Technometrics
A Journal of Statistics for the Physical, Chemical, and Engineering Sciences

Technometrics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/utch20

Ranking Features to Promote Diversity: An Approach Based on Sparse Distance Correlation

Andi Wang, Juan Du, Xi Zhang & Jianjun Shi

To cite this article: Andi Wang, Juan Du, Xi Zhang & Jianjun Shi (2022) Ranking Features to Promote Diversity: An Approach Based on Sparse Distance Correlation, Technometrics, 64:3, 384-395, DOI: 10.1080/00401706.2021.2020171

To link to this article: https://doi.org/10.1080/00401706.2021.2020171

TECHNOMETRICS
2022, VOL. 64, NO. 3, 384–395
https://doi.org/10.1080/00401706.2021.2020171

Taylor & Francis
Taylor & Francis Group

Ranking Features to Promote Diversity: An Approach Based on Sparse Distance Correlation

Andi Wang[a], Juan Du[b,c], Xi Zhang[d], and Jianjun Shi[e]

[a]The Polytechnic School, Arizona State University, Mesa, AZ; [b]Guangzhou HKUST Fok Ying Tung Research Institute, Guangzhou, China; [c]The Hong Kong University of Science and Technology, Guangzhou, China; [d]Peking University, Beijing, China; [e]Georgia Institute of Technology, Atlanta, GA

ABSTRACT

The improvement of sensing technology enables features of process variables to be collected during the fabrication of products. This article develops an automatic tool for process feature rankings based on these data. Based on the sensing data characteristics and the need of manufacturing system analysis, we propose two rules of the feature ranking scheme: assessing general dependency between each individual process feature and the quality variable, and satisfying a diversity rule. Specifically, we propose a feature ranking scheme based on the sparse distance correlation (SpaDC) that satisfies these two rules. Theoretical properties of the proposed algorithm are investigated. Simulation studies and two real-case studies from semiconductor manufacturing applications demonstrate that the SpaDC method ranks the features effectively given these two ranking rules.

ARTICLE HISTORY
Received March 2020
Accepted October 2021

KEYWORDS
Distance correlation; Feature ranking; Manufacturing process; Sparsity

1. Introduction

In the most manufacturing processes, multiple process variables (e.g., temperature, pressure, speed, and vibration signals) are collected and converted to process features (e.g., mean, variance, and natural frequency of those process variables) to characterize the conditions of the manufacturing processes. Ranking the process features based on the dependency relationship with the quality variable is highly demanded (Vakharia et al. 2016; Shao et al. 2013). Based on the ranking result, the practitioners can analyze the root causes of quality variations, quickly identify the leading features for process design improvements, prioritize the resources for quality improvement, and optimize the sensor placement for monitoring the key process features.

In traditional statistical quality control, identifying the major factors from personnel, machines, materials, methods, and environments is mainly based on experiential knowledge. Fishbone diagrams and Pareto charts (Montgomery 2007; Abidin et al. 2011) have been widely used as standard methods to identify the leading factors of a process. At the present time, the advanced sensing technologies are widely used in manufacturing processes and generate a large amount of process data during system operations. By retrospective analysis of the dependent relationship between the product quality variable and the process features obtained from the manufacturing processes, we aim at developing algorithms to perform process feature ranking.

An automatic feature ranking shall be stipulated by certain rules based on inherent characteristics of the process. First, as many sensors are installed along the entire production line, the size of total process features is usually large. However, the root causes that lead to the process faults and disturbance in a given time period is quite limited among all the potential failures. Furthermore, each root cause typically affects multiple process variables simultaneously, resulting in both dependency and redundancy among the process features (Abidin et al. 2011). Second, a specific product quality issue only involves a few disturbances, and thus numerous process features may be weakly, or even not related to the specific quality variable. Third, the dependency relationship between the quality variable and the process features, and the dependency relationship among the process features themselves are complex and ambiguous: they may be nonlinearly related, or certain features may relate to the variance of the quality variable instead of its mean.

These characteristics of the manufacturing process spawn two specific rules of the feature ranking procedure. First, the ranking should be based on a general dependency, given the complex and ambiguous relationship between process features and the quality variable. This general dependency measure shall take both linear and nonlinear dependency relationships into accounts between the process features and quality variables—including the nonlinear relationship between individual process features and the quality variables, as well as the relationship between the individual process features and the variance of the quality variable. Second, since many process features are associated with few root causes, the ranking procedure shall satisfy the diversity rule—a process feature shall be prioritized if it is not correlated to other features which have already been deemed as strongly related to the quality variable. The diversity rule is proposed to address the fact that one root cause that occurred will impact a set of process features that are dependent to each other. With the diversity rule, only one feature within a bunch of dependent features according to each root cause is prioritized and thereby encourages a small number of leading

CONTACT Xi Zhang xi.zhang@pku.edu.cn College of Engineering, Peking University, Beijing, P. R. China, 100871.

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/TECH.

features to cover all potential root causes that relate to the quality variable. If the diversity rule is not satisfied, then the highly ranked features may entirely relate to the major root cause, whereas other process features showing less dependency to the quality variable, but related to other minor root causes, are neglected. In this way, the highly ranked features cannot represent the necessary information for all root causes. Thus, a ranking scheme without the consideration of the diversity rule may deliver misleading results for the objectives. As we will see from the literature review, few existing feature ranking methods consider diversity or discrepancy of features. However, this goal is usually achieved by traditional quality tools like fishbone charts, as they intrinsically consider the difference of items therein.

In this article, we develop a feature ranking scheme that satisfies the diversity rule. This scheme also *partially* satisfies the following rule of general dependency: the ranks can reflect the dependency between the quality variable and individual process features. Here, we have to emphasize that the general dependency is referred as both linear and nonlinear dependency between individual process features and quality variables, rather than only the linear dependency. It should be noted that the dependency between the quality variable and the interactions of those process features is not the focus on this study and it cannot be identified by our proposed method. The ranking method is originated from the distance correlation, where we incorporated a new distance metric with the weights on features. To rank the features, we formulate an optimization problem by maximizing the weighted distance correlation while maintaining a certain degree of weight sparsity. This optimization problem is essentially a conic quadratic programming problem (Ben-Tal and Nemirovski 2001) and thus can be solved efficiently. This method is named as the Sparse Distance Correlation (SpaDC) method. As discussed above, it is suitable for retrospective analysis of the process data collected from manufacturing systems for identifying the leading features related to the variation of the quality variable.

The remainder of this article is organized as follows. Section 2 reviews the related literatures on feature ranking and general dependency measures. Section 3 introduces the proposed SpaDC method. Section 4 investigates the theoretical properties of SpaDC, provides an illustration of how it works, and discusses some characteristics of the method. Section 5 validates the method via the simulation studies. Section 6 presents two applications of SpaDC: one involves ranking 24 process features in the epitaxy stage of a solar cell manufacturing process, and the other involves ranking over one thousand overlay measurements in a lithography process. Section 7 concludes this article. Proofs are provided in the supplementary materials.

2. Literature Review

The problem of feature ranking and selection has been studied in the literature for a long time. Most feature selection methods are developed with a statistical model that associates the features with the responses. For linear models, methods such as stepwise regression (Weisberg 2005) and Lasso (Tibshirani 1996) can be used for feature ranking. Grömping (2006) introduced the R package "relaimpo," which provides six different assessments for the relative importance of regressors in a linear model, either based on the regression coefficients and their standard errors, or the decomposition of R^2 statistic. Choi et al. (2020) discussed how ridge regression could also help to infer the importance of variables, and the ranking result is evaluated by concordance score, with the comparison with Lasso and the elastic net regression. They discovered that when the pairwise correlations among the features have a large variation, the ridge regression has improved ranking performance. However, these ranking procedures are based on linear models between inputs and outputs and thus only aim for designated situations. Although predictive methods such as ensemble models (Friedman, Hastie, and Tibshirani 2001) tackle broader relationships between inputs and outputs, feature ranking approaches based on predictive models are not desirable in general. For one thing, these feature ranking rules require reconfiguration once the process changes. Furthermore, predictive models cannot capture certain relationships between process features and quality variables, for example, the case in which the variance of the quality variable is dependent with the process features.

Except for the model-based feature ranking procedure, there are also ranking methods based on general dependency indices. General dependency indices are the extensions of Pearson correlation coefficients that not only measure the correlation between variables but also take the general dependency of random variables into account. Examples of general dependency indices include mutual information (Steuer et al. 2002), distance correlation (Lyons 2013, Székely, Rizzo, and Bakirov 2007), and Hilbert–Schmidt Independence Criterion (HSIC) (Gretton, Herbrich, et al. 2005, Gretton, Smola, et al. 2005). Among them, the mutual-information-based method requires the estimation of the marginal and joint densities of the random variables and thus is difficult to be calculated efficiently. Distance correlation has received much attention in recent years. The distance correlation originates from energy distance (Székely and Rizzo 2013), a technique that characterizes the difference between distributions using pairs of observations. It was used by Huling and Mak (2020) to balance the distributions of covariates for estimating causal effects based on observational data. The distance correlation and the HSIC have been shown to be equivalent (Sejdinovic et al. 2013).

General dependency indices can be used for feature ranking. In the literature, Song et al. (2012) established an HSIC-based stepwise feature selection method, which can also be used for feature ranking. Li, Zhong, and Zhu (2012) and Kong, Wang, and Wahba (2015) developed a feature screening method by selecting a threshold to remove the features with a small distance correlation of the response variable. Yenigün and Rizzo (2015) proposed a stepwise variable selection method using distance correlation for regression modeling. However, these ranking procedures do not take diversity rule into consideration.

Recently, the diversity of the features has been proposed in Christidis et al. (2020). They propose to aggregate estimators in linear regression to form an overall fit. To achieve high accuracy of the prediction, they suggest that the groups of features used by these estimators should be different. In essence, nonoverlapping groups of features are encouraged as they provide unique information for the predictor of interest. Although this idea is similar to the diversity rule of feature ranking, the diversity in this article is based on the general dependency between individual features, and our goal is feature ranking instead of predictive modeling. The diversity rule of feature ranking is also similar to the minimal-redundancy-maximal-relevance (mRMR) criterion (Peng, Long, and Ding 2005), which adopts a stepwise procedure and selects the mth feature as the one most relevant to the output and most irrelevant with the previous $m-1$ features. However, mRMR is based on mutual information criterion, which relies on the density estimation and thus involves high computational complexity. Instead, the SpaDC method is based on the distance correlation from each pair of features, which can be calculated efficiently using the method proposed in Huo and Székely (2016).

3. Sparse Distance Correlation (SpaDC) Ranking Procedure

Let $\boldsymbol{X} = \left(X_1, \ldots, X_p\right)$ be the p-dimensional process features, and let Y be the associated quality variable. When n products are fabricated from the manufacturing system, the features are formatted into a data matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$. Let $\mathbf{X} = \left[\mathbf{x}_1, \ldots, \mathbf{x}_p\right] = \begin{bmatrix} \mathbf{x}^{(1)\top} \\ \vdots \\ \mathbf{x}^{(n)\top} \end{bmatrix}$, where $\mathbf{x}_i$ represents the ith process feature of all products and $\mathbf{x}^{(j)\top}$ represents all process features obtained from sample j. The quality indices of these n products are denoted as $\mathbf{y} = \left(y^{(1)}, \ldots, y^{(n)}\right)^\top \in \mathbb{R}^{n \times 1}$. From the data $\mathbf{X}$ and $\mathbf{y}$, we aim to obtain the ranks of the features that satisfy the diversity rule.

3.1. Distance Correlation

Our feature ranking procedure is based on distance correlation (Székely and Rizzo 2004). Distance correlation is an energy statistic (Székely and Rizzo 2017), which is a function of distances between all pairs of samples. As introduced in Section 2, it is a general dependency measure and can identify linear and nonlinear dependency relationships.

Let random vector $(\boldsymbol{X}, Y)$ follow an arbitrary joint distribution $F_{\boldsymbol{X},Y}$. The distance covariance and distance correlation between $\boldsymbol{X}$ and Y are defined based on two prescribed distance metrics $d_{\boldsymbol{X}}(\cdot,\cdot)$ and $d_Y(\cdot,\cdot)$ of space $\mathbb{R}^p$ and $\mathbb{R}$ respectively (Lyons 2013). With these distance metrics, the *population distance covariance* for $(\boldsymbol{X}, Y)$ is defined as the square root of

$$V^2(\boldsymbol{X},Y) = \mathbb{E}[(d_{\boldsymbol{X}}(\boldsymbol{X}^{(1)},\boldsymbol{X}^{(2)}) - \bar{d}_{\boldsymbol{X}}(\boldsymbol{X}^{(1)}) - \bar{d}_{\boldsymbol{X}}(\boldsymbol{X}^{(2)}) + \bar{\bar{d}}_{\boldsymbol{X}})$$
$$\times (d_Y(Y^{(1)},Y^{(2)}) - \bar{d}_Y(Y^{(1)}) - \bar{d}_Y(Y^{(2)}) + \bar{\bar{d}}_Y)].$$

Here $\bar{d}_{\boldsymbol{X}}(\cdot) = \mathbb{E}_{X_1}\left[d_{\boldsymbol{X}}\left(\cdot,\boldsymbol{X}^{(1)}\right)\right]$ and $\bar{\bar{d}}_{\boldsymbol{X}} = \mathbb{E}_{X_1,X_2}\left[d_{\boldsymbol{X}}\left(\boldsymbol{X}^{(1)},\boldsymbol{X}^{(2)}\right)\right]$. $\left(\boldsymbol{X}^{(1)},Y^{(1)}\right)$ and $\left(\boldsymbol{X}^{(2)},Y^{(2)}\right)$ are two independent samples from the distribution $F_{\boldsymbol{X},Y}$. The function $\bar{d}_Y(\cdot)$ and the quantity $\bar{\bar{d}}_Y$ are defined similarly.

Based on $V^2(\boldsymbol{X},Y)$, the squared-distance correlation between random vector $\boldsymbol{X}$ and Y is defined as $R^2(\boldsymbol{X},Y) = \frac{V^2(\boldsymbol{X},Y)}{\sqrt{V^2(\boldsymbol{X},\boldsymbol{X})V^2(Y,Y)}}$ if $V(\boldsymbol{X},\boldsymbol{X})\,V(Y,Y) > 0$. Under certain condition of $d_{\boldsymbol{X}}(\cdot,\cdot)$ and $d_Y(\cdot,\cdot)$ (Lyons 2013), the value of $R^2(\boldsymbol{X},Y)$ can be regarded as a dependency measure between $\boldsymbol{X}$ and Y, as $\leq R^2(\boldsymbol{X},Y) \leq 1$ and $R^2(\boldsymbol{X},Y) = 0$ if and only if $\boldsymbol{X}$ and Y are independent.

From the observed samples $\mathbf{X} \in \mathbb{R}^{n \times p}$ and $\mathbf{y} \in \mathbb{R}^{n \times 1}$, $V(\boldsymbol{X},Y)$ and $R^2(\boldsymbol{X},Y)$ are estimated with the following procedure. First, calculate the pairwise distance $a_{kl} = d_X\left(\mathbf{x}^{(k)},\mathbf{x}^{(l)}\right)$, and obtain $A_{kl} = a_{kl} - \bar{a}_{k\cdot} - \bar{a}_{\cdot l} + \bar{a}_{\cdot\cdot}$ where $\bar{a}_{k\cdot} = \frac{1}{n}\sum_{l=1}^n a_{kl}$, $\bar{a}_{\cdot l} = \frac{1}{n}\sum_{k=1}^n a_{kl}$, and $\bar{a}_{\cdot\cdot} = \frac{1}{n^2}\sum_{k,\,l=1}^n a_{kl}$. Similarly, calculate B_{kl} based on $b_{kl} = d_Y\left(y^{(k)},y^{(l)}\right)$. The sample distance covariance is defined as

$$V_n^2(\mathbf{X},\mathbf{y}) = \frac{1}{n^2}\sum_{k,\,l=1}^{n} A_{kl}B_{kl}. \tag{1}$$

The squared sample distance correlation $\hat{R}_n^2(\mathbf{X},\mathbf{y})$ is defined analogously as $R_n^2(\mathbf{X},\mathbf{y}) = \frac{V_n^2(\mathbf{X},\mathbf{y})}{\sqrt{V_n^2(\mathbf{X},\mathbf{X})V_n^2(\mathbf{y},\mathbf{y})}}$ when $V_n(\mathbf{X},\mathbf{X})$ $V_n(\mathbf{y},\mathbf{y}) > 0$.

Evidently, $R_n^2(\mathbf{X},\mathbf{y})$ and $V_n^2(\mathbf{X},\mathbf{y})$ are consistent estimators to their population counterparts. Through their sampling distributions, these statistics can be used to test the general independence between $\boldsymbol{X}$ and Y. The effectiveness of the distance-based method in detecting general relationships has been validated in the literature (Simon et al. 2014).

3.2. Distance Covariance Based on the Weighted l_1-Distance Metric

To facilitate feature ranking, we assign a weight $\beta_i \geq 0$ to each process feature X_i, and perform the ranking based on regularization path of $\boldsymbol{\beta} = \left(\beta_1, \ldots, \beta_p\right)^\top$ when maximizing a weighted sample distance correlation between $\boldsymbol{X}$ and $\mathbf{y}$. To calculate the weighted sample distance correlation from the dataset, we define the following $\boldsymbol{\beta}$-weighted ℓ_1-distance between features $\mathbf{x}$ and $\mathbf{x}'$:

$$d_{\boldsymbol{\beta}}\left(\mathbf{x},\mathbf{x}'\right) = \sum_{i=1}^{p} \beta_i \left|x_i - x_i'\right|. \tag{2}$$

The weighted ℓ_1-distance metric is used here because it leads to a convex formulation of the optimization problem as we shall see later, though this distance metric cannot directly identify the dependence between Y and the interaction effects of x_i's. With the Euclidean distance metric on the domain of y and the $\boldsymbol{\beta}$-weighted ℓ_1-distance on the domain of $\mathbf{x}$, the weighted sample distance covariance and the weighted sample distance correlation can be directly derived from Equation (1). The detailed derivation is given in Supplementary material A.

$$V_{n,\boldsymbol{\beta}}^2(\mathbf{X},\mathbf{y}) = \mathbf{d}_n^\top\boldsymbol{\beta};\, R_{n,\boldsymbol{\beta}}^2(\mathbf{X},\mathbf{y}) \propto \frac{V_{n,\boldsymbol{\beta}}^2(\mathbf{X},\mathbf{y})}{\sqrt{V_{n,\boldsymbol{\beta}}^2(\mathbf{X},\mathbf{X})}} = \frac{\mathbf{d}_n^\top\boldsymbol{\beta}}{\sqrt{\boldsymbol{\beta}^\top\mathbf{F}_n\boldsymbol{\beta}}}.$$

Here, the jth element of vector $\mathbf{d}_n$ is $d_{n,j} = V_n\left(\mathbf{x}_j,\mathbf{y}\right)$, and the (i,j)-element of $\mathbf{F}_n$ is $[\mathbf{F}_n]_{ij} = V_n\left(\mathbf{x}_i,\mathbf{x}_j\right)$, where $\mathbf{x}_i \in \mathbb{R}^{n\times 1}\mathbf{y} \in \mathbb{R}^{n\times 1}$. The function $V_n(\cdot,\cdot)$ is defined according to Equation (1) by evaluating the sample distance covariance of two $n\times 1$ vectors, using Euclidian distance metric $d(u,v) = |u-v|$ for both domains. Notably, $\mathbf{d}_n$ and $\mathbf{F}_n$ are calculated from the sample distance covariance between each process feature and the quality variable, and each pair of features respectively. The readers should be aware that this step indicates that the dependency between Y and $\mathrm{int}\left(X_i,X_j\right)$ cannot be identified, if Y is independent with both X_i and X_j. The fast computation procedure Huo and Székely (2016) can be employed to calculate entries of $\mathbf{d}_n$ and $\mathbf{F}_n$. We note that $V_{\boldsymbol{\beta}}\left(\mathbf{X},Y\right)=0$ if and only if each feature X_i is independent of Y for every i corresponding to $\beta_i>0$, as shown in Supplementary material B.

3.3. Formulating the Optimization Problem

We assume that the features $\mathbf{x}_1,\ldots,\mathbf{x}_p$ are scaled to have $V_n\left(\mathbf{x}_i,\mathbf{x}_i\right)=1$ for $i=1,\ldots,p$. We formulate the following optimization problem to achieve feature ranking:

$$\max_{\boldsymbol{\beta}} \mathbf{d}_n^\top \boldsymbol{\beta}$$

$$\text{subject to } \boldsymbol{\beta}^\top \mathbf{F}_n \boldsymbol{\beta}=1, \sum_{i=1}^{p}\beta_i \le c; \beta_i \ge 0 \text{ for all } i=1,\ldots,p. \quad (3)$$

In this problem, we aim to find a sparse weight vector $\boldsymbol{\beta}$ that leads to the maximum weighted sample distance correlation $R^2_{n,\beta}\left(\mathbf{X},\mathbf{y}\right) \propto \frac{\mathbf{d}_n^\top\boldsymbol{\beta}}{\sqrt{\boldsymbol{\beta}^\top\mathbf{F}_n\boldsymbol{\beta}}}$. The denominator of $R^2_{n,\beta}\left(\mathbf{X},\mathbf{y}\right)$ is restricted to 1, and the constraint $\sum_{i=1}^{p}\beta_i \le c$ is applied to encourage sparsity of $\boldsymbol{\beta}$ for key feature ranking. The parameter c controls the level of regularization, and the positive elements of the solution $\boldsymbol{\beta}\left(c\right)$ specify a subset of features that relate to Y. Considered that Problem (3) is not a convex optimization problem due to the constraint $\boldsymbol{\beta}^\top\mathbf{F}_n\boldsymbol{\beta}=1$, it is further relaxed to the following convex optimization problem:

$$\min_{\boldsymbol{\beta}} -\boldsymbol{\beta}^\top \mathbf{d}_n$$

$$\text{s.t. } \boldsymbol{\beta}^\top \mathbf{F}_n \boldsymbol{\beta} \le 1; \sum_{i=1}^{p}\beta_i \le c; \beta_i \ge 0 \text{ for all } i=1,\ldots,p. \quad (4)$$

Proposition 1 gives a result on the validity of the relaxation and the uniqueness of the solution.

Proposition 1. With probability 1, all elements in vector $\mathbf{d}_n$ have different values and $\mathbf{F}_n$ is positive definite. As a result,

i. If Problem (3) is feasible, then Problem (4) has a unique optimal solution;
ii. If Problem (3) is not feasible, then at most one element of $\boldsymbol{\beta}\left(c\right)$ is nonzero, and the optimal solution of (4) is also unique.

The proof is given in Supplementary material C.

Problem (4) can be transformed to a standard form of a conic quadratic programming problem (Ben-Tal and Nemirovski 2001), as detailed in Supplementary material D. Therefore, it can be solved efficiently with existing interior-point convex optimization solver. In Section 4, we shall see that this problem leads to diversity, the intriguing property that is critical for feature ranking.

3.4. Feature Ranking With Distance Correlation Criteria

The SpaDC method ranks the features by solving Problem (4) with different values of regularization parameter c. According to Proposition 1, the solution to Problem (4) is unique, and we denoted it by $\boldsymbol{\beta}\left(c\right)$. Let $\mathcal{J}\left(c\right)=\left\{i: \left[\boldsymbol{\beta}\left(c\right)\right]_i>0\right\}$ be the set of nonzero elements of $\boldsymbol{\beta}\left(c\right)$. As c increases from 0 to a larger number, some elements among $\boldsymbol{\beta}\left(c\right)$ enter $\mathcal{J}\left(c\right)$ and the features are ranked based on the sequence of their first appearance in it. Specifically, each feature X_i is associated with a threshold

$$T_i=\inf\left\{c: i\in\mathcal{J}\left(c\right)\right\}. \quad (5)$$

The features $X_1,\ldots,X_p$ are then ranked by sorting $T_1,\ldots,T_p$.

To implement the above idea, we first need to calculate all possible sets $\mathcal{J}\left(c\right)$ for a series of values $c\geq$. A direct approach is to construct a regularization path $\left\{\boldsymbol{\beta}\left(c\right): c\geq 0\right\}$. However, there is no existing algorithm for it. The presence of quadratic constraints of $\boldsymbol{\beta}$ makes Problem (4) essentially different from those with well-studied regularization paths (Efron et al. 2004; Hastie et al. 2004; Rosset and Zhu 2007; Tibshirani and Taylor 2011). As an alternative approach, we need to evaluate $\boldsymbol{\beta}\left(c_j\right)$ for a series of values c_j, $j=1,\ldots,J$ to form a dictionary $\mathcal{D}=\left\{\left(c_j,\boldsymbol{\beta}\left(c_j\right)\right): j=1,\ldots,J\right\}$. With such a dictionary $\mathcal{D}$, we can obtain $\tilde{T}_i=\min\left\{c_j: i\in\mathcal{J}\left(c_j\right), j=1,\ldots,J\right\}$, by which we rank features $\left\{X_i, i=1,\ldots,p\right\}$.

There are two specific implementations to obtain $\mathcal{D}$. One implementation is to adopt a bisection search algorithm. Using Proposition 2, we can effectively limit the values of c's for which Problem (4) needs to be solved.

Proposition 2. Let $\mathbf{F}_n$ be positive definite and all elements of $\mathbf{d}_n$ be different.

i. Problem (3) is not feasible if $c<1$, and it is feasible when $c\geq 1$.
ii. $\mathcal{J}\left(c\right)=\mathcal{J}\left(\sqrt{p}\right)$ for $c>\sqrt{p}$.
iii. If $1\le c_1<\tilde{c}<c_2\le\sqrt{p}$ and $\mathcal{J}\left(c_1\right)=\mathcal{J}\left(c_2\right)$, $\mathcal{J}\left(\tilde{c}\right)=\mathcal{J}\left(c_1\right)=\mathcal{J}\left(c_2\right)$.

The proof of Proposition 2 is given in Supplementary material E. Statements (i) and (ii) of Proposition 2 specify that Problem (4) only needs to be solved for $c\in\left[1,\sqrt{p}\right]$ and statement (iii) indicates that if the solution of Problem (4) at c_1c_2 shows that $\mathcal{J}\left(c_1\right)=\mathcal{J}\left(c_2\right)$, then solving Problem (4) again for $c\in\left(c_1,c_2\right)$ is unnecessary. With Proposition 2, we implemented a bisection search algorithm (Algorithm 1) to determine the ranks of all features. According to Proposition 2, the exploration starts with $c_{\min}=1$ and $c_{\max}=\sqrt{p}$ in Step 1. In Step 2, the subroutine "Search_Interval" finds all the possible $\mathcal{J}\left(c\right)$'s according to $c\in\left(c_1,c_2\right)$, by evaluating if the middle point $\tilde{c}$ satisfies $\mathcal{J}\left(\tilde{c}\right)=\mathcal{J}\left(c_1\right)$ or $\mathcal{J}\left(\tilde{c}\right)=\mathcal{J}\left(c_2\right)$, and recursively exploring the subintervals $\left(c_1,\tilde{c}\right)$ if $\mathcal{J}\left(\tilde{c}\right)\neq\mathcal{J}\left(c_1\right)$ and the subinterval $\left(\tilde{c},c_2\right)$ if $\mathcal{J}\left(\tilde{c}\right)\neq\mathcal{J}\left(c_2\right)$.

Besides the bisection method, a warm-start strategy, motivated by Friedman et al. (2007), is another implementation, especially suitable if there are many process features while we are only interested in obtaining the ranks of the leading r features. This algorithm is summarized in Algorithm 2. In this procedure, we start with $c=1$. In every step, we solve the optimization Problem (4) at $c=1+k\delta$ using the interior point method, by setting $1+(k-1)\,\delta$ as the initial value. If the nonzero elements of $\boldsymbol{\beta}\,(1+k\delta)$ and $\boldsymbol{\beta}\,(1+(k-1)\,\delta)$ are different, then the rank of a new feature is obtained.

Algorithm 1 Bisection search for ranking the features

```
1. Initiate $c_{\min}= 1$, $c_{\max} = \sqrt{p}$ and calculate $\mathcal{J}(c_{\min})$ and $\mathcal{J}\,(c_{\max})$. Initiate the dictionary $\mathcal{D} = \{(c_{\min}, \mathcal{J}(c_{\min})), (c_{\max}, \mathcal{J}(c_{\max}))\}$; Set $K_{\max}$, the maximum levels of recursion.
2. Call Search_Interval($c_{\min} c_{\max}, \mathcal{J}\,(c_{\min})\,,\ \mathcal{J}\,(c_{\max})$) to add entries into $\mathcal{D}$. .
3. Calculate $k_i = \min\{c\colon \{c; \mathcal{J}\,(c)\} \in \mathcal{D}; i \in \mathcal{J}\,(c)\}$. Then the rank of the features is determined by the ascending order of $k_i$, $i = 1,\ldots,p$.
```

Subroutine `Search_Interval`$(c_1,c_2,\mathcal{J}\,(c_1)\,,\mathcal{J}\,(c_2)\,,K)$

```
1. Let $c = (c_1 + c_2)\,/2$, and calculate $\mathcal{J}\,(c)$. If $\mathcal{J}\,(c) \neq \mathcal{J}\,(c_1)$ and $\mathcal{J}\,(c) \neq \mathcal{J}\,(c_2)$, write $\{c : \mathcal{J}\,(c)\}$ to the dictionary $\mathcal{D}$;
2. If $K \geq K_{\max}$ return;
3. If $\mathcal{J}\,(c) \neq \mathcal{J}\,(c_1)$, call Search_Interval ($c_1, c, \mathcal{J}\,(c_1)\,,\ \mathcal{J}\,(c), K+1$);
4. If $\mathcal{J}\,(c) \neq \mathcal{J}\,(c_2)$, call Search_Interval ($c, c_2, \mathcal{J}\,(c)\,,\ \mathcal{J}\,(c_2), K+1$);
```

In practice, features i and i' may share the same rank if we observe $k_i = k_{i'}$ in Step 3 of Algorithm 1 or if we find that $\boldsymbol{\beta}\,(1+k\delta)$ contains two or more nonzero elements than $\boldsymbol{\beta}\,(1+(k-1)\,\delta)$ when implementing the warm-start strategy. We regard such tied features with the same priority. For some features, the solved weight will always be 0 no matter how we increase c. These features are regarded as having the least importance with respect to Y. The ties may be either caused by a small search depth $K_{\max}$, a large step δ, or some inherent reasons related to the ranking procedure that will be elaborated in Section 4.3.

4. Theoretical Properties and Discussions

In this section, we first investigate the theoretical properties of the SpaDC method. We show that under certain conditions, the features dependent with Y are ranked before the independent ones and that the diversity rule can be achieved. The explanation of how SpaDC satisfies the diversity rule and how the ties are generated are illustrated using a three-feature demonstration. Finally, we discuss applicable scenarios of the proposed methods.

4.1. Theoretical Properties

Let us assume that Y is dependent with some of the features $X_1,\ldots,X_m$ and independent with the other features $X_{m+1},\ldots,X_p$. Proposition 3 states that the probability that $\mathcal{J}\,(c) = \{1,\ldots,m\}$ for a $c > 0$ will converge to 1 as the sample size $n \to \infty$, under certain conditions.

Proposition 3. Let $\boldsymbol{X} = \left(\boldsymbol{X}_1^\top, \boldsymbol{X}_2^\top\right)^\top$, where $\boldsymbol{X}_1 = (X_1,\ldots, X_m)^\top \in \mathbb{R}^m$ and $\boldsymbol{X}_2 = \left(X_{m+1},\ldots,X_p\right)^\top \in \mathbb{R}^{p-m}$. X_i is independent with Y if and only if $i > m$. Assume that $\mathrm{E}\,|X_i|^{2v} < \infty$ for all $i = 1,\ldots,p$ and $\mathrm{E}\,|Y|^{2v} < \infty$ for an even number $v \geq 2$. Let $A\,(\boldsymbol{X}_n,\boldsymbol{Y}_n)$ indicate the event that for a c, $\mathcal{J}\,(c) = \{1,\ldots,m\}$. Let $\left[V\left(X_i,X_j\right)\right]_{p\times p} := \mathbf{F} = \begin{bmatrix} \mathbf{F}_{11} & \mathbf{F}_{12} \\ \mathbf{F}_{21} & \mathbf{F}_{22} \end{bmatrix}$, the population counterpart of $\mathbf{F}_\mathrm{n}$, and let $\mathbf{d} = \begin{bmatrix} \mathbf{d}_1 \\ \mathbf{d}_2 \end{bmatrix}$ be the population counterpart of $\mathbf{d}_\mathrm{n}$. If the vector $\mathbf{d}_1$ belongs to the interior of the cone spanned by vectors $\mathbf{F}_{11}^{(1)}\mathbf{F}_{11}^{(2)},\ldots,\mathbf{F}_{11}^{(m)}\mathbf{1}_m$, where $\mathbf{F}_{11}^{(1)},\ldots,\mathbf{F}_{11}^{(\mathrm{m})}$, are the columns of $\mathbf{F}_{11}$, $\mathbf{1}_m = (1,\ldots,1)^\top \in \mathbb{R}^m$, we have $\mathrm{P}\,(\mathrm{A}\,(\boldsymbol{X}_n,\mathbf{Y}_n)) = 1 - O\left(n^{1-v}\right)$.

The proof of Proposition 3 is given in Supplementary material F. Proposition 3 points out that the probability that there exists a c such that "$\mathcal{J}\,(c)$ contains exactly the dependent features" goes to 1 when $n \to \infty$.

The statement in Proposition 3 relies on the condition that vector $\mathbf{d}_1$ belongs to the interior of the cone spanned by vectors $\mathbf{F}_{11}^{(1)}\mathbf{F}_{11}^{(2)},\ldots,\mathbf{F}_{11}^{(m)}\mathbf{1}_m$. In general, it holds when the dependency of features among $\boldsymbol{X}_1$ is weak, because the cone spanned by $\left[\mathbf{F}_{11}^{(1)},\ldots,\mathbf{F}_{11}^{(m)},\mathbf{1}\right]$ has a large range. Especially, if all features in $\boldsymbol{X}_1$ are independent, the cone is simply $\mathbb{R}_+^m$, so any $\mathbf{d}_1 > 0$ must lay in this cone, and so the statement of Proposition 3 holds.

Despite the implication of Proposition 3, the SpaDC method does not simply rank the features based on their sample distance correlation with Y like Li, Zhong, and Zhu (2012). The following proposition illustrates how SpaDC method achieves the diversity rule, and it will be illustrated intuitively in Section 4.2.

Proposition 4. Let $\mathbf{F} = \left[V\left(X_i,X_j\right)\right]_{p\times p}$ and $\mathbf{d} = [V\,(X_i, Y)]_{p\times 1}$, where $V\,(\cdot,\cdot)$ is the distance covariance based on the univariate Euclidean metrics. Write $\mathbf{F}$ and $\mathbf{d}$ in the following block-wise form:$\mathbf{F} = \begin{pmatrix} \mathbf{F}_{11} & \mathbf{f}_1 & \mathbf{F}_{12} \\ \mathbf{f}_1^\top & \mathrm{f}_{11} & \mathbf{f}_2^\top \\ \mathbf{F}_{12}^\top & \mathbf{f}_2 & \mathbf{F}_{22} \end{pmatrix}$; $\mathbf{d} = \begin{pmatrix} \mathbf{d}_1 \\ \mathrm{d}_* \\ \mathbf{d}_2 \end{pmatrix}$ where $\mathbf{F}_{11} \in \mathbb{R}^{m\times m}, \mathbf{f}_1 \in \mathbb{R}^{m\times 1}, \mathbf{F}_{12} \in \mathbb{R}^{m\times(p-m-1)}, \mathbf{f}_2 \in \mathbb{R}^{(p-m-1)\times 1}$, and $\mathbf{F}_{22} \in \mathbb{R}^{(p-m-1)\times(p-m-1)}$. The following statements hold:

1. If the probability that "for a $c > 1$, Problem (4) has a solution $\left(\boldsymbol{\beta}_{1,n}^\top\, 0\right)^\top$ with $\boldsymbol{\beta}_{1,n} > 0$" goes to 1 when $n \to \infty$, $\mathbf{d}_1 = \mathbf{F}_{11}\boldsymbol{\gamma}_1 + \mu\mathbf{1}$ for a $\boldsymbol{\gamma}_1 \geq 0$ and $\mu \geq 0$.
2. Assume that the condition in Statement (i) holds. Under additional assumption that $\mathbf{f}_1 = \mathbf{0}$, $\mathbf{F}_{12}^\top\boldsymbol{\gamma}_1 + d_*\mathbf{f}_2 > \mathbf{d}_2$, and $d_* > \mu$, the probability that there exists a c' such that "$m+1 \in \mathcal{J}\left(c'\right)$ and $m+2,\ldots,p \notin \mathcal{J}\left(c'\right)$" goes to 1 when $n \to \infty$

Proposition 4 indicates that the probability that "$J\,(c)$ is an increasing set sequence as c increases, whereas X_{m+1} is not ranked before the features $X_{m+2},\ldots,X_p$" goes to zero. The proof of this proposition is given in Supplementary material G. In Proposition 4, Statement (i) guarantees that the probability of selecting features $X_1,\ldots, X_m$ goes to 1 when $n \to \infty$. Statement (ii) gives the critical condition that with a high probability, $\mathcal{J}\left(c'\right) = \{1,\ldots,m+1\}$ for a c'. Except for the condition that X_{m+1} is independent with the features $X_{m+2},\ldots,X_p$, the following situations help to satisfy the assumptions in Statement (ii):

- X_{m+1} is strongly dependent with Y and the rest of the features $X_{m+2}, \ldots, X_p$ (i.e., the values of d_* and elements in $\mathbf{f}_2$ are large).
- The dependency between each of $X_{m+2}, \ldots, X_p$ and Y is small (i.e., the values of elements in $\mathbf{d}_2$ are small).
- The dependency between $X_{m+2}, \ldots, X_p$ and certain members in $X_1, \ldots, X_m$ is strong (i.e., the values of elements in $\mathbf{F}_{12}^{\top}\boldsymbol{\gamma}_1$ are large).

The last situation indicates the diversity rule.

4.2. A Graphical Illustration for SpaDC Method

In this section, we aim at acquiring an in-depth understanding of the SpaDC method using geometric illustrations of Problem (4) for cases involving three features X_1, X_2, X_3. An example of the geometric illustration for Problem (4) is shown in Figure 1(a). In this 3D figure, the three axes denote the decision variables β_1, β_2, and β_3. The 3D shape is the intersection of the ellipsoid $\boldsymbol{\beta}^{\top}\mathbf{F}_n\boldsymbol{\beta} \leq 1$, the half-space $\beta_1 + \beta_2 + \beta_3 \leq c_0$, and the first octant $\{\boldsymbol{\beta} : \beta_1, \beta_2, \beta_3 \geq 0\}$. The grayscale of this 3D shape's surface illustrates the negative objective value, $\boldsymbol{\beta}^{\top}\mathbf{d}_n$, where a large value is indicated by the light shade, and a small value is indicated by the dark shade. The solution $\boldsymbol{\beta}(c_0)$ at the current value $c_0 = 1.28$ is marked by the thick solid dot, and path of $\boldsymbol{\beta}(c)$ when c changes from 1 to $c_0 = 1.28$ is illustrated by the solid black curve.

The first case, shown through the three figures in Figure 1(b), aims to demonstrate how the diversity rule is satisfied. In this case, we let X_1 and X_3 be strongly dependent, X_1 and X_2 be independent, and X_2 and X_3 be independent. Therefore, we set $\mathbf{F}_n(1,3) = 0.5$, $\mathbf{F}_n(1,2) = 0$ and $\mathbf{F}_n(2,3) = 0$, leading to the special curvature of the ellipsoid. The feature X_1 is strongly related to Y with $d_{n,1} = 0.6$, whereas X_2 and X_3 have a similar degree of relatedness with Y, that is, $d_{n,2} = d_{n,3} = 0.4$. The top plot in Figure 1(b) shows that when $c_0 = 1$, $\boldsymbol{\beta}(c_0) = (1,0,0)^{\top}$ and only $\beta_1 > 0$. The middle plot shows that $\boldsymbol{\beta}(c)$ moves along the bottom plane $\beta_3 = 0$ and β_2 becomes non-zero when $c > 1$, so X_2 ranks the second. When c becomes even larger ($c > 1.38$), all β_1, β_2, and β_3 become positive, and when $c > 1.42$, the half-space constraint $\beta_1 + \beta_2 + \beta_3 \leq c$ becomes inactive, as shown in the bottom plot. Therefore, X_1 ranks the first, X_2 ranks the second, and X_3 ranks the third. In this example, we see that although X_2 and X_3 have a similar degree of dependency with

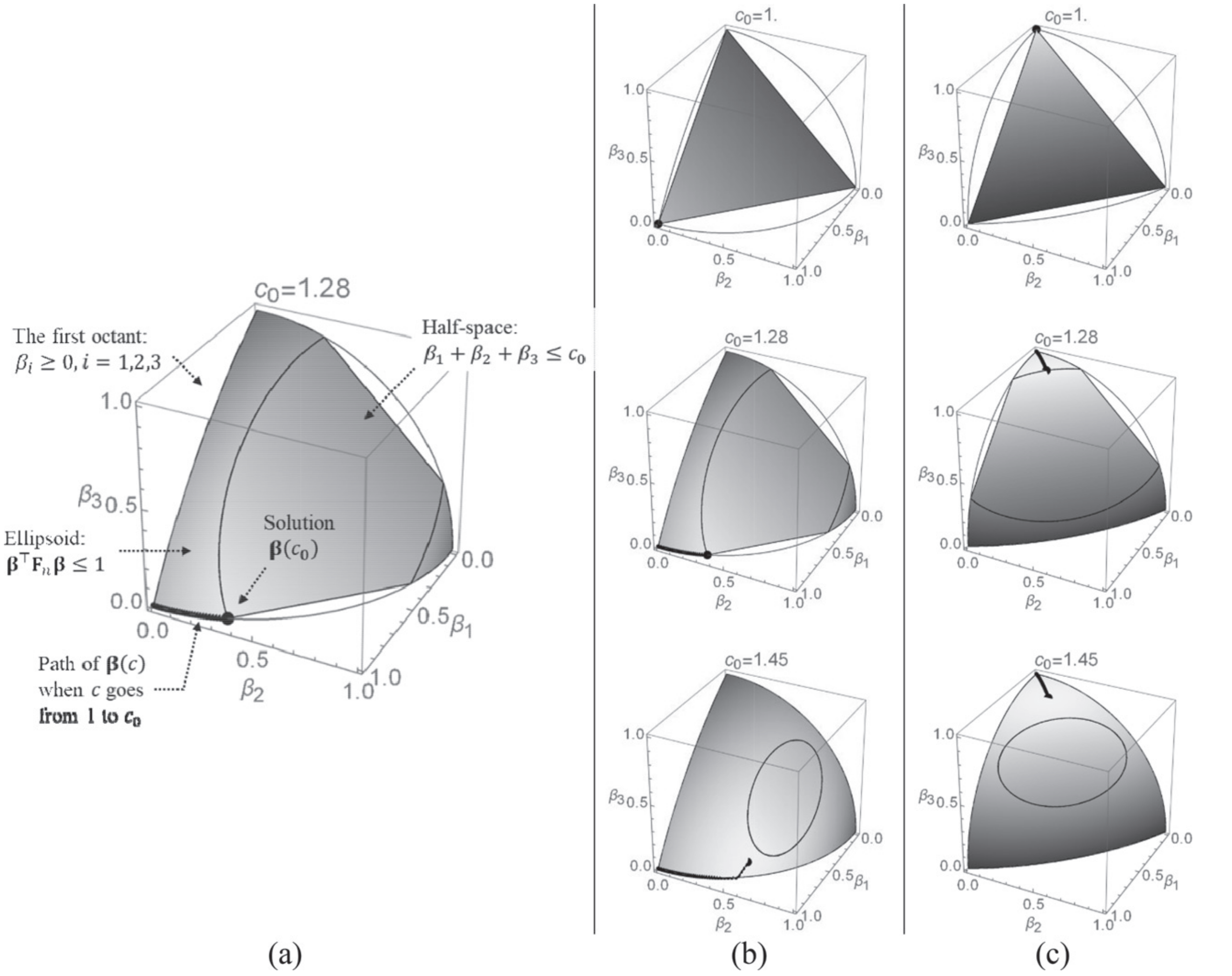
c₀=1.28
The first octant:
$\beta_i \geq 0, i = 1,2,3$
Half-space:
$\beta_1 + \beta_2 + \beta_3 \leq c_0$
Ellipsoid:
$\boldsymbol{\beta}^{\top}\mathbf{F}_n\boldsymbol{\beta} \leq 1$
Solution
$\boldsymbol{\beta}(c_0)$
Path of $\boldsymbol{\beta}(c)$
when c goes
from 1 to c_0
1.0
β_3 0.5
0.0
0.0
0.5
β_2
1.0
0.0
0.5 β_1
1.0
c₀=1.
c₀=1.28
c₀=1.45
(a)
(b)
(c)

Figure 1. A geometric illustration of the optimization Problem (4), and the path of $\boldsymbol{\beta}(c)$.

Y, the ranking result shows that X_2 ranks before X_3. In essence, the curvature of the ellipsoid surface driven by the dependency relationship among features warps the path of $\boldsymbol{\beta}(c)$ and helps to achieve the diversity rule.

The second case, shown in Figure 1(c), aims to illustrate how the ties in the features are generated. In this case, the features X_1 and X_2 are weakly dependent to Y with $d_{n,1} = d_{n,2}= 0.15$, whereas the feature X_3 is strongly dependent to Y with $d_{n,3}= 0.5$. X_1 and X_2 are dependent and $\mathbf{F}_n(1,2) = 0.5$, whereas each of them is independent with X_3. The top plot shows that when $c_0= 1$, $\boldsymbol{\beta}(c_0) = (0,0,1)^\top$. The path of $\boldsymbol{\beta}(c)$ in the middle plot shows that when c increases from 1, the first and the second element of $\boldsymbol{\beta}(c)$ become positive simultaneously, until the constraint $\beta_1 + \beta_2 + \beta_3 \leq c_0$ becomes inactive. Therefore, X_3 ranks first, and X_1X_2 tie at the second place. From this example, we can see that the ties can be inherent to the optimization problem. This example also shows a typical situation where ties happen: the features are dependent with each other and not much dependent with Y. Although the SpaDC procedure gathers the features into ordered groups with tying features, the authors do not regard it as a disadvantage for the proposed method in engineering practice, as the groups indicate different degrees of importance of features. As the features in the early groups tend to be more related to the quality variable and not dependent on each other, the tying features also provide useful information for the process.

In supplementary material H, we provide an animation to illustrate the path of $\boldsymbol{\beta}(c)$ for the above two cases with interpretations.

4.3. Discussion

In this section, we discuss the computational complexity of the SpaDC procedure and discuss one limitation of the SpaDC algorithm on identifying the interaction effect of the features.

Computational Complexity. The overall computational time for the SpaDC involves two parts: (i) the calculation time of $\mathbf{d}_n$ and $\mathbf{F}_n$ and (ii) the computational time for solving Problem (4) with a series of c's. The vector $\mathbf{d}_n$ and matrix $\mathbf{F}_n$ involve $p(p+1)/2$ values of sample distance covariances. Using the method of Huo and Székely (2016), each of these elements can be calculated with $O(n \log n)$ floating point operations, and the computation of different elements can be performed in parallel. For the second-order cone programming, the computation time for each $\boldsymbol{\beta}(c)$ is $O(p^3 \log(1/\epsilon))$ for calculating a solution $\boldsymbol{\beta}(c)$ with ϵ-accuracy (Ben-Tal and Nemirovski 2001).

Interaction Effects: The SpaDC method is essentially based on the statistics of $\mathbf{d}_n$ and $\mathbf{F}_n$, the pairwise sample distance covariance between features and the sample distance covariance between $\mathbf{x}_i$ and $\mathbf{y}$. For this reason, it cannot identify the dependency between Y and the *interaction effects* of two or more feature X_i's. A fundamental reason is that the weighted ℓ_1-distance metric is not a strong negative type (Lyons 2013) and thus the induced distance covariance cannot account for all dependency relationships between $\boldsymbol{X}$ and Y, though it facilitates a convex formulation of the optimization problem. Our target in the article is to conduct feature ranking for root-cause trace in manufacturing processes, where the main effects of process features are more important.

5. Simulation Studies

In this section, we compare the SpaDC method with six existing feature selection and ranking methods in the literature. We aim to validate that our scheme ranks the dependent features prior to the independent ones and meanwhile it satisfies the diversity rule.

5.1. Existing Benchmarks and General Settings

Six existing feature selection and ranking methods are used in the simulation study for benchmarking purposes. Yenigün and Rizzo (2015) proposed a stepwise variable selection method for a regression model based on the distance correlation of the residuals. This method, which is called the YR method in short, derives a variable ranking method directly, because a forward-selection procedure naturally gives an order of the variables. Li, Zhong, and Zhu (2012) proposed a feature screening method through ranking the features $X_1, \ldots, X_p$ according to the individual relationship with Y, and the ranking scheme is called as LZZ in our simulation study. We also included the LMG method (Lindeman 1980) implemented with the R package "relaimpo" (Grömping 2006) in our comparison study, which ranks features based on the R^2 statistics of linear models. Three feature ranking methods in our comparison are based on predictive models. Two of them are based on linear models, that is, the Lasso and adaptive Lasso methods (Zou 2006, Huang, Ma, and Zhang 2008). We used the MATLAB package "penalized" to compute the regularization paths (McIlhagga 2016) for ranking the process features. The last method is based on feature importance indices of the random forest model (Altmann et al. 2010), and we abbreviate it as the RF method.

In the next two subsections, we will consider five settings where the features are independent and dependent. Under each setting, we generally follow the procedure in Yenigün and Rizzo (2015) and generate the datasets $(\mathbf{X}, \mathbf{y})$ by repeating the procedure 1000 times. Six competing methods are applied to these 1000 datasets, generating 1000 sequences of the corresponding features. For $i = 1, \ldots p$, we count the number of times that each feature is ranked as the ith one. When a tie of r features appears in a ranked feature sequence, each feature in this tie is then counted as $1/r$ replication on every tied rank. For example, assume that feature X_1 is ranked as the first feature; X_2 and X_3 are tied at the second feature in one replication. For X_1, this replication is counted as one replication ranked as the first feature. For X_2 and X_3, half replication is counted as the second feature, and half is counted as the third feature. Finally, the ranking distribution for each feature is calculated.

5.2. Simulation With Independent Features

In the first three settings, the number of features to be ranked is $p= 8$, and they are independent of each other.

- Setting 1: Let $X_1, \ldots, X_8 \sim N(0,1)$, and $Y = |X_1| + X_2^2 + X_3 + \varepsilon$, where $\varepsilon \sim N(0,1)$. A total of 100 samples are generated from $(\mathbf{X}, Y)$.
- Setting 2: Let $X_1, \ldots, X_8 \sim N(0,1)$ and $Y = \log(4 + \sin(2X_1) + \sin(X_2) + X_3^2 + X_4 + 0.1) + \varepsilon$, where $\varepsilon \sim N(0, 0.1^2)$. The sample size is 500.

- Setting 3: Let Y be dependent on three variables $X_1, \ldots, X_3$ with $Y = Z\left(4 - X_1^2 - X_2^2 - X_3^2\right) + \varepsilon$, where $X_1, X_2, X_3 \sim \text{Unif}(-1, 1)$, $\varepsilon \sim N\left(0, 0.1^2\right)$, $Z = +1$ or -1 with probability of 0.5, and Z is independent with $X_1, \ldots X_8$. Here, Y has an equal probability of being positive or negative. A total of 500 samples are generated from $(\boldsymbol{X}, Y)$.

The ranking results of Settings 1–3 are given in Supplementary material I, where we count the proportion of runs where each variable X_j is ranked to the 1st to the 8th place using each individual method. From the results of Setting 1, we find that all methods rank X_3 at the first place most of the time. SpaDC, LZZ, and YR usually rank X_1 and X_2 at the second place and the third place. However, Lasso, Adaptive Lasso, and LMG tend to rank X_1 and X_2 to the first three places less often, and RF ranks X_2 even fewer to the top-three. This is because Lasso and AdpLasso only capture the linear relationship, and RF is not as sensitive to nonlinear dependency relationships as distance correlation-based methods.

For the nonlinear relationship specified in Setting 2, SpaDC, LZZ, and YR rank $X_1, \ldots, X_4$ ahead of $X_5, \ldots, X_8$ in the most replications. However, the Lasso, AdpLasso, LMG, and RF rank features X_3 to the 5th to 8th places most of the time. Hence, the schemes of the SpaDC, LZZ, and YR are more likely to rank the dependent features before the irrelevant ones when nonlinear dependency exists.

The results of Setting 3 show that the methods based on general dependency measures tend to rank features dependent with the quality variable before the independent ones. However, the ranking methods based on predictive models (Lasso, adaptive Lasso, LMG, and RF) cannot deliver such performance, because the features X_1, X_2, X_3 influence the variance of Y instead of its mean. We further discusse and compare the situation where the process variables and quality variable are linearly dependent and where the process variables impact the variance of the quality variable in Supplementary Material H.

5.3. Simulation With Dependent Features

In the next two settings, we investigate the situation where the features are dependent. We focus on testing if the diversity rule is satisfied.

- Setting 4: Three features, $\boldsymbol{X} = (X_1, X_2, X_3)^\top$ are generated, and Y represents the quality variable. $(X_1, X_2, X_3, Y)^\top$ jointly follows a multivariate normal distribution with zero mean and the following covariance structure: $\Sigma_{\boldsymbol{X},Y} = \begin{pmatrix} 1 & 0 & \rho & M \\ 0 & 1 & 0 & m \\ \rho & 0 & 1 & m \\ M & m & m & 1 \end{pmatrix}$

Here, the feature X_1 and Y are strongly correlated $(\text{corr}(X_1, Y) = M = 0.6)$, while (X_2, X_3) and Y are weakly correlated ($m = \text{corr}(X_2, Y) = \text{corr}(X_3, Y) = 0.2 < M$). Among the three features, X_2 is independent with X_1, while X_3 is correlated with X_1 with a correlation coefficient of $\rho = 0.1, 0.2, \ldots, 0.6$. A total of 1000 samples are generated in this setting.

- Setting 5: A total of six features $X_1, \ldots, X_6$ are generated, and Y represents the quality variable. $(X_1, \ldots, X_6, Y)^\top$ jointly follows a multivariate normal distribution with zero mean and the following covariance structure $\Sigma_{\boldsymbol{X},Y} = \begin{pmatrix} 1 & \rho_{12} & 0 & 0 & 0 & 0 & M \\ \rho_{12} & 1 & 0 & 0 & 0 & 0 & m \\ 0 & 0 & 1 & \rho_{34} & 0 & 0 & M \\ 0 & 0 & \rho_{34} & 1 & 0 & 0 & m \\ 0 & 0 & 0 & 0 & 1 & \rho_{56} & m \\ 0 & 0 & 0 & 0 & \rho_{56} & 1 & m \\ M & m & M & m & m & m & 1 \end{pmatrix}$

With this structure, the features $X_1, \ldots, X_6$ can be divided into three correlated groups: $[X_1, X_2]$ $[X_3, X_4]$ and $[X_5, X_6]$. Y is strongly correlated with X_1 and X_3, with $M = 0.6$. Meanwhile, Y is weakly correlated with the rest, that is, $m = 0.2$. The parameters ρ_{12}, ρ_{34}, and ρ_{56} are equal, and they are selected from four values, that is, 0.4, 0.5, 0.6, and 0.7. A total of 1000 samples are generated from $(\boldsymbol{X}, Y)$.

Figure 2 (the left panel) illustrates the distribution of the ranks for X_1, X_2, and X_3 in Setting 4 through line charts. Here, the row indicates the methods, and the column indicates the variables $X_1, \ldots, X_3$. Each subplot describes the proportion of runs (y-axis) that this feature is ranked as the first (black line), the second (red line), and the third (blue line) place using one method, when ρ (x-axis) varies from 0.1 to 0.6. According to the results of Setting 4, feature X_1 is always ranked as the first one. When ρ is 0.1, the frequencies that the ranks of X_2 and X_3 in SpaDC are distributed at the second and the third places are very close, as can be observed from the panel corresponding to X_2 and X_3 for the SpaDC method. However, when ρ increases from 0.1 to 0.6, X_2 and X_3 are more inclined to be ranked in the second and third places by the SpaDC method, respectively. This situation has not been observed in the other four methods. Recall that X_2 is independent with X_1, and thus prioritized to the second place. Therefore, SpaDC tends to prioritize the features that are independent of others leading features to meet the diversity rule, whereas other methods do not.

Recall that in Setting 5, Y is strongly correlated with X_1 and X_3, with $M = 0.6$. As expected, the results show that X_1 and X_3 are ranked in the first two places in most of the 1000 replications for methods in comparison. Figure 2 (the right panel) shows how each method ranks other features to the *third* place among these replications: the y-axis of each subplot represents the proportion one method rank X_j to the third place, $j = 2$, 4, 5, 6, and the x-axis illustrates $\rho = \rho_{12} = \rho_{34} = \rho_{56}$. When ρ increases from 0.4 to 0.7, the SpaDC method constantly ranks X_5 or X_6 in the third place with more replications than X_2 or X_4. As ρ increases, the gap becomes much larger, and X_5 or X_6 is always ranked to the third place following X_1 and X_3 when $\rho = 0.7$. This trend has not been observed in the LZZ or YR method. Recall that X_5 and X_6 are the features that are always independent with X_1 and X_3. When X_2 and X_4 are increasingly dependent with X_1 or X_3, in more cases X_5 or X_6 thus ranks the third place due to the diversity rule. The result thus demonstrates that SpaDC meets the diversity rule when the relationship between $\boldsymbol{X}$ and Y becomes more complex. However, the other four methods do not have such properties.

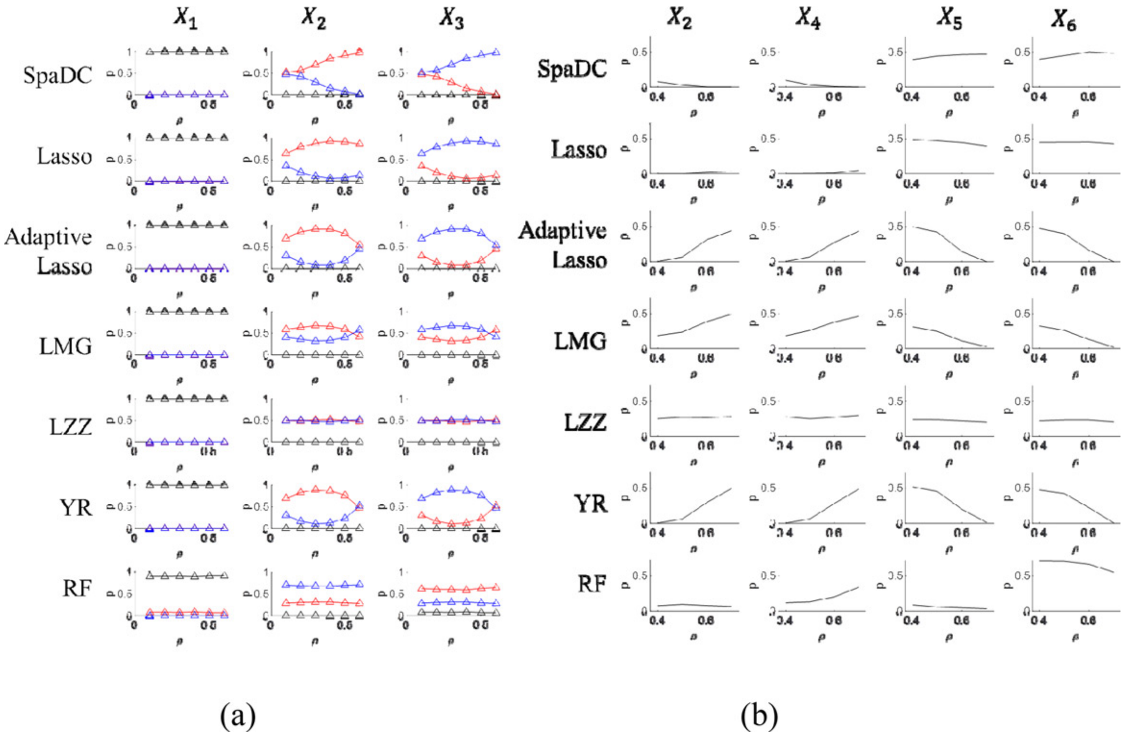

Figure 2. (a) The comparison results of setting 4, when the correlation ρ between X_1 and X_3 changes from 0.1 to 0.6. Black, red and blue lines correspond to the percentage of cases where the feature is ranked first, second, and third. (b) The comparison results of setting 5. The percentage of replications that X_i is ranked as the third feature when ρ varies between 0.4 and 0.7.

In conclusion, according to the results of the first three simulation settings, the SpaDC method is similar to the YR or LZZ method when the process features are independent of each other. Compared with the schemes based on linear models (i.e., Lasso, Adaptive Lasso, and LMG) and the random forest, the ranking schemes based on general dependency can capture the nonlinear dependency between the features $\boldsymbol{X}$ and the quality variable Y as well as the case where the features $\boldsymbol{X}$ affect the variance of Y. The simulations with Settings 4 and 5 further illustrate that the SpaDC method is superior to the YR and LZZ methods as it satisfies the diversity rule.

6. Case Studies

In this section, we validate the SpaDC method using two real examples. One is the data analysis of a solar cell manufacturing process. The other one is the analysis of overlay data from a lithography process.

6.1. *Epitaxy Process in Solar Cell Manufacturing*

In this case study, we investigate an epitaxy process in solar cell manufacturing (McEvoy, Castaner, and Markvart 2012). In this process, wafer substrates and the bases of solar cells, are sequentially loaded into a reaction chamber. On the top of each wafer, three thin films are deposited sequentially. During the epitaxy process of each wafer, three *in situ* process variables (i.e., the reflectance of the films and two temperature variables within the chamber), are measured as three time series. These three time series are, respectively, transformed into 18 and 6 indices through the feature extraction process detailed in Du, Zhang, and Shi (2018). In the end, 24 features are generated from the epitaxy process, and we denote them as $\boldsymbol{X} = (X_1, \ldots, X_{24})$.

The solar conversion efficiency (SCE), denoted as Y, is one of critical quality metrics in solar cell manufacturing processes. However, it must be individually tested offline after completing the entire fabrication. Since SCE is closely related to the epitaxy process, practitioners may rank the process features obtained from the epitaxy process based on their relationships with the SCE. With this ranking, they can monitor a small number of leading features during the manufacturing process and respond quickly once detect the process changes without waiting for the SCE inspections on the final product.

In this case study, 50 samples of $(\boldsymbol{X}, Y)$ are collected and the features are ranked using the SpaDC method. The ranking results show that 24 features are ranked as $X_1, X_8, X_{23}, X_{22}, X_{24}, X_{20}, X_{11}, X_{10}$, followed by all the rest of the features tied together. From the bisection algorithm, the values of c's within the dictionary are 1.0, 1.6177, 1.7188, 2.0781, 2.2466, 2.2578, 2.3926, and 4.8990. The strongest dependency between X_1 and Y is validated by the seven known follow-up test samples acquired after these 50 samples. Figure 3(a) shows that the final quality of the first two follow-up samples is in control and the last five follow-up samples are with a shifted mean. We check the individual control charts that monitor each feature and find that only X_1 exhibits an abrupt change during the last five samples, as shown in Figure 3(b). This result shows that the SpaDC method ranks X_1 correctly as the first feature.

We may interpret the results of the other leading features through Figure 4, which illustrates the sample distance correlation between each pair of features $(\mathbf{F}_n)$ and the sample distance correlation between each feature and the SCE $(\mathbf{d}_n)$. Here, X_i corresponds to the row i or column i of the left figure, and its rank

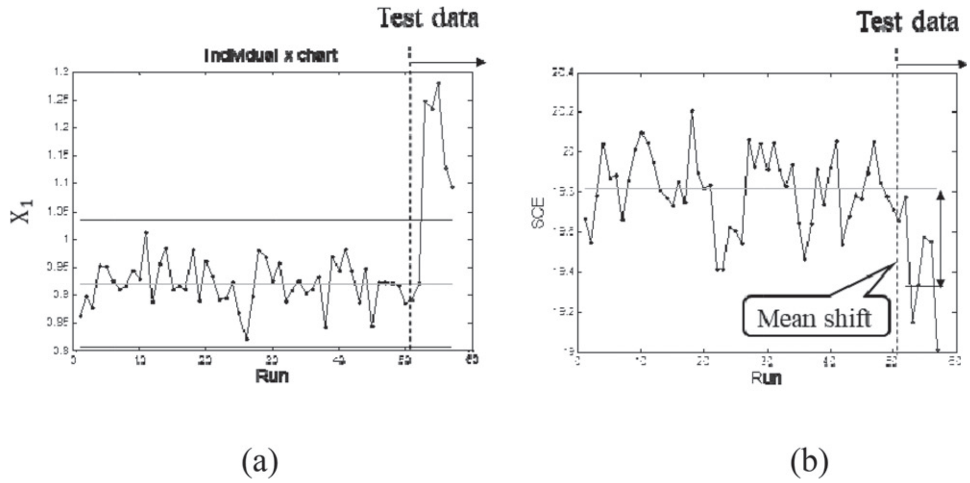

Figure 3. (a) Control chart for X_1; (b) Control chart for SCE.

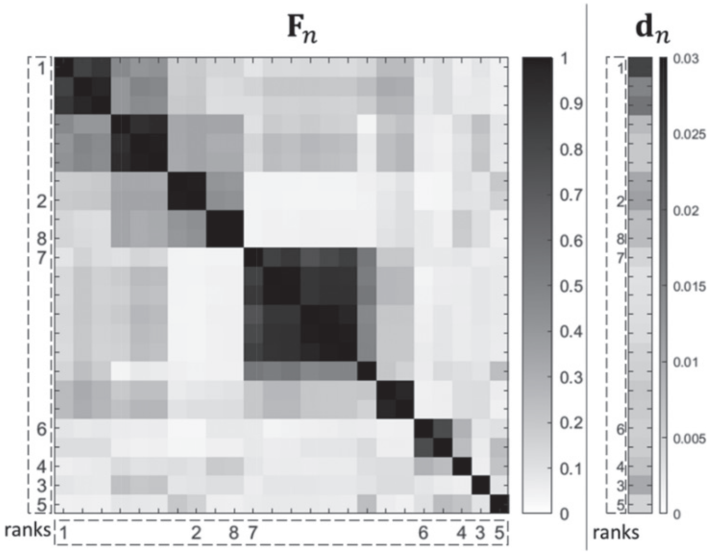

Figure 4. Distance correlations between the two process features ($\mathbf{F}_n$, left) and the distance correlation between each process feature and its quality variable ($\mathbf{d}_n$, right). The rows/columns of the matrix (left) and the element of the vector (right) correspond to features $X_1, \ldots, X_{24}$. The numbers marked at the left side and the bottom of the matrix and the numbers at the left side of the vector are the features' ranks.

is marked at the left side and the bottom of the matrix $\mathbf{F}_n$ and at the left side of the vector $\mathbf{d}_n$. We observe that the features $X_1, X_8, X_{23}, X_{22}, X_{24}, X_{20}, X_{11}, X_{10}$ (whose ranks are marked as numbers at the sides of the matrix $\mathbf{F}_n$ or vector $\mathbf{d}_n$ in the figure) are all moderately dependent on Y, and we can observe that their sequence confirms with the magnitude of $\mathbf{d}_n$. From $\mathbf{F}_n$, we identify that the remaining 16 tied features (without marked numbers) relate to the former features or barely dependent on Y. Although the pairwise distance correlation values shown in Figure 4 facilitate interpretation of the results, the ranks of these features cannot be obtained directly.

6.2. Lithography Process in Semiconductor Manufacturing

In a lithography process, the geometric pattern of one layer of microstructures is projected from a reticle onto the wafer surface through an exposure system. The overlay measurements of a wafer refer to the displacement error on the wafer of this projection process, and they are regarded as the most important process variables for lithography. The overlay measurements of an entire wafer are in the form of 2D overlay error map, as illustrated in Figure 5(a). In this figure, the x-y plane represents the surface of the wafer, and each vector at a point represents the displacement error of the printed geometric pattern at this location. Therefore, the desired appearance of an overlay error map is that all vectors are short and random.

The root causes in the lithography lead to specific patterns of the overlay error map. In this case study, we use a simulation testbed to generate the overlay errors for 1000 wafers. For the purpose of illustrating the diversity rule, we only generate root causes that lead all vectors within a region to shift simultaneously along a random direction. Such root causes of the lithography process include local bumps of chucking and local lens distortion. Specifically, we consider four root causes that affect the overlay error in four fixed regions, as illustrated in the shaded regions in Figure 5(b). The measurements of overlay vectors are taken on a 21 × 21 grid. Each overlay vector is described by two real values, and thus the entire overlay vector field contains over 842 features of the overlay vectors. For each wafer, the quality variable is the sum of the magnitudes of the

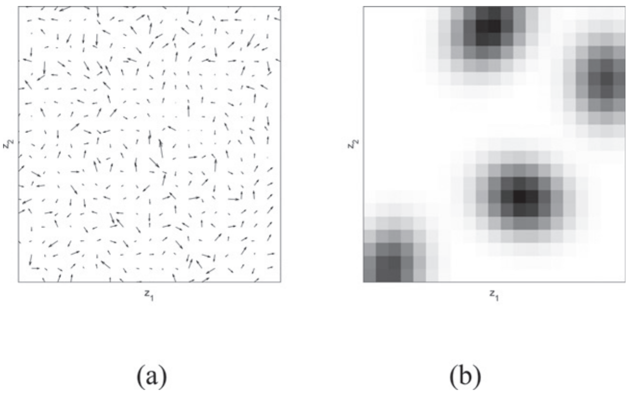

Figure 5. (a) the overlay error on one sample wafer, where the overlay error (arrows) is measured on many locations on the wafer plane (z_1, z_2) (b) the locations of the defects. In each shaded region, the overlay vector has a simultaneous shifting trend due to one root cause.

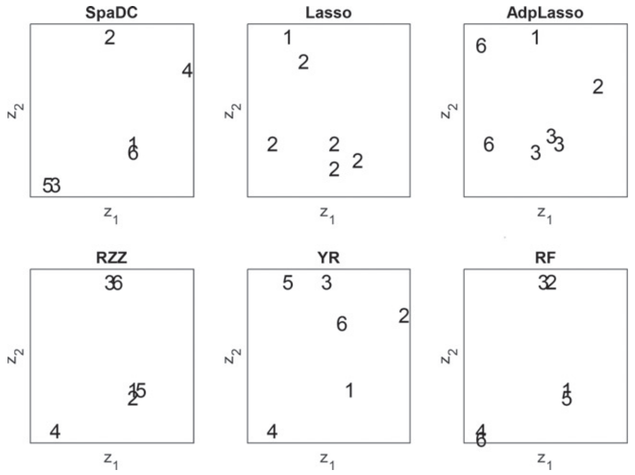

Figure 6. The ranking results of the six methods. Each feature is either an *x* or a *y* component of the displacement vector measured at one point on the wafer. The numbers indicate the rank of the top features, and the locations of the numbers represent their corresponding points on the wafer.

underlying shifts in four regions obtained from the testbed. We thereby obtained data matrix $\boldsymbol{X} \in \mathbb{R}^{1000\times 842}$ and $\mathbf{y} \in \mathbb{R}^{1000}$.

To automatically reveal the root causes of the overlay process, we rank all overlay features based on their relationship with this quality variable. The quality variable is nonlinearly related to the individual overlay measurement as a large magnitude of shift may cause either positive or negative value of features. Meanwhile, the overlay vectors corresponding to each root cause are significantly correlated.

We applied six different methods as comparison, namely, SpaDC, Lasso, Adaptive Lasso, RZZ, YR and RF, in ranking the overlay features. The leading six features are marked in Figure 6 at the location of the error vector with their rank numbers. For the SpaDC method, the first six features are obtained with the values c =1.00, 1.05, 1.15, 1.20, 1.55, and 1.65. We can see that the features with ranks 1–4 correspond to the four root causes or regions of defects. It means that the SpaDC method finds the leading features in each of the four regions corresponding to four independent root causes. Also, no features in the regions without potential root causes (i.e., without shadows) are included among the top-ranked features. It indicates that the SpaDC method is able to identify all potential root causes from the leading features, instead of only prioritizing features from one or two dominant zones of specific root causes.

This goal is also achieved by the YR method, although YR does not satisfy the diversity rule in Settings 4 and 5 of the simulation studies. Compared with them, RZZ and RF both miss one root cause (the region at the upper-right part of the wafer) within the six leading features. We found that the feature of this missing root cause is ranked as numbers 7 and 9, respectively. The underlying reason is that RF and RZZ do not take the diversity rule into consideration, and thus several leading features are dependent and correspond to the same root cause. Finally, the leading features from Lasso and Adaptive Lasso do not correspond to the regions affected by the root causes. This is mainly due to the nonlinearity between the features and the quality variable, while Lasso and Adaptive Lasso are more suitable for linear models. This case study shows that the SpaDC method prioritizes the dependent features and simultaneously satisfies the diversity rule.

7. Conclusions

In this article, we develop an automatic method to rank the process features based on their relationship with the quality variable. Based on the characteristics of the process data, we proposed two ranking rules to guarantee that the leading features provide useful information for process improvement (i) the ranking method should be based on general-dependency measure; and (ii) the ranking scheme considers diversity rule. We further proposed SpaDC ranking scheme, which takes both rules into consideration. The theoretical investigation and a graphic illustration of the SpaDC indicate that it indeed satisfies the ranking rules. The method is further validated through the simulation and real case studies of semiconductor manufacturing processes.

The SpaDC method may be further improved and extended in two aspects. One potential aspect is to find distance metrics that both enable feature ranking and take the interaction effect among features into consideration. Another desirable option is to extend the formulation of SpaDC to accommodate scalable computation with a large number of features.

Supplementary Materials

The article contains supplementary materials, including the proofs, the video illustrations, additional simulation results, further discussions, and computer codes for reproduction.

References

Abidin, K., Lee, K., Ibrahim, I., and Zainudin, A. (2011), “Problem Analysis at a Semiconductor Company: A Case Study on IC Packages,” *Journal of Applied Sciences*, 1–8. [384]

Altmann, A., Toloşi, L., Sander, O., and Lengauer, T. (2010), “Permutation Importance: A Corrected Feature Importance Measure,” *Bioinformatics* 26, 1340–1347. [390]

Ben-Tal, A., and Nemirovski, A. (2001), *Lectures on Modern Convex Optimization: Analysis, Algorithms, and Engineering Applications* (Vol. 2). SIAM. [385,387,390]

Choi, N.-H., Shedden, K., Xu, G., Zhang, X., and Zhu, J. (2020), “Comment: Ridge Regression, Ranking Variables and Improved Principal Component Regression,” *Technometrics*, 62, 451–455. [385]

Christidis, A.-A., Lakshmanan, L., Smucler, E., and Zamar, R. (2020), "Split Regularized Regression," *Technometrics*, 62, 330–338. [385]

Du, J., Zhang, X., and Shi, J. (2018), "A Condition Change Detection Method for Solar Conversion Efficiency in Solar Cell Manufacturing Processes," *IEEE Transactions on Semiconductor Manufacturing*, 32, 82–92. [392]

Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R. (2004), "Least Angle Regression," *The Annals of Statistics*, 32, 407–499. [387]

Friedman, J., Hastie, T., Höfling, H., and Tibshirani, R. (2007), "Pathwise Coordinate Optimization." *The Annals of Applied Statistics*, 1, 302–332. [387]

Friedman, J., Hastie, T., and Tibshirani, R. (2001), *The Elements of Statistical Learning* (Vol. 1), New York: Springer Series in Statistics. [385]

Gretton, A., Herbrich, R., Smola, A., Bousquet, O., and Schölkopf, B. (2005), "Kernel Methods for Measuring Independence," *Journal of Machine Learning Research* 6, 2075–2129. [385]

Gretton, A., Smola, A. J., Bousquet, O., Herbrich, R., Belitski, A., Augath, M., Murayama, Y., Pauls, J., Schölkopf, B., and Logothetis, N. K. (2005), "Kernel Constrained Covariance for Dependence Measurement," Paper read at AISTATS. [385]

Grömping, U. (2006), "Relative Importance for Linear Regression in R: The Package Relaimpo," *Journal of Statistical Software* 17, 1–27. [385,390]

Hastie, T., Rosset, S., Tibshirani, R., and Zhu, J. (2004), "The Entire Regularization Path for the Support Vector Machine," *Journal of Machine Learning Research*, 5, 1391–1415. [387]

Huang, J., Ma, S., and Zhang, C.-H. (2008), "Adaptive Lasso for Sparse High-Dimensional Regression Models," *Statistica Sinica*, 1603–1618. [390]

Huling, J. D., and Mak, S. (2020), "Energy Balancing of Covariate Distributions," arXiv:2004.13962. [385]

Huo, X., and Székely, G. J. (2016), "Fast Computing for Distance Covariance," *Technometrics,* 58, 435–447. [386,387,390]

Kong, J., Wang, S., and Wahba, G. (2015), "Using Distance Covariance for Improved Variable Selection with Application to Learning Genetic Risk Models," *Statistics in Medicine* 34, 1708–1720. [385]

Li, R., Zhong, W., and Zhu, L. (2012), "Feature Screening Via Distance Correlation Learning," *Journal of the American Statistical Association* 107, 1129–1139. [385,388,390]

Lindeman, R. H. (1980), *Introduction to Bivariate and Multivariate Analysis*, Glenview, IL: Scott, Foresman. [390]

Lyons, R. (2013), "Distance Covariance in Metric Spaces," *The Annals of Probability*, 41, 3284–3305. [385,386,390]

McEvoy, A. J., Castaner, L., and Markvart, T. (2012), *Solar Cells: Materials, Manufacture and Operation*, Waltham, MA: Elsevier. [392]

McIlhagga, W. H. (2016), "Penalized: A MATLAB Toolbox for Fitting Generalized Linear Models with Penalties," *Journal of Statistical Software*, 72, 1–21. [390]

Montgomery, D. C. (2007), *Introduction to Statistical Quality Control*, Hoboken, NJ: Wiley. [384]

Peng, H., Long, F., and Ding, C. (2005), "Feature Selection Based on Mutual Information Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27, 1226–1238. [386]

Rosset, S., and Zhu, J. (2007), "Piecewise Linear Regularized Solution Paths," *The Annals of Statistics*, 35, 1012–1030. [387]

Sejdinovic, D., Sriperumbudur, B., Gretton, A., and Fukumizu, K. (2013), "Equivalence of Distance-Based and Rkhs-Based Statistics in Hypothesis Testing," *The Annals of Statistics*, 2263–2291. [385]

Shao, C., Paynabar, K., Kim, T. H., Jin, J. J., Hu, S. J., Spicer, J. P., ... Abell, J. A. (2013), "Feature Selection for Manufacturing Process Monitoring Using Cross-validation," *Journal of Manufacturing Systems*, 32, 550–555. [384]

Simon, N., and Tibshirani, R. (2014), "Comment on" Detecting Novel Associations in Large Data Sets" by Reshef Et Al, Science Dec 16, 2011." arXiv:1401.7645. [386]

Song, L., Smola, A., Gretton, A., Bedo, J., and Borgwardt, K. (2012), "Feature Selection Via Dependence Maximization," *Journal of Machine Learning Research* 13, 1393–1434. [385]

Steuer, R., Kurths, J., Daub, C. O., Weise, J., and Selbig, J. (2002), "The Mutual Information: Detecting and Evaluating Dependencies between Variables," *Bioinformatics* 8, S231–S240. [385]

Székely, G. J., and Rizzo, M. L. (2004), "Testing for Equal Distributions in High Dimension," *InterStat*, 5, 1249–1272. [386]

——— (2013), "Energy Statistics: A Class of Statistics Based on Distances." *Journal of Statistical Planning and Inference*, 143, 1249–1272. [385]

——— (2017), "The Energy of Data," *Annual Review of Statistics and Its Application*, 4, 447–479. [386]

Székely, G. J., Rizzo, M. L., and Bakirov, N. K. 2007. "Measuring and Testing Dependence by Correlation of Distances," *Annals of Statistics* 35, 2769–2794. [385]

Tibshirani, R. (1996), "Regression Shrinkage and Selection Via the Lasso," *Journal of the Royal Statistical Society*, Series B, 267–288. [385]

Tibshirani, R. J., and Taylor, J. (2011), "The Solution Path of the Generalized Lasso," *The Annals of Statistics* 39, 1335–1371. [387]

Vakharia, V., Gupta, V. K., and Kankar, P. K. (2016), "A Comparison of Feature Ranking Techniques for Fault Diagnosis of Ball Bearing," *Soft Computing*, 20, 1601–1619. [384]

Weisberg, S. (2005), *Applied Linear Regression* (Vol. 528), Hoboken, NJ: Wiley. [385]

Yenigün, C. D., and Rizzo, M. L. (2015), "Variable Selection in Regression Using Maximal Correlation and Distance Correlation," *Journal of Statistical Computation and Simulation* 85, 1692–1705. [385,390]

Zou, H. (2006), "The Adaptive Lasso and Its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429. [390]