# Clustered Discriminant Regression for High Dimensional Data Feature Extraction and its Applications in Healthcare and Additive Manufacturing

Bo Shen, *Student Member, IEEE,*, Weijun Xie, Zhenyu (James) Kong, *Member, IEEE*

***Abstract*—The recent increase in applications of high dimensional data poses a severe challenge to data analytics such as supervised classification, particularly for online applications. To tackle this challenge, efficient and effective methods for feature extraction are critical for the performance of classification analysis. The objective of this work is to develop a new supervised feature extraction method for high dimensional data. It is achieved by developing a Clustered Discriminant Regression (CDR) to extract informative and discriminant features for high dimensional data. In CDR, the variables are clustered into different groups or sub-spaces, within which feature extraction is performed separately. The CDR algorithm, which is a greedy approach, is implemented to obtain the solution towards optimal feature extraction. One numerical study is performed to demonstrate the performance of the proposed method for variable selection. Three case studies using healthcare and additive manufacturing datasets are accomplished to demonstrate the classification performance of the proposed methods for real-world applications. The results clearly show that the proposed method is superior over the existing method for high dimensional data feature extraction.**

***Note to Practitioners*—This paper forwards a new supervised feature extraction method termed Clustered Discriminant Regression (CDR). This method is highly effective for classification analysis of high dimensional data such as images or videos, where the number of variables is much larger than the number of samples. In our case studies on healthcare and additive manufacturing, the performance of classification analysis based on our method is superior over the existing feature extraction methods, which is confirmed by using various popular classification algorithms. For image classification, our method with elaborately selected classification algorithms can outperform convolutional neural network (CNN). In addition, the computation efficiency of the proposed method is also promising, which enables its online applications such as advanced manufacturing process monitoring and control.**

***Index Terms*—Discriminant regression (DR), Clustering, Variable selection, Classification analysis, Greedy algorithm, Healthcare, Additive manufacturing (AM).**

## NOMENCLATURE

| | |
|---|---|
| $N$ | Number of training samples |
| $P$ | Number of variables |
| $K$ | Number of clusters |
| $L$ | Number of class labels |

| | |
|---|---|
| $i, j, n$ | Index of variables, clusters, and samples |
| $\mathbf{X}$ | $\mathbb{R}^{N\times P}$ data matrix |
| $\mathbf{X}_i$ | $i$th column of $\mathbf{X}$ |
| $\mathbf{x}_n$ | $n$th row of $\mathbf{X}$ |
| $t_n$ | Label of $\mathbf{x}_n$, $t_n \in \{1, \ldots, L\}$ |
| $\mathbf{Y}$ | Label indicator matrix $\mathbf{Y} \in \mathbb{R}^{N\times L}$ |
| $\mathbf{Y}_t$ | $t$th column of $\mathbf{Y}$ |
| $G_j$ | The set of index of variables in $j$th cluster |
| $G_j^*$ | The optimal set of index of variables in $j$th cluster |
| $\mathcal{G}$ | The partition (or clustering) result $\{G_1, \ldots, G_K\}$ |
| $\lambda_j$ | Regularization coefficient for $j$th cluster |
| $\mathbf{X}_{G_j}$ | Data matrix only consists of variables index $G_j$ |
| $\mathbf{\Phi}_{G_j}$ | $\mathbf{I}_N - \mathbf{X}_{G_j}(\mathbf{X}_{G_j}^\top \mathbf{X}_{G_j} + N\lambda_j \mathbf{I}_{\lvert G_j\rvert})^{-1}\mathbf{X}_{G_j}^\top$ |
| $g_j$ | Cardinality constrain for $j$th cluster |
| $\mathcal{W}$ | The set of projection matrices from each cluster |
| $\mathbf{W}_{G_j}$ | Projection matrix for $j$th cluster |
| $\lVert\cdot\rVert_F$ | The Frobenius norm of a matrix |
| $\mathrm{Tr}(\cdot)$ | Trace of a matrix |
| $\mathbf{I}_P$ | $P \times P$ identicial matrix |
| $\mathbf{F}_{G_j}$ | Feature extracted from $j$th cluster |
| $\mathbf{F}$ | Final feature set $\mathbf{F} = (\mathbf{F}_{G_1}, \mathbf{F}_{G_2}, \ldots, \mathbf{F}_{G_K})$ |
| $\sigma_{\max}(\cdot)$ | Largest eigenvalue of a matrix |
| $\sigma_{\min}(\cdot)$ | Smallest eigenvalue of a matrix |
| $\theta_g$ | $\theta_g = \max_{\lvert G\rvert=g} \sigma_{\max}(\mathbf{X}_G\mathbf{X}_G^\top)$ |
| $g$ | $g = \max_{j\in[K]} g_j$ |
| $\underline{\lambda}$ | $\underline{\lambda} = \min_{j\in[K]} \lambda_j$ |
| $\overline{\lambda}$ | $\overline{\lambda} = \max_{j\in[K]} \lambda_j$ |
| $\beta$ | $\beta = \max_{j\in[K]} \frac{\theta_{g_j}}{N\lambda_j}$ |
| $\underline{\theta}$ | $\underline{\theta} = \min_{G\in[P], \lvert G\rvert \geq P-\sum_{j=1}^K g_j} \sigma_{\min}(\mathbf{X}_G\mathbf{X}_G^\top)$ |
| $\gamma$ | $\gamma = \frac{\underline{\theta}}{K(N\overline{\lambda}+\theta_1)(1+\beta)^2} \log(\frac{P+1}{P+1-\sum_{j=1}^K g_j})$ |
| $\delta$ | Assignment matrix $\delta \in \{0,1\}^{P\times K}$ |
| $v^G$ | Objective value from greedy algorithm |
| $v^*$ | Optimal objective value |
| $G_j^U$ | $G_j^U = G_j \cup G_j^*$ |
| $\theta_{\max}$ | $\theta_{\max} = \max_{j\in[K]} \theta_{\lvert G_j \setminus G_j^*\rvert}$ |
| $\sigma_{\min}$ | $\sigma_{\min} = \min_{j\in[K]} \sigma_{\min}(\mathbf{X}_{G_j^U}^\top \mathbf{X}_{G_j^U}))$ |

*Corresponding author: Zhenyu (James) Kong*.

B. Shen, W. Xie, Z. Kong are with the Grado Department of Industrial and Systems Engineering, Virginia Tech, Blacksburg, VA 24061 USA e-mail: (boshen@vt.edu; wxie@vt.edu; zkong@vt.edu).

## I. Introduction

As one of the most important techniques in machine learning, classification analysis has played a crucial role in solving a large variety of real-world problems, such as image processing [1], text mining [2], quality control [3], etc. For example, in the control system of additive manufacturing (AM), classification analysis is one key component to identify different defects in process monitoring. Meanwhile, with the rapid development of data acquisition technology, high dimensional data is prevalent in AM, such as image/video data [3], point cloud [4], etc. The increase in dimensionality of sensor data poses a severe challenge to many existing classification algorithms for both accuracy of the analysis and computational efficiency [5], especially for the ill-posed case where the number of samples, namely, $N$, is much less than the number of features, namely, $P$. This ill-posed case, namely, $P \gg N$, is very common in some applications such as high dimensional genetic data where the number of samples (patients) is limited [6], manufacturing process monitoring data where data under abnormal or failure conditions is relatively hard to capture [7]. In order to address this challenge, it is necessary to apply feature extraction methods to the raw data before executing classification algorithms.

Discriminant regression (DR) [8] is an effective supervised feature extraction method for the case of ill-posed high dimensional data where $P \gg N$. Consider a data set that consists of $N$ samples $\{(\mathbf{x}_n, t_n)\}_{n=1}^N$ where $\mathbf{x}_n \in \mathbb{R}^P$, and $t_n \in \{1, 2, \ldots, L\}$ denotes the class label of the $n$-th sample. The data matrix $\mathbf{X}$ can be written as $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_N]^\top$. DR follows the general framework of multivariate linear regression (MLR) [9] with a specific class indicator matrix as response variables, which has the following form

$$\mathbf{Y} = \mathbf{X}\mathbf{W} + \mathbf{E}, \tag{1}$$

where $\mathbf{Y} \in \mathbb{R}^{N\times L}$ is the response matrix, $\mathbf{W} \in \mathbb{R}^{P\times L}$ is the projection matrix, and $\mathbf{E} \in \mathbb{R}^{N\times L}$ is the random error matrix. One typical way to encode $\mathbf{Y} \in \{0,1\}^{N\times L}$, which apply a vector valued class code for each of the sample, is as follows,

$$\mathbf{Y}_{nt} = \begin{cases} 1 & \text{if } t_n = t \\ 0 & \text{Otherwise.} \end{cases} \tag{2}$$

The classic DR determines the projection matrix $\mathbf{W}_{\text{cDR}} \in \mathbb{R}^{P\times L}$ via the minimization of the least square error,

$$\mathbf{W}_{\text{cDR}} = \arg\min_{\mathbf{W}} \frac{1}{N}\|\mathbf{Y} - \mathbf{X}\mathbf{W}\|_F^2. \tag{3}$$

The optimal $\mathbf{W}$ is given by

$$\mathbf{W}_{\text{cDR}} = (\mathbf{X}\mathbf{X}^\top)^+\mathbf{X}\mathbf{Y}, \tag{4}$$

where $(\cdot)^+$ denotes Moore–Penrose inverse. As a result of least square form of regression, various regularization techniques can be readily incorporated into the formulation to improve model sparsity and generalization ability. Taking the form of multivariate ridge regression [10], the penalized DR (pDR) [11] problem aims to solve the below problem

$$\mathbf{W}_{\text{pDR}} = \arg\min_{\mathbf{W}} \frac{1}{N}\|\mathbf{Y} - \mathbf{X}\mathbf{W}\|_F^2 + \lambda\|\mathbf{W}\|_F^2, \tag{5}$$

where $\lambda > 0$ is the regularization coefficient. The optimization problem (5) has the following unique solution [10]:

$$\mathbf{W}_{\text{pDR}} = (\mathbf{X}\mathbf{X}^\top + N\lambda\mathbf{I}_P)^{-1}\mathbf{X}\mathbf{Y}. \tag{6}$$

In DR, $\mathbf{X}\mathbf{W}_{\text{cDR}}$ and $\mathbf{X}\mathbf{W}_{\text{pDR}}$ are the extracted low dimensional feature matrices, where each sample is in $\mathbb{R}^L$. In other words, the dimension of extracted features is upper bounded by $L$.

Under mild conditions, the solutions of (4) and (6) have been shown to be equivalent to the solutions from linear discriminant analysis (LDA) [8] and regularized linear discriminant analysis (RLDA) [11], respectively. The superiority of the DR formulation over LDA/RLDA formulations is that it can lead to efficient and scalable implementations due to the least square formulation. However, DR suffers from severe problems when tackling the case of ill-posed high dimensional data. First, correlation between input variables is not adequately addressed in the process of feature extraction. In ill-posed high dimensional data, strong empirical correlation between variables or near linear dependence among a few variables is often encountered in many applications, which remains as notorious problems [12]. Second, DR may result in significant information loss in the form of underfitting. The underfitting problem is due to the dimension of extracted features is upper bounded by the number of class labels, i.e., $L$, which is typically a small number compared with the dimension of input data [8], [13].

The objective of this study is to address the challenges of correlated variables and underfitting discussed above. To achieve this objective, a new supervised feature extraction method, namely, clustered discriminant regression (CDR) is proposed in this paper. The novelty of CDR is twofold here.

(i) An idea to selectively group variables into different clusters, and then perform feature extraction from variables in each cluster. Accordingly, a unified formulation, which simultaneously optimizes clustering, variable selection and fitting error, is proposed.

(ii) An equivalent form of the proposed formulation to the Convex Integer Program (CIP) is proven. Based upon this, an efficient greedy algorithm is implemented to solve the corresponding CIP formulation. In addition, the theoretical guarantee of the CDR algorithm is also provided in the paper.

The remainder of this paper is organized as follows. A brief review of related research work is provided in Section II. The proposed formulation and its equivalent CIP formulation are introduced in Section III. The proposed algorithm and its theoretical guarantee are provided in Section IV, followed by the numerical study in Section V and real-world case studies in Section VI for testing and validation of the proposed approach. Finally, the conclusions and future work are discussed in Section VII.

## II. Research Background

In Section II-A, the review on feature extraction for classification is introduced. Related research about the techniques of clustering, variable selection in regression is reviewed briefly

in Section II-B. Afterwards, the research gaps of the existing work are identified in Section II-C.

### A. Feature Extraction for Classification

For the ill-posed high dimensional data, it is necessary to apply feature extraction methods as a pre-processing step of classification. In the area of bioinformatics analysis, the ill-posed high dimensional data is very common [6]. In the literature of bioinformatics analysis, there are a number of feature extraction techniques. Among them, kernel principal component analysis (KPCA) [14], local linear embedding (LLE) [15], and ISOMAP [16] have been widely applied and proven their success in bioinformatics analysis. These unsupervised feature extraction methods explore the relationship of the input data $\mathbf{X}$, however, they do not utilize the data of class labels. On the other hand, supervised feature extraction aims to build a connection between extracted features and class labels and thus has the potential to improve the classification accuracy. LDA is one of the most widely used methods in supervised feature extraction [17], which involves a computational intensive eigenvalue problem. However, LDA has the singularity problem of sample covariance matrix in the case of the ill-posed high dimensional data. In order to solve the singularity problem, RLDA [18] is proposed by adding an additional identical matrix to the sample covariance matrix to stabilize the sample covariance estimator. It has been shown that LDA and RLDA have a very close relationship with DR and pDR, respectively. Specifically, Sun *et al.* [13] found that DR is equivalent to LDA under mild conditions, leading to a scalable and efficient implementation. Based on a carefully designed class indicator matrix $\mathbf{Y}$, it has been proven that pDR is equivalent to RLDA [11]. The limitation of DR and pDR is that the number of extracted features is upper bounded by the number of class labels, namely, $L$, which causes loss of important information.

### B. Clustering and Variable Selection in Regression

In the case of ill-posed high dimensional data regression, it is very common for the variables (predictors) to be highly correlated [19], [20]. In the literature, there are two types of strategies reviewed below that aim to obtain a robust estimation by exploring the correlation among variables.

In the first type of strategy, clustering, the cluster-based estimation has very good predictive potential and yields more stable results [12], [21], [22]. For instance, to reveal groups of genes which act together and whose collective expression is strongly associated with an outcome variable of interest for microarray data, Dettling and Bühlmann [21] presented *Pelora*, an algorithm based on penalized logistic regression analysis. This method does not explicitly take correlation-structure among the variables into account and still exhibits difficulties when groups of variables are nearly linearly dependent. Accordingly, Bühlmann *et al.* [12] proposed a cluster group lasso, which identifies groups among the variables using (for instance) hierarchical clustering, and then applies the group lasso to the resulting groups. Furthermore, Witten *et al.* [22] proposed the cluster elastic net (CEN) that simultaneously estimates regression coefficients and clusters variables in a multivariate regression model, to identify sets that have the highest correlation with the response.

Variable selection, the second type of strategy, is an important approach in improving the performance of regression [9]. The advantage of variable selection in our application is that it can remove the redundant information. Li and Li [23] proposed to develop a network-constrained regularization procedure for fitting linear-regression models and variable selection, where the variables in the regression model are genomic data with graphical structures. Furthermore, the same authors considered the general problem of regression analysis when the predictor variable are nodes on a graph under the same procedure [24]. Huang *et al.* [25] proposed a penalized method for variable selection and estimation that explicitly incorporates the correlation patterns among predictors. This method is based on a combination of the minimax concave penalty and Laplacian quadratic associated with a graph as the penalty function. In [26], Xue and Qu proposed a new Semi-standard PArtial Covariance (SPAC) which reduces correlation effects from other predictors while incorporating the magnitude of coefficients.

Often times, these two strategies work together to achieve better performance. For instance, octagonal shrinkage and clustering algorithm for regression (OSCAR) [27] was proposed to simultaneously select variables while clustering them into predictive clusters. The technique is based on penalized regression with a geometrically intuitive penalty function that shrinks some coefficients to exactly zero. Penalized adaptive clustering and sparsity (PACS) [28] encouraged correlated variables to take on identical coefficient estimates via the use of a novel penalty function that can be interpreted as an octagonal constraint region. Robust versions of PACS (RPACS) [29] has been proposed by replacing the least squares and non robust weights in PACS with MM estimation and robust weights depending on robust correlations instead of person correlation, respectively.

### C. Research Gap Identification

In the case of ill-posed high dimensional data, there are two problems. Namely, (1) the correlation between variables is not properly explored and (2) the underfitting is caused by the dimension of class labels. Although various techniques are introduced to address the first problem including variables/coefficients clustering and variable selection in Section II-B, they cannot mitigate the underfitting problem as introduced in Section II-A. In addition, they cannot control the cardinality for variable selection. Therefore, this paper seeks to address these research gaps by devising a new supervised feature extraction method, namely, the clustered discriminant regression.

## III. Proposed Method

In this section, a new method for supervised feature extraction using pDR, namely, CDR, is proposed in Section III-A. It integrates clustering variables and selecting variables together with pDR into one formulation. This formulation is a mixed integer problem. In order to derive the solution for our formulation, an equivalent Convex Integer Program

formulation, which can be solved in a more efficient way, is proposed in Section III-B. Some theoretical results of the algorithm performance can also be achieved.

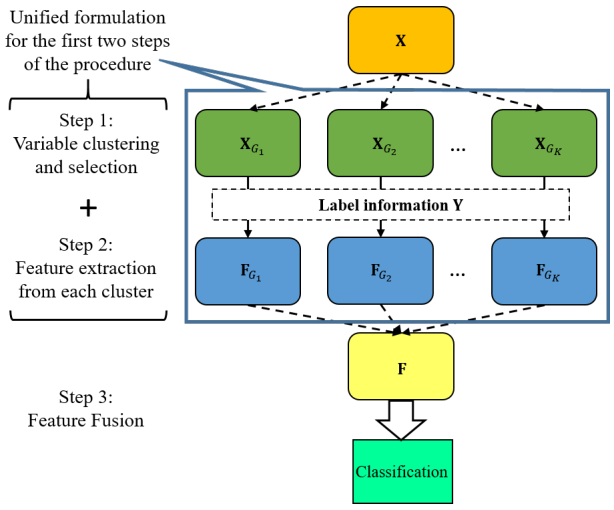



Fig. 1. Framework of the proposed method

### A. Unified Formulation for the First Two Steps Procedure

As discussed in Section II-B, it is beneficial to integrate clustering and variable selection to address the correlation problem. For reader's convenience, the pDR model (5) is mentioned again here

$$\mathbf{W}_{\text{cDR}} = \arg\min_{\mathbf{W}} \frac{1}{N}\|\mathbf{Y} - \mathbf{X}\mathbf{W}\|_F^2.$$

Based on the model of pDR, our proposed clustered discriminant regression takes advantages of clustering and variable selection, which consists of three steps summarized in Fig. 1. Considering the index set $\{1, \ldots, P\}$ for the variables in $\mathbf{X}$, the three steps procedure of clustered discriminant regression follows,

**Step 1** Variable clustering and selection: find a partition $\mathcal{G}$ where $\cup_{j=1}^{K} G_j \subset \{1, \ldots, P\}$ and $G_r \cap G_l = \emptyset, \forall r \neq l$, to separate $\mathbf{X}$ into disjoint clusters $\mathbf{X}_{G_1}, \mathbf{X}_{G_2}, \ldots, \mathbf{X}_{G_K}$.

**Step 2** Feature extraction from each cluster: apply pDR on $\mathbf{X}_{G_j}$'s to obtain the projection matrices $\mathcal{W} = \{\mathbf{W}_{G_1}, \mathbf{W}_{G_2}, \ldots, \mathbf{W}_{G_K}\}$.

**Step 3** Construct final feature set $\mathbf{F}$ by serial feature fusion [30] for classification analysis:

$$\mathbf{F} = (\mathbf{F}_{G_1}, \mathbf{F}_{G_2}, \ldots, \mathbf{F}_{G_K}),$$

where $\mathbf{F}_{G_j} = \mathbf{X}\mathbf{W}_{G_j}$ is the extracted feature from $j$th cluster.

For the testing dataset, the same scheme of **Step 3** can be applied to construct the final feature since the clustering $\mathcal{G}$ and the projection matrices $\mathcal{W}$ have already been identified in **Step 1** and **Step 2**.

However, there is still some room for improvement in the above procedure. Namely, **Step 1** and **Step 2** are in a sequential manner, they have two independent objective functions, where the clustering and variable selection result from **Step 1** are independent of the feature extraction in **Step 2**. As a result, the feature extraction is not synchronized together with the clustering and variable selection since the objective function for clustering and variable selection is not related to the class labels, while the class labels are the target of feature extraction in our work. To address this limitation, a unified formulation is proposed to integrate **Step 1** and **Step 2**. The formulation, aims to optimize the clustering and variable selection $\mathcal{G}$ and pDR projection matrices $\mathcal{W}$ simultaneously under the cardinality constrains assigned to each cluster for variable selection, is as follows.

$$\begin{aligned} \min_{\mathcal{G},\mathcal{W}} \quad & \sum_{j=1}^{K} \frac{1}{N}\|\mathbf{Y} - \mathbf{X}\mathbf{W}_{G_j}\|_F^2 + \lambda_j \|\mathbf{W}_{G_j}\|_F^2 \\ \text{subject to} \quad & |G_j| \leq g_j, j = 1, \ldots, K, \end{aligned} \tag{7}$$

where $\|\mathbf{W}_{G_j}\|_F^2 = \text{Tr}(\mathbf{W}_{G_j}^\top \mathbf{W}_{G_j})$ represents the ridge regularization term to avoid over-fitting issue, $\mathbf{W}_{G_j}(i,) = \mathbf{0}, \forall i \notin G_j$ for the variables not in $j$th cluster, $g_j$ is the maximum number of variables can be selected in $j$th cluster, and $\lambda_j$ is regularization coefficient. The regularization term in formulation (7) can also replaced by lasso regularization term $\|\mathbf{W}_{G_j}\|_1$ [31]. The reason to use ridge regression term instead of lasso is that given the clustering result $\mathcal{G}$, ridge based method can result in close form solutions, leading to a more tractable and scalable formulation (see Theorem 1 in Section III-B), while the lasso based method needs iterative algorithm to achieve a solution. Therefore, lasso based method is computational expensive and not scalable. Our formulation looks like a ridge regularization except the summation sign. In our formulation, variables in each cluster $G_j$ are selected to have maximum correlation with the label response $\mathbf{Y}$, satisfying the cardinality constrains. There are different definitions of indicator matrix $\mathbf{Y} \in \mathbb{R}^{N \times L}$ in the literature. The one showed in (2) is most common used while there are also many other variants of label indicator matrices defined in [11], [13]. Our proposal (7) differs from previous work in various aspects. First, the summation of least square error is not only the loss function for pDR and variable selection but also the similarity measurement for clustering. Second, the constrains in the formulation is the cardinality enforced in each cluster for the purpose of variable selection. The formulation (7) is an integrated one but also brings challenges since it is nonconvex mixed integer problem, which is intractable without further development. Therefore, an equivalent tractable formulation of (7) is developed in next Section III-B.

### B. Equivalent Formulation

As mentioned above, the optimization problem (7) is a nonconvex mixed integer optimization problem. The usual way to solve this type of problem is big-$\mathcal{M}$ method, which indeed tries to solve for the optimal projection matrices $\mathcal{W}$ in (7) directly [32]. In this paper, an equivalent convex integer program formulation is derived in order to avoid the calculation of continuous variables, namely, $\mathcal{W}$, which enables us to have a more efficient way to implement the greedy algorithm [33] with theoretical guarantee. To reformulate the problem (7) as a convex integer program, the main idea is

to separate the optimization in (7) into two steps: (i) first, optimize over $\mathcal{W}$ by fixing clustering result of variables and satisfying the cardinality constrains for each cluster, and (ii) then select the best combination of variables based on the cardinality constrains. After the first step, it turns out that it is a convex integer program. For a given clustering result $\mathcal{G}$, $\mathbf{W}_{G_j}$ can be obtained independently using multivariate ridge regression. Similar as (6), each $\mathbf{W}_{G_j}$ has the following close form solution:

$$\mathbf{W}_{G_j} = \begin{cases} (\mathbf{X}_{G_j}^\top \mathbf{X}_{G_j} + N\lambda_j \mathbf{I}_{|G_j|})^{-1} \mathbf{X}_{G_j}^\top \mathbf{Y} & i \in G_j \\ \mathbf{0} & i \notin G_j. \end{cases} \tag{8}$$

Assuming that the optimal object value of formulation (7) is $v^*$. By substituting (8) in the objective function of (7), the problem (7) is equivalent to

$$\begin{aligned} v^* = \min_{\mathcal{G}} \frac{1}{N} \sum_{j=1}^{K} \mathrm{Tr}\left(\mathbf{Y}^\top \mathbf{\Phi}_{G_j} \mathbf{Y}\right) \\ \text{subject to} \quad |G_j| \le g_j, j = 1, \ldots, K, \end{aligned} \tag{9}$$

where $\mathbf{\Phi}_{G_j} = \mathbf{I}_N - \mathbf{X}_{G_j}(\mathbf{X}_{G_j}^\top \mathbf{X}_{G_j} + N\lambda_j \mathbf{I}_{|G_j|})^{-1} \mathbf{X}_{G_j}^\top$. Now the only decision variable of problem (9) is $\mathcal{G}$. Based on the result from [32], [34], the problem in (9) can be formulated into a CIP formulation, which is described as follows.

**Theorem 1.** *The formulation (7) is equivalent to the following convex integer program optimization problem:*

$$\begin{aligned} \min_{\delta} \quad & \sum_{j=1}^{K} \lambda_j \sum_{t=1}^{L} \mathbf{Y}_t^\top [N\lambda_j \mathbf{I}_N + \sum_{i=1}^{P} \delta_{ij} \mathbf{X}_i \mathbf{X}_i^\top]^{-1} \mathbf{Y}_t \\ \text{subject to} \quad & \sum_{j=1}^{K} \delta_{ij} = 0 \text{ or } 1, \\ & \sum_{i=1}^{P} \delta_{ij} \le g_j, \\ & \delta_{ij} \in \{0,1\} \forall i, j. \end{aligned} \tag{10}$$

*Proof:* See proof in APPENDIX A. ∎

Based on the equivalent formulation (10), the decision variable is reduced to the assignment matrix $\delta \in \{0,1\}^{P \times K}$ only. Therefore, the algorithm to solve this formulation is readily to be developed in next section.

## IV. Proposed CDR Algorithm

In this section, the CDR algorithm to solve (10) is proposed, which is a greedy approach as shown in Fig. 2. In addition, the performance guarantee of the CDR algorithm is also provided. The greedy approach, i.e., forward selection, has been commonly used for the best subset selection [33]. The idea of the greedy approach is to select a variable which minimizes the marginal decrement of objective value in (10) while the constrains are satisfied. For a given selected subset $G \in [P]$ and an index $i \notin G$, the marginal objective value difference by adding $i$ to $G$ can be computed explicitly via the Sherman-Morrison formula [35] as below:

$$\mathbf{y}^\top [\mathbf{A}_G + \mathbf{X}_i \mathbf{X}_i^\top]^{-1} \mathbf{y} - \mathbf{y}^\top \mathbf{A}_G^{-1} \mathbf{y} = -\frac{(\mathbf{y}^\top \mathbf{A}_G^{-1} \mathbf{X}_i)^2}{1 + \mathbf{X}_i^\top \mathbf{A}_G^{-1} \mathbf{X}_i}$$

**Algorithm 1** CDR algorithm

**Input:** $\mathbf{X}, K, g_1, \ldots, g_K$.
**Output:** $v^G, G_1, \ldots, G_K$.

**Initialization:** set $\{G_j = \emptyset\}_{j=1}^K, G = \emptyset$, $g_1, \cdots, g_K$, $\mathbf{A}_{G_j} = N\lambda_j \mathbf{I}_N$.
**for** iter $= 1, \cdots, g$ **do**
  **for** $j = 1, \cdots, K$ **do**
    **if** $g_j < g$ **then**
      **Break**
    **end if**
    Let $i^* \in \underset{i \in [P] \setminus G}{\operatorname{argmin}} \{ -\frac{\sum_{t=1}^{L} (\mathbf{Y}_t^\top \mathbf{A}_{G_j}^{-1} \mathbf{X}_i)^2}{1 + \mathbf{X}_i^\top \mathbf{A}_{G_j}^{-1} \mathbf{X}_i} \}$.
    $G_j = G_j \cup \{i^*\}$, $G = G \cup \{i^*\}$.
    $\mathbf{A}_{G_j} = \mathbf{A}_{G_j} + \mathbf{X}_{i^*} \mathbf{X}_{i^*}^\top$ and $\mathbf{A}_{G_j}^{-1} = \mathbf{A}_{G_j}^{-1} - \frac{\mathbf{A}_{G_j}^{-1} \mathbf{X}_{i*} \mathbf{X}_{i*}^\top \mathbf{A}_{G_j}^{-1}}{1 + \mathbf{X}_{i*}^\top \mathbf{A}_{G_j}^{-1} \mathbf{X}_{i*}}$.
    $v^G \leftarrow \sum_{j=1}^{K} \lambda_j \sum_{t=1}^{L} \mathbf{Y}_t^\top \mathbf{A}_{G_j}^{-1} \mathbf{Y}_t$.
  **end for**
**end for**

Fig. 2. Procedure of the proposed algorithm for CDR

$$\mathbf{A}_{G \cup \{i\}}^{-1} = \mathbf{A}_G^{-1} - \frac{\mathbf{A}_G^{-1} \mathbf{X}_i \mathbf{X}_i^\top \mathbf{A}_G^{-1}}{1 + \mathbf{X}_i^\top \mathbf{A}_G^{-1} \mathbf{X}_i},$$

where $\mathbf{A}_G = N\lambda \mathbf{I}_N + \sum_{i \in G} \mathbf{X}_i \mathbf{X}_i^\top$ is the positive definite matrix based on subset $G$. This formula avoids heavy computation for matrix inverse at each iteration. It motivates us an efficient implementation of the CDR algorithm, which is detailed in Fig. 2. In our proposed algorithm, at each iteration, $\{\mathbf{A}_{G_j} \mathbf{X}_i\}_{i \in [P]}$ $\{\mathbf{X}_i^\top \mathbf{A}_{G_j} \mathbf{X}_i\}_{i \in [P]}$ and $\{\mathbf{Y}_t^\top \mathbf{A}_{G_j} \mathbf{X}_i\}_{i \in [P], t \in [L]}$ need to be kept track of, which has space complexity $O(NPL)$. In addition, updating them from one iteration to another costs $O(NPL)$ operations per iteration. In total, the space and time complexity of the algorithm are $O(NPLK)$ and $O(NPLKg)$, respectively, where $g = \max_j g_j$ is the maximum cardinality among all clusters.

Next, we are going to investigate the CDR solution and to prove that it can be very close to the true optimal, in particular when the regularization coefficient $\lambda_j$ is not too small. At the beginning, define $\theta_g$ to be the largest eigenvalue of all the matrices $\mathbf{X}_G \mathbf{X}_G^\top$ with $|G| = g$, i.e.,

$$\theta_g = \max_{|G| = g} \sigma_{\max}(\mathbf{X}_G \mathbf{X}_G^\top),$$

for each $G \subset [P]$. Accordingly, $\theta_1 \le \theta_2 \le \cdots \le \theta_P$, and by default, let $\theta_0 = 0$. Define $\underline{\lambda} = \min_j \lambda_j$, $\overline{\lambda} = \max_j \lambda_j$, and $\beta = \max_j \frac{\theta_{g_j}}{N\lambda_j}$. Our main result of near-optimality of the CDR algorithm is stated as below. That is, if $P \ge N + \sum_{j=1}^{K} g_j$, then the objective value of CDR algorithm will be quite close to any optimal value from (7) as $\lambda_j$ grows.

**Lemma 1.** *Suppose that $P \ge N + \sum_{j=1}^{K} g_j$, then the algorithm output, i.e., $v^G$, of the CDR algorithm is bounded by*

$$v^* \le v^G \le (1+\beta)(1-\gamma) v^*,$$

*where* $\gamma = \frac{\underline{\theta}}{K(N\underline{\lambda}+\theta_1)(1+\beta)^2}\log(\frac{P+1}{P+1-\sum_{j=1}^{K} g_j})$, *and* $\underline{\theta}$ *is defined as*

$$\underline{\theta} = \min_{G\in[P],|G|\geq P-\sum_{j=1}^{K} g_j} \sigma_{\min}(\mathbf{X}_G\mathbf{X}_G^\top).$$

*Proof:* See proof in APPENDIX B. ∎

In addition, remark that $\mathbf{W}_{G_j}$ of the CDR algorithm can be compute as (8) after obtaining the assignment matrix $\delta$. In the next theorem, the derived estimator from the CDR algorithm is showed to be quite close to an optimal solution set $\{\mathbf{W}_{G_j^*}\}_{j=1}^K$.

**Theorem 2.** *Let* $\{\mathbf{W}_{G_j^*}\}_{j=1}^K$ *be the optimal solution set to* (7) *with selected set* $\{G_1^*, G_2^*, \ldots, G_K^*\}$ *and* $\{\mathbf{W}_{G_j}\}_{j=1}^K$ *be the solution set from CDR algorithm with selected sets* $\{G_1, G_2, \ldots, G_K\}$. *If* $P \geq N + \sum_{j=1}^{K} g_j$, *then*

$$\sum_{j=1}^{K} \|\mathbf{W}_{G_j} - \mathbf{W}_{G_j^*}\|_F \leq \frac{2K\sqrt{N\theta_{\max}v^*}}{N\underline{\lambda}+\sigma_{\min}} + \sqrt{\frac{\alpha KNv^*}{N\underline{\lambda}+\sigma_{\min}}}, \tag{11}$$

*where* $G_j^U = G_j \cup G_j^*$, $\theta_{\max} = \max_{j\in[K]} \theta_{|G_j\setminus G_j^*|}$, $\sigma_{\min} = \min_{j\in[K]} \sigma_{\min}(\mathbf{X}_{G_j^U}^\top \mathbf{X}_{G_j^U})$, *and*

$$\alpha = (1+\beta)(1-\gamma) - 1.$$

*Proof:* See proof in APPENDIX C. ∎

If $P < N + \sum_{j=1}^{K} g_j$, then $\underline{\theta} = 0$ based on its definition. However, the condition $P \geq N + \sum_{j=1}^{K} g_j$ is quite mild in the case of high dimensional regression. Theorem 2 gives the upper bound of error between the CDR estimator and global optimal estimator. It also shows asymptotic optimality of the CDR algorithm. When $\underline{\lambda}$ increases, $\alpha$ is closer to 0. As a result, the CDR estimator becomes more accurate because the error term in (11) becomes smaller when $\underline{\lambda}$ increases. Specially, the error term vanishes if $\underline{\lambda} \to +\infty$.

## V. Numerical Study

To evaluate the performance of the proposed CDR algorithm, numerical illustration of CDR algorithm on synthetic data is conducted in this section. All results in this section are the average results of ten repetitions for comparison. The CPU of the computer used in the experiments is an Intel® Xeon® Processor E3-1220 v3 (Quad Core, 3.10 GHz Turbo, 8MB).

In this numerical study, the clustering and variables selection result $\mathcal{G}$ is assumed to be known in order to evaluate the performance of the CDR algorithm by how good the algorithm can identify the right support of $\mathcal{W}$. In Section V-A, the procedure of data generation for evaluation of our algorithm is introduced. The numerical results are summarized in Section V-B.

### A. Simulation Data Generation

In order to simulate the clustering effect in variables, the data matrix $\mathbf{X}$ is generated from a multivariate normal distribution, where the covariance matrix has two blocks structure. Each sample (row) in $\mathbf{X}$ is drawn independent identically distributed (i.i.d.) from the multivariate normal distribution; that is

$$\mathbf{x}_n \sim N(0, \mathbf{\Sigma}), \forall n \in [N].$$

**Algorithm 2** Data generation

**Input:** $\mathbf{X}$; Random select $G_1 \in [P/2]$ and $G_2 \in [P] \setminus [P/2]$ with $|G_1| = g_1, |G_2| = g_2$.
**Output:** $\mathbf{Y}, \mathbf{W}_1, \mathbf{W}_2$.

1: **Initialization:** Generate $\mathbf{Y}$ following (1) $\mathbf{Y} = \mathbf{X}\mathbf{W}_1 + \mathbf{E}$, where the values of nonzero coefficients in $\mathbf{W}_1$ are drawn i.i.d. from the uniform distribution $Unif(2.9, 3)$ and $\mathbf{E}(n,t) \sim N(0, \sigma^2)$.
2: **Main iteration:**
3: **while** $\|\mathbf{Y} - \mathbf{X}\mathbf{W}_1\| + \|\mathbf{Y} - \mathbf{X}\mathbf{W}_2\| > \epsilon$ **do**
4: $\quad \mathbf{Y} \leftarrow Binarize(\mathbf{Y}_1)$, where the $Binarize(\cdot)$ is the row operator to set the maximum value in a row to be 1 and other locations to be 0.
5: $\quad \mathbf{W}_1, \mathbf{W}_2$ can be updated through (4).
6: $\quad \mathbf{Y}_1 \leftarrow \mathbf{X}\mathbf{W}_1$.
7: **end while**

Fig. 3. Procedure of the proposed algorithm for data generation

Followed from [27], [32], [34], the covariance matrix $\mathbf{\Sigma}$ can be parametrized by the correlation coefficient $\rho_1, \rho_2 \in [0, 1)$ as $\mathbf{\Sigma}_{ij} = \rho_1^{|i-j|}$, if $i, j \leq P/2$; $\mathbf{\Sigma}_{ij} = \rho_2^{|i-j|}$, if $i, j > P/2$; 0, otherwise. For simplicity, the case $L = 2$ and $K = 2$ is considered in our numerical experiments. For each block, projection matrix $\mathbf{W}_j$ is generated to have the same binary response $\mathbf{Y}$ to simulate our situation. The response $\mathbf{Y} \in \{0, 1\}^{N\times 2}$ and projection matrices $\mathbf{W}_1 \in \mathbb{R}^{P\times 2}$ and $\mathbf{W}_2 \in \mathbb{R}^{P\times 2}$ are generated synthetically in a iterative way, which is described in Fig. 3. To control the signal-to-noise ratio (SNR) [34], the value of $\sigma^2$ can be chosen by satisfying

$$SNR = \frac{var(\mathbf{X}\mathbf{W}_1(,1)) + var(\mathbf{X}\mathbf{W}_1(,2))}{\sigma^2} = 9,$$

where $\mathbf{W}_1(,j)$ represents the $j$th column of matrix $\mathbf{W}_1$.

TABLE I
Numerical performance of CDR algorithm

| $P$ | Cardinality $(g_1, g_2)$ | Samples.Size $(N)$ | CPU. Time (Seconds) | Avg. False Alarm Rate |
|---|---|---|---|---|
| 2000 | (5,5) | 250 | 5.21 | 18.0% |
| | | 500 | 13.64 | 10.0% |
| | | 1000 | 40.96 | 0.0% |
| | (10,10) | 250 | 9.24 | 27.0% |
| | | 500 | 29.11 | 22.0% |
| | | 1000 | 79.46 | 13.0% |
| | (15,15) | 250 | 15.57 | 41.3% |
| | | 500 | 43.08 | 28.0% |
| | | 1000 | 138.10 | 18.7% |
| | (20,20) | 250 | 19.23 | 52.5% |
| | | 500 | 49.99 | 32.0% |
| | | 1000 | 158.50 | 22.0% |

### B. Numerical Results

The performance of the CDR algorithm is evaluated by comparing the support of estimated $\mathbf{W}$ using the CDR algorithm and the actual support of $\mathbf{W}$ in the data generation process. *Avg. False Alarm Rate* is used here for the performance evaluation. In our numerical experiments, there are different combinations of $(g_1, g_2), N$ used in the following cases, where $(g_1, g_2) \in \{(5,5), (10,10), (15,15), (20,20)\}$

and $N \in \{250, 500, 1000\}$. For all the testing instances, tuning parameters $\rho_1 = 0.8$, $\rho_2 = 0.2$, $\lambda_1 = \lambda_2 = 0.1$, and $P = 2000$ are used for data simulation. The experiment results are listed in TABLE I. It can be clearly seen that as $(g_1, g_2)$ grows, *Avg. False Alarm Rate* increases for the same $N$. When $(g_1, g_2) = (5, 5)$ and $N = 1000$, the CDR algorithm can detect all the correct variables. Under the same $(g_1, g_2)$, *Avg. False Alarm Rate* decreases as $N$ increases. It implies that the CDR algorithm's ability to identify right variables improves as the sample size $N$ increases. This can be justified by the inequality (11) in Theorem 2, the first term in the right hand side of inequality is the major error for the upper bound because it is quadratic order of $\sqrt{\frac{K}{N\underline{\lambda}+\sigma_{\min}}}$ when $\frac{K}{N\underline{\lambda}+\sigma_{\min}} > 1$. When all the variables can be identified correctly, this first error term vanishes because $\theta_{\max} = 0$ in (11). If more variables can be detected correctly, the first error term is smaller because $\theta_{\max}$ becomes smaller. Even though it is very hard to calculate $\mathcal{W}^*$, the results in TABLE I show the ability of CDR algorithm to recover the supports of true $\mathcal{W}^*$ and to estimate the true $\mathcal{W}^*$.

## VI. Real-World Case Studies

In this section, the proposed method is applied to three real-world case studies for classification. Section VI-A: Arcene Cancer Dataset from UCI [36] is used. The data consists of cancer patients and healthy patients. Section VI-B: Fused Filament Fabrication (FFF) AM process [3]. The data is acquired from a digital microscope at sampling rate of 1 Hz. Section VI-C: Electron Beam Melting (EBM) Metal AM process. High resolution CCD camera is used to capture images of printed parts. In all experiments, all the pixels in an image are vectorized for convenience of operation. pDR, KPCA, ISOMAP, and LLE are selected as benchmarks for comparison. To evaluate the classification performance of different feature extraction methods, KNN [37], Naïve Bayes Classifier (NBC) [38], SVM [39], LDA [17] and nerual network (NN) [40] are selected as classification algorithms. All the $\mathbf{Y}$'s are encoded as representation in (2) for simplicity. $\lambda_j$'s are selected through cross validation. All results in this section are the average results of ten repetitions for comparison.

Since the data in Sections VI-B and VI-C are AM images, it is necessary to compare with popular methods in deep learning. Specifically, convolutional neural network (CNN) [41] has been applied to AM and shown its success in [42]–[44]. Therefore, we compare our method with CNN regarding the classification performance. The structure of CNN used in this paper contains five convolutional layers and one fully-connected layer, which is selected so that it has best classification performance. For $i$th convolutional layer ($i = 1, \ldots, 5$), the number of filters is $2^{i+2}$ with size $3 \times 3$, batch normalization and max pooling are used together with the ReLU activation function. For the final fully-connected layer, it uses the Softmax activation function for classification. The maximum number of epochs is set to 20, which makes CNN converges for all the cases.

### A. Arcene Cancer Classification

This case study demonstrates the classification performance on Arcene Cancer Dataset [36]. This is a mass-spectrometric dataset, which contains 7000 continuous input variables. In addition, there are 200 records in which 88 of 200 are cancer patients, 112 of 200 are healthy patients. Therefore, this is a two-class classification problem.

In our experiment, 60%, 30% and 20% of patients from each class are selected as the training dataset. The remaining is treated as the testing dataset. For our method, $1 \sim 20$ clusters are conducted in the experiments for classification. For all these cases, 3500 out of 7000 variables are selected as equally spaced as possible for different clusters (if 3 clusters, cluster 1 to 2 have 1167 variables and cluster 3 has 1166 variables). The best performance among different number of clusters represents our method, namely, CDR(Best). In addition to pDR, KPCA, ISOMAP, and LLE are selected as feature extraction methods for comparison since they have been widely applied in the literature of bioinformatics analysis [6] where $P \gg N$. The average F-score is reported in TABLE II.

For all the cases, our method shows the best classification results except that LLE has the best classification performance in one case. The proposed CDR can not only improve over pDR but also outperform popular feature extraction methods in bioinformatics analysis. For comparison with pDR, CDR has different degrees of improvement over pDR for different classification algorithms. For NBC, SVM, and NN, there is a significant improvement on the performance of classification. When the percentage of training sample decreases, there is a decreasing trend of classification performance. However, our proposed CDR can still maintain high classification accuracy.

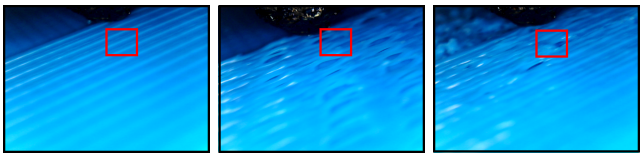


Fig. 4. (a) Normal; (b) Under-fill caused by fan; (c) Under-fill caused by feed rate (Red boxes are Region of interest) [3]

### B. Polymer Additive Manufacturing Process

In this subsection, a polymer AM dataset is used to explore the effect of number of clusters on different classification algorithms. AM, which is also called 3D printing, is the process of joining materials to make objectives from 3D models, usually in a layer by layer fashion [3].

In this work, a desktop fused filament fabrication (FFF) 3D printer, namely, a Hyrel System 30M 3D printer, is used to print a test artifact, which is a cuboid of dimensions 2 in.$\times$2 in.$\times$2 in. The machine has an extruder with 0.5 mm nozzle diameter and uses acrylonitrile butadiene styrene (ABS) with a diameter of 1.75 mm as filament for printing. In reality, the FFF printing process frequently occurred defects such as voids, over-fill, under-fill, etc. Therefore, sensor based

TABLE II
CLASSIFICATION RESULTS OF ARCENE CANCER DATA FOR DIFFERENT TRAINING TESTING RATIO USING DIFFERENT CLASSIFICATION ALGORITHMS

| (training sample %, testing sample %) | Feature extraction | KNN ($k=3$) | NBC | LDA | SVM | NN |
|---|---|---|---|---|---|---|
| $(60\%, 40\%)$ | KPCA | 0.6620 | 0.6531 | 0.6798 | 0.6798 | 0.6620 |
| | ISOMAP | 0.7341 | 0.6620 | 0.7002 | 0.7256 | 0.7040 |
| | LLE | 0.8108 | 0.6907 | **0.8620** | 0.7999 | 0.8522 |
| | pDR | 0.8068 | 0.4190 | 0.8316 | 0.5093 | 0.7046 |
| | CDR(Best) | **0.8719** | **0.8648** | 0.8421 | **0.8665** | **0.8897** |
| $(30\%, 70\%)$ | KPCA | 0.6505 | 0.6454 | 0.5976 | 0.6505 | 0.6378 |
| | ISOMAP | 0.6838 | 0.6306 | 0.6580 | 0.7070 | 0.6870 |
| | LLE | 0.7670 | 0.7063 | 0.7967 | 0.7689 | 0.7860 |
| | pDR | 0.7850 | 0.3802 | 0.7836 | 0.5915 | 0.7829 |
| | CDR(Best) | **0.7988** | **0.7725** | **0.8109** | **0.8076** | **0.8195** |
| $(20\%, 80\%)$ | KPCA | 0.6671 | 0.6270 | 0.6804 | 0.6559 | 0.6604 |
| | ISOMAP | 0.6525 | 0.5846 | 0.6257 | 0.6501 | 0.6529 |
| | LLE | 0.7092 | 0.6841 | 0.7068 | 0.7004 | 0.7424 |
| | pDR | 0.6748 | 0.5129 | 0.7182 | 0.5691 | 0.6755 |
| | CDR(Best) | **0.7363** | **0.7122** | **0.7412** | **0.7186** | **0.7512** |

Note: The bold numbers are the best performance for each training testing ratio.
CDR(Best) represents the best performance of CDR where $K = 1 \sim 20$.

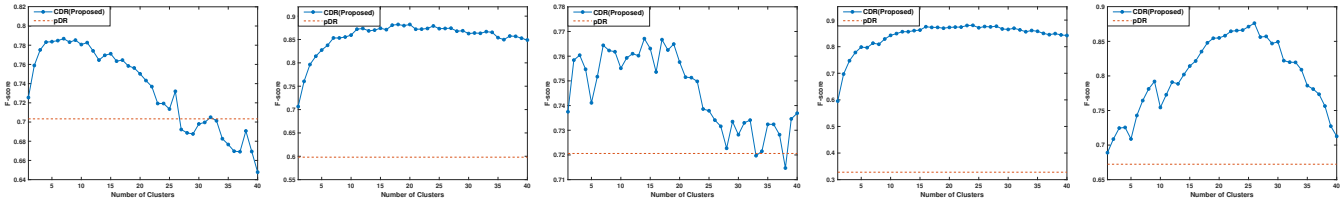
TABLE III
CLASSIFICATION RESULTS FOR POLYMER AM DATA USING DIFFERENT CLASSIFICATION ALGORITHMS

| Classification algorithm | Feature Extraction | Accuracy | Precision | Recall | F-score | CPU. Time (Seconds) |
|---|---|---|---|---|---|---|
| KNN ($k=3$) | KPCA | 0.6475 | 0.6626 | 0.6320 | 0.6367 | 0.0628 |
| | ISOMAP | 0.6888 | 0.7055 | 0.6716 | 0.6773 | 0.1025 |
| | LLE | 0.6455 | 0.6805 | 0.6300 | 0.6347 | 0.1142 |
| | pDR | 0.7380 | 0.7727 | 0.6949 | 0.7033 | **0.0210** |
| | CDR(Best) | **0.8010** | **0.8320** | **0.7694** | **0.7876** | 0.0508 |
| NBC | KPCA | 0.3832 | 0.3825 | 0.3811 | 0.3786 | 0.0478 |
| | ISOMAP | 0.7784 | 0.7769 | 0.7741 | 0.7690 | 0.0781 |
| | LLE | 0.8142 | 0.8126 | 0.8097 | 0.8044 | 0.0870 |
| | pDR | 0.6664 | 0.6298 | 0.5985 | 0.5981 | **0.0160** |
| | CDR(Best) | **0.8904** | **0.8912** | **0.8880** | **0.8822** | 0.0200 |
| LDA | KPCA | 0.6475 | 0.6626 | 0.6320 | 0.6367 | 0.0508 |
| | ISOMAP | 0.7077 | 0.7239 | 0.6775 | 0.6902 | 0.0830 |
| | LLE | 0.7552 | 0.7564 | 0.7238 | 0.7385 | 0.0924 |
| | pDR | 0.7512 | 0.7801 | 0.7095 | 0.7206 | **0.0170** |
| | CDR(Best) | **0.7865** | **0.8045** | **0.7530** | **0.7671** | 0.0380 |
| SVM | KPCA | 0.6406 | 0.6589 | 0.6254 | 0.6285 | 0.2348 |
| | ISOMAP | 0.8689 | 0.8578 | 0.8469 | 0.8520 | 0.3832 |
| | LLE | 0.7586 | 0.7513 | 0.7473 | 0.7438 | 0.2469 |
| | pDR | 0.4954 | 0.7240 | 0.2808 | 0.3280 | **0.0785** |
| | CDR(Best) | **0.8974** | **0.8918** | **0.8832** | **0.8803** | 0.0906 |
| NN | KPCA | 0.6317 | 0.6381 | 0.6155 | 0.6172 | 0.0568 |
| | ISOMAP | 0.7586 | 0.7636 | 0.7356 | 0.7412 | 0.0927 |
| | LLE | 0.7198 | 0.7252 | 0.7021 | 0.7033 | 0.1033 |
| | pDR | 0.7137 | 0.7261 | 0.6702 | 0.6723 | **0.0190** |
| | CDR(Best) | **0.8983** | **0.8982** | **0.8749** | **0.8763** | 0.0358 |
| CNN (End-to-End) | | 0.8552 | 0.8435 | 0.8396 | 0.8404 | 0.0102 |

Note: The bold numbers are the best performance for each classification algorithm.
CDR(Best) represents the best performance of CDR where $K = 1 \sim 40$.

classification is often used for online monitoring the process of manufacturing. In this paper, high quality images of part surface being printed are collected by two digital microscopes at sampling frequency of 1Hz, which are installed at the two sides of the extruder of 3D printer. Images for normal, under-fill feed rate, and under-fill fan as shown in Fig. 4 can be collected by setting up the machine parameters, which has been studied through design of experiments in [3]. For each image, a region of interest (ROI) is utilized, which is cropped from the original image ($640 \times 480$ pixels) to a smaller region ($80 \times 80$ pixels) right below the nozzle as shown in Fig. 4. The image in ROI is transformed into a vector with 6400 variables for testing different algorithms. For each variable, it represents pixel value ranging from $0 \sim 255$. In total, there are 305 normal images, 197 under-fill feed rate images and 153 under-fill fan images.

In our experiments, 3200 out of 6400 variables are selected as equally spaced as possible for different clusters (if 6 clusters, cluster 1 to 5 have 533 variables and cluster 6 has 534 variables). The average statics for accuracy, precision, recall and F-score, is used as the final result to verify the effectiveness of classification. The results summarized in TABLE III, where CDR(best) represents the best performance of our method by ranging the number of clusters from 1 to 40. It shows that our method outperforms pDR, KPCA, ISOMAP, and LLE for all the five classification algorithms. Especially

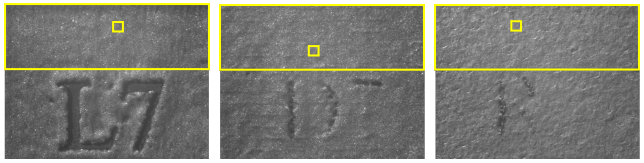



(a) KNN (b) NBC (c) LDA (d) SVM (e) NN

Fig. 5. Classification performance on Poly AM data for different algorithms: (a) KNN; (b) NBC; (c) LDA; (d) SVM; (e) NN.

NBC, SVM, and NN, there is a significant improvement on the performance of pDR. For comparison with CNN, the classification performance of our method with different classifiers varies. Our proposed CDR with classifiers NBC, SVM, and NN, can achieve better classification performance than CNN does. The average computational time for constructing feature and classification of a single observation is less than 0.1 seconds (equivalent to 10 Hz) for all the classification algorithms in CDR(Best). It can be observed in TABLE III that although it is not as fast as the benchmark method, pDR, it is sufficient for online defect diagnosis of FFF processes since the computational speed (10 Hz) is faster than the sampling rate for the images (1 Hz) during the printing process.

On the other hand, the effect of number of clusters on the performance of classification is an interesting and valuable problem to explore since it is also a quite important problem in the clustering literature. However, the results in TABLE III cannot reflect the effect of number of clusters. Accordingly, the classification results based on F-score are plotted in Fig. 5 via number of clusters. For most of cases, the F-score is higher than the pDR (red dash line) corresponding to each classification algorithm as shown in Fig. 5. For all these classification algorithms, there is an increasing trend at the beginning in terms of F-score. After reaching a specific number of clusters (depends on the classification algorithm), the classification performance starts to decrease. This case study demonstrates the power of clustering due to the additional information it brings in.




(a) Raster (b) Dehoff (c) Random

Fig. 6. (a) Raster; (b) Dehoff; (c) Random (Yellow square boxes are selected random samples); L7, D7, and R are the markers to identify samples using different scan strategies.

### C. Metal Additive Manufacturing Process

In this subsection, metal AM data from Electron Beam Melting (EBM) process is used to evaluate classification performance on different training/testing ratio and different number of variables to be selected. In the printing process, the machine ARCAM Q10 plus is applied to print samples with dimensions 15mm×15mm×25mm using Ti-6Al-4V powder. In EDM, there are three different scan strategies, i.e., Dehoff, raster and random [45] during the printing process. By using different scan strategies, it provides different surface patterns of the printed samples. After printing three different samples by Dehoff, raster and random scan strategies, a high resolution CCD camera is used to capture very detailed 2D information about the top surface (15mm×15mm) quality. The image data shown in Fig. 6 are the surface patterns of Dehoff, raster and random, respectively. The objective of classification in this case study is to identify the scan strategy based on the surface patterns. For each image in Fig. 6, the size is 1448×1928. To obtain multiple training samples, 120 images of size 60 by 60 are randomly selected without overlap from the upper part of each image (500×1928) since bottom part with letter and number has lots of defects such as porosity. In total, there are 120 images of Raster pattern, 120 images of Dehoff pattern, and 120 images of Random pattern available for experiments. For each image 60 by 60, it is transformed into a vector with 3600 variables, where each variable represents pixel value ranging from 0 to 255, for testing different algorithms.

In our experiment, 70%, 60% and 50% of images from each pattern are selected as the training dataset. The remaining is treated as the testing dataset. For our method, $1 \sim 20$ clusters are conducted in the experiments for classification. For all these cases, 1600 out of 3600 variables are selected as equally spaced as possible for different clusters. The best performance among different number of clusters represents our method, namely, CDR(Best). The average F-score is reported in TABLE IV. The classification result implies that our method shows better classification result for different training phases. Specifically, all classification algorithms have different degrees of improvement over pDR. NBC and SVM exhibit significant improvement, which is consistent with the results in section VI-B. The classification results of NBC and SVM for pDR are poor because the dimension of extracted features is three, which caused severe underfitting. For comparison with CNN, our method with all classifiers can achieve similar performance with CNN. For the case of 70% training samples, our method with KNN, NBC, LDA, and SVM can achieve better performance than CNN; For the case of 60% training samples, our method with KNN, NBC, LDA, and SVM can achieve better performance than CNN; For the case of 50% training samples, our method with KNN, LDA can achieve better performance than CNN.

The TABLE IV only shows the classification results using 1600 out of 3600 variables. In order to explore the effect of total number of variables to be used on classification

TABLE IV
CLASSIFICATION RESULTS OF METAL AM DATA FOR DIFFERENT TRAINING TESTING RATIO USING DIFFERENT CLASSIFICATION ALGORITHMS

| (training sample %, testing sample %) | CNN (End-to-End) | Feature extraction | KNN ($k=3$) | NBC | LDA | SVM | NN |
|---|---|---|---|---|---|---|---|
| $(70\%, 30\%)$ | 0.9729 | KPCA | 0.5000 | 0.5000 | 0.5000 | 0.5000 | 0.5000 |
| | | ISOMAP | 0.9307 | 0.9218 | 0.9296 | 0.9207 | 0.9014 |
| | | LLE | 0.9238 | 0.9144 | 0.9205 | 0.9245 | 0.9030 |
| | | pDR | 0.9324 | 0.2438 | 0.9095 | 0.3210 | 0.6164 |
| | | CDR(Best) | **0.9963** | **0.9944** | **0.9944** | **0.9824** | **0.9432** |
| $(60\%, 40\%)$ | 0.9709 | KPCA | 0.4865 | 0.4759 | 0.4920 | 0.5240 | 0.5123 |
| | | ISOMAP | 0.9249 | 0.9068 | 0.8814 | 0.9163 | 0.9089 |
| | | LLE | 0.9125 | 0.9017 | 0.9054 | 0.8928 | 0.8802 |
| | | pDR | 0.9270 | 0.2493 | 0.9007 | 0.3534 | 0.5586 |
| | | CDR(Best) | **0.9839** | **0.9841** | **0.9860** | **0.9799** | **0.9508** |
| $(50\%, 50\%)$ | 0.9789 | KPCA | 0.4725 | 0.4824 | 0.5059 | 0.4520 | 0.4987 |
| | | ISOMAP | 0.9051 | 0.9008 | 0.8749 | 0.8721 | 0.8722 |
| | | LLE | 0.8469 | 0.8944 | 0.8755 | 0.8591 | 0.8500 |
| | | pDR | 0.9228 | 0.2510 | 0.8952 | 0.3926 | 0.5138 |
| | | CDR(Best) | **0.9832** | **0.9693** | **0.9866** | **0.9423** | **0.9432** |

Note: The bold numbers are the best performance for each training testing ratio.
CDR(Best) represents the best performance of CDR where $K = 1 \sim 20$.

TABLE V
CLASSIFICATION RESULTS OF METAL AM DATA VIA DIFFERENT NUMBER OF SELECTED VARIABLES

| $\sum_{j=1}^{K} g_j$ | $K$ | KNN ($k=3$) | NBC | LDA | SVM | NN |
|---|---|---|---|---|---|---|
| 3200 | 6 | 0.9944 | 0.0654 | 0.9935 | 0.3921 | 0.7012 |
| | 8 | 0.9953 | 0.1796 | 0.9879 | 0.3331 | 0.7484 |
| | 10 | 0.9907 | 0.6169 | 0.9860 | 0.4851 | 0.7814 |
| | 12 | 0.9907 | 0.9879 | 0.9879 | 0.7579 | 0.7873 |
| 1600 | 6 | 0.9851 | 0.9257 | 0.9785 | 0.7081 | 0.8088 |
| | 8 | 0.9869 | 0.9888 | 0.9822 | 0.8939 | 0.8420 |
| | 10 | 0.9916 | 0.9842 | 0.9683 | 0.9777 | 0.8758 |
| | 12 | 0.9889 | 0.9889 | 0.9730 | 0.9843 | 0.8831 |
| 800 | 6 | 0.9832 | 0.9776 | 0.9655 | 0.9592 | 0.8689 |
| | 8 | 0.9841 | 0.9841 | 0.9794 | 0.9814 | 0.9033 |
| | 10 | 0.9907 | 0.9897 | 0.9805 | 0.9860 | 0.9253 |
| | 12 | 0.9889 | 0.9916 | 0.9898 | 0.9889 | 0.9591 |
| 400 | 6 | 0.9795 | 0.9814 | 0.9794 | 0.9804 | 0.8361 |
| | 8 | 0.9814 | 0.9832 | 0.9833 | 0.9860 | 0.8613 |
| | 10 | 0.9842 | 0.9898 | 0.9851 | 0.9889 | 0.8522 |
| | 12 | 0.9814 | 0.9917 | 0.9804 | 0.9907 | 0.8524 |

performance, for $70\%$ training samples, the cases of 3200, 1600, 800, and 400 variables are conducted for classification. For each case, 6, 8, 10, 12 clusters are selected for the consideration of clustering effect. TABLE V summarizes the results of different classification algorithms. For KNN and LDA, the classification results are very similar for different number of variables and clusters. For NBC, SVM and NN, the classification results are improved when the number of variables is decreased. It shows that selecting proper number of variables can remove redundant information to improve the performance of classification. This case study shows that our method preserves classification performance for different training/test ratio and the variable selection scheme in our method is very effective.

## VII. CONCLUSION

In this paper, a new supervised feature extraction method is developed for high dimensional data to address two problems arisen in DR. The proposed CDR considers clustering, variable selection together with minimizing the fitting error. To access the solutions efficiently, a greedy algorithm for CDR is implemented with performance guarantee. The numerical study shows that the CDR algorithm is able to recover the true supports and find near-optimal solutions. The classification results of the Arcene Cancer Dataset demonstrate that our algorithm outperforms pDR and popular methods in bioinformatics analysis, namely, KPCA, LLE and ISOMAP. More importantly, two case studies based on additive manufacturing process images illustrate that the effectiveness of clustering and variable selection on classification in real-world applications. For AM image classification, our proposed method with elaborately selected classification algorithms can achieve better classification performance than CNN.

Additionally, there are still some aspects of CDR that deserve further investigations. First, as the experimental results show that the number of clusters can affect the performance of classification. Therefore, finding the relationship among number of clusters and the performance of classification is one of the next steps of research. Second, the cardinality $g_j$ assigned to $j$th cluster is also critical to the classification performance. Thereafter, how to estimate the cardinality for each cluster can be further investigated. Third, the parallel version of CDR algorithm needs further research in order to improve the computation efficiency for different requirements of applications.

## APPENDIX A
## PROOF OF THEOREM 1

*Proof:* The objective function in (9) can be rewritten as,

$$\begin{aligned} &\frac{1}{N}\sum_{j=1}^{K}\mathrm{Tr}\left(\mathbf{Y}^{\top}\mathbf{\Phi}_{G_j}\mathbf{Y}\right) \\ =&\frac{1}{N}\sum_{j=1}^{K}\sum_{t=1}^{L}\mathbf{Y}_t^{\top}\mathbf{\Phi}_{G_j}\mathbf{Y}_t, \end{aligned} \tag{12}$$

where $\mathbf{Y}_t$ is $t$th column of $\mathbf{Y}$. The first equality is from the definition of trace. Next,

$$\begin{aligned}
&\mathbf{\Phi}_{G_j}\\
=&\mathbf{I}_N - \mathbf{X}_{G_j}(\mathbf{X}_{G_j}^\top\mathbf{X}_{G_j} + N\lambda_j\mathbf{I}_{|G_j|})^{-1}\mathbf{X}_{G_j}^\top\\
=&\mathbf{I}_N + \frac{1}{N\lambda_j}\big[\mathbf{X}_{G_j}\mathbf{X}_{G_j}^\top\mathbf{X}_{G_j}(\mathbf{X}_{G_j}^\top\mathbf{X}_{G_j} + N\lambda_j\mathbf{I}_{|G_j|})^{-1}\mathbf{X}_{G_j}^\top\\
&-\mathbf{X}_{G_j}(\mathbf{X}_{G_j}^\top\mathbf{X}_{G_j} + N\lambda_j\mathbf{I}_{|G_j|})(\mathbf{X}_{G_j}^\top\mathbf{X}_{G_j} + N\lambda_j\mathbf{I}_{|G_j|})^{-1}\mathbf{X}_{G_j}^\top\big]\\
=&\mathbf{I}_N - \frac{1}{N\lambda_j}\mathbf{X}_{G_j}\mathbf{X}_{G_j}^\top\\
&+\frac{1}{N\lambda_j}\mathbf{X}_{G_j}\mathbf{X}_{G_j}^\top\mathbf{X}_{G_j}(\mathbf{X}_{G_j}^\top\mathbf{X}_{G_j} + N\lambda_j\mathbf{I}_{|G_j|})^{-1}\mathbf{X}_{G_j}^\top\\
=&\mathbf{I}_N - \frac{1}{N\lambda_j}\mathbf{X}_{G_j}\mathbf{X}_{G_j}^\top\mathbf{\Phi}_{G_j}.
\end{aligned} \tag{13}$$

The first equality is the definition of $\mathbf{\Phi}_{G_j}$. The second, third and fourth equalities are straightforward mathematical algebra. Move the second term in the most right hand side of (13) to the most left, then

$$(\mathbf{I}_{|G_j|} + \frac{1}{N\lambda_j}\mathbf{X}_{G_j}\mathbf{X}_{G_j}^\top)\mathbf{\Phi}_{G_j} = \mathbf{I}_N.$$

Therefore, the below equality can be observed,

$$\begin{aligned}
\mathbf{\Phi}_{G_j} &= N\lambda_j[N\lambda_j\mathbf{I}_N + \mathbf{X}_{G_j}\mathbf{X}_{G_j}^\top]^{-1}\\
&= N\lambda_j[N\lambda_j\mathbf{I}_N + \sum_{i=1}^{P}\delta_{ij}\mathbf{X}_i\mathbf{X}_i^\top]^{-1}.
\end{aligned} \tag{14}$$

This equality (14) is also observed in [32], [34]. Combined (14) with (12), the objective function in (10) is valid. The binary variable $\delta_{ij} = 1$ if $i \in G_j$ and 0, otherwise. The constrains in (9) ensure that each variable is selected at most once into a cluster and the cardinality assigned to each cluster. ■

## Appendix B
## Proof of Lemma 1

*Proof:* Suppose that $\{G_1^*, G_2^*, \ldots, G_K^*\}$ be the optimal solution for the CIP problem. According to the definition of $\theta_{g_j}$, the inequality $N\lambda_j\mathbf{I}_N + \sum_{i\in G_j^*}\mathbf{X}_i\mathbf{X}_i^\top \preceq (N\lambda_j + \theta_{g_j})\mathbf{I}_N$ holds, thus,

$$\begin{aligned}
v^* &= \sum_{j=1}^{K}\lambda_j\sum_{t=1}^{L}\mathbf{Y}_t^\top[N\lambda_j\mathbf{I}_N + \sum_{i=1}^{I}\delta_{ij}\mathbf{X}_i\mathbf{X}_i^\top]^{-1}\mathbf{Y}_t\\
&\geq \sum_{j=1}^{K}\frac{\lambda_j}{N\lambda_j + \theta_{g_j}}\sum_{t=1}^{L}\|\mathbf{Y}\|_F^2\\
&\geq \frac{K}{1+\beta}\frac{\|\mathbf{Y}\|_F^2}{N}.
\end{aligned} \tag{15}$$

According to Step 3 of Algorithm in Fig. 2, for any given $G_j, |G_j| < g_j$, and $\mathbf{A}_{G_j} = N\lambda_j\mathbf{I}_N + \sum_{i\in G_j}\mathbf{X}_i\mathbf{X}_i^\top$. For any $i \notin [P]\setminus G$, it can be observed that

$$\begin{aligned}
&\sum_{t=1}^{L}\mathbf{Y}_t^\top[\mathbf{A}_{G_j} + \mathbf{X}_i\mathbf{X}_i^\top]^{-1}\mathbf{Y}_t - \mathbf{Y}_t^\top\mathbf{A}_{G_j}^{-1}\mathbf{Y}_t\\
=&-\frac{\sum_{t=1}^{L}(\mathbf{Y}_t^\top\mathbf{A}_{G_j}^{-1}\mathbf{X}_i)^2}{1 + \mathbf{X}_i^\top\mathbf{A}_{G_j}^{-1}\mathbf{X}_i}.
\end{aligned} \tag{16}$$

Suppose that $\sum_{j=1}^{K}|G_j| = \tau \geq 0$, using the equality (16), the CDR value $v^\tau$ is bounded as (17),

$$v^\tau \leq \Big(K - \frac{\underline{\theta}}{(N\overline{\lambda} + \theta_1)(1+\beta)^2}\sum_{i\in[\tau]}\frac{1}{P+1-i}\Big)\frac{\|\mathbf{Y}\|_F^2}{N}. \tag{17}$$

If $\sum_{j=1}^{K}|G_j| = 0$, then (17) holds trivially. Suppose that $\sum_{j=1}^{K}|G_j| = \tau \geq 0$, the inequality (17) holds. Let $\sum_{j=1}^{k}|G_j| = \tau + 1$, the following inequality can be obtained by induction,

$$\begin{aligned}
v^{\tau+1} &= v^\tau + \min_{i\in[P]\setminus(\cup_{j=1}^{K}G_j)}\lambda_j\sum_{t=1}^{L}(\mathbf{Y}_t^\top[\mathbf{A}_{G_j} + \mathbf{X}_i\mathbf{X}_i^\top]^{-1}\mathbf{Y}_t\\
&\quad - \mathbf{Y}_t^\top\mathbf{A}_{G_j}^{-1}\mathbf{Y}_t)\\
&\leq v^\tau + \frac{\lambda_j}{P-\tau}\sum_{i\in[P]\setminus(\cup_{j=1}^{K}G_j)}\sum_{t=1}^{L} -\frac{(\mathbf{Y}_t^\top\mathbf{A}_{G_j}^{-1}\mathbf{X}_i)^2}{1 + \mathbf{X}_i^\top\mathbf{A}_{G_j}^{-1}\mathbf{X}_i}\\
&\leq v^\tau - \frac{N\lambda_j^2}{(P-\tau)(N\overline{\lambda} + \theta_1)}\times\\
&\quad \sum_{t=1}^{L}\mathbf{Y}_t^\top\mathbf{A}_{G_j}^{-1}(\mathbf{X}_{[P]\setminus(\cup_{j=1}^{k}G_j)}\mathbf{X}_{[P]\setminus(\cup_{j=1}^{k}G_j)}^\top)\mathbf{A}_{G_j}^{-1}\mathbf{Y}_t\\
&\leq v^\tau - \frac{N\lambda_j^2\underline{\theta}}{(P-\tau)(N\overline{\lambda} + \theta_1)(N\lambda_j + \theta_{g_j})^2}\|\mathbf{Y}\|_F^2\\
&\leq \frac{K}{N}\|\mathbf{Y}\|_F^2\\
&\quad - \frac{\underline{\theta}}{(N\overline{\lambda} + \theta_1)(1+\beta)^2}\sum_{i\in[\tau+1]}\frac{1}{P+1-i}\frac{\|\mathbf{Y}\|_F^2}{N},
\end{aligned}$$

where the first inequality is due to minimum value is no larger than the average of all the remaining values, the second inequality is because $\mathbf{A}_{G_j} \succeq N\lambda_j\mathbf{I}_N$, $\|\mathbf{X}_i\|_2^2 \leq \theta_1$ and the definition of $\overline{\lambda}$, and the third inequality is due to the fact that $\mathbf{A}_{G_j} \preceq (N\lambda_j + \theta_{g_j})\mathbf{I}_N$, $\underline{\theta} \leq \sigma_{\min}\big(\mathbf{X}_{[P]\setminus G}\mathbf{X}_{[P]\setminus G}^\top\big)$. The last inequality is because of the induction and definition of $\beta$. According to the inequality $\sum_{i\in[\tau]}\frac{1}{P+1-i} \geq \int_0^\tau\frac{1}{P+1-x}dx = \log(\frac{P+1}{P+1-\tau})$, then

$$v^G \leq K(1-\gamma)\frac{\|\mathbf{Y}\|_F^2}{N}. \tag{18}$$

Combine (18) with (15), the following holds,

$$v^* \leq v^G \leq (1+\beta)(1-\gamma)v^*.$$

■

There are something need to be notified about Theorem 1.

(i) The condition $P \geq N + \sum_{j=1}^{K}g_j$ is quite mild under this paper's setting, however, it cannot guarantee $\underline{\theta} > 0$. Based on the definition, $\underline{\theta} = \min_{G\in[P],|G|=P-\sum_{j=1}^{K}g_j}\sigma_{\min}(\mathbf{X}_G\mathbf{X}_G^\top)$. On the other hand, if the matrix $\mathbf{X}$ is said to satisfy $(P - \sum_{j=1}^{K}g_j)$-restricted isometry property (RIP) with restricted isometry constant $\delta_{P-\sum_{j=1}^{K}g_j} \in (0,1)$ , then we have

$$\underline{\theta} \geq 1 - \delta_{P-\sum_{j=1}^{K}g_j}$$

according to [46].

(ii) Given that $\underline{\theta} > 0$, if $\underline{\lambda}$ is larger, the objective value of CDR approach becomes closer to the true optimal value.
(iii) When $K = 1, L = 1$, our result reduces to the result in [34].

## APPENDIX C
## PROOF OF THEOREM 2

*Proof:* Consider the CDR estimator $\mathbf{W}_{G_j}$ from (8),

$$
\begin{aligned}
&\frac{1}{N}\|\mathbf{Y}-\mathbf{X}\mathbf{W}_{G_j}\|_F^2+\lambda_j\|\mathbf{W}_{G_j}\|_F^2\\
&-\frac{1}{N}\|\mathbf{Y}-\mathbf{X}\mathbf{W}_{G_j^*}\|_F^2-\lambda_j\|\mathbf{W}_{G_j^*}\|_F^2\\
&=-2\,\mathrm{Tr}\left((\mathbf{W}_{G_j}-\mathbf{W}_{G_j^*})^\top\Big[\frac{1}{N}\mathbf{X}(\mathbf{Y}-\mathbf{X}\mathbf{W}_{G_j^*})-\lambda_j\mathbf{W}_{G_j^*}\Big]\right)\\
&+\mathrm{Tr}\left((\mathbf{W}_{G_j}-\mathbf{W}_{G_j^*})^\top\Big[\frac{1}{N}\mathbf{X}^\top\mathbf{X}+\lambda_j\mathbf{I}_P\Big](\mathbf{W}_{G_j}-\mathbf{W}_{G_j^*})\right)\\
&=2\,\mathrm{Tr}\left((-\mathbf{W}_{G_j}+\mathbf{W}_{G_j^*})^\top\Big[\frac{1}{N}\tilde{\mathbf{X}}^j(\mathbf{Y}-\mathbf{X}\mathbf{W}_{G_j^*})\Big]\right)\\
&+\mathrm{Tr}\left((\mathbf{W}_{G_j}-\mathbf{W}_{G_j^*})^\top\Big[\frac{1}{N}\mathbf{X}^\top\mathbf{X}+\lambda_j\mathbf{I}_P\Big](\mathbf{W}_{G_j}-\mathbf{W}_{G_j^*})\right)\\
&\geq-\frac{2}{N}\|\mathbf{W}_{G_j}-\mathbf{W}_{G_j^*}\|_F\|\tilde{\mathbf{X}}^j(\mathbf{Y}-\mathbf{X}\mathbf{W}_{G_j^*})\|_F\\
&+\big(\lambda_j+\frac{\sigma_{\min}(\mathbf{X}_{G_j^U}^\top\mathbf{X}_{G_j^U})}{N}\big)\|\mathbf{W}_{G_j}-\mathbf{W}_{G_j^*}\|_F^2\\
&\geq-\frac{2}{N}\sqrt{\theta_{|G_j\setminus G_j^*|}}\|\mathbf{Y}-\mathbf{X}\mathbf{W}_{G_j^*}\|_F\|\mathbf{W}_{G_j}-\mathbf{W}_{G_j^*}\|_F\\
&+\big(\underline{\lambda}+\frac{\sigma_{\min}(\mathbf{X}_{G_j^U}^\top\mathbf{X}_{G_j^U})}{N}\big)\|\mathbf{W}_{G_j}-\mathbf{W}_{G_j^*}\|_F^2\\
&\geq-\sqrt{\frac{4\theta_{|G_j\setminus G_j^*|}v^*}{N}}\|\mathbf{W}_{G_j}-\mathbf{W}_{G_j^*}\|_F\\
&+\big(\underline{\lambda}+\frac{\sigma_{\min}(\mathbf{X}_{G_j^U}^\top\mathbf{X}_{G_j^U})}{N}\big)\|\mathbf{W}_{G_j}-\mathbf{W}_{G_j^*}\|_F^2,
\end{aligned}
$$

where $\tilde{\mathbf{X}}^j$ is defined as follows,

$$
\begin{cases}\tilde{\mathbf{X}}^j_{G_j\setminus G_j^*}=\mathbf{X}_{G_j\setminus G_j^*}\\ \tilde{\mathbf{X}}_i=\mathbf{0}\quad\text{if } i\notin G_j\setminus G_j^*.\end{cases}
$$

The first equality is mathematical algebra and the second one is due to the optimal condition of $\mathbf{W}_{G_j^*}$, where $-\frac{1}{N}\mathbf{X}_{G_j^*}^\top(\mathbf{Y}-\mathbf{X}_{G_j^*}\mathbf{W}_{G_j^*})+\lambda_j\mathbf{W}_{G_j^*}=\mathbf{0}$. The first inequality is due to the Cauchy inequality $\mathrm{Tr}(\mathbf{AB})\leq\|\mathbf{A}\|_F\|\mathbf{B}\|_F$, the second inequality is because of $\|\tilde{\mathbf{X}}^j(\mathbf{Y}-\mathbf{X}\mathbf{W}_{G_j^*})\|_F\leq\sqrt{\theta_{|G_j\setminus G_j^*|}}\|\mathbf{Y}-\mathbf{X}\mathbf{W}_{G_j^*}\|_F$, and the last inequality is due to $\|\mathbf{Y}-\mathbf{X}\mathbf{W}_{G_j^*}\|_F\leq\sqrt{Nv^*}$.

Take the summation over the above inequality, then

$$
\begin{aligned}
\alpha v^*&\geq\sum_{j=1}^K\frac{1}{N}\|\mathbf{Y}-\mathbf{X}\mathbf{W}_{G_j}\|_F^2+\lambda_j\|\mathbf{W}_{G_j}\|_F^2\\
&-\frac{1}{N}\|\mathbf{Y}-\mathbf{X}\mathbf{W}_{G_j^*}\|_F^2-\lambda_j\|\mathbf{W}_{G_j^*}\|_F^2\\
&\geq-\sqrt{\frac{4\theta_{\max}v^*}{N}}\sum_{j=1}^K\|\mathbf{W}_{G_j}-\mathbf{W}_{G_j^*}\|_F\\
&+\big(\underline{\lambda}+\frac{\sigma_{\min}}{N}\big)\sum_{j=1}^K\|\mathbf{W}_{G_j}-\mathbf{W}_{G_j^*}\|_F^2\\
&\geq-\sqrt{\frac{4\theta_{\max}v^*}{N}}\sum_{j=1}^K\|\mathbf{W}_{G_j}-\mathbf{W}_{G_j^*}\|_F\\
&+\frac{N\underline{\lambda}+\sigma_{\min}}{NK}\big(\sum_{j=1}^K\|\mathbf{W}_{G_j}-\mathbf{W}_{G_j^*}\|_F\big)^2.
\end{aligned}
$$

The first inequality is the result from Theorem 1, the second inequality is due to the definition of $\theta_{\max}$ and $\sigma_{\min}$, and the last inequality is because of Cauchy inequality $(\sum_{j=1}^K\|\mathbf{W}_{G_j}-\mathbf{W}_{G_j^*}\|_F)^2\leq K\sum_{j=1}^K\|\mathbf{W}_{G_j}-\mathbf{W}_{G_j^*}\|_F^2$. Accordingly,

$$
\sum_{j=1}^K\|\mathbf{W}_{G_j}-\mathbf{W}_{G_j^*}\|_F\leq\frac{2K\sqrt{N\theta_{\max}v^*}}{N\underline{\lambda}+\sigma_{\min}}+\sqrt{\frac{\alpha KNv^*}{N\underline{\lambda}+\sigma_{\min}}}.
$$

The above inequality is because any solution of the following quadratic inequality $ax^2-bx-c\leq0$ with $a,b,c>0$ is upper bounded by $\frac{b}{a}+\sqrt{\frac{c}{a}}$. ∎


## ACKNOWLEDGMENT

Research reported in this publication was partially supported by the Office of Naval Research under Award Number N00014-18-1-2794.


## REFERENCES


[1] R. M. Haralick, K. Shanmugam, and I. H. Dinstein, "Textural features for image classification," *IEEE Transactions on systems, man, and cybernetics*, vol. 44, no. 6, pp. 610–621, 1973.
[2] C. C. Aggarwal and C. Zhai, *Mining text data*. Springer Science & Business Media, 2012.
[3] C. Liu, A. C. C. Law, D. Roberson, and Z. J. Kong, "Image analysis-based closed loop quality control for additive manufacturing with fused filament fabrication," *Journal of Manufacturing Systems*, vol. 51, pp. 75–86, 2019.
[4] S. Chowdhury and S. Anand, "Artificial neural network based geometric compensation for thermal deformation in additive manufacturing processes," in *ASME 2016 11th International Manufacturing Science and Engineering Conference*. American Society of Mechanical Engineers Digital Collection, 2016.
[5] L. Yu and H. Liu, "Feature selection for high-dimensional data: A fast correlation-based filter solution," in *Proceedings of the 20th international conference on machine learning (ICML-03)*, 2003, pp. 856–863.
[6] Z. M. Hira and D. F. Gillies, "A review of feature selection and feature extraction methods applied on microarray data," *Advances in bioinformatics*, vol. 2015, 2015.
[7] P. K. Rao, J. P. Liu, D. Roberson, Z. J. Kong, and C. Williams, "Online real-time quality monitoring in additive manufacturing processes using heterogeneous sensors," *Journal of Manufacturing Science and Engineering*, vol. 137, no. 6, 2015.
[8] J. Ye, "Least squares linear discriminant analysis," in *Proceedings of the 24th international conference on Machine learning*. ACM, 2007, pp. 1087–1093.
[9] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*. Springer series in statistics New York, 2001, vol. 1.
[10] Y. Haitovsky, "On multivariate ridge regression," *Biometrika*, vol. 74, no. 3, pp. 563–570, 1987.
[11] Z. Zhang, G. Dai, C. Xu, and M. I. Jordan, "Regularized discriminant analysis, ridge regression and beyond," *Journal of Machine Learning Research*, vol. 11, no. Aug, pp. 2199–2228, 2010.
[12] P. Bühlmann, P. Rütimann, S. van de Geer, and C.-H. Zhang, "Correlated variables in regression: clustering and sparse estimation," *Journal of Statistical Planning and Inference*, vol. 143, no. 11, pp. 1835–1858, 2013.
[13] L. Sun, S. Ji, and J. Ye, "A least squares formulation for a class of generalized eigenvalue problems in machine learning," in *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM, 2009, pp. 977–984.

[14] F. Reverter, E. Vegas, and J. M. Oller, "Kernel-pca data integration with enhanced interpretability," *BMC systems biology*, vol. 8, no. S2, p. S6, 2014.

[15] S. Chao and C. Lihui, "Feature dimension reduction for microarray data analysis using locally linear embedding," in *Proceedings Of The 3rd Asia-Pacific Bioinformatics Conference*. World Scientific, 2005, pp. 211–217.

[16] K. Dawson, R. L. Rodriguez, and W. Malyj, "Sample phenotype clusters in high-density oligonucleotide microarray data sets are revealed using isomap, a nonlinear algorithm," *Bmc Bioinformatics*, vol. 6, no. 1, p. 195, 2005.

[17] S. Mika, G. Ratsch, J. Weston, B. Scholkopf, and K.-R. Mullers, "Fisher discriminant analysis with kernels," in *Neural networks for signal processing IX: Proceedings of the 1999 IEEE signal processing society workshop (cat. no. 98th8468)*. Ieee, 1999, pp. 41–48.

[18] J. H. Friedman, "Regularized discriminant analysis," *Journal of the American statistical association*, vol. 84, no. 405, pp. 165–175, 1989.

[19] M. R. Segal, K. D. Dahlquist, and B. R. Conklin, "Regression approaches for microarray data analysis," *Journal of Computational Biology*, vol. 10, no. 6, pp. 961–980, 2003.

[20] D. J. Balding, "A tutorial on statistical methods for population association studies," *Nature reviews genetics*, vol. 7, no. 10, p. 781, 2006.

[21] M. Dettling and P. Bühlmann, "Finding predictive gene groups from microarray data," *Journal of Multivariate Analysis*, vol. 90, no. 1, pp. 106–131, 2004.

[22] D. M. Witten, A. Shojaie, and F. Zhang, "The cluster elastic net for high-dimensional regression with unknown variable grouping," *Technometrics*, vol. 56, no. 1, pp. 112–122, 2014.

[23] C. Li and H. Li, "Network-constrained regularization and variable selection for analysis of genomic data," *Bioinformatics*, vol. 24, no. 9, pp. 1175–1182, 2008.

[24] ——, "Variable selection and regression analysis for graph-structured covariates with an application to genomics," *The annals of applied statistics*, vol. 4, no. 3, p. 1498, 2010.

[25] J. Huang, S. Ma, H. Li, and C.-H. Zhang, "The sparse laplacian shrinkage estimator for high-dimensional regression," *Annals of statistics*, vol. 39, no. 4, p. 2021, 2011.

[26] F. Xue and A. Qu, "Variable selection for highly correlated predictors," *arXiv preprint arXiv:1709.04840*, 2017.

[27] H. D. Bondell and B. J. Reich, "Simultaneous regression shrinkage, variable selection, and supervised clustering of predictors with oscar," *Biometrics*, vol. 64, no. 1, pp. 115–123, 2008.

[28] D. B. Sharma, H. D. Bondell, and H. H. Zhang, "Consistent group identification and variable selection in regression with correlated predictors," *Journal of Computational and Graphical Statistics*, vol. 22, no. 2, pp. 319–340, 2013.

[29] A. Alkenani and T. R. Dikheel, "Robust group identification and variable selection in regression," *Journal of Probability and Statistics*, vol. 2017, 2017.

[30] J. Yang, J.-y. Yang, D. Zhang, and J.-f. Lu, "Feature fusion: parallel strategy vs. serial strategy," *Pattern recognition*, vol. 36, no. 6, pp. 1369–1381, 2003.

[31] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 1996.

[32] D. Bertsimas, B. Van Parys *et al.*, "Sparse high-dimensional regression: Exact scalable algorithms and phase transitions," *The Annals of Statistics*, vol. 48, no. 1, pp. 300–323, 2020.

[33] A. Das and D. Kempe, "Approximate submodularity and its applications: subset selection, sparse approximation and dictionary selection," *The Journal of Machine Learning Research*, vol. 19, no. 1, pp. 74–107, 2018.

[34] W. Xie and X. Deng, "The ccp selector: Scalable algorithms for sparse ridge regression from chance-constrained programming," *arXiv preprint arXiv:1806.03756*, 2018.

[35] J. Sherman and W. J. Morrison, "Adjustment of an inverse matrix corresponding to a change in one element of a given matrix," *The Annals of Mathematical Statistics*, vol. 21, no. 1, pp. 124–127, 1950.

[36] I. Guyon, S. Gunn, A. Ben-Hur, and G. Dror, "Result analysis of the nips 2003 feature selection challenge," in *Advances in neural information processing systems*, 2005, pp. 545–552.

[37] J. M. Keller, M. R. Gray, and J. A. Givens, "A fuzzy k-nearest neighbor algorithm," *IEEE transactions on systems, man, and cybernetics*, vol. 10, no. 4, pp. 580–585, 1985.

[38] A. McCallum, K. Nigam *et al.*, "A comparison of event models for naive bayes text classification," in *AAAI-98 workshop on learning for text categorization*, vol. 752. Citeseer, 1998, pp. 41–48.

[39] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, S. Y. Philip *et al.*, "Top 10 algorithms in data mining," *Knowledge and information systems*, vol. 14, no. 1, pp. 1–37, 2008.

[40] M. T. Hagan, H. B. Demuth, and M. Beale, *Neural network design*. PWS Publishing Co., 1997.

[41] S. Lawrence, C. L. Giles, A. C. Tsoi, and A. D. Back, "Face recognition: A convolutional neural-network approach," *IEEE transactions on neural networks*, vol. 8, no. 1, pp. 98–113, 1997.

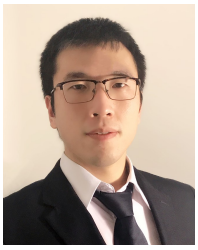
[42] X. Qi, G. Chen, Y. Li, X. Cheng, and C. Li, "Applying neural-network-based machine learning to additive manufacturing: current applications, challenges, and future perspectives," *Engineering*, vol. 5, no. 4, pp. 721–729, 2019.

[43] H. Elwarfalli, D. Papazoglou, D. Erdahl, A. Doll, and J. Speltz, "In situ process monitoring for laser-powder bed fusion using convolutional neural networks and infrared tomography," in *2019 IEEE National Aerospace and Electronics Conference (NAECON)*. IEEE, 2019, pp. 323–327.

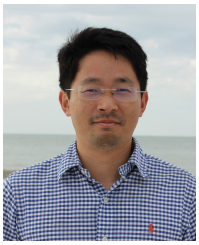
[44] Y. Banadaki, N. Razaviarab, H. Fekrmandi, and S. Sharifi, "Toward enabling a reliable quality monitoring system for additive manufacturing process using deep convolutional neural networks," *arXiv preprint arXiv:2003.08749*, 2020.

[45] M. M. Kirka, Y. Lee, D. A. Greeley, A. Okello, M. J. Goin, M. T. Pearce, and R. R. Dehoff, "Strategy for texture management in metals additive manufacturing," *Jom*, vol. 69, no. 3, pp. 523–531, 2017.

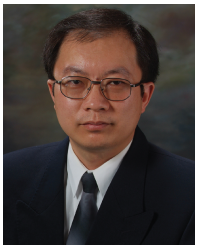
[46] E. J. Candes and T. Tao, "Decoding by linear programming," *IEEE transactions on information theory*, vol. 51, no. 12, pp. 4203–4215, 2005.

**Bo Shen** (S'19) received his B.S. degree in Statistics from University of Science and Technology of China, Hefei, China, in July 2017. He is currently working on his Ph.D. in Industrial and Systems Engineering at Virginia tech, Blacksburg, VA. His research interests include optimization in machine learning, data analytics in smart manufacturing.



**Weijun Xie** is Assistant Professor in the Department of Industrial and Systems Engineering at Virginia Tech. Dr. Xie obtained his Ph.D. in Operations Research at Georgia Institute of Technology in August 2017. His research interests are in theory and applications of stochastic, discrete, and convex optimization. His papers have received multiple awards, including Honorable Mention in George Nicholson Student Paper Competition at INFORMS 2017, Third Place in Junior Faculty Interest Group Paper Competition at INFORMS 2018, Runner-up of Dupacova-Prekopa Best Student Paper Prize in Stochastic Programming at ICSP 2019.



**Zhenyu (James) Kong** (M'02) received his B.S. and M.S. degrees in mechanical engineering from Harbin Institute of Technology, Harbin, China, in 1993 and 1995, respectively, and his Ph.D. degree from the Department of Industrial and System Engi-neering, University of Wisconsin–Madison, Madison, WI, USA, in 2004. He is currently a professor with the Grado Department of Industrial and Systems Engineering, Virginia Tech, Blacksburg, VA, USA. His research interests include sensing and analytics for smart manufacturing, and modeling, synthesis, and diagnosis for large and complex manufacturing systems.