta_2^{(1)} \ldots, \theta_{P_1}^{(1)}, \theta_1^{(2)}, \theta_2^{(2)} \ldots, \theta_{P_2}^{(2)}, \ldots, \theta_1^{(p)},\ \theta_2^{(\mathrm{p})} \ldots, \theta_{P_p}^{(p)})^T$ is the shrinkage vector (with each element nonnegative) to encourage variable selection, and $\theta_k^{(j)}$ is the shrinkage factor for the kth feature in the jth group; the "·" stands for element-wise multiplication; and $\tilde{\boldsymbol{\beta}}$ is an initial estimate of the model coefficients, which can be estimated by maximum likelihood estimation. If $\theta_k^{(j)} = 1$, the corresponding coefficient $\beta_k^{(j)}$ is estimated as the initial estimate. When $\theta_k^{(j)} = 0$, the corresponding coefficient shrinks to zero, and the predictor is not selected in the model. To perform variable selection with the hierarchical group structure shown in Table 1, some adjustments have to be made to the approach. Specifically, we design two levels of constraints and minimize the negative log-likelihood through the following optimization problem:

$$\begin{aligned}
&\min \mathrm{L}\left(\boldsymbol{\beta}\right) = -\log\left\{\prod\nolimits_{i=1}^{n}\left[p(\boldsymbol{x}_i)^{y_i}\left(1-p\left(\boldsymbol{x}_i\right)\right)^{1-y_i}\right]\right\},\\
&\text{subject to}\ :\ \beta_k^{(j)} = \theta_k^{(j)}\tilde{\beta}_k^{(j)},\ \theta_k^{(j)} \geq 0,\ \forall j,\ k,\\
&\qquad\qquad \sum_{k=1}^{P_j}\theta_k^{(j)} \leq \gamma_j,\ 0 \leq \gamma_j \leq P_j,\\
&\qquad\qquad \sum_{j=1}^{p}\gamma_j \leq M,\ 0 \leq M \leq P,
\end{aligned} \tag{2}$$

where γ_j is the shrinkage factor for the jth group and $\boldsymbol{\gamma} = (\gamma_1,\ \gamma_2,\ \ldots,\ \gamma_p)^T$ is the shrinkage vector for different groups. The optimization problem determines the optimal $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ to minimize the objective function. In this optimization problem, we have several constraints. $\beta_k^{(j)} = \theta_k^{(j)}\tilde{\beta}_k^{(j)},\ \theta_k^{(j)} \geq 0,\ \forall j,\ k$ are the constraints for NNG to encourage general variable selection. The first level of constraints $\sum_{k=1}^{P_j}\theta_k^{(j)} \leq \gamma_j,\ 0 \leq \gamma_j \leq P_j$ controls the number of features selected within the group. The upper limit of γ_j is set to be P_j, which is the number of coefficients in each group. The second level of constraints $\sum_{j=1}^{p}\gamma_j \leq M,\ 0 \leq M \leq P$ controls the number of groups selected. The upper limit of M is set to be P, which is the total number of coefficients. These upper limits are recommended to be used if no prior knowledge on variable importance is available. The intuition behind these selections is to allow the least squares estimation of the model coefficients in the feasible region (i.e., when $\theta_k^{(j)} = 1$ for all k and j). If the group level shrinkage γ_j becomes zero, then all feature coefficients in the jth group will be zero, which indicates that the jth process variable is not significant and *vice versa*. If the feature level shrinkage $\theta_k^{(j)}$ becomes zero, then the kth feature in the jth group will not be significant and *vice versa*. Here M is a tuning parameter that can be selected based on the BIC, the validation data set, or Cross Validation (CV; Hastie *et al.* (2009)).

To facilitate fast computation for Equation (2), we adopt a similar approach to that of Deng and Jin (2015) and use a second-order Taylor expansion at the current estimate of $\boldsymbol{\beta}$ to approximate the objective function and update this approximation iteratively. After Taylor expansion, the objective function has a quadratic form as shown in Equation (3):

$$\min \mathrm{L}\left(\boldsymbol{\beta}\right) = 1/2\left(\tilde{\boldsymbol{Y}} - \boldsymbol{X}\boldsymbol{\beta}\right)^T\boldsymbol{W}\left(\tilde{\boldsymbol{Y}} - \boldsymbol{X}\boldsymbol{\beta}\right), \tag{3}$$

where $\boldsymbol{W} = \text{diag}(p(\boldsymbol{x}_1)(1-p(\boldsymbol{x}_1)), \ldots, p(\boldsymbol{x}_n)(1-p(\boldsymbol{x}_n)))$ is an $n \times n$ diagonal matrix and $\tilde{\boldsymbol{Y}} = \boldsymbol{X}\tilde{\boldsymbol{\beta}} + \boldsymbol{W}^{-1}(\boldsymbol{Y} - \boldsymbol{p})$, $\boldsymbol{X} = (\boldsymbol{x}_1,\ \ldots, \boldsymbol{x}_n)^T$, $\boldsymbol{Y} = (y_1, \ldots, y_n)^T$, $\boldsymbol{p} = (p(\boldsymbol{x}_1), \ldots,\ p(\boldsymbol{x}_n))^T$. This quadratic programming guarantees a global optimum and a brief derivation is provided in the Appendix. In this way, our method can simultaneously select the significant groups and features with all computational issues having been addressed. The optimal solution to minimize Equation (3) can be obtained by following Algorithm 1.

Algorithm 1.

Step 1. Compute the initial estimate $\tilde{\boldsymbol{\beta}}$, choose the range of tuning parameter M, and set the initial values for $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$.

Step 2. Solve for the $\boldsymbol{\beta}$ with the objective functions defined in Equation (3) and denote the current $\boldsymbol{\beta}$ as $\boldsymbol{\beta}^j$ at the jth iteration.

Step 3. Check the convergence. The problem converges if $||\boldsymbol{\theta}^j - \boldsymbol{\theta}^{j-1}|| < \delta$. If not, update $\tilde{\boldsymbol{\beta}}$=$\boldsymbol{\beta}^j$ and go back to Step 2. δ is a predetermined threshold; e.g., $\delta = 10^{-3}$. ▲

Some practical suggestions for the initial value selection in Algorithm 1 are provided as follows. First, the initial estimates should not contain many zero terms. In our problem, the ridge regression coefficients are used as initial estimates. Such initial

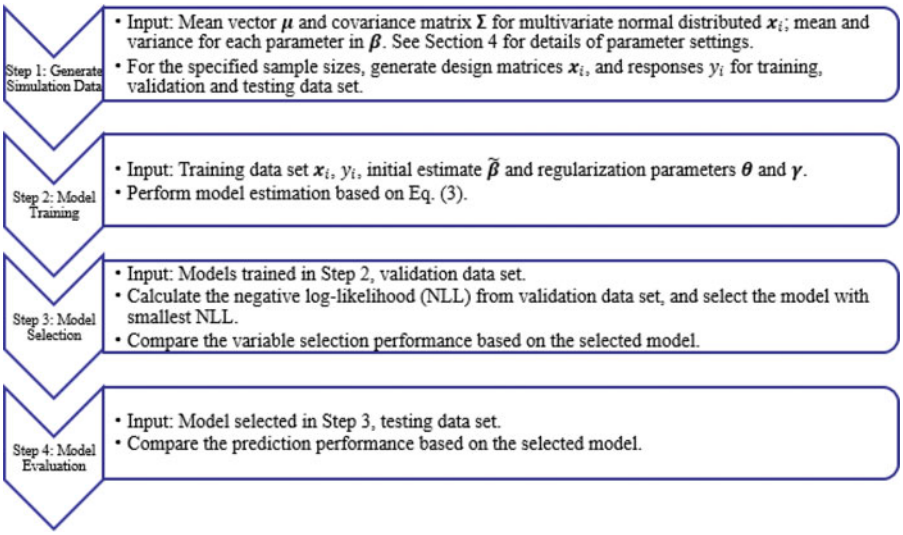

Figure 3. Illustration of the simulation procedure.

estimates are also recommended by Yuan and Lin (2006) and Makalic and Schmidt (2011). Second, the tuning parameter M varies from a small value (e.g., 0.1) to the total number of coefficients under study. Third, due to the quadratic approximation in Equation (2), the optimization will reach the global optimum. The initial values of $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ will not affect the optimal solutions. The initial values of $\boldsymbol{\theta}$ and $\boldsymbol{\gamma}$ in this work are both set to one.

4. Simulations

To evaluate the prediction and variable selection performance of the proposed method, we conducted simulations under different scenarios. For each scenario, the simulation procedure followed the steps listed in Fig. 3.

In the simulation, the response y_i followed the binominal distribution

$$y_i = \begin{cases} 1 & \text{w.p. } p(\boldsymbol{x}_i) \\ 0 & \text{w.p. } 1 - p(\boldsymbol{x}_i) \end{cases}, \tag{4}$$

where $p(\boldsymbol{x}_i) = e^{\boldsymbol{x}_i{}^T\boldsymbol{\beta}}/(1 + e^{\boldsymbol{x}_i{}^T\boldsymbol{\beta}})$ and "w.p." stands for "with probability." The predictors followed a multivariate normal distribution with mean vector $\boldsymbol{\mu} = (\mathbf{0}, \mathbf{0}, \ldots, \mathbf{0})$ and covariance matrix

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\rho}_{11} & \boldsymbol{\tau}_{12} & \ldots & \boldsymbol{\tau}_{1p} \\ \boldsymbol{\tau}_{12} & \boldsymbol{\rho}_{22} & \ldots & \boldsymbol{\tau}_{2p} \\ \ldots & \ldots & \ldots & \ldots \\ \boldsymbol{\tau}_{1p} & \boldsymbol{\tau}_{2p} & \ldots & \boldsymbol{\rho}_{pp} \end{bmatrix},$$

which were used to represent the wavelet coefficients of functional process variables. $\boldsymbol{\rho}_{ii}$ is the covariance matrix within a group and $\boldsymbol{\tau}_{ij}$ is the covariance matrix among groups. The number of groups was set to be four and the number of features in each group was set to be five. In total, we had 20 predictors. To evaluate the performance of the proposed method, we tested its performance by varying sample size, correlation structure, and sparsity of predictors.

Specifically, we denoted the sample sizes for training data sets, validation data sets, and testing data sets as n_{tr}, n_{va}, and n_{te}; we chose n_{tr} to be 20, 100, 200 and set $n_{va} = n_{tr}$ and $n_{te} = 2n_{tr}$. These training, validation, and testing data sets were generated from the same model as shown in Equation (4). The covariance matrix of predictors within and among groups was set to be

$$\boldsymbol{\rho} = \begin{bmatrix} 1 & \rho^{|i-j|} & \ldots & \rho^{|i-j|} \\ \rho^{|i-j|} & 1 & \ldots & \rho^{|i-j|} \\ \ldots & \ldots & \ldots & \ldots \\ \rho^{|i-j|} & \rho^{|i-j|} & \ldots & 1 \end{bmatrix}$$

$$\text{and } \tau = \begin{bmatrix} \tau & \tau^{|i-j|+1} & \ldots & \tau^{|i-j|+1} \\ \tau^{|i-j|+1} & \tau & \ldots & \tau^{|i-j|+1} \\ \ldots & \ldots & \ldots & \ldots \\ \tau^{|i-j|+1} & \tau^{|i-j|+1} & \ldots & \tau \end{bmatrix},$$

respectively, where i and j are the row and column indices of the matrix $\boldsymbol{\rho}$ and $\boldsymbol{\tau}$. Two levels of correlation were selected for ρ and τ, and there were four combinations for the correlation structure. Specifically, the within-group correlation coefficient ρ was set to be zero or 0.6, and between-group correlation coefficient τ was set to be zero or 0.3. The sparsity (denoted as S) represents the proportion of significant predictors in the underlying model, and it was set to be 10% or 40%. The coefficient for a significant predictor $\beta_k^{(j)}$ was taken to follow the normal distribution $N(\mu_j,\ 0.1)$ and $\mu_j = 1,\ 1.3,\ 1.6,\ 1.9$, respectively, for the four groups of coefficients. In summary, there were three levels of sample size, four combinations of covariance structure, and two levels of sparsity. In total, 24 scenarios of simulation settings were evaluated.

We compared our proposed method with Logistic Regression (LR), Lasso, Ridge, NNG, GrpLasso, and HLasso methods for the binary response prediction. We used the training data set to obtain the regression models and used the validation data set for the tuning parameter selection. The model with the selected tuning parameter was used to compare variable selections. We used a threshold to determine whether the coefficient was significant or not. If the magnitude (absolute value) of the coefficient was larger than the threshold, then the corresponding predictor was considered to be significant. Specifically, the threshold was set to be 10^{-6}. Then we compared the misclassification errors of the testing data set (called "testing error") for the proposed

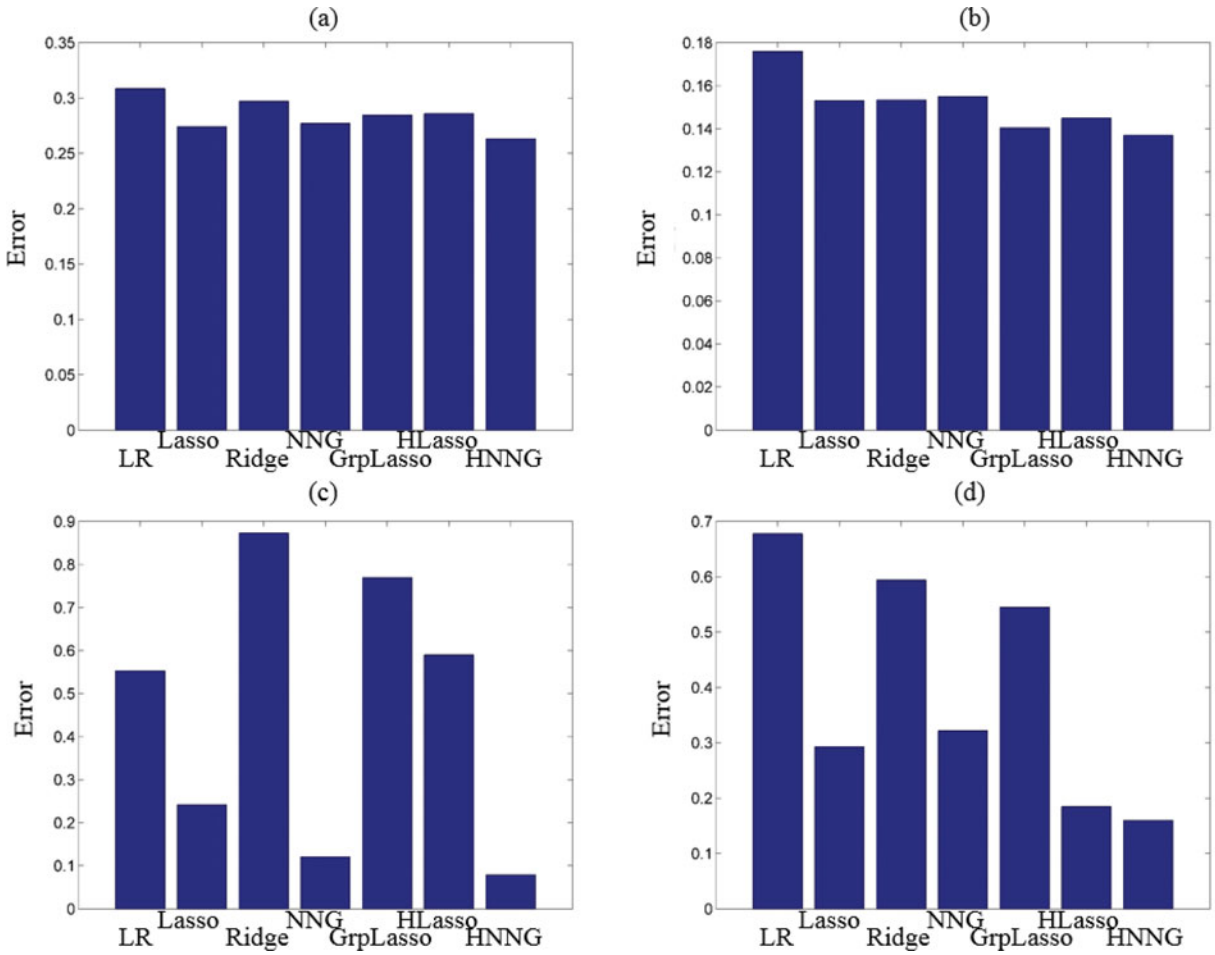

Figure 4. (a) Average testing errors over 50 replications under $n_{tr} = 100$, $S = 0.1$, $\rho = 0.6$, $\tau = 0$; (b) Average testing errors over 50 replications under $n_{tr} = 100$, $S = 0.4$ $\rho = 0.6$, $\tau = 0$; (c) Average overall variable selection errors over 50 replications under $n_{tr} = 100$, $S = 0.1$, $\rho = 0.6$, $\tau = 0$; (d) Average overall variable selection errors over 50 replications under $n_{tr} = 100$, $S = 0.4$, $\rho = 0.6$, $\tau = 0$.

model and all benchmark models. The above modeling process was repeated 50 times for each scenario. Figure 4 shows some simulation results (testing errors and overall variable selection errors) when the training sample size was 100 and $\rho = 0.6$, $\tau = 0$. More detailed simulation results (such as testing errors, Type I variable selection errors, Type II variable selection errors, and overall variable selection errors) as well as their definitions are described in the online Supplemental Material A. In Fig. 4, the bars represent the average errors over 50 simulation replicates under the same setting. The horizontal axis represents the benchmark methods and the proposed HNNG methods. Testing error is the error for the testing data. The overall variable selection error was calculated as the percentage of total incorrectly selected variables in the final estimated model among all predictors.

The simulation results are summarized as follows. When the sample size is small, GrpLasso has the best prediction performance, but HNNG is comparable, especially when the sparsity is small. When the sample size becomes larger, the performance of HNNG is among the best. For variable selection performance, Lasso, NNG, and HLasso perform well in variable selection when the sample size is small, but HNNG is comparable. When the sample size becomes larger, GrpLasso can identify the important features, but the corresponding Type II error (i.e., percentage of insignificant variables being selected in the final estimated model) is large, since it selects all features in a significant group. HLasso performs well when the sparsity is large. HNNG has comparable Type I variable selection error (i.e., percentage of significant variables not being selected in the final estimated model) and performs best for the Type II variable selection error under most settings. The overall variable selection performance of HNNG is among the best. The proposed method also has good variable selection performance for moderate sample size when the underlying model is sparse.

In summary, our proposed method outperforms the benchmark methods in terms of prediction performance when the sample size is large or the underlying model is sparse. The proposed method can also eliminate insignificant predictors and outperforms the benchmark methods in terms of variable selection under the above situations. This is mainly because the HNNG can capture the hierarchical variable structure and can be easily formulated as linear constraints in the optimization problem.

5. Case study

We further used the proposed method to analyze real data from a CZ process for single-crystal growth. Fourteen ingots (nine conforming ingots and five polycrystalline ingots) grown from the same furnace were used in the modeling study. We

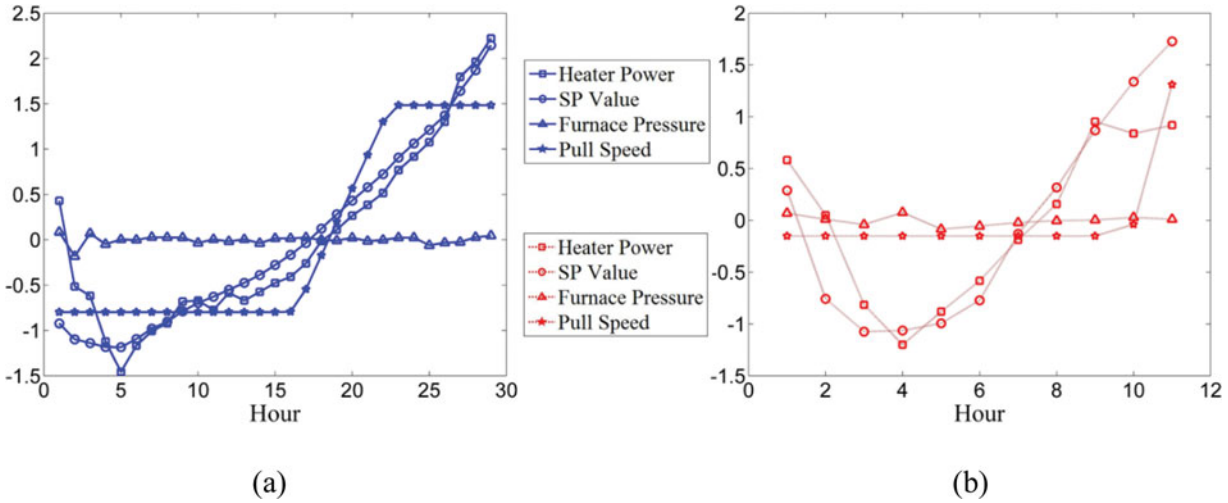

Figure 5. Selected standardized process variables in a CZ process (a) a conforming batch and (b) a batch containing polycrystalline defects.

selected four key process variables based on the built-in PID control algorithms used in the process: (i) heater power, which is the power supplied to the furnace to change the temperature gradient in the furnace; (ii) SP value, which is the temperature measurement performed by a thermocouple near the heater; (iii) pull speed, which is the pulling speed of the crystal; and (iv) furnace pressure, which is the pressure measurement in the furnace. These process variables need to be jointly controlled. For instance, if the thermal gradient at the interface is too large, the residual stress in the ingot will be large and the defect density will increase (Voronkov, 1982; Sinno *et al.*, 2000). On the other hand, if the thermal gradient is too small, the silicon melt will solidify at a slow rate and the corresponding growth speed will be slow. In addition, the larger the thermal gradient, the larger the ingot diameter tends to be, whereas a higher pulling speed leads to smaller ingot diameter. As a result, the thermal gradient and pulling speed should be jointly adjusted in order to obtain a target ingot diameter.

Figure 5 shows a few standardized process variables of a conforming batch and a polycrystalline batch. Each point in the figure represents the average of measurement over an hour. The sampling rate of the process variables is one measurement per minute. Notice that the growth time of the polycrystalline batch is shorter than that of the conforming batch, because the process has to be stopped once polycrystalline defects are observed (the polycrystalline defects were recorded by an operator at around the 11th hour in this example). From Fig. 5, it is clear that the key process variables are functional variables, and it is hard to directly distinguish between the polycrystalline batch and the conforming batch using these averaged measurements. Thus, it is necessary to look into the detailed features of the measurements and predict the occurrence of the polycrystalline defects in a timely manner.

The selected process variables were standardized and then truncated into 15-minute windows. For each ingot, we selected the window of the first 15-minute data points as the first sample and labeled the window based on the quality of the ingot for that period of time. Then we selected the window of the next 15 minutes of data points as the second sample and labeled it. Thus, we partitioned the whole data set into windows. After the truncation, these windows were regarded as separate samples modeled by Equation (1). In this case, we can predict if the ingot becomes polycrystalline every 15 minutes. This is a significant improvement over the current practice, where polycrystalline defects are detected by visual inspections performed by experienced operators. For each window, we performed wavelet analysis for each process variable with Daubechies 4 (db4) wavelet basis (Jensen and La Cour-Harbo, 2001). The number of dilations was selected to be four, which is the maximum number of dilations allowed in a 15-minute window. Interested readers can refer to Ganesan *et al.* (2004) for information on how to select the number of dilations. As a result, we processed the raw data and turned it into 108 features as predictors and 435 samples for use in the modeling study.

To evaluate the prediction performance, we used a leave-one-out CV approach. In iterations, we used the data of all 15-minute windows from 13 out of 14 ingots to estimate the model and perform variable selection. Then we evaluated the classification error based on the data of all 15-minute windows of the ingot that were not used in the training of the model (i.e., the left-out ingot). The average classification error of these left-out ingots is called the "CV Error" and it was used to evaluate the prediction performance of the model. In the evaluation, the

Table 2. CV error in the case study.

	LR	*Lasso*	*Ridge*	*NNG*	*GrpLasso*	*HLasso*	*HNNG*
Overall classification error	0.0785	0.0958	0.0805	0.0824	0.0671	0.0728	**0.0632**
Type I classification error	0.0581	0.0710	0.0409	0.0516	0.0366	0.0538	**0.0323**
Type II classification error	0.2456	0.2983	0.4035	0.3333	0.3158	**0.2281**	0.3158

Models with the smallest overall, Type I and Type II classification errors are highlighted in bold.

Table 3. Variable selection results in the case study.

	LR	*Lasso*	*Ridge*	*NNG*	*GrpLasso*	*HLasso*	*HNNG*
Average number of groups selected	4	4	4	3	2	1	2
Average number of features selected	60	8.4285	17.5714	9.1429	28.9286	27	5.5714

predicted binary response was compared with the real quality response labeled by a domain expert. The tuning parameter M was selected using the BIC.

The overall classification error, Type I classification error, and Type II classification error are summarized in Table 2. The overall classification error was defined as the percentage of total misclassified samples. The Type I classification error was defined as the percentage of conforming samples classified as polycrystalline samples, and the Type II classification error was defined as the percentage of polycrystalline samples classified as conforming samples. The cut-off probability for the logistic regression prediction was selected to be 0.5. The Receiver Operating Characteristic Curve and corresponding Area under the Curve values over different cut-off probabilities are investigated (Bradley, 1997); see details in online Supplemental Material B. The selection of the cut-off probability influences the errors, and other cut-off probabilities can be selected based on one's needs. In Table 2, the model with the best prediction performance is highlighted in bold. We conclude that the proposed method has the smallest overall classification error and Type I classification error. In summary, our proposed method can successfully identify polycrystalline defects while maintaining the smallest overall error. Note that HNNG has a larger Type II classification error than HLasso and is comparable to Lasso, NNG, and GrpLasso. One possible reason is that the sample sizes of the two classes are unbalanced. Specifically, the number of conforming samples is 378, and the number of nonconforming samples is 57. The variable selection results are summarized in Table 3. The proposed method selects a moderate number of groups while it has the smallest number of features selected. The coefficients selected by HNNG come from the coarse levels of heater power and SP value, which implies that the changes in thermal field are responsible for polycrystalline defects in the production process considered in the case study. The detailed information of the selected local features is available in online Supplemental Material C.

6. Conclusions and future research

A crystal growth process is the first step in the semiconductor manufacturing industry; however, the crystal can suffer from polycrystalline defects. In current practice, a large number of polycrystalline ingots are discarded, and a lot of energy and time is wasted in the rework stage.

With abundant observational data now being available, we proposed a logistic regression model with HNNG-based variable selection to extract important features from functional process variables. The method encourages variable selection in a hierarchical group structure for a binary response, where each group represents a functional process variable and each predictor in the group is a wavelet coefficient reflecting local time and frequency information. The model performance was compared with benchmark methods, such as Lasso, NNG, GrpLasso, and HLasso, when sample size, correlation structure, and sparsity of predictors were varied. The proposed method was shown to be better than benchmark methods in terms of prediction and variable selection, when the sample size was large or the underlying model was sparse. The proposed method also performed well for a real data set from a crystal growth process.

In future research, weighted logistic regression can be tried to attack the problem of unbalanced classes. The proposed method will be generalized to multinomial responses. The relationships between successive samples and the observational data from other crystal growth phases can be used in the modeling of polycrystalline defects. One idea to predict the binary response using process data from previous samples is to form a historical functional regression model, in which the temporal relationship is embedded in the model structure (Malfait and Ramsay, 2003). The selected feature can also be used for process monitoring and automatic process control.

Acknowledgements

The authors thank Liang Zhu and Jun Zhang for providing the background information used in the data set. The authors also thank the Editor and the anonymous reviewers for their constructive comments that helped to improve this article.

Notes on contributors

Hongyue Sun received a B.E. degree in Mechanical Engineering and Automation from the Beijing Institute of Technology in 2012 and an M.S. degree in Statistics from Virginia Tech in 2015. Currently, he is working toward a Ph.D. degree in the Grado Department of Industrial and Systems Engineering at Virginia Tech. His research interests include engineering-driven data fusion for manufacturing system quality control and functional data analysis. He is a member of INFORMS, IIE, and ASME.

Xinwei Deng is an Assistant Professor in the Department of Statistics at Virginia Tech. He received his Ph.D. degree in Industrial Engineering from Georgia Tech and his bachelor's degree in Mathematics from Nanjing University, China. His research interests are in statistical modeling and analysis of massive data, including high-dimensional classification, graphical model estimation, interface between experimental design and machine learning, and statistical approaches to nanotechnology. He is a member of INFORMS and ASA.

Kaibo Wang is a Professor in the Department of Industrial Engineering, Tsinghua University, Beijing, China. He received his B.S. and M.S. degrees in Mechatronics from Xi'an Jiaotong University, Xi'an, China, and his Ph.D. in Industrial Engineering and Engineering Management from the Hong Kong University of Science and Technology, Hong Kong. He has published papers in journals such as *Journal of Quality Technology*, *IIE Transactions*, *Quality and Reliability Engineering International*, *International Journal of Production Research*, and others. His research is devoted to statistical quality control and data-driven complex system modeling, monitoring, diagnosis, and control, with a special emphasis on the integration of engineering knowledge and statistical theories for solving real-world problems.

Ran Jin received a Ph.D. degree in Industrial Engineering from Georgia Tech, master's degrees in Industrial Engineering and in Statistics, both from the University of Michigan, and a bachelor's degree in Electronic Engineering from Tsinghua University. He is an Assistant Professor at the Grado Department of Industrial and Systems Engineering at Virginia Tech. His

research interests are in engineering-driven data fusion for manufacturing system modeling and performance improvements, such as the integration of data mining methods and engineering domain knowledge for multistage system modeling and variation reduction and sensing, modeling, and optimization based on spatial correlated responses. He is a member of INFORMS, IIE, and ASME.

References

Akaike, H. (1974) A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **19**(6), 716–723.

Avci, M. and Yamacli, S. (2010) Neural network reinforced point defect concentration estimation model for Czochralski-grown silicon crystals. *Mathematical and Computer Modelling*, **51**, 857–862.

Bradley, A.P. (1997) The use of the area under the ROC Curve in the evaluation of machine learning algorithms. *Pattern Recognition*, **30**(7), 1145–1159.

Breiman, L. (1995) Better subset regression using the nonnegative garrote. *Technometrics*, **37**(4), 373–384.

Brown, R.A., Wang, Z. and Mori, T. (2001) Engineering analysis of microdefect formation during silicon crystal growth. *Journal of Crystal Growth*, **225**(2–4), 97–109.

Bukkapatnam, S.T.S., Nichols, J.M., Seaver, M., Trickey, S.T. and Hunter, M. (2005) A wavelet-based, distortion energy approach to structural health monitoring. *Structural Health Monitoring*, **4**(3), 247–258.

Deng, X. and Jin, R. (2015) QQ models: joint modeling for quantitative and qualitative quality responses in manufacturing systems. *Technometrics*, **57**(3), 320–331.

Derby, J.J. and Brown, R.A. (1986) Thermal-capillary analysis of Czochralski and liquid encapsulated Czochralski crystal growth: I. Simulation. *Journal of Crystal Growth*, **74**(3), 605–624.

Dhanaraj, G., Byrappa, K., Prasad, V. and Dudley, M. (2010) *Springer Handbook of Crystal Growth*, Springer, Berlin, Germany.

Fan, J. and Li, R. (2001) Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, **96**(456), 1348–1360.

Fischer, B., Friedrich, J., Jung, T., Hainke, M., Dagner, J., Führer, T. and Schwesig P. (2005) Modeling of industrial bulk crystal growth—state of the art and challenges. *Journal of Crystal Growth*, **275**(1–2), 240–250.

Fisher, G., Seacrist, M.R. and Standley, R.W. (2012) Silicon crystal growth and wafer technologies. *Proceedings of the IEEE*, **100**, 1454–1474.

Führer, T. and Jung, T. (2004) Use of genetic algorithms for the development and optimization of crystal growth processes. *Journal of Crystal Growth*, **266**, 229–238.

Ganesan, R., Das, T.K. and Venkataraman, V. (2004) Wavelet-based multiscale statistical process monitoring: a literature review. *IIE Transactions*, **36**(9), 787–806.

Hastie, T., Tibshirani, R. and Friedman, J. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, New York, NY.

Huang, J., Breheny, P. and Ma, S. (2012) A selective review of group selection in high-dimensional models. *Statistical Science*, **27**(4), 481–499.

Huang, J., Ma, S., Xie, H. and Zhang, C. (2009) A group bridge approach for variable selection. *Biometrika*, **96**(2), 339–355.

Jensen, A. and La Cour-Harbo, A. (2001) *Ripples In Mathematics: The Discrete Wavelet Transform*, Springer, Berlin, Germany.

Jeong, M.K., Lu, J.C., Huo, X., Vidakovic, B. and Di, C. (2006) Wavelet-based data reduction techniques for process fault detection. *Technometrics*, **48**(1), 26–40.

Jin, J. and Shi, J. (1999) Feature-preserving data compression of stamping tonnage information using wavelets. *Technometrics*, **41**(4), 327–339.

Jin, J. and Shi, J. (2001) Automatic Feature extraction of waveform signals for in-process diagnostic performance improvement. *Journal of Intelligent Manufacturing*, **12**(3), 257–268.

Jin, R. and Deng, X. (2015) Ensemble modeling for data fusion in manufacturing process scale-up. *IIE Transactions*, **47**(3), 203–214.

Li, J., Shi, J. and Chang, T.S. (2007) On-line seam detection in rolling processes using snake projection and discrete wavelet transform. *Journal of Manufacturing Science and Engineering*, **129**(5), 926–933.

Mahajan, S. (2000) Defects in semiconductors and their effects on devices. *Acta Materialia*, **48**(1), 137–149.

Makalic, E. and Schmidt, D. (2011) Logistic regression with the nonnegative garrote. in *AI 2011: Advances in Artificial Intelligence*, Wang, D. and Reynolds, M. (eds), Springer, Berlin, Germany, pp. 82–91.

Malfait, N. and Ramsay, J.O. (2003) The historical functional linear model. *Canadian Journal of Statistics*, **31**(2), 115–128.

Mallat, S.G. (1989) A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **11**(7), 674–693.

Meier, L., van de Geer, S. and Bühlmann, P. (2008) The group lasso for logistic regression. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **70**(1), 53–71.

Miller, A.J. (2002) *Subset Selection in Regression*, Chapman & Hall/CRC, Boca Raton, FL.

Miyano, T. and Shintani, A. (1993) Nonlinear analysis of complexities in striations of Czochralski silicon crystals. *Applied Physics Letters*, **63**, 3574–3576.

Müller, G. (2002) Experimental analysis and modeling of melt growth processes. *Journal of Crystal Growth*, **237–239**(3), 1628–1637.

Paynabar, K. and Jin, J. (2011) Characterization of non-linear profiles variations using mixed-effect models and wavelets. *IIE Transactions*, **43**(4), 275–290.

Paynabar, K., Jin, J. and Reed, M.P. (2015) Hierarchical non-negative garrote for group variable selection. *Technometrics*, **57**(4), 514–523.

Ratcliffe, S.J., Heller, G.Z. and Leader, L.R. (2002) Functional data analysis with application to periodically stimulated foetal heart rate data. II: Functional logistic regression. *Statistics in Medicine*, **21**(8), 1115–1127.

Schwarz, G. (1978) Estimating the dimension of a model. *The Annals of Statistics*, **6**(2), 461–464.

Shintani, A., Miyano, T. and Hourai, M. (1995) A novel apporach to the characterization of growth striations in Czochralski silicon crystals. *Journal of the Electrochemical Society*, **142**, 2463–2469.

Sinno, T., Dornberger, E., Ammon, W.V., Brown, R.A. and Dupret, F. (2000) Defect engineering of Czochralski single-crystal silicon. *Materials Science and Engineering: R: Reports*, **28**(5–6), 149–198.

Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, **58**(1), 267–288.

Voronkov, V.V. (1982) The mechanism of swirl defects formation in silicon. *Journal of Crystal Growth*, **59**(3), 625–643.

Yuan, M. and Lin, Y. (2006) Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **68**(1), 49–67.

Yuan, M. and Lin, Y. (2007) On the non-negative garrotte estimator. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, **69**(2), 143–161.

Zhang, J., Li, W., Wang, K. and Jin, R. (2014) Process adjustment with an asymmetric quality loss function. *Journal of Manufacturing Systems*, **33**(1), 159–165.

Zhao, P., Rocha, G. and Yu, B. (2009) The composite absolute penalties family for grouped and hierarchical variable selection. *The Annals of Statistics*, **37**(6A), 3468–3497.

Zhou, N.F. and Zhu, J. (2010) Group variable selection via a hierarchical lasso and its oracle property. *Statistics and its Interface*, **3**, 557–574.

Zhou, S. and Jin, J. (2005) An unsupervised clustering method for cycle-based waveform signals in manufacturing processes. *IIE Transactions*, **37**, 569–584.

Zhu, L., Dai, C., Sun, H., Li, W., Jin, R. and Wang, K. (2014) Curve monitoring for a single-crystal ingot growth process, in *Proceedings of the Fifth International Asia Conference on Industrial Engineering and Management Innovation*, Atlantis Press, Xi'an, China, pp. 227–232.

Zou, H. and Hastie, T. (2005) Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, **67**(2), 301–320.

Zulehner, W. (1983) Czochralski growth of silicon. *Journal of Crystal Growth*, **65**(1–3), 189–213.

Appendix

The approximation of Equation (2) by quadratic programming with second-order Taylor expansion is briefly summarized here; see Deng and Jin (2015) for details. The log-likelihood function can be written as

$$\begin{aligned} \mathrm{L}(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \left(y_i \log p(\boldsymbol{x}_i) + \left(1 - y_i\right) \log \left(1 - p(\boldsymbol{x}_i)\right)\right) \\ &= \sum_{i=1}^{n} \left(y_i \log \frac{p(\boldsymbol{x}_i)}{1 - p(\boldsymbol{x}_i)} + \log \left(1 - p(\boldsymbol{x}_i)\right)\right) \\ &= \sum_{i=1}^{n} \left(y_i \boldsymbol{x}_i \boldsymbol{\beta} + \log \left(1 - p(\boldsymbol{x}_i)\right)\right) \\ &= \sum_{i=1}^{n} \left(y_i \boldsymbol{x}_i \boldsymbol{\beta} - \log \left(1 + e^{\boldsymbol{x}_i \boldsymbol{\beta}}\right)\right). \end{aligned}$$

The first- and second-order derivatives of the log-likelihood function are

$$\begin{aligned} \frac{\partial \mathrm{L}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \sum\nolimits_{i=1}^{n} \left(y_i \boldsymbol{x}_i - \frac{e^{\boldsymbol{x}_i \boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}_i \boldsymbol{\beta}}} \boldsymbol{x}_i\right) \\ &= \sum\nolimits_{i=1}^{n} \left(y_i - p(\boldsymbol{x}_i; \boldsymbol{\beta})\right) \boldsymbol{x}_i = \boldsymbol{X}^T \left(\boldsymbol{y} - \boldsymbol{p}\right), \end{aligned}$$

$$\frac{\partial^2 L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} = - \sum\nolimits_{i=1}^{n} \left(\boldsymbol{x}_i \boldsymbol{x}_i^T p(\boldsymbol{x}_i; \boldsymbol{\beta}) \left(1 - p(\boldsymbol{x}_i; \boldsymbol{\beta})\right)\right) = -\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X},$$

where $\boldsymbol{X}$ is an $n \times p$ matrix, $\boldsymbol{y}$ and $\boldsymbol{p}$ are $n \times 1$ vectors, and $\boldsymbol{W} = \mathrm{diag}(p(\boldsymbol{x}_1; \boldsymbol{\beta})(1 - p(\boldsymbol{x}_1; \boldsymbol{\beta})), \ldots, p(\boldsymbol{x}_n; \boldsymbol{\beta})(1 - p(\boldsymbol{x}_n; \boldsymbol{\beta})))$ is an $n \times n$ diagonal matrix.

The second-order Taylor expansion at the initial estimator $\tilde{\boldsymbol{\beta}}$ is

$$\begin{aligned} \mathrm{L}(\boldsymbol{\beta}) &= \mathrm{L}\left(\tilde{\boldsymbol{\beta}}\right) + \left(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}\right)^T \boldsymbol{X}^T \left(\boldsymbol{y} - \boldsymbol{p}\right) \\ &\quad - \frac{1}{2} \left(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}\right)^T \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X} \left(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}\right) \\ &= C_1 - \frac{1}{2} \boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X} \boldsymbol{\beta} + \boldsymbol{\beta}^T \boldsymbol{X}^T \boldsymbol{W} \left(\boldsymbol{X} \tilde{\boldsymbol{\beta}} + \boldsymbol{W}^{-1} \left(\boldsymbol{y} - \boldsymbol{p}\right)\right) \\ &= C_2 - \frac{1}{2} \left(\tilde{\boldsymbol{y}} - \boldsymbol{X} \boldsymbol{\beta}\right)^T \boldsymbol{W} \left(\tilde{\boldsymbol{y}} - \boldsymbol{X} \boldsymbol{\beta}\right), \end{aligned}$$

where $\tilde{\boldsymbol{y}} = \boldsymbol{X} \tilde{\boldsymbol{\beta}} + \boldsymbol{W}^{-1} (\boldsymbol{y} - \boldsymbol{p})$ is a constant.