Logistic Regression for Crystal Growth Process Modeling through Hierarchical Nonnegative Garrote based Variable Selection

Hongyue Sun[1], Xinwei Deng[2], Kaibo Wang[3], and Ran Jin[1]

[1]Grado Department of Industrial and Systems Engineering, Virginia Tech., Blacksburg, VA 24061, USA

[2]Department of Statistics, Virginia Tech., Blacksburg, VA 24061, USA

[3]Department of Industrial Engineering, Tsinghua University, Beijing 100084, 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 paper, 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, which 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.

[Supplemental materials are available for this article. Go to the publisher's online edition of *IIE Transactions* for the supplemental materials.]

Keywords: Crystal Growth, Logistic Regression, Polycrystalline, Process Modelling, Variable Selection

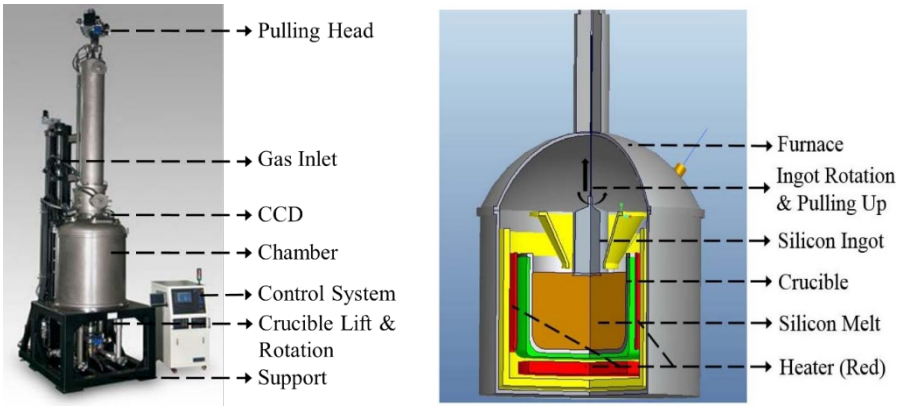
1. Introduction

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 are grown by Czochralski crystal growth processes (CZ processes) in industry (Fisher *et al.*, 2012). A successful CZ process is maintained at 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. By jointly controlling of thermal gradient and pulling speed, the ingot grows to the desired diameter. Afterwards, the ingot is slowly pulled upwards and rotated simultaneously. Such pulling and rotation process will last 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 part 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 Figure 1 (a) (Zhu *et al.*, 2014). Inside the furnace, the structure and operation conditions in the hot zone are critical for the ingot growth (Figure 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 would result in great waste of energy, time and cost. The quality defects include microscopic defects and macroscopic defects (Dhanaraj *et al.*, 2010). Examples of microscopic defects are void, interstitial, dislocation, etc., which will affect the electronic and mechanical properties of the downstream products (Mahajan, 2000). The macroscopic defects are more severe and may cause the failure of the entire growth process. In such a situation, the manufacturer has to discard the nonconforming segments of the ingot, or re-melt the material and repeat the growth process, which leads to further waste. Among these

macroscopic defects, polycrystalline defect is the most frequently observed type. Polycrystalline defect refers to the phenomenon that the desired monocrystalline ingot becomes polycrystalline. Once a segment of the ingot becomes polycrystalline, the entire segment will be discarded (Zhang *et al.*, 2014). Thus, it is critical to reduce this type of quality defect during the manufacturing. In the literature, defects analysis in crystal growth mainly focuses on microscopic defects (Voronkov, 1982; Sinno *et al.*, 2000; Brown *et al.*, 2001; Dhanaraj *et al.*, 2010). In this paper, we focus on polycrystalline defect modeling during the body growth phase, since the majority of polycrystalline defect appears in this phase.



(a) Furnace (b) Internal Structure of the Chamber (Hot Zone)

Figure 1. A Schematic of a Crystal Growth Furnace

(Redrawn from Zhu *et al*. (2014) and Zhang *et al*. (2014), with authors' permission)

To model the polycrystalline defect, we use a binary variable as the indicator for polycrystalline defect, and propose a logistic regression model to model the binary quality variable (response) with the functional process variables (predictors). Engineering perceptions suggest that the features of the process variables should be captured, because sudden changes of the process variables are potential root causes for polycrystalline defect. Therefore, we adopt wavelet analysis for each functional process variable. Wavelet analysis is selected because it performs well in extracting features from local time and frequency (Mallat, 1989). Thus, all the wavelet coefficients of a functional process variable form a group of features. In this paper, the wavelet coefficients of a process variable are called "features" or "local features" and

one 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 nonnegative garrote (HNNG) based variable selection is used.

The nonnegative 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 NNG for logistic regression models. The 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 paper, 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 variable selection for significant groups and features simultaneously. 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 part of the paper 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 thermal field distribution of the system for equipment design. Such models were typically based on partial differential equations (PDE) describing 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-process relationships. In most cases, these simulation models were solved offline by finite element methods. The performance of simulation models depends on the engineering assumptions, boundary conditions and accuracy of the material properties characterization. These models can be hardly adopted for polycrystalline modeling with potential online prediction requirements. Another category of models focused on microscopic defects, which typically modeled the distribution of the microscopic defects with process variables. Voronkov (1982) concluded that the ratio of crystal pulling speed and 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 fault ring were also modeled. Comprehensive reviews for defect modeling were provided by Sinno *et al.* (2000) and Brown *et al.* (2001). However, these models focused on the microscopic defects, and there were limited engineering-driven models to predict the polycrystalline defect quantitatively.

Researchers also attempted to model the CZ processes by using statistics, optimization or data mining methods. For instance, time series analysis for the dynamic properties of striations in the ingot was explored (Miyano and Shintani, 1993; Shintani *et al.*, 1995). Back-propagation, regularization and perceptron neural networks were used for the ingot striations pattern predictions. In addition, a genetic algorithm, coupled with thermal PDE, was used for the optimization of CZ furnace heat shield configuration (Fühner and Jung, 2004). For another example, Avci and Yamacli used artificial neural

network (ANN) to modify the PDE describing defect concentration (Avci and Yamacli, 2010). Such a method yielded good defect concentration prediction accuracy.

To model a binary quality variable with 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. Functional logistic regression model can also be used to link the binary response and functional predictors (Ratcliffe *et al.*, 2002). In this paper, we adopt the latter approach. To improve the model performance as well as interpretability, different kinds of variable selection methods have been proposed. These methods include subset and stepwise regression (Miller, 2002), Akaike information criterion (AIC) (Akaike, 1974), Bayesian information criterion (BIC) (Schwarz, 1978), Lasso (Tibshirani, 1996), non-negative garrote (NNG) (Breiman, 1995), smoothly clipped absolute deviation (Fan and Li, 2001), and elastic net (Zou and Hastie, 2005). For variable selection with group structure, the penalization methods introduced above may not perform well. To address the group variable structure, Yuan and Lin (2006) proposed Group Lasso. Zhao *et al.* (2009) proposed the flexible composite absolute penalties. Meier *et al.* (2008) studied the group variable selection for logistic regression via Group Lasso (GrpLasso). Though these methods usually have 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.* (2014).

To deal with the hierarchical variable selection problem, Huang *et al.* (2009) proposed Group Bridge (GrpBridge). However, GrpBridge penalty is not always differentiable and tends to be inconsistent for feature selection (Huang *et al.*, 2012). Zhou and Zhu (2010) proposed Hierarchical Lasso (HLasso), which penalizes the coefficients by two levels of L_1 penalty. Paynabar *et al.* (2014) claimed that Zhou and Zhu's method may fall into a local optimum. They proposed a hierarchical NNG for group variable selection in linear regression by firstly identifying the important groups, and then the important features

within the selected groups in two steps. They demonstrated that hierarchical NNG performed well in prediction and variable selection for linear regression models. In this paper, we will explore the hierarchical variable selection for a logistic regression via HNNG. The advantage of HNNG is that it can select important groups and features simultaneously in one step. Besides, the hierarchical NNG method by Paynabar *et al.* (2014) focused on linear regression models, while 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 of 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 to the force signal in a stamping process for data compression. Jin and Shi (2001) further adopted wavelet analysis for fault diagnosis 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).

3. The Proposed Method

3.1. Overview of the Proposed Method

The overview of the proposed method is shown in Figure 2. Based on the proportional-integral-derivative (PID) control loops of the CZ process, the potentially important process variables are selected for the modelling. 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.

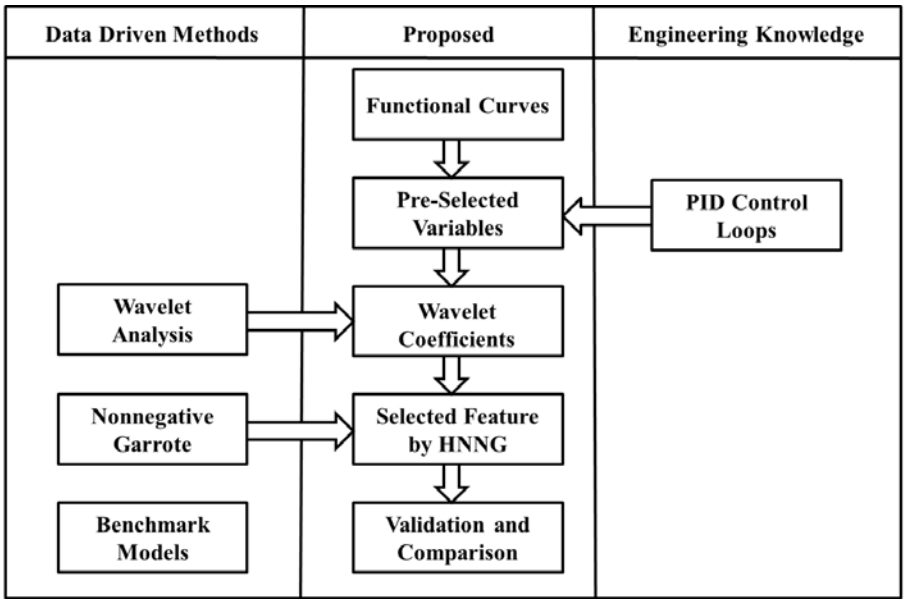


Figure 2. Overview of the Proposed Method

3.2. Data Structure

Assuming that we have p functional process variables to be modeled, and the number of dilations in wavelet analysis is set to be m. After wavelet decomposition, we have m levels of detail 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 j-th 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 will be P predictors with 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
$x_{1,1}\ x_{2,1}\ \ldots\ x_{p_1,1}$	$x_{1,2}\ x_{2,2}\ \ldots\ x_{p_2,2}$	…	$x_{1,m}\ x_{2,m}\ \ldots\ x_{p_m,m}$	$x_{1,c}\ x_{2,c}\ \ldots\ x_{p_c,c}$

3.3. HNNG based Logistic Regression Model

The logistic regression model has the form illustrated in Eq. (1),

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

where y_i is the binary response for the *i*-th 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 *i*-th sample is polycrystalline (i.e., $y_i = 1$); $\boldsymbol{x}_i = \left(\boldsymbol{x}_{1,i}{}^T,\ \boldsymbol{x}_{2,i}{}^T, \dots,\ \boldsymbol{x}_{p,i}{}^T\right)^T = (x_{1,1,i}, x_{2,1,i}, \dots, x_{P_1,1,i}, x_{1,2,i}, x_{2,2,i}, \dots, x_{P_2,2,i}, \dots, x_{1,p,i}, x_{2,p,i}, \dots, x_{P_p,p,i})^T$ is the predictor vector for the *i*-th sample, where $x_{k,j,i}$ is the *k*-th feature in the *j*-th group for the *i*-th sample. In the above notations, there are p groups of process variables and P_j features in each process variable. $\boldsymbol{\beta} = (\beta_1{}^{(1)},\ \beta_2{}^{(1)}, \dots, \beta_{P_1}{}^{(1)}, \beta_1{}^{(2)}, \beta_2{}^{(2)}, \dots, \beta_{P_2}{}^{(2)}, \dots, \beta_1{}^{(p)},\ \beta_2{}^{(p)}, \dots,\ \beta_{P_p}{}^{(p)})^T$ is model coefficient vector with $\beta_k{}^{(j)}$ the coefficient for the *k*-th feature in the *j*-th group.

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

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

where γ_j is the shrinkage factor for the *j*-th group; and $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_p)^T$ is the shrinkage vector for different groups. The optimization problem will determine 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 *j*-th group will be zero, which indicates that the *j*-th process variable is not significant, vice versa. If the feature level shrinkage $\theta_k^{(j)}$ becomes zero, then the *k*-th feature in the *j*-th group will not be significant, vice versa. Here M is a tuning parameter which can be selected by BIC, the validation data set, or cross validation (CV) (Hastie *et al.*, 2009).

To facilitate fast computation for Eq. (2), we adopt a similar approach to 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 quadratic form shown in Eq. (3),

$$\min \mathrm{L}(\boldsymbol{\beta}) = {}^1\!/_2 \left(\tilde{\boldsymbol{Y}} - \boldsymbol{X}\boldsymbol{\beta}\right)^T \boldsymbol{W} \left(\tilde{\boldsymbol{Y}} - \boldsymbol{X}\boldsymbol{\beta}\right), \tag{3}$$

where $\boldsymbol{W} = \mathrm{diag}\left(p(\boldsymbol{x}_1)\left(1 - p(\boldsymbol{x}_1)\right), \ldots, p(\boldsymbol{x}_n)\left(1 - p(\boldsymbol{x}_n)\right)\right)$ is an $n \times n$ diagonal matrix; and $\tilde{\boldsymbol{Y}} = \boldsymbol{X}\tilde{\boldsymbol{\beta}} + \boldsymbol{W}^{-1}(\boldsymbol{Y} - \boldsymbol{p})$, $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$, $\boldsymbol{Y} = (y_1, \ldots, y_n)^T$, $\boldsymbol{p} = \left(p(\boldsymbol{x}_1), \ldots, p(\boldsymbol{x}_n)\right)^T$. This

quadratic programming guarantees a global optimum and a brief derivation is provided in the Appendix. In this way, our method can select the significant groups and features simultaneously with computational issues addressed. The optimal solution to minimize Eq. (3) can be obtained by following Algorithm 1.

Algorithm 1.

Step 1. Compute the initial estimate $\widetilde{\boldsymbol{\beta}}$, 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 Eq. (3), and denote the current $\boldsymbol{\beta}$ as $\boldsymbol{\beta}^j$ at the *j*-th iteration;

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

Some practical suggestions for the initial values 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 estimates are also recommended by Yuan and Lin (Yuan and Lin, 2006) and Makalic and Schmidt (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 of Eq. (2), the optimization will reach to 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 set as 1's.

4. Simulation

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

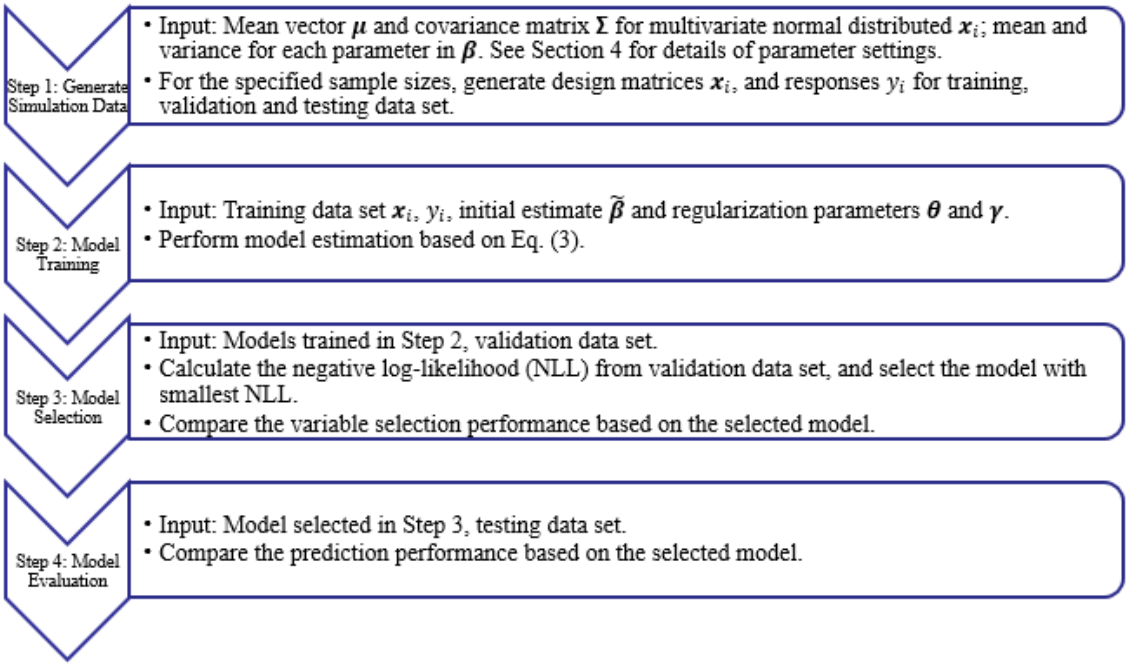


Figure 3. Illustration of the Simulation Procedure

In the simulation, the response y_i follows binominal distribution,

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

where $p(\boldsymbol{x}_i) = \frac{e^{\boldsymbol{x}_i{}^T\boldsymbol{\beta}}}{1+e^{\boldsymbol{x}_i{}^T\boldsymbol{\beta}}}$ and "$w.p.$" stands for "with probability". The predictors follow multivariate normal distribution with mean vector $\boldsymbol{\mu}$ =(**0**, **0**, …, **0**) and covariance matrix $\boldsymbol{\Sigma}$=$\begin{bmatrix} \boldsymbol{\rho}_{11} & \boldsymbol{\tau}_{12} & \cdots & \boldsymbol{\tau}_{1p} \\ \boldsymbol{\tau}_{12} & \boldsymbol{\rho}_{22} & \cdots & \boldsymbol{\tau}_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ \boldsymbol{\tau}_{1p} & \boldsymbol{\tau}_{2p} & \cdots & \boldsymbol{\rho}_{pp} \end{bmatrix}$, which are 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 is set to be 4 and the number of features in each group is set to be 5. In total, we have 20 predictors. To evaluate the performance of the proposed method, we test its performance by varying sample size, correlation structure and sparsity of predictors.

Specifically, denote the sample sizes for training data sets, validation data sets and testing data sets as n_{tr}, n_{va}, and n_{te}, we choose 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 are generated from the same model as shown in Eq. (4). The covariance matrix of predictors within and among groups are set to be $\boldsymbol{\rho} = \begin{bmatrix} 1 & \rho^{|i-j|} & \dots & \rho^{|i-j|} \\ \rho^{|i-j|} & 1 & \dots & \rho^{|i-j|} \\ \dots & \dots & \dots & \dots \\ \rho^{|i-j|} & \rho^{|i-j|} & \dots & 1 \end{bmatrix}$ and $= \begin{bmatrix} \tau & \tau^{|i-j|+1} & \dots & \tau^{|i-j|+1} \\ \tau^{|i-j|+1} & \tau & \dots & \tau^{|i-j|+1} \\ \dots & \dots & \dots & \dots \\ \tau^{|i-j|+1} & \tau^{|i-j|+1} & \dots & \tau \end{bmatrix}$, respectively, where i and j are the row and column indices of the matrix $\boldsymbol{\rho}$ and $\boldsymbol{\tau}$. Two levels of correlation are selected for ρ and τ, and there are four combinations for the correlation structure. Specifically, the within group correlation coefficient ρ is set to be 0 and 0.6, and between group correlation coefficient τ is set to be 0 and 0.3. The sparsity (denoted as S) represents the proportion of significant predictors in the underlying model, and is set to be 10% and 40%. The coefficient for a significant predictor $\beta_k^{(j)}$ follows 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 are 3 levels of sample size, 4 combinations of covariance structure, and 2 levels of sparsity. In total, 24 scenarios of simulation settings are evaluated.

We compare our proposed method with logistic regression (LR), Lasso, Ridge, NNG, GrpLasso, and HLasso methods for the binary response prediction. We use the training data set to obtain the regression models, and use the validation data set for the tuning parameter selection. The model with the selected tuning parameter is used for variable selection comparison. We use a threshold to determine whether the coefficient is significant or not. If the magnitude (absolute value) of the coefficient is larger than the threshold, then the corresponding predictor is considered as significant. Specifically, the threshold is set to be 10^{-6}. Then we compare misclassification errors of the testing data set (called "testing error") for the proposed model and all benchmark models. The above modeling process is repeated 50 times for each scenario. Figure 4 shows some simulation results (testing errors and overall variable selection errors) when the training sample size is 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 Supplemental Material A. In Figure 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. Overall variable selection error is 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 underling 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.

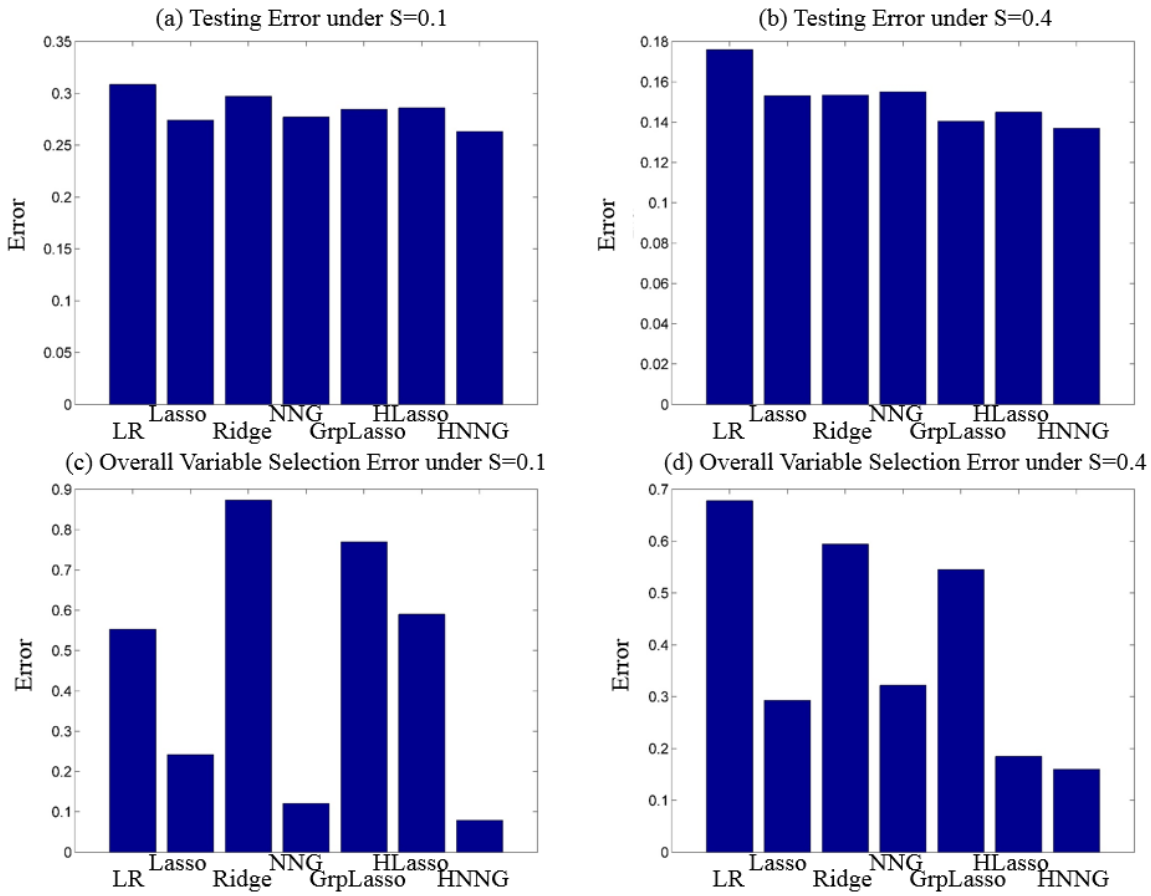


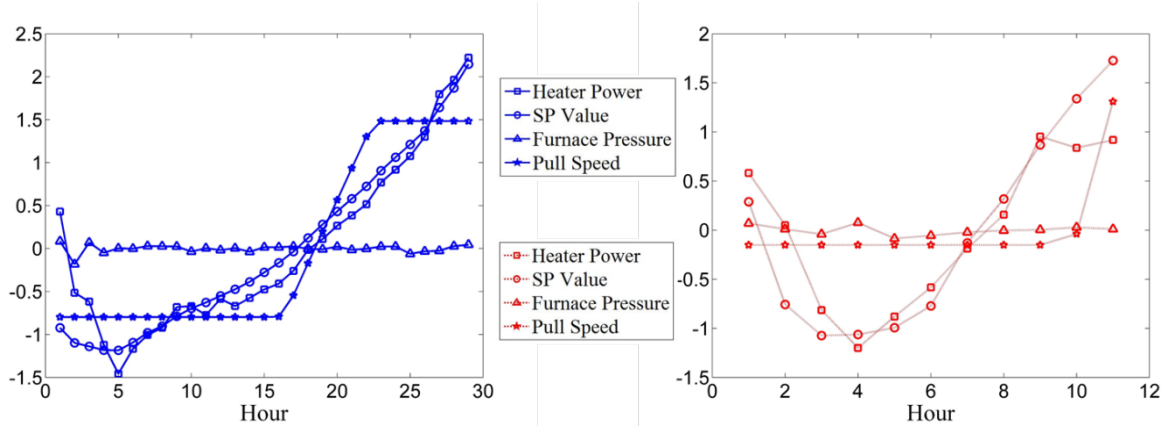
Figure 4. (a) Average Testing Errors over 50 Replications under n_{tr}=100 S=0.1 $= 0.6$ $\tau = 0$; (b) Average Testing Errors over 50 Replications under n_{tr}=100 S=0.4 $\rho = 0.6$ $\tau = 0$; (c) Average Overall Variable Selection Errors over 50 Replications under n_{tr}=100 S=0.1 $\rho = 0.6$ $\tau = 0$; (d) Average Overall Variable Selection Errors over 50 Replications under n_{tr}=100 S=0.4 $\rho = 0.6$ $\tau = 0$

5. Case Study

We further use the proposed method to analyze the real data from a CZ process for single-crystal growth. 14 ingots (9 conforming ingots and 5 polycrystalline ingots) grown from the same furnace are used for the modeling. We select four key process variables based on the process built-in PID control algorithms: (1) heater power, which is the power supplied to the furnace to affect the temperature gradient in the furnace, (2) SP value, which is the temperature measurement by a thermocouple near the heater, (3) pull speed, which is the pulling speed of the crystal, and (4) furnace pressure, which is the pressure measurement of

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; while higher pulling speed leads to smaller ingot diameter. As a result, the thermal gradient and pulling speed should be jointly adjusted for obtaining 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 1 measurement per minute. Notice that growth time of the polycrystalline batch is shorter than the conforming batch, because the process has to be stopped once polycrystalline defect is observed (the polycrystalline defect was recorded by an operator at around the 11th hour in this example). From Figure 5, it is clear that the key process variables are functional variables and it is hard to distinguish between the polycrystalline batch and the conforming batch directly from these averaged measurements. Thus, it is necessary to look into the detailed features of the measurements, and predict the polycrystalline in a timely manner.



(a) A Conforming Batch (b) A Polycrystalline Batch

Figure 5. Selected Standardized Process Variables in a CZ Process

The selected process variables are standardized and then truncated into 15-minute windows. For each ingot, we select the window of the first 15-minute data points as the first sample, and label the window based on the quality of the ingot for that period of time. Then we select the window of the next 15-minute data points as the second sample, and label it. Thus, we can partition the whole batch of the data set into windows. After the truncation, these windows are regarded as separate samples modeled by Eq. (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 the polycrystalline defect is detected by visual inspections performed by experienced operators. For each window, we perform wavelet analysis for each process variable with Daubechies 4 (db4) wavelet basis (Jensen and Ia Cour-Harbo, 2001). The number of dilations is selected to be four, which is the maximum available dilations for a 15-minute window. Interested readers can refer to Ganesan *et al.* on how to select the number of dilations (Ganesan *et al.*, 2004). As a result, we process the raw data and turn it into 108 features as predictors and 435 samples in the modeling.

To evaluate the prediction performance, we use leave-one-out CV. In iterations, we use the data of all the 15-minute windows from 13 out of 14 ingots to estimate the model and perform variable selection. Then we evaluate the classification error based on the data of all the 15-minute windows of the ingot that is not used for model training (i.e., the left-out ingot). The average classification error of these left-out ingots is called “CV Error” and is used for model prediction performance evaluation. In the evaluation, the predicted binary response is compared with the real quality response labeled by domain expert. The tuning parameter M is selected by BIC.

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

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

The overall classification error, Type I classification error and Type II classification error are summarized in Table 2. The overall classification error is defined as the percentage of total misclassified samples. Type I classification error is defined as the percentage of conforming samples classified as polycrystalline samples, and Type II classification error is defined as the percentage of polycrystalline samples classified as conforming samples. The cut-off probability for the logistic regression prediction is selected to be 0.5. The Receiver Operating Characteristic Curve (ROC) and corresponding Area under the Curve (AUC) values over different cut-off probabilities are investigated (Bradley, 1997), see details in Supplemental Material B. The selection of cut-off probability will influence 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 defect while maintain the smallest overall error. Note that HNNG has 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 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 defect in the production for case study. The detailed information of the selected local features is available in Supplemental Material C.

6. Conclusions and Future Research

A crystal growth process is the first step in semiconductor manufacturing industry, which suffers from the polycrystalline defect. In current practice, a huge amount of polycrystalline ingots are discarded, and a lot of energy and time are wasted in the rework stage.

With abundant observational data available, we propose a logistic regression model with HNNG based variable selection to extract important features from functional process variables. The method encourages variable selection in 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 is compared with benchmark methods, such as Lasso, NNG, GrpLasso and HLasso, when sample size, correlation structure and sparsity of predictors are varied. The proposed method is better than benchmark methods in terms of prediction and variable selection, when the sample size is large or the underlying model is sparse. The proposed method also performs well for the real data set from a crystal growth process.

In future research, weighted logistic regression can be tried to attack the unbalanced class problem. 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 for polycrystalline defect modeling. 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 acknowledge Mr. Liang Zhu and Mr. Jun Zhang for their help to provide the background information of the data set. The authors thank the editors and the anonymous reviewers for their constructive comments for this paper.

Notes on Contributors

Hongyue Sun received the B.E. degree in Mechanical Engineering and Automation from the Beijing Institute of Technology in 2012, and M.S. degree in Statistics from Virginia Tech. in 2015. Currently, he is working toward the Ph.D. degree in 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 problems from real industry.

Ran Jin received his Ph.D. degree in Industrial Engineering from Georgia Tech, his master's degree in Industrial Engineering and in Statistics, both from the University of Michigan, and his 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, Heidelberg, Berlin.

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., Fühner, 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.

Fühner, 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., Ma, S., Xie, H., and Zhang, C. (2009) A Group Bridge Approach for Variable Selection. *Biometrika*, **96**(2), 339-355.

Huang, J., Breheny, P., and Ma, S. (2012) A Selective Review of Group Selection in High-Dimensional Models. *Statistical Science,* **27**(4), 481-499.

Jensen, A. and Ia Cour-Harbo, A. (2001) *Ripples In Mathematics: the Discrete Wavelet Transform*, Springer, Heidelberg, Berlin.

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, R., and Deng, X. (2015) Ensemble Modeling for Data Fusion in Manufacturing Process Scale-up. *IIE Transactions*, **47**(3), 203-214.

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.

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, Heidelberg, Berlin, 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.

Müller, G. (2002) Experimental Analysis and Modeling of Melt Growth Processes. *Journal of Crystal Growth*, **237–239, Part 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. (2014) Hierarchical Non-Negative Garrote for Group Variable Selection. Accepted by *Technometrics*.

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 5th International Asia Conference on Industrial Engineering and Management Innovation (IEMI2014)*, Atlantis Press, 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 Eq. (2) by quadratic programming with second-order Taylor expansion is briefly summarized here, see Deng and Jin (2015) for details. The log-likelihood function,

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

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

$$\begin{aligned} \frac{\partial \mathrm{L}(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= \sum_{i=1}^{n} \left(y_i \boldsymbol{x}_i - \frac{e^{\boldsymbol{x}_i \boldsymbol{\beta}}}{1 + e^{\boldsymbol{x}_i \boldsymbol{\beta}}} \boldsymbol{x}_i \right) = \sum_{i=1}^{n} \left(y_i - p(\boldsymbol{x}_i; \boldsymbol{\beta}) \right) \boldsymbol{x}_i = \boldsymbol{X}^T (\boldsymbol{y} - \boldsymbol{p}), \\ \frac{\partial^2 L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} &= -\sum_{i=1}^{n} \left(\boldsymbol{x}_i \boldsymbol{x}_i^T p(\boldsymbol{x}_i; \boldsymbol{\beta}) \left(1 - p(\boldsymbol{x}_i; \boldsymbol{\beta})\right) \right) = -\boldsymbol{X}^T \boldsymbol{W} \boldsymbol{X}, \end{aligned}$$

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

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

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

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