SCIENTIFIC REPORTS

OPEN

Received: 27 July 2016
Accepted: 15 November 2016
Published: 14 December 2016

A Novel Information-Theoretic Approach for Variable Clustering and Predictive Modeling Using Dirichlet Process Mixtures

Yun Chen[1] & Hui Yang[2]

In the era of big data, there are increasing interests on clustering variables for the minimization of data redundancy and the maximization of variable relevancy. Existing clustering methods, however, depend on nontrivial assumptions about the data structure. Note that nonlinear interdependence among variables poses significant challenges on the traditional framework of predictive modeling. In the present work, we reformulate the problem of variable clustering from an information theoretic perspective that does not require the assumption of data structure for the identification of nonlinear interdependence among variables. Specifically, we propose the use of mutual information to characterize and measure nonlinear correlation structures among variables. Further, we develop Dirichlet process (DP) models to cluster variables based on the mutual-information measures among variables. Finally, orthonormalized variables in each cluster are integrated with group elastic-net model to improve the performance of predictive modeling. Both simulation and real-world case studies showed that the proposed methodology not only effectively reveals the nonlinear interdependence structures among variables but also outperforms traditional variable clustering algorithms such as hierarchical clustering.

Predictive modeling extracts useful information and patterns from the data to drive decisions or actions. For example, insurance companies have gathered a vast amount of data in their data warehouses[1]. The objective of the predictive model is not only to improve the pricing or marketing process, but also to analyze profitability, fraud, catastrophe, and other insurance operations. In the 21st century, wireless sensing, electronic health records, and health Internet of Things are increasingly adopted to assist in the process of clinical decision making[2–4]. This amount of information from multiple sources provides numerous variables for the contemplated predictive model.

When a predictive model involves large amounts of variables (i.e., explanatory or response variables), researchers are confronted with the need to reduce the number of variables in order to build the compact model. To some extent, the variables are unknown to be redundant or relevant to the objective of predictive models but rather need to be tested with real-world data. In addition, when there is an enormous amount of variables, it becomes difficult to find out the relationship between variables. If the model building involves too many variables, it will impact the model compactness and efficiency. There is also a possibility to increase the model sensitivity to noises and overfit the data with many variables. The model parameters are not stable when variables are highly correlated. It is even more difficult to explain the physical meanings of the predictive model when there are many variables. Finally, model building with a large amount of variables is computationally expensive and could take indefinite time for the exhaustive search. An intermediate approach to the exhaustive search may also be time-consuming and some combinations of variables could be overseen.

Data clustering is an unsupervised method to group data samples into homogeneous clusters, while variable clustering is to detect subsets of homogeneous variables and then cluster them into the same group, in which variables have stronger interrelations to each other than to those in other groups. As shown in Fig. 1a, data clustering groups data samples into clusters and each sample has two values, e.g., (0.26, −0.09) in the 2-dimensional

[1]School of Mechanical Engineering, Jiangsu University of Science and Technology, Zhenjiang, China. [2]Complex Systems Monitoring, Modeling and Control Laboratory, The Pennsylvania State University, University Park, PA, USA. Correspondence and requests for materials should be addressed to H.Y. (email: huy25@psu.edu)

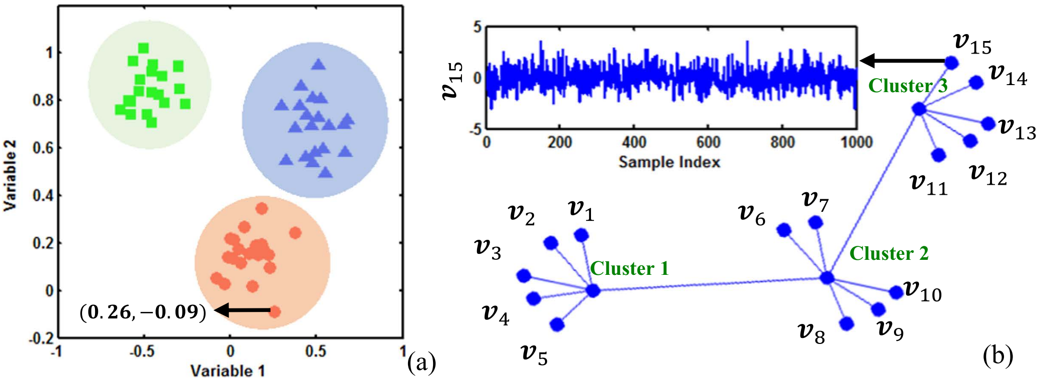
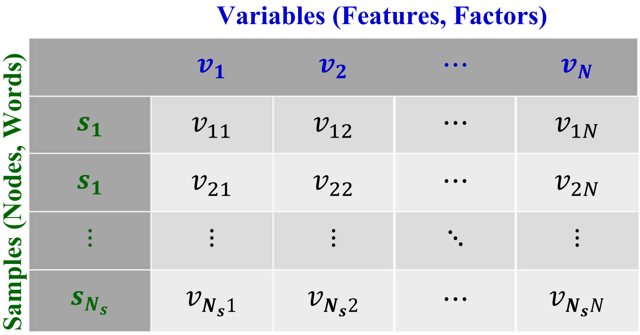


Figure 1. (**a**) Data clustering with each point representing a data sample; (**b**) Variable clustering with each point representing a variable.



Variables (Features, Factors)

Samples (Nodes, Words)

	v_1	v_2	$\cdots$	v_N
s_1	v_{11}	v_{12}	$\cdots$	v_{1N}
s_1	v_{21}	v_{22}	$\cdots$	v_{2N}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
s_{N_s}	v_{N_s1}	v_{N_s2}	$\cdots$	v_{N_sN}

Figure 2. Data in the table form, where variables are in columns and samples are in rows.

space, where X-axis is the value of variable 1, and Y-axis is the value of variable 2. Data samples are clustered based on the similarity measure, e.g., Euclidean distance. However, variable clustering is different from data clustering. Figure 1b illustrates the clustering results of 15 variables, each of which has 1000 data samples. For example, the variable $\boldsymbol{v}_{15}$ represents a series of 1000 samples. Notably, each point in Fig. 1b is a variable instead of a data sample.

Variable clustering uncovers natural groups of objects (variables, features, or factors) in a multivariate dataset. The hierarchical clustering (HC)[5], a generic clustering procedure, sequentially merges pairs of clusters that share common characteristics based on similarity measures. HC procedures generate a nested set of partitions, also called hierarchy. The choice of the similarity measure plays an important role in the clustering process because it indirectly defines the structure of the clusters. This choice is not only guided by problems to solve, but also restricted to commonly used measures, such as the Euclidean distance or Pearson's correlation coefficient. However, nonlinear interdependence among variables cannot be adequately captured by linear correlation. Further, we cannot relocate the variables once the merge is done for two closest clusters, because HC is not a dynamic approach. There is no adaptive step for two variables to make modifications in the later stage if they are 'incorrectly' clustered at the early stage.

In this paper, we develop a new methodology of information theoretic approach for variable clustering and predictive modeling. The proposed approach investigates both redundancy and relevancy among variables. Specifically, nonlinear interdependence structures are measured among variables. Further, we introduced nonparametric Dirichlet process to cluster embedded variables with their probability distributions. Finally, orthonormalized variables were integrated with group elastic net models to improve the performances of predictive models. Both simulation and real-world case studies demonstrate that the proposed methodology not only outperforms traditional variable clustering algorithms such as hierarchical clustering, but also effectively identifies nonlinear interdependence structures among variables and further improves the performance of predictive modeling.

Research Background

Clustering Analysis.

When "clustering" is used in the literature, it is referred to be "data clustering" for most of the time. The approach of data clustering groups data samples into homogeneous subsets, in which data samples are closer to each other in the same cluster than to other clusters. As shown in Fig. 2, data clustering is more concerned about the samples that are rows (i.e., $\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_{N_s}$) in the table format of a dataset but variable clustering focus on the variables in the columns (i.e., $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_N$). The variables, $\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots, \boldsymbol{v}_N$, are also known as features or factors, where $\boldsymbol{v}_i = (v_{1i}, v_{2i}, \ldots, v_{N_s i})^T, i = 1, 2, \ldots, N$, N is the number of variables and N_s is the number of samples respectively. The samples, $\boldsymbol{s}_1, \boldsymbol{s}_2, \ldots, \boldsymbol{s}_{N_s}$, are also called nodes in the network or words in the text, where $\boldsymbol{s}_j = (v_{j1}, v_{j2}, \ldots, v_{jN})^T, j = 1, 2, \ldots, N_s$. It may be noted that big data often brings a large number of variables that can be bigger than the number of samples, i.e., $N > N_s$. Complex interdependence structures among variables significantly challenge the traditional framework of predictive modeling. As such, variable clustering to delineate homogeneous groups of variables is urgently needed.

In recent years, community detection in network analysis receives increasing interests in data clustering. Network-based methods cluster nodes with strong connections into a community. For example, mixed membership stochastic blockmodels (MMSB)[6] were proposed to discover complex network structure in a variety of applications, e.g., large-scale protein interaction network and social network. The MMSB develops a novel class of latent variable models for relational data, and assumes each variable belongs to multiple communities/clusters rather than a single community/cluster. Joint Gamma process Poisson factorization (J-GPPF)[7] was developed to jointly model sparse networks with large size and side information. Infinite edge partition model[8] was introduced to not only study overlapping communities and inter-community interactions but also predict missing edges. However, community detection groups nodes that represent data samples (e.g., proteins), rather than variables, into communities by considering the unweighted or weighted edges between nodes.

In addition, topic models are widely used for data clustering in the field of text mining. Topic models are statistical models for discovering topics that occur in a collection of documents with a large number of words (i.e., data samples in rows of table-form data in Fig. 2). Latent Dirichlet allocation (LDA)[9] was first introduced as an unsupervised model to cluster documents in the topic space. LDA assumes the topic distribution to have a Dirichlet prior and maximizes the likelihood (or posterior probability) of the document collection. It may also be noted that supervised topic models with side information (e.g., document categories or review rating scores) were proposed to find latent topics and provide more predictive power than regression on unsupervised LDA features. For example, supervised latent Dirichlet allocation (sLDA)[10] introduced the real-valued document rating as regression response and jointly modeled the documents and response by maximizing the joint likelihood. Maximum entropy discrimination LDA (MedLDA)[11,12] proposed a unified constrained optimization framework that solves problems of dimensionality reduction and max-margin classification using features in the reduced-dimension space. Topic models formulate statistical models based on the intuition that specific words would appear more or less frequently in the document for a given topic. However, variable clustering does not hold this intuition. As such, topic models address special clustering problems in text mining that are different from other general data clustering or variable clustering problems.

Moreover, many previous approaches group a dataset into co-clusters (or biclusters), which are subsets of data samples exhibit similar behaviors across a subset of variables, or vice versa. Co-clustering approaches have been widely used in a variety of applications such as biological gene expression data[13] and text mining[14,15]. Notably, a simultaneous co-clustering and learning (SCOAL)[16] framework was proposed to generalize co-clustering and construct predictive models simultaneously. The SCOAL co-cluster the entire dataset into subsets of samples and variables such that each subset can be well characterized by a predictive model. However, the whole data set is divided into multiple subsets that capture incomplete data information. These subsets are then used to construct multiple predictive models rather than one model. In addition, nonlinear correlations among variables were not fully utilized in traditional co-clustering approaches. Instead, nonlinear predictive models were usually introduced to account for data nonlinearity, which also brings a large number of parameters.

Hierarchical Clustering. Variable clustering is the task to group homogeneous variables into the same category, in which variables have stronger interrelations than to those in other groups. Variable clustering considers the interdependence structure among variables, e.g., correlation. The Pearson's correlation[17] between variables $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$ is

$$\rho_{\boldsymbol{v}_1\boldsymbol{v}_2} = \frac{\mathrm{cov}(\boldsymbol{v}_1, \boldsymbol{v}_2)}{\sigma_{\boldsymbol{v}_1}\sigma_{\boldsymbol{v}_2}} = \frac{E[(\boldsymbol{v}_1 - \mu_{\boldsymbol{v}_1})(\boldsymbol{v}_2 - \mu_{\boldsymbol{v}_2})]}{\sigma_{\boldsymbol{v}_1}\sigma_{\boldsymbol{v}_2}} \tag{1}$$

where $\mathrm{cov}(\boldsymbol{v}_1, \boldsymbol{v}_2)$ is the covariance between $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$, $\sigma_{\boldsymbol{v}_1}$ and $\sigma_{\boldsymbol{v}_2}$ are variances of $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$, $\mu_{\boldsymbol{v}_1}$ and $\mu_{\boldsymbol{v}_2}$ are means of $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$, E is the expectation. However, the Pearson's correlation only measures the linear relationship between variables $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$.

In the literature, Pearson's correlation was integrated with hierarchical clustering (HC) for variable clustering[5]. There are two ways to perform HC procedures - the agglomerative way and the divisive way. For example, agglomerative HC defines each variable as a singleton cluster in the first step. Then, two closest clusters with smallest dissimilarity measure are merged into one cluster. The merging process recursively moves up along the hierarchy until the stopping criterion is satisfied, e.g., the maximum number of clusters or the maximum group-average (GA) dissimilarity. The criterion of group average measures the average intergroup dissimilarity between two clusters, i.e.,

$$D_{GA}(C_m, C_n) = \frac{1}{N_{C_m}N_{C_n}}\sum_{\boldsymbol{v}_i\in C_m}\sum_{\boldsymbol{v}_j\in C_n} D_{\boldsymbol{v}_i\boldsymbol{v}_j} \tag{2}$$

where N_{C_m} and N_{C_n} are the sizes of cluster C_m and C_n, $D_{\boldsymbol{v}_i\boldsymbol{v}_j}$ is the dissimilarity between variables $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$, which is usually calculated as $1 - \rho_{\boldsymbol{v}_i\boldsymbol{v}_j}$.

Here, a motivating example is shown to evaluate the performance of HC with Pearson's correlation for variable clustering. Two clusters of variables are generated as follows:

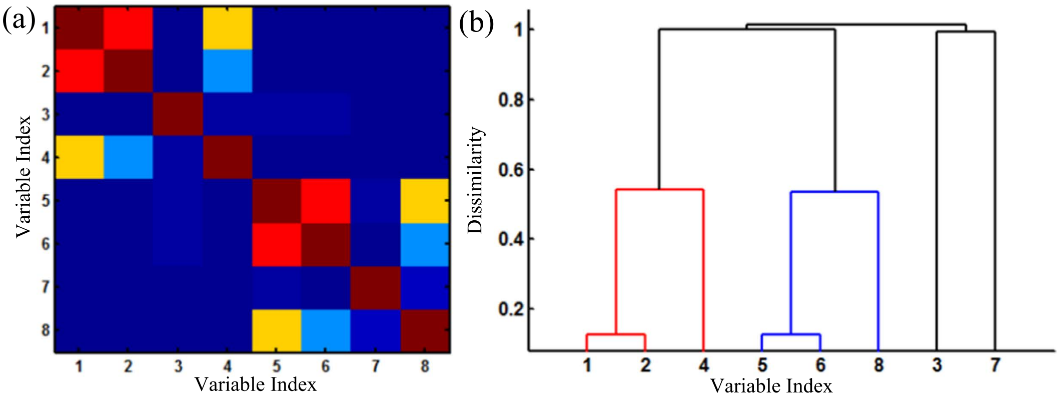


Figure 3. Illustration of Pearson's correlation (**a**) and hierarchical clustering (**b**).

$$\textbf{Cluster 1}: \{v_1, v_2 = \sin(v_1), v_3 = |v_1|, v_4 = {v_1}^4\};$$

$$\textbf{Cluster 2}: \{\boldsymbol{v}_5, \boldsymbol{v}_6 = \sin(\boldsymbol{v}_5), \boldsymbol{v}_7 = |\boldsymbol{v}_5|, \boldsymbol{v}_8 = {\boldsymbol{v}_5}^4\}.$$

where $\boldsymbol{v}_1$ and $\boldsymbol{v}_5$ are independent standard normal variables. In the cluster 1, variable $\boldsymbol{v}_1$ has linear correlation with variable $\boldsymbol{v}_2$ and nonlinear correlation with variables $\boldsymbol{v}_3$ and $\boldsymbol{v}_4$. The cluster 2 has the similar situation. Figure 3a shows the correlation matrix of these eight variables. The red color represents a high correlation, while the blue color indicates no interrelationships. It may be noted that the correlation matrix effectively detects the linear correlation between variables $\boldsymbol{v}_1$ and $\boldsymbol{v}_2$, $\boldsymbol{v}_5$ and $\boldsymbol{v}_6$. However, nonlinear correlations are not well captured. Figure 3b shows the hierarchical clustering results based on the Pearson's correlation. Variables $\boldsymbol{v}_1$, $\boldsymbol{v}_2$ and $\boldsymbol{v}_4$ are clustered in the same cluster, and variables $\boldsymbol{v}_5$, $\boldsymbol{v}_6$ and $\boldsymbol{v}_8$ are clustered in another cluster. However, hierarchical clustering failed to cluster variable $\boldsymbol{v}_3$ into Cluster 1, and variable $\boldsymbol{v}_7$ into Cluster 2. This is mainly due to the fact that nonlinear correlations among variables are not fully considered. Very little work has been done to cluster a large number of variables with complex structures of nonlinear interdependences. Thus, we propose a new methodology that integrates information theoretic approach with Dirichlet process mixtures for variable clustering and predictive modeling.

Research Methodology.

In this section, we will first characterize nonlinear correlation (i.e., mutual information) among variables and then embed variables in the lower-dimensional space. Second, we introduce the nonparametric Dirichlet process (DP) to derive self-organizing clusters of homogeneous variables with specific consideration of nonlinear interdependence. Finally, we orthonormalize variables in each cluster and then integrate them with group elastic-net model to improve the performance of predictive modeling.

Mutual Information based Embedding of Variables.

First, mutual information is characterized and quantified among variables. Traditionally, such interrelationships are estimated with linear methods such as Pearson's correlation. As aforementioned, Pearson's correlation, a second-order quantity, evaluates merely linear dependency among data and is limited in the ability to represent the variable-to-variable dissimilarities. Therefore, we propose to characterize the variable-to-variable dissimilarity matrix using mutual information and further embed variables into low-dimensional feature vectors that preserve the dissimilarity distances among variables.

Mutual information[18] quantifies both linear and nonlinear interdependence between two variables $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$, i.e., *i*th and *j*th columns in Fig. 2. Although there are various measures that capture nonlinear correlations among variables, mutual information has the advantage to equitably quantify statistical associations between two variables that is insensitive to the form of the underlying function[19], where equitability means that the statistic gives similar scores to equally noisy relationships of different types[20]. In other words, mutual information has an attractive feature to provide an equitable measure of association between two variables that is insensitive to the form of the underlying function[19]. It may be noted that mutual information was introduced to cluster nonlinear structures among data samples (e.g., feature vectors of a gene, a company and a movie) by formulating a tradeoff function among average similarity and information carried by the cluster identities[21]. However, this information-theoretic approach considers nonlinear correlation structures among data samples, rather than variables, by introducing mutual information as a similarity measure. Moreover, the number of clusters was pre-defined in order to solve the tradeoff function.

The mutual information is defined as:

$$MI(\boldsymbol{v}_i, \boldsymbol{v}_j) = \sum_{v_{ik}\in\boldsymbol{v}_i}\sum_{v_{jl}\in\boldsymbol{v}_j} p(v_{ik}, v_{jl})\log\left(\frac{p(v_{ik}, v_{jl})}{p(v_{ik})p(v_{jl})}\right) \quad (3)$$

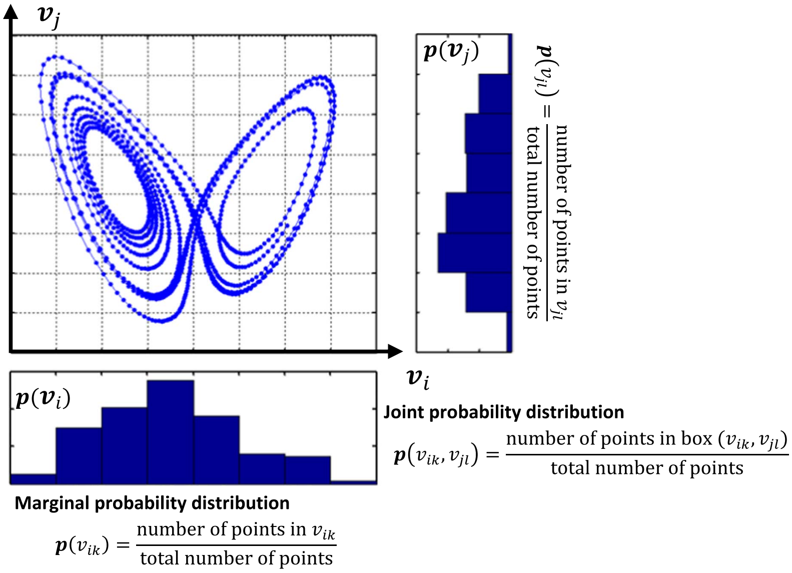

$\boldsymbol{v}_j$
$\boldsymbol{p}(\boldsymbol{v}_j)$
$\boldsymbol{p}(v_{jl}) = \frac{\text{number of points in } v_{jl}}{\text{total number of points}}$
$\boldsymbol{v}_i$
$\boldsymbol{p}(\boldsymbol{v}_i)$
Joint probability distribution
$\boldsymbol{p}(v_{ik}, v_{jl}) = \frac{\text{number of points in box } (v_{ik}, v_{jl})}{\text{total number of points}}$
Marginal probability distribution
$\boldsymbol{p}(v_{ik}) = \frac{\text{number of points in } v_{ik}}{\text{total number of points}}$

Figure 4. An illustration for the computation of mutual information.

where $p(v_{ik}, v_{jl})$ is the joint probability distribution, $p(v_{ik})$ and $p(v_{jl})$ are marginal probabilities. Figure 4 shows the practical implementation to compute the mutual information with the scatter plot of two variables $\boldsymbol{v}_i$ and $\boldsymbol{v}_j$, and the marginal histogram for each variable. Marginal probabilities $p(v_{ik})$ and $p(v_{jl})$ are computed as the number of points in v_{ik} and v_{jl} divided by the total number of points in the 2-dimensional space. While the joint probability $p(v_{ik}, v_{jl})$ is computed as the number of points in box (v_{ik}, v_{jl}) divided by the total number of points in the space. In practice, large box size will lead to an accurate estimation of average probability, but a flat estimation of joint probability $p(v_{ik}, v_{jl})$. As such, this will underestimate the mutual information $MI(\boldsymbol{v}_i, \boldsymbol{v}_j)$. In contrast, small box size estimates the joint probability $p(v_{ik}, v_{jl})$ in small scales but brings significant variations, which overestimate the mutual information $MI(\boldsymbol{v}_i, \boldsymbol{v}_j)$. In the present investigation, we choose the number of bins as $\sqrt{N_S}/2$[21], where N_S is the sample size.

Once the mutual information is computed for each pair of variables, the dissimilarity matrix among variables will be generated. It may be noted that the mutual information is inversely proportional to the dissimilarity. Therefore, we define $\delta_{ij} = 1/MI(\boldsymbol{v}_i, \boldsymbol{v}_j)$ as the dissimilarity measure between ith and jth variables in $N \times N$ dissimilarity matrix Δ. Further, an embedding algorithm is developed to transform the dissimilarity matrix into low-dimensional feature vectors that preserve the variable-to-variable dissimilarity matrix. Let $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$ denote the ith and jth feature vectors. The objective function is formulated as:

$$\min \sum_{i<j} (\|\boldsymbol{y}_i - \boldsymbol{y}_j\| - \delta_{ij});\ i, j \in [1, \boldsymbol{N}] \tag{4}$$

where $\|\cdot\|$ is the Euclidean norm. The Gram matrix B is firstly reconstructed from the dissimilarity matrix Δ in order to solve this optimization problem:

$$B = -\frac{1}{2} H \Delta^{(2)} H \tag{5}$$

where $H = I - N^{-1}11^T$ is the centering matrix, I is the identity matrix with size N and 1 is a column vector with N ones. The $\Delta^{(2)}$ is a squared matrix and each element in $\Delta^{(2)}$ is δ_{ij}^2. Then the element b_{ij} in matrix B is:

$$b_{ij} = -\frac{1}{2}\left[\delta_{ij}^2 - \frac{1}{N}\sum_{k=1}^{N}\delta_{ik}^2 - \frac{1}{N}\sum_{k=1}^{N}\delta_{kj}^2 + \frac{1}{N^2}\sum_{g=1}^{N}\sum_{h=1}^{N}\delta_{gh}^2\right] \tag{6}$$

Due to the property of Gram matrix, it is defined as the scalar product $B = YY^T$, where the matrix Y minimizes the aforementioned objective function. It is known that Gram matrix B is decomposed as:

$$B = V\Lambda V^T = V\sqrt{\Lambda}\sqrt{\Lambda}V^T \tag{7}$$

where V is a matrix of eigenvectors and Λ is a diagonal matrix of eigenvalues. Then, the matrix of feature vectors is obtained as: $Y = [\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_N] = V\sqrt{\Lambda}$. As such, each variable is embedded as a feature vector in the low-dimensional network that preserves the dissimilarity matrix.

Dirichlet Process for Variable Clustering. Furthermore, we propose to cluster low-dimensional feature vectors that are embedded from variables. Although K-Means clustering is the most popular algorithm for data

```
Initialization:
    $\theta_1 \sim G_0$    //$1^{st}$ table has value drawn from base probability measure
    $K = 1$    //total number of occupied tables so far
    $n = 1$    //total number of customers arrived so far
    $c_1 = 1$    //$1^{st}$ customer sits at the $1^{st}$ table
    $m_1 = 1$    //number of customers seated at 1st table
While  $n \le N$
    $c_n = \begin{cases} k & \text{with probability } n_k/(n-1+\alpha)\text{, for } k = 1,2,..,K \\ K+1 & \text{with probability } \alpha/(n-1+\alpha)\text{, for a new table} \end{cases}$
    If  $c_n = k$    //nth customer sits at the kth table
        $m_k = m_k + 1$
    Else $c_n = K + 1$    //nth customer sits at first unoccupied table
        $\theta_{K+1} \sim G_0$
        $m_{K+1} = 1$
        $K = K + 1$
    End
End
```

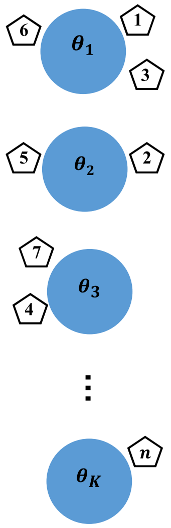


Figure 5. Algorithm and illustration of Chinese Restaurant Process.

clustering[22], it has several drawbacks. First, it is a parametric model and the number of clusters needs to be predefined. For clusters that are not well separated, this may not be straightforward. Second, K-Means algorithm needs to recalculate the objective function for assigning a cluster label to a new variable. Third, the results of K-Means clustering are not unique due to the recalculation of objective function. Therefore, we introduced the nonparametric Dirichlet process (DP) models to cluster variables[23,24]. DP models partition the vector space into local clusters, and assign cluster labels for new observations according to the assignment probability derived from the mean and covariance of each cluster, with each one following a multivariate Gaussian distribution.

The Chinese Restaurant Process (CRP) is an effective representation of DP, which visualizes the clustering effects more explicitly. Figure 5 shows the algorithm and illustration of CRP. Suppose a restaurant has potentially infinite many tables $k = 1, 2, \ldots$, and each table has value θ_k drawn from base probability measure G_0. Customers are indexed by $n = 1, 2, \ldots, N$ as they arrive, while indicator variables $c_n = k$ denotes that the nth customer choose to sit at the kth table. The tables are chosen according to the following random process:

1. The first customer always chooses the first table.
2. The nth customer chooses an existing kth table with probability $m_k/(n - 1 + \alpha)$, and a new table with probability $\alpha/(n - 1 + \alpha)$.

where $\alpha > 0$ is a concentration parameter, and m_k denotes the number of customers seated at the kth table. From the conditional probability distribution above, we can see that a customer is more likely to sit at a table if there are already many people sitting there. However, a customer will sit at a new table with the probability proportional to α.

This CRP provides an effective representation for the inference in Dirichlet process mixture models (DPMM). In DPMM, the distribution of indicator variables $c_1, c_2, \ldots, c_N$ given mixing proportions $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_K)$ is multinomial

$$p(c_1, c_2, \ldots, c_N|\boldsymbol{\pi}) = \prod_{k=1}^{K} \pi_k^{m_k} \tag{8}$$

where $m_k = \sum_{i=1}^{N} \delta(c_i, k)$ is the number of data points in kth cluster and $\sum m_k = N$. Since the Dirichlet distribution is conjugate to the multinomial, we can assume mixing proportions $\boldsymbol{\pi}$ for K clusters have a Dirichlet prior

$$p(\boldsymbol{\pi}|\alpha) = p(\pi_1, \pi_2, \ldots, \pi_K|\alpha) \sim Dir\left(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K}\right) = \frac{\Gamma(\alpha)}{\Gamma\left(\frac{\alpha}{K}\right)^K} \prod_k \pi_k^{\frac{\alpha}{K}-1} \tag{9}$$

Then, integrating out the mixing proportions gives:

$$p(c_1, c_2, \ldots, c_N|\alpha) = \int p(\boldsymbol{c}|\boldsymbol{\pi}) p(\boldsymbol{\pi}|\alpha) d\boldsymbol{\pi} = \frac{\Gamma(\alpha)}{\Gamma(\alpha + N)} \prod_{k=1}^{K} \frac{\Gamma(\alpha/K + m_k)}{\Gamma(\alpha/K)} \tag{10}$$

If the total number of clusters K is finite, then the probability of nth data point belongs to kth cluster given all other data points and concentration parameter α is

$$p(c_n = k | c_{-n}, \alpha) = \frac{m_{-n,k} + \alpha/K}{n - 1 + \alpha} \tag{11}$$

where c_{-n} denotes all indices except n, and $m_{-n,k} = \sum_{i \neq n} \delta(c_i, k)$ is the number of data points in the kth cluster for assigning the first $(n - 1)$ data points. If K is infinite as $K \to \infty$, we can update the posterior indicator distribution using Gibbs sampling as:

$$\begin{aligned} p(c_n = k | c_{-n}, \alpha) &= \frac{m_{-n,k}}{n - 1 + \alpha} \\ p(c_n \neq c_i \ \forall\ i \neq n | c_{-n}, \alpha) &= \frac{\alpha}{n - 1 + \alpha} \end{aligned} \tag{12}$$

The distribution for a new variable $\boldsymbol{y}_*$ within a mixture cluster follows normal distribution

$$p(\boldsymbol{y}_* | c_* = k, \mu_k, \Sigma_k) \sim N(\mu_k, \Sigma_k) \tag{13}$$

where the parameters μ_k and Σ_k are the mean and the covariance for cluster k. As a result, the weight for each cluster is obtained as

$$\vartheta_k = p(c_* = k | \boldsymbol{y}_*) = \frac{p(\boldsymbol{y}_* | c_* = k) p(c_* = k)}{\sum_{k=1}^{K} p(\boldsymbol{y}_* | c_* = k) p(c_* = k)} \tag{14}$$

Due to the nonparametric nature of DP, the shape as well as the number of clusters need not be known a priori. Therefore, DP clusters are derived from characteristics inherent to data.

Predictive Modeling with Clustered Variables. Although the Dirichlet process clusters variables into different groups, the variables in each group are similar to each other and thus bring the redundant information. It is necessary to delineate the structure of latent variables hidden in each cluster. As such, homogeneous variables in each cluster are orthonormalized before predictive modeling. Assume we have K clusters and there are M_k variables, i.e., $\boldsymbol{v}_{k1}, \boldsymbol{v}_{k2}, \ldots, \boldsymbol{v}_{kM_k}$, in the k-th cluster. Then, the redundant information within original variables $(\boldsymbol{v}_{k1}, \boldsymbol{v}_{k2}, \ldots, \boldsymbol{v}_{kM_k})$ is minimized by transforming them into the orthonormal set of new variables $(\boldsymbol{w}_{k1}, \boldsymbol{w}_{k2}, \ldots, \boldsymbol{w}_{kM_k})$ in each cluster using the Gram-Schmidt orthonormalization (GSO). The procedure begins by normalizing $\boldsymbol{v}_{k1}$,

$$\boldsymbol{x}_{k1} = \boldsymbol{v}_{k1}; \ \ \boldsymbol{w}_{k1} = \frac{\boldsymbol{x}_{k1}}{\|\boldsymbol{x}_{k1}\|} \tag{15}$$

where $\boldsymbol{w}_{k1}$ is the normalized variable of $\boldsymbol{v}_{k1}$. Then, we orthogonalize and normalize the second vector $\boldsymbol{v}_{k2}$ as,

$$\boldsymbol{x}_{k2} = \boldsymbol{v}_{k2} - \langle \boldsymbol{v}_{k2}, \boldsymbol{w}_{k1} \rangle \boldsymbol{w}_{k1}; \ \boldsymbol{w}_{k2} = \frac{\boldsymbol{x}_{k2}}{\|\boldsymbol{x}_{k2}\|} \tag{16}$$

where $\boldsymbol{w}_{k2}$ is the second orthonormalized vector. The process is recursively updated to get the m-th orthogonal vector $\boldsymbol{x}_{km}$

$$\boldsymbol{x}_{km} = \boldsymbol{v}_{km} - \sum_{i=1}^{m-1} \langle \boldsymbol{v}_{km}, \boldsymbol{w}_{ki} \rangle \boldsymbol{w}_{ki}; \ \boldsymbol{w}_{km} = \frac{\boldsymbol{x}_{km}}{\|\boldsymbol{x}_{km}\|} \tag{17}$$

where $\boldsymbol{w}_{km}$ is the m-th orthonormalized vector. Further, we leverage orthonormalized variables in each cluster to develop a group elastic-net model[25], which achieves the model sparsity by the group-level and individual-level selection of features. The elastic net criterion is defined as:

$$L(\lambda_1, \lambda_2, \boldsymbol{\beta}) = |\boldsymbol{z} - w\boldsymbol{\beta}|^2 + \lambda_2 |\boldsymbol{\beta}|^2 + \lambda_1 |\boldsymbol{\beta}|_1 \tag{18}$$

where λ_1 and λ_2 are non-negative real values. The elastic net estimator $\hat{\boldsymbol{\beta}}$ is to minimize the equation

$$\hat{\boldsymbol{\beta}} = \underset{\beta}{\operatorname{argmin}} \{L(\lambda_1, \lambda_2, \boldsymbol{\beta})\} \tag{19}$$

Solving $\hat{\boldsymbol{\beta}}$ is equivalent to the optimization problem

$$\begin{aligned} &\hat{\boldsymbol{\beta}} = \underset{\beta}{\operatorname{argmin}} |\boldsymbol{z} - w\boldsymbol{\beta}|^2 \\ &s.\ t.\ \ (1 - \gamma)|\boldsymbol{\beta}|_1 + \gamma |\boldsymbol{\beta}|^2 \leq \lambda \end{aligned} \tag{20}$$

where $\gamma = \frac{\lambda_2}{\lambda_1 + \lambda_2}$ and λ is the tuning parameter.

Cluster #	1st variable	2nd variable	3rd variable	4th variable	5th variable
1	$\boldsymbol{v}_1$	$\boldsymbol{v}_2 = \lvert \boldsymbol{v}_1 \rvert$	$\boldsymbol{v}_3 = \boldsymbol{v}_1{}^2$	$\boldsymbol{v}_4 = \boldsymbol{v}_1{}^3$	$\boldsymbol{v}_5 = \boldsymbol{v}_1{}^4$
2	$\boldsymbol{v}_6$	$\boldsymbol{v}_7 = \lvert \boldsymbol{v}_6 \rvert$	$\boldsymbol{v}_8 = \boldsymbol{v}_6{}^2$	$\boldsymbol{v}_9 = \boldsymbol{v}_6{}^3$	$\boldsymbol{v}_{10} = \boldsymbol{v}_6{}^4$
3	$\boldsymbol{v}_{11}$	$\boldsymbol{v}_{12} = \boldsymbol{v}_{11}(t+3)$	$\boldsymbol{v}_{13} = \boldsymbol{v}_{11}(t+5)$	$\boldsymbol{v}_{14} = \boldsymbol{v}_{11}(t+7)$	$v_{15} = v_{11}(t+9)$
4	$\boldsymbol{v}_{16}$	$\boldsymbol{v}_{17} = \boldsymbol{v}_{16}(t+10)$	$\boldsymbol{v}_{18} = \boldsymbol{v}_{16}(t+20)$	$\boldsymbol{v}_{19} = \boldsymbol{v}_{16}(t+30)$	$\boldsymbol{v}_{20} = \boldsymbol{v}_{16}(t+40)$

Table 1. Four cluster of simulated variables. Where $\boldsymbol{v}_1$ and $\boldsymbol{v}_6$ are independent standard normal variables, $\boldsymbol{v}_{11}$ is a nonlinear variable sampled from logistic map $\boldsymbol{v}_{11}(n+1) = 3.8\boldsymbol{v}_{11}(n)(1 - \boldsymbol{v}_{11}(n))$, $\boldsymbol{v}_{16}$ is a second-order autoregressive variable that is nonlinearly coupled with x_{Lorenz}, $\boldsymbol{v}_{16}(n) = 1.095\boldsymbol{v}_{16}(n - 1) - 0.4\boldsymbol{v}_{16}(n - 2) + 0.7\varepsilon_n + 0.3x_{Lorenz}^2$, where ε_n is Gaussian noise, x_{Lorenz} is the x-component of a Lorenz system: $x' = 10(y - x)$, $y' = x(28 - z) - y$, $z' = xy - \frac{8}{3}z$ with time step 0.01. Each variable has a sample size of 1000.

To develop the group elastic-net model for logistic regression, we define $h_\beta(\boldsymbol{w}, i)$ as the probability for z$_i$ being a success (i.e., $z_i = 1$) and thus $1 - h_\beta(\boldsymbol{w}, i)$ is the probability for z_i being a failure (i.e., $z_i = 0$), where $\boldsymbol{\beta} = (\beta_0, \beta_{11} \ldots, \beta_{KM_K})^T$ is the coefficient vector. Then we have

$$\log\left(\frac{h_\beta(\boldsymbol{w}, i)}{1 - h_\beta(\boldsymbol{w}, i)}\right) = \boldsymbol{\beta}^T \boldsymbol{W} \tag{21}$$

and

$$h_\beta(\boldsymbol{w}, i) = \frac{1}{1 + \exp[-(\beta_0 + (\sum_{k=1}^{K}\sum_{m=1}^{M_k} w_{km}(i)\beta_{km}))]} \tag{22}$$

The likelihood function given observations ($\boldsymbol{w}(i)$, z_i) is

$$\prod_{i=1}^{N} h_\beta(\boldsymbol{w}, i)^{z_i}(1 - h_\beta(\boldsymbol{w}, i))^{1-z_i} \tag{23}$$

Taking the logarithm for equation, we have

$$L(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{\beta}) = \sum_{i=1}^{N}[z_i \ \log(h_\beta(\boldsymbol{w}, i)) + (1 - z_i)\log(1 - h_\beta(\boldsymbol{w}, i))] \tag{24}$$

Therefore, we derive the group elastic-net model as:

$$\begin{aligned} &\max_{\beta}\sum_{i=1}^{N}[z_i \ \log(h_\beta(\boldsymbol{w}, i)) + (1 - z_i)\log(1 - h_\beta(\boldsymbol{w}, i))]^2 \\ &s.\ t.\ \ \sum_{k=1}^{K}\sum_{m=1}^{M_k}(\gamma\beta_{km}{}^2 + (1 - \gamma)|\beta_{km}|) \le \lambda \end{aligned} \tag{25}$$

where γ and λ are penalization parameters, the logistic function $h_\beta(\boldsymbol{w}, i)$ is used in the likelihood function because of the binary responses. The proposed approach will be evaluated and validated using experimental studies. The details are shown in the next section.

Experimental Materials and Results

In this section, we evaluate and validate the proposed methodology using both simulation and real-world case studies.

Simulation Study. First, a simulation study is shown to evaluate the performance of the proposed methodology for variable clustering. We simulate four clusters of variables in Table 1 as follows.

Figure 6a shows the matrices of Pearson's correlations among variables that are computed from the simulation data set. Notably, the linear correlation in Fig. 6a cannot fully identify the nonlinear interdependence among simulated variables. Figure 6b shows that the HC cannot delineate the cluster structure of variables. This is mainly due to the fact that Pearson's correlation is limited in the ability to detect nonlinear interdependence structures among variables.

Figure 7a shows the mutual information based correlation matrix among variables that are computed from the simulated data set. The red color represents a higher nonlinear correlation, while the blue color indicates no interrelationships. Figure 7a shows significant nonlinear correlation within the simulated clusters. Also, variables from different clusters have little interrelationship. If we use Dirichlet process to cluster variables based on low-dimensional vectors embedded from the dissimilarity matrix of mutual information, four clusters of variables are distinctly separated in the space (see Fig. 7b). The simulation study shows that Dirichlet process models effectively cluster these 20 variables into 4 groups and identifies the underlying cluster structures of variables.

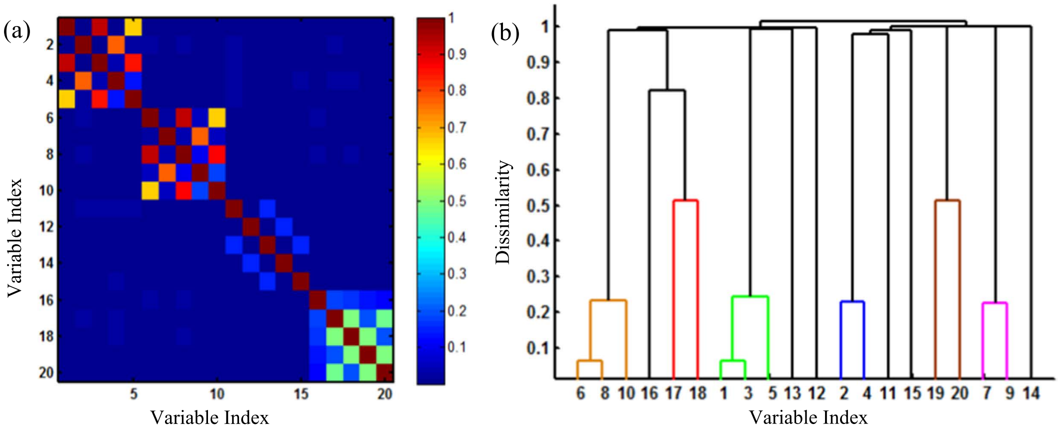


Figure 6. (**a**) Matrix of Pearson's correlation; (**b**) Hierarchical clustering of simulated variables.

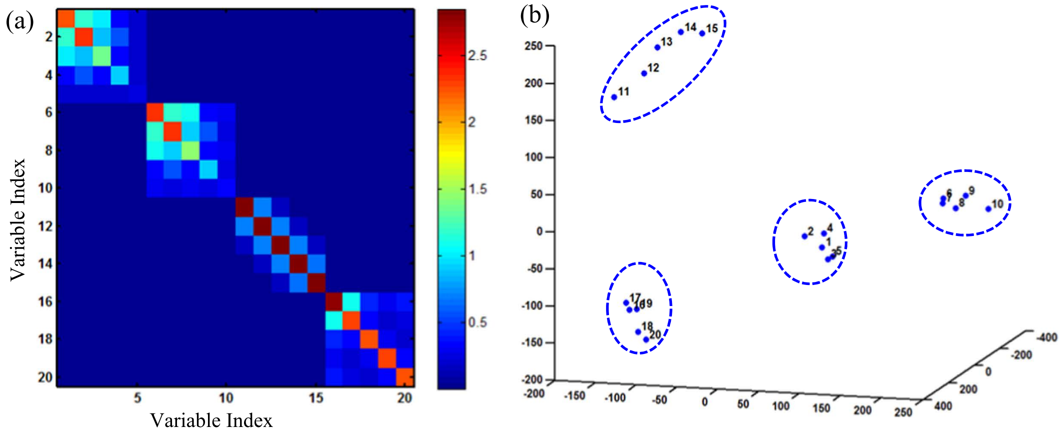


Figure 7. (**a**) Mutual information based dissimilarity matrix; (**b**) Dirichlet process clustering of simulated variables.

Real-world Case Study. In the previous work, we characterized and represented 3-dimensional vectorcardiogram (VCG) signals using a sparse basis function model[26]. This sparse representation not only reduces large amounts of data to a limited number of model parameters, but also preserves the signal information. As opposed to the original data, this present paper will utilize parameters in basis function models as explanatory variables to further predict the myocardial infarctions. VCG signals are represented by L superposed basis functions in order to capture intrinsic characteristics of cardiac electrical activity as:

$$\theta(t) = \omega_0 + \sum_{j=1}^{L} \omega_j \psi_j((t - \varphi_j)/\sigma_j) + \varepsilon \tag{26}$$

where φ_j and σ_j are shifting and scaling factors, $\psi_j(\cdot)$ are basis functions, and ω_j are weight factors, respectively. The objective is to optimize the sparse representation of 3D VCG signals:

$$\operatorname{argmin}\left[\|\theta(t) - \omega_0 - \sum_{j=1}^{L} \omega_j \psi_j((t - \varphi_j)/\sigma_j)\|^2, \{\boldsymbol{\omega}, \boldsymbol{\varphi}, \boldsymbol{\sigma}, \boldsymbol{\psi}, L\}\right] \tag{27}$$

In order to identify a compact set of basis functions that minimize the representation error, the number of basis functions L is minimized and basis functions $\boldsymbol{\psi}$ are optimally placed. Model parameters $\boldsymbol{\omega}$, $\boldsymbol{\varphi}$, $\boldsymbol{\sigma}$ are adaptively estimated by "best matching" projections of VCG signals onto a dictionary of nonlinear basis functions. The optimization algorithms of a sparse basis function representation for spatiotemporal VCG signals were detailed in our previous work[26].

In this present study, model parameters, i.e., weight, shifting, scaling factors and residuals, are extracted from the sparse basis function representation of VCG signals, and then are further utilized as explanatory variables for the identification of cardiac disorders (i.e., myocardial infarctions). The parameter set is $\{\boldsymbol{\omega}_{3\times L}, \boldsymbol{\phi}_{3\times L}, \boldsymbol{\sigma}_{3\times L}\}$ for L basis functions because there are 3 channels of signals in 3-lead VCG. Our previous study[26] showed that modeling performance is greater than 99.9% goodness-of-fit with a parsimonious set of 20 basis functions for a variety of

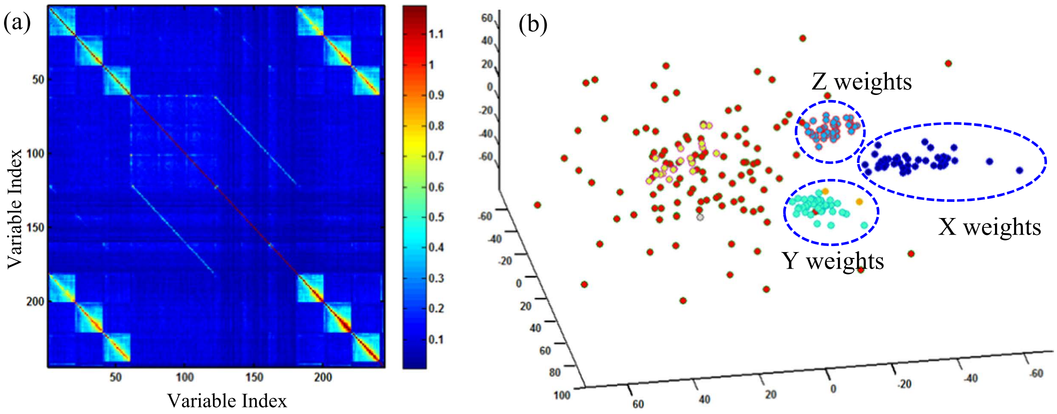


Figure 8. (**a**) Dissimilarity matrix based on mutual information measured between variables; (**b**) Dirichlet process clustering of model-based parametric features.

cardiac conditions. Hence, a total of 180 model parameters is adaptively estimated from the 3D VCG trajectory. In addition, we add other parameters in this present investigation, the overall feature matrix is:

$$\boldsymbol{F} = \{\boldsymbol{\omega}_{3\times 20},\ \boldsymbol{\varphi}_{3\times 20},\ \boldsymbol{\sigma}_{3\times 20},\ |\boldsymbol{\omega}|_{3\times 20},\ \boldsymbol{RSS}_{3\times 1},\ RR_{1\times 1}\} \tag{28}$$

where $|\boldsymbol{\omega}|_{3\times 20}$ are absolute values of weights, describing amplitudes of each basis function and indicating local strengths of a heartbeat. The residual sum of squares $\boldsymbol{RSS}_{3\times 1}$ measure the discrepancy between model representation and VCG signals in each channel. The heart rate $RR_{1\times 1}$ characterizes temporal beat-to-beat variations of cardiac electrical activity. Therefore, these 244 parameter-based features are used to represent the details of original VCG signals. Notably, the high-dimensional VCG signals are reduced into a parsimonious set of model parameters using the sparse representation without losing clinically important information.

A total of 388 (79 controls and 309 infarctions) 3-lead VCG signals, available in the PhysioNet Database[27], are used in this investigation. These signals were digitized at 1 kHz sampling rate with a 16-bit resolution over a range of 16.384 mV. Our previous study showed that most of model-driven parameters (146 over 244 features) are statistically significant between healthy controls and diseased conditions, i.e., Kolmogorov-Smirnov (K-S) statistics are greater than critical value 0.17[28]. In addition, weight factors yield larger K-S statistics than other parametric features. However, the "curse of dimensionality" as well as the overfitting problems come out with a large number of predictors for the predictive modeling. Therefore, the lasso-penalized logistic regression model was utilized to shrink the number of predictors and further identify cardiac disorders (i.e., myocardial infarctions) in our previous study[28].

Nonetheless, our previous study[28] focused on the relevancy between predictor and response variables, without specifically considering nonlinear interdependence structures among predictor variables. Prior research showed that the collinearity (i.e., large correlation between variables) leads to stability problems in predictive models (i.e., increased variances of estimation)[29]. The present paper further investigates the nonlinear correlations between variables and then identifies the cluster structures of variables for improving the predictive performance. Figure 8a shows the visualization of information-based dissimilarity matrix measured among variables. It may be noted that six groups of variables have stronger nonlinear relationships, i.e., $\boldsymbol{\omega}_{3\times 20}$ and $|\boldsymbol{\omega}|_{3\times 20}$ as the weights and absolute weights of X, Y and Z-axis directions. However, few, if any, previous work has explicitly considered such relationships among variables in the process of predictive modeling. Moreover, weight factors $\boldsymbol{\omega}_{3\times 20}$ also have strong nonlinear correlation with the variables of absolute weights $|\boldsymbol{\omega}|_{3\times 20}$. Without taking these nonlinear interrelationships into account, predictive models are sensitive to extraneous noises and are limited in the ability to provide an effective prediction of myocardial infarctions.

Figure 8b shows the nonparametric Dirichlet process for variable clustering of model-based parametric features. As shown in Fig. 8b, the Dirichlet process cluster all the variables into five groups based on the embedding features from the variable-to-variable dissimilarity matrix of mutual information. Three clusters are shown to be significant, i.e., weight and absolute weight variables of X, Y and Z-axis respectively. As a result, homogeneous variables are clustered into subset communities. It may be noted that the result of variable clustering is consistent with the prior knowledge and the variable-to-variable dissimilarity matrix of mutual information.

Figure 9 shows the results of variable clustering by our proposed algorithm and the information-based clustering. Note that there are 244 variables represented as color markers, and each marker with the same color represents the same cluster. Each row denotes a type of variables. For example, the first row of 20 markers is weight factors $\boldsymbol{\omega}_{X1:20}$ in the X-dimension of VCG signals. Figure 9a shows the clustering results for MI-DP clustering (also see Fig. 8b), while Fig. 9b shows the clustering results for the information-based clustering. It may be noted that information-based clustering was designed to cluster data samples rather than variables. We modified the original algorithm in ref. 21 for variable clustering. Because information-based clustering[21] needs to predefine the number of clusters, we therefore use the same number of clusters identified by our proposed algorithm. Note that Fig. 9 shows there are slight differences in clustering results by MI-DP and information-based clustering. Figure 9b shows that a small portion of Y weights is not accurately clustered by information-based clustering. In

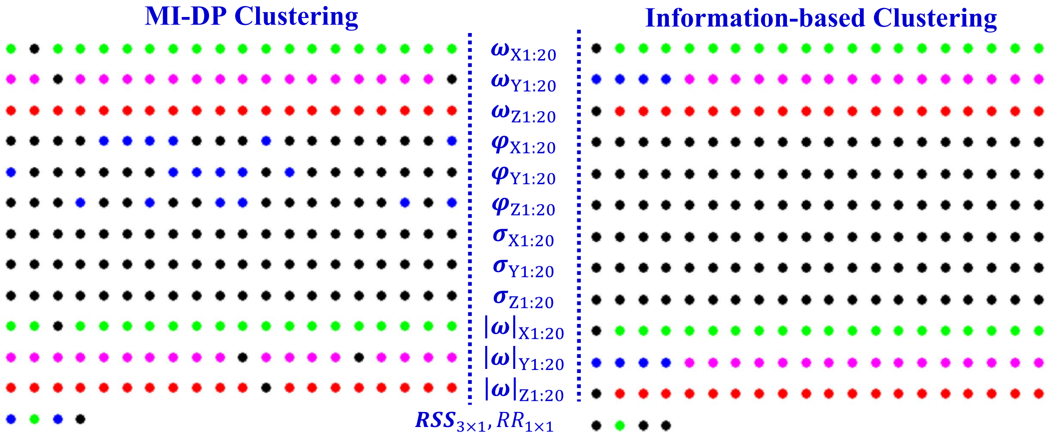


Figure 9. The results of variable clustering by (**a**) MI-DP clustering and (**b**) Information-based clustering with color coding.

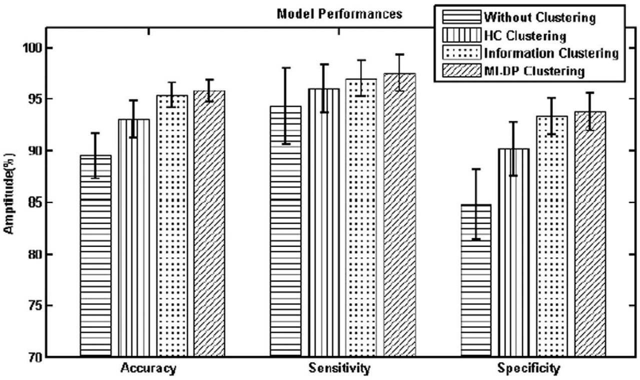


Figure 10. The comparison of averages and standard deviation of prediction performances in the real-world case study. "Without clustering": lasso-penalized logistic regression model; "HC clustering": group elastic-net model with hierarchical clustering using linear correlation measured between variables; "Information clustering": information-based clustering[21]; "MI-DP clustering": group elastic-net model with variable clustering using mutual information and Dirichlet Process Mixtures.

addition, some variables such as shifting and scaling factors, and residuals are grouped together and cannot be well separated. As such, information-based clustering yields slightly inferior performance of predictive modeling in comparison with the proposed MI-DP approach (also see Fig. 10).

Figure 10 shows the comparison of prediction performances of different clustering procedures in the real-world case study. "Without clustering" represents the results from the lasso-penalized logistic regression model in our previous study[28]. "HC clustering" denotes the hierarchical clustering with linear correlation measured between variables. "Information clustering" is the information-based clustering from the literature[21]. "MI-DP clustering" is the proposed information theoretic approach for variable clustering using mutual information and Dirichlet Process Mixtures. As shown in Fig. 10, MI-DP clustering yields better performance than "Without clustering". Note that MI-DP clustering improves the predictive accuracy from 89.50% to 95.84%, the sensitivity is improved from 94.33% to 97.56%, and the specificity is increased from 84.80% to 93.78%. In addition, MI-DP clustering yields smaller standard deviations of performance metrics (i.e., accuracy, sensitivity, and specificity) than "without clustering". Similarly, the results of MI-DP clustering are better than "HC clustering" (i.e., accuracy 93.07%, sensitivity 96.05% and specificity 90.18%) and "Information clustering" (i.e., accuracy 95.38%, sensitivity 97.02% and specificity 93.33%). Experimental results showed that MI-DP clustering effectively delineates the nonlinear correlation structures among variables and further derive homogeneous groups of variables, thereby improving the prediction performance.

Discussion and Conclusions

Advanced sensing and real-time data acquisition bring the proliferation of big data. This provides an unprecedented opportunity to move forward data-driven knowledge discovery. However, it is common that big data involves large amounts of variables with complex interdependence structures, which brings significant challenges on traditional modeling strategies. To tackle these challenges, variable selection and variable clustering are widely used in the literature. Nonetheless, variable selection focuses primarily on the relevancy between predictors and response variables, but does not explicitly consider the redundancy among variables. The variable clustering, on the other hand, focuses on the linear relevancy between variables. There is a need to develop new methodologies to improve the effectiveness and efficiency of variable clustering and predictive analytics.

The computational complexity of MI-DP clustering consists of three components, namely measure of mutual information, low-dimensional embedding, and DP variable clustering. First, mutual information is measured among $N(N-1)/2$ pairs of variables. The computational complexity for one pair of variables is $o((\# \text{ of bins})^2)$, i.e., $o(\sqrt{N_S}/2)$. Hence, the complexity is approximately $o(\sqrt{N_S}N(N-1)/4)$. Second, the complexity of low-dimensional embedding is shown to be $o(N\sqrt{N})$ in the literature[30]. Third, the Dirichlet process allocates each variable to a cluster with a computational complexity of $o(N)$. In the present case studies, there are not significant challenges in computational complexity. However, it is worth mentioning that a new research direction is to design efficient algorithms to compute the pairwise mutual information (MI) between all pairs of variables, which will significantly improve the performance of MI-DP approach for big data applications.

This paper presents a new information-theoretic approach for variable clustering and predictive modeling using Dirichlet process mixtures. This new methodology investigates both redundancy and relevancy among variables for improving the performance of predictive modeling. Both simulation and real-world case studies demonstrate that the proposed MI-DP clustering algorithm not only outperforms traditional methods (i.e., lasso-penalized variable selection and classical hierarchal clustering), but also identifies nonlinear interdependence structures among variables and further improves the performance of predictive modeling. The new methodology of MI-DP variable clustering is generally applicable for predictive modeling in many disciplines that involve a large number of highly-redundant variables. In the future work, we will also consider the integration of our proposed MI-DP clustering algorithm with co-clustering approach to investigate the nonlinear interdependence among subsets of both samples and variables.

References

1. Verhoef, P. C. & Donkers, B. Predicting customer potential value an application in the insurance industry. *Decision Support Systems* **32,** 189–199 (2001).
2. Chen, Y. & Yang, H. Sparse Modeling and Recursive Prediction of Space-time Dynamics in Stochastic Sensor Network. *IEEE Transactions on Automation Science and Engineering* **13,** 215–226 (2016).
3. Chen, Y. & Yang, H. *Heterogeneous postsurgical data analytics for predictive modeling of mortality risks in intensive care units* (Proceedings of 2014 IEEE Engineering in Medicine and Biology Society Conference (EMBC), 2014).
4. Yang, H. & Kundakcioglu, E. Healthcare Intelligence: Turning Data into Knowledge. *Intelligent Systems, IEEE* **29,** 54–68 (2014).
5. Ward, J. H. Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association* **58,** 236–244 (1963).
6. Airoldi, E. M., Blei, D. M., Fienberg, S. E. & Xing, E. P. Mixed Membership Stochastic Blockmodels. *Journal of Machine Learning Research* **9,** 1981–2014 (2008).
7. Acharya, A. *et al.* In ECML PKDD (eds Appice, A. *et al.*) 283–299 (2015).
8. Zhou, M. Infinite Edge Partition Models for Overlapping Community Detection and Link Prediction. *In Proceedings of AISTATS* **38** (2015).
9. Blei, D. M., Ng, A. Y. & Jordan, M. I. Latent Dirichlet Allocation. *Journal of Machine Learning Research* **3,** 993–1022 (2003).
10. Blei, D. M. & McAuliffe, J. D. Supervised topic models. *Advances in Neural Information Processing Systems* (*NIPS*), 121–128 (2007).
11. Zhu, J., Ahmed, A. & Xing, E. P. MedLDA: Maximum Margin Supervised Topic Models. *Journal of Machine Learning Research* **13,** 2237–2278 (2012).
12. Zhu, J., Chen, N., Perkins, H. & Zhang, B. Gibbs Max-Margin Topic Models with Fast Sampling Algorithms. *In Proceedings of ICML* **28** (2013).
13. Cheng, Y. & Church, G. M. *Biclustering of Expression Data*. *In Proceedings of ISMB* **8,** 93–103 (2000).
14. Dhillon, I. S. Co-clustering documents and words using bipartite spectral graph partitioning. *In Proceedings of ACM SIGKDD*, 269–274 (2001).
15. Dhillon, I. S., Mallela, S. & Modha, D. S. Information-theoretic co-clustering. *In Proceedings of ACM SIGKDD*, 83–89 (2003).
16. Deodhar, M. & Ghosh, J. SCOAL: A Framework for Simultaneous Co-Clustering and Learning from Complex Data. *ACM Transactions on Knowledge Discovery from Data* **4,** 11–31 (2010).
17. Pearson, K. Notes on regression and inheritance in the case of two parents. *Proceedings of the Royal Society of London* **58,** 240–242 (1895).
18. Fraser, A. M. & Swinney, H. L. Independent coordinates for strange attractors from mutual information. *Phys. Rev. A* **33,** 1134–1140 (1986).
19. Kinney, J. B. & Atwal, G. S. Equitability, mutual information, and the maximal information coefficient. *PNAS* **111,** 3354–3359 (2014).
20. Reshef, D. N. *et al.* Detecting Novel Associations in Large Data Sets. *Science* **334,** 1518–1524 (2011).
21. Slonim, N., Atwal, G. S., Tkačik, G. & Bialek, W. Information-based clustering. *PNAS* **102,** 18297–18302 (2005).
22. Jain, A. K. Data clustering: 50 years beyond K-means. *Pattern Recognition Letters* **31,** 651–666 (2010).
23. Le, T. Q., Cheng, C., Sangasoongsong, A., Wongdhamma, W. & Bukkapatnam, S. T. S. Wireless wearable multisensory suite and real-time prediction of obstructive sleep apnea episodes. *IEEE Journal of Translational Engineering in Health and Medicine* **1,** 2700109 (2013).
24. Blei, D. M. & Jordan, M. I. Variational inference for Dirichlet process mixtures. *Bayesian Analysis* **1,** 121–144 (2006).
25. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. R. Statist. Soc. B* **67,** 301–320 (2005).
26. Liu, G. & Yang, H. Multiscale adaptive basis function modeling of spatiotemporal cardiac electrical signals. *IEEE Journal of Biomedical and Health Informatics* **17,** 484–492 (2013).
27. Goldberger, A. L. *et al.* PhysioBank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals. *Circulation* **23,** e215–e220 (2000).
28. Liu, G., Kan, C., Chen, Y. & Yang, H. Model-driven parametric monitoring of high-dimensional nonlinear functional profiles. *Automation Science and Engineering* (*CASE*)*, 2014 IEEE International Conference*. 722–727 (2014).
29. Nas, T. & Mevik, B. H. Understanding the collinearity problem in regression and discriminant analysis. *Journal of Chemometrics* **15,** 412–426 (2001).
30. Morrison, A., Ross, G. & Chalmers, M. Fast multidimensional scaling through sampling, springs and interpolation. *Information Visualization* **2,** 68–77 (2003).

Acknowledgements

This work is supported in part by the National Science Foundation (CMMI-1646660, CMMI-1617148, CMMI-1619648, and IOS-1146882). The authors also thank Harold and Inge Marcus Career Professorship (HY) for additional financial support.

Author Contributions

H.Y. conceived the study and contributed to the design of the study, data collection, data interpretation, and revised the manuscript. Y.C. contributed to the development of algorithms, evaluated the data, performed the data analysis, and drafted the manuscript. All authors read and approved the final manuscript.

Additional Information

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Chen, Y. and Yang, H. A Novel Information-Theoretic Approach for Variable Clustering and Predictive Modeling Using Dirichlet Process Mixtures. *Sci. Rep.* **6**, 38913; doi: 10.1038/srep38913 (2016).

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.