Ann Oper Res
DOI 10.1007/s10479-017-2442-2

Self-organizing network for variable clustering

Gang Liu[1] · **Hui Yang**[2]

Abstract Advanced sensing and internet of things bring the big data, which provides an unprecedented opportunity for data-driven knowledge discovery. However, it is common that a large number of variables (or predictors, features) are involved in the big data. Complex interdependence structures among variables pose significant challenges on the traditional framework of predictive modeling. This paper presents a new methodology of self-organizing network to characterize the interrelationships among variables and cluster them into homogeneous subgroups for predictive modeling. Specifically, we develop a new approach, namely nonlinear coupling analysis to measure variable-to-variable interdependence structures. Further, each variable is represented as a node in the complex network. Nonlinear-coupling forces move these nodes to derive a self-organizing topology of the network. As such, variables are clustered into sub-network communities. Results of simulation experiments demonstrate that the proposed method not only outperforms traditional variable clustering algorithms such as hierarchical clustering and oblique principal component analysis, but also effectively identifies interdependent structures among variables and further improves the performance of predictive modeling. Additionally, real-world case study shows that the proposed method yields an average sensitivity of 96.80% and an average specificity of 92.62% in the identification of myocardial infarctions using sparse parameters of vectorcardiogram representation models. The proposed new idea of self-organizing network is generally applicable for predictive modeling in many disciplines that involve a large number of highly-redundant variables.

Keywords Self-organizing network · Variable clustering · Predictive modeling · Nonlinear coupling analysis · Myocardial infarction · Vectorcardiogram

✉ Hui Yang
huy25@psu.edu

[1] Arbor Research Collaborative for Health, Ann Arbor, MI 48104, USA

[2] Harold and Inge Marcus Department of Industrial and Manufacturing Engineering, Pennsylvania State University, University Park, PA 16802, USA

Published online: 27 February 2017

1 Introduction

Predictive analytics leverages information and patterns extracted from large amounts of data to predict outcomes and drive decisions. It is widely used in a variety of disciplines, e.g., business, healthcare, and manufacturing. In business, online retailers use predictive analytics to stratify customers and then recommend potential products to them (Neslin et al. 2006). In healthcare, professionals use predictive analytics to extract information from clinical data and exploit data-driven patterns for better medical decision making (Chen and Yang 2014; Yang and Kundakcioglu 2014). In manufacturing, predictive analytics helps estimate the degradation trajectory so as to prevent potential failures of manufacturing equipment and defective products (Ding et al. 2006). Indeed, predictive analytics is critical to increasing the profits of a company, improving the health of our society, and enhancing the performance of manufacturing systems.

More specifically, advanced biosensing and health information technology bring big data in biomedical science and health care, which presents a gold mine in the 21st century to advance data-driven personalized medicine and medical decision making. For example, electrocardiogram (ECG) is noninvasive and always obtainable with extremely portable equipment. As a result, real-time ECG monitoring in days, months and even years generates large amounts of data. This provides an unprecedented opportunity for the early prediction of cardiac disorders. Nonetheless, big data also poses significant challenges for human experts to inspect ECG signals visually. New analytical methods and tools are urgently needed to realize the full potential of big data for predicting cardiac risks and improving healthcare outcomes.

However, it is common that big data often involve a large number of variables (or predictors, features). High dimensionality and complex structures among variables also pose significant challenges on traditional methodologies in predictive analytics. Realizing the full potentials of big data for predictive analytics hinges upon the development of new methodologies that effectively handle the high dimensionality and complex variable-to-variable interrelationships. A large number of variables brings the issue of "curse of dimensionality" in predictive analytics. When the dimensionality increases, large amounts of training data are required to learn within predictive models. It may be noted that "curse of dimensionality" increases mean squared errors and the bias of predicted responses (Friedman 1997). On the other hand, complex interdependence structures among variables (e.g., collinearity—a higher correlation >0.90) will lead to more sensitive estimations of parameters in predictive models (i.e., increased variances of estimation) (Nas and Mevik 2001). Also, complex systems often exhibit nonlinear coupling and synchronization behaviors (Wang et al. 2009). As such, there are nonlinear interdependence structures among process variables of complex systems. Linear and nonlinear redundancies among variables impact the performance of predictive analytics.

In the literature, variable selection and variable clustering are widely used to address these challenges. For example, generalized linear models are often integrated with shrinkage methods to optimize model sparsity and improve the prediction accuracy. Examples of shrinkage and selection methods include best-subset selection (Narendra and Fukunaga 1977), ridge regression (Hoerl and Kennard 1970), LASSO (Tibshirani 1996), least angle regression (Efron et al. 2004) and elastic net (Zou and Hastie 2005). However, most of existing methods focus on the relevancy between predictors and response variables. Interdependent structures among predictors are often overlooked, or not explicitly investigated. On the other hand, variable clustering depends on similarity measurements between variables

such as linear correlation or mutual information. Note that linear correlation cannot capture nonlinear interdependences among variables. Mutual information characterizes linear and nonlinear correlation, but requires the stationarity assumption (Fraser and Swinney 1986). Latent-variable methods are also proposed for variable clustering, e.g., oblique principal component clustering (OPCC) (Lee et al. 2008). However, oblique rotation and principal component analysis are based on linear projections of variables. Nonlinear interdependences among variables are not explicitly considered. It should be noted that variable clustering mainly focuses on the redundancy among variables but neglects the relevancy between predictors and response variables. New methodologies that integrate variable clustering with variable selection to improve effectiveness and efficiency of predictive analytics are urgently needed.

In this paper, we develop a new methodology of self-organizing network to investigate both redundancy and relevancy among variables for improving the performance of predictive modeling. Specifically, we propose to characterize and model nonlinear interdependence structures among variables. Further, these variables are embedded as nodes in a complex network. Nonlinear-coupling forces move these nodes to derive a self-organizing topology of the network. As such, variables are grouped into sub-network communities in the space. Experimental results in both simulation studies and real-world case studies demonstrate that the proposed methodology not only outperforms traditional variable clustering algorithms such as hierarchal clustering and oblique principal component analysis, but also effectively identifies interdependent structures among variables and further improves the performance of predictive modeling.

The remainder of this paper is organized as follows: Sect. 2 reviews the research background; Sect. 3 presents the methodology; Sect. 4 contains experimental design and results of simulation study; Sect. 5 shows the results of a real-world case study that extracts model parameters from vectorcardiogram (VCG) signals for the identification of myocardial infarctions; and Sect. 6 includes the conclusions arising out of this investigation.

2 Research background

2.1 Variable clustering versus data clustering

Table 1 shows the data structure in the traditional table form, where samples are in rows and variables are in columns. In the literature, "clustering" is often referred to be "data clustering." Data clustering focuses on the samples in rows (i.e., $s_1, s_2, \ldots, s_N$) that share similar properties or patterns, where N is the number of samples. In other words, data clustering groups data samples into homogeneous subsets, in which data samples are closer to each

Table 1 The data structure with samples in rows and variables in columns

Samples	Variables (features, factors)			
	x_1	x_2	$\cdots$	x_P
s_1	x_{11}	x_{12}	$\cdots$	x_{1P}
s_1	x_{21}	x_{22}	$\cdots$	x_{2P}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
s_N	x_{N1}	x_{N2}	$\cdots$	x_{NP}

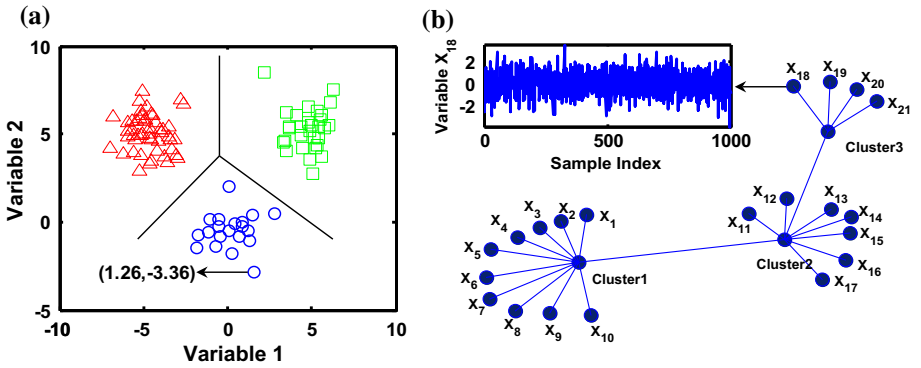

Fig. 1 **a** Data clustering with each point representing a data sample and **b** variable clustering with each point representing a variable

other in the same cluster than other clusters. However, variable clustering is more concerned about the variables in the columns (i.e., $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_P$), where P is the number of variables. Big data often brings a large number of variables that may be bigger than the number of samples, i.e., $P > N$. Complex interdependence structures among variables call upon the development of variable clustering 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 separates samples into clusters, where each point represents a data sample, X-axis is the dimension of variable 1, and Y-axis is the dimension of variable 2. Here, each data point has two coordinates, e.g., $(1.26, -3.36)$ that represent a data sample in the 2-dimensional space. Data samples are clustered based on the similarity measure, e.g., Euclidean distance. Data clustering is an unsupervised method to group data samples into homogeneous clusters based on similarity measures. However, Fig. 1b illustrates the clustering results of 21 variables, each of which has 1000 data samples. For example, the variable X_{18} represents a series of 1000 samples. Notably, each point in Fig. 1b is a variable instead of a data sample. Variable clustering considers the interdependence structure among variables, e.g., correlation or mutual information.

2.2 Hierarchical clustering

In the literature, both Pearson's correlation and mutual information were integrated with hierarchical clustering (HC) (Ward 1963) for variable clustering. The Pearson's correlation ρ_{x_i,x_j} measures the relationship of linear interdependence between x_1 and x_2. Mutual information MI_{x_1,x_2} characterizes and quantifies linear and nonlinear correlations between variables (Fraser and Swinney 1986), i.e.,

$$MI_{x_1,x_2} = \sum_{x_2} \sum_{x_1} Pr\,(x_1, x_2) \log \left(\frac{Pr\,(x_1, x_2)}{Pr\,(x_1)\, Pr\,(x_2)} \right) \tag{1}$$

The HC procedure is either done in an agglomerative way or in a divisive way. For example, each variable is a singleton cluster in the first step of agglomerative HC. Then, two closest clusters are merged into one cluster. The recursive merging continues to move 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 intergroup dissimilarity as the average dissimilarity $D_{GA}\left(C_i, C_j\right)$ between two clusters, i.e.,

$$D_{GA}\left(C_i, C_j\right) = \frac{1}{N_{C_i} N_{C_j}} \sum_{x_i \in C_i} \sum_{x_j \in C_j} D_{x_i x_j} \tag{2}$$

where N_{C_i} and N_{C_j} are the number of variables in the cluster C_i and C_j, $D_{x_i x_j}$ is the dissimilarity between variables x_i and x_j, which is usually calculated as $1 - \rho_{x_i, x_j}$ or $1 - MI_{x_i, x_j}$.

It should be noted that linear correlation cannot adequately capture nonlinear interdependence among variables. Most importantly, both linear correlation and mutual information measure symmetric interdependence between variables. In other words, two variables $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ can be placed interchangeably without impacting the correlation measures. In fact, HC is only applicable when dissimilarity measures are symmetric. However, it is not uncommon that the interdependence structure between two variables are asymmetric, e.g., $Pr(\boldsymbol{x}_1|\boldsymbol{x}_2) \neq Pr(\boldsymbol{x}_2|\boldsymbol{x}_1)$. In other words, information transfer between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ is not necessarily symmetric. The presence of nonlinear and asymmetric interdependence structures poses a significant challenge for variable clustering. Further, HC is not a dynamic approach. In other words, we cannot relocate the variables once the merge is done for two closest clusters. If two variables are 'incorrectly' clustered at the early stage, there is no adaptive step in the later stage to make corrections. An effective algorithm should allow the re-examination of clustering results in every step of the hierarchy.

2.3 Oblique principal component clustering

Also, latent-variable methods such as OPCC (Lee et al. 2008) are widely used for variable clustering. Suppose $\boldsymbol{X_{n\times p}} = \left[\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots \boldsymbol{x}_p\right], \boldsymbol{x}_i = [x_{i1}, x_{i2}, \ldots x_{in}]^T$ is the data matrix of n rows representing n samples and p columns representing p variables. Without the loss of generality, we standardize the variables in data matrix $\boldsymbol{X}$ to have zero mean and unit standard deviation. Principal component analysis (PCA) transforms the data matrix into the orthogonal space, where a sparse set of q ($q \leq p$) principal components (PCs) preserve most of information in original data (Yang and Chen 2014). These PCs are latent variables which are linear projections of original variables. The k-th ($k = 1, 2, \ldots, q$) PC is calculated as

$$\boldsymbol{X}\boldsymbol{z}_k = z_{k1}\boldsymbol{x}_1 + z_{k2}\boldsymbol{x}_2 + \cdots + z_{kp}\boldsymbol{x}_p = \sum_{i=1}^{p} z_{ki}\boldsymbol{x}_i \tag{3}$$

where $\boldsymbol{z}_k = \left[z_{k1}, z_{k2}, \ldots, z_{kp}\right]^T$ is the k-th eigenvector with a unity norm. The eigenvector $\boldsymbol{z}_k$ is derived by maximizing the variance of the k-th PC (i.e., $\boldsymbol{X}\boldsymbol{z}_k$), while meeting with the constraints: (1) eigenvectors are orthogonal to each other; (2) PCs are ordered according to the magnitude of variances. The PCs can be derived by

$$\begin{aligned} &\underset{\boldsymbol{z}_k}{\text{argmax}}\, var\left(\sum_{i=1}^{p} \boldsymbol{z}_{ki}\boldsymbol{x}_i\right) = \underset{\boldsymbol{z}_k}{\text{argmax}}\, \boldsymbol{z}_k^{\boldsymbol{T}} \boldsymbol{\Sigma} \boldsymbol{z}_k \\ &s.t.\ \ \boldsymbol{z}_k^{\boldsymbol{T}} \boldsymbol{z}_{k'} = 0\boldsymbol{z},\ {}_k^{\boldsymbol{T}}\boldsymbol{z}_k = 1\left(k' = 1, \ldots, k-1\right) \end{aligned} \tag{4}$$

where $\boldsymbol{\Sigma} = \boldsymbol{X}^T\boldsymbol{X}$ is the covariance matrix of $\boldsymbol{X}$.

Although PCA orthogonalizes the variables and tackles the multicollinearity issue, it is limited in the capability to interpret data matrix in the original input space. Such an interpretation is critical to cluster variables in the input space. Therefore, the OPCC method

was further developed to enhance the interpretability of principal components and identify the cluster structure of variables. The OPCC method rotates the eigenvector matrix $\boldsymbol{Z}$ to obtain a new one $\boldsymbol{B} = \boldsymbol{Z\Omega}$ that has a simpler structure. In other words, oblique rotation is aimed at obtaining a sparse matrix $\boldsymbol{B}$ in which most of the elements are close to 0. The Varimax criterion is to find the oblique rotation matrix $\boldsymbol{\Omega}$ that maximizes the function:

$$\max_{\boldsymbol{\Omega}} \sum_{i=1}^{p} \left[\sum_{j=1}^{q} b_{ij}^4 - \left(\sum_{j=1}^{q} b_{ij}^2 \right)^2 \right] \tag{5}$$

where b_{ij} is the element in the i-th row and j-th column of new loading matrix $\boldsymbol{B}$. See details of oblique rotation, for example, in Kaiser (1958). After the oblique rotation, the simple structure of $\boldsymbol{B}$ facilitates the identification of cluster structures of variables. The steps for OPCC are as follows:

(i) Perform principal component analysis of all variables and find the first two PCs.
(ii) Oblique rotation of eigenvectors, $\boldsymbol{Z}$, to obtain the $\boldsymbol{B}$.
(iii) Calculate the linear correlation between all variables and the rotated components, and then assign each variable to one of the two clusters based on the higher squared correlation.
(iv) Repeat the binary split for each cluster.
(v) Stop the recursive split when the second eigenvalue is less than 1.

2.4 Simulation study

Both OPCC and HC methods are based on linear transformation and are limited in their capability to handle nonlinear interdependences among variables. Here, we show a motivating example to evaluate the performance of HC and OPCC for variable clustering. Four clusters of variables are generated as follows:

Cluster 1: $\{\boldsymbol{x}_1, \boldsymbol{x}_2 = |\boldsymbol{x}_1|, \boldsymbol{x}_3 = \boldsymbol{x}_1^2, \boldsymbol{x}_4 = \boldsymbol{x}_1^3, \boldsymbol{x}_5 = \boldsymbol{x}_1^4\}$;
Cluster 2: $\{\boldsymbol{x}_6, \boldsymbol{x}_7 = |\boldsymbol{x}_6|, \boldsymbol{x}_8 = \boldsymbol{x}_6^2, \boldsymbol{x}_9 = \boldsymbol{x}_6^3, \boldsymbol{x}_{10} = \boldsymbol{x}_6^4\}$;
Cluster 3: $\{\boldsymbol{x}_{11}, \boldsymbol{x}_{12} = \boldsymbol{x}_{11}(t+3), \boldsymbol{x}_{13} = \boldsymbol{x}_{11}(t+5), \boldsymbol{x}_{14} = \boldsymbol{x}_{11}(t+7), \boldsymbol{x}_{15} = \boldsymbol{x}_{11}(t+9)\}$;
Cluster 4: $\{\boldsymbol{x}_{16}, \boldsymbol{x}_{17} = \boldsymbol{x}_{16}(t+10), \boldsymbol{x}_{18} = \boldsymbol{x}_{16}(t+20), \boldsymbol{x}_{19} = \boldsymbol{x}_{16}(t+30), \boldsymbol{x}_{20} = \boldsymbol{x}_{16}(t+40)\}$.

where x_1 and x_6 are independent standard normal variables, x_{11} is a nonlinear variable sampled from logistic map $x_{11}(n+1) = 3.8x_{11}(n)(1 - x_{11}(n))$, x_{16} is a second-order autoregressive variable that is nonlinearly coupled with x_{Lorenz}, $x_{16}(n) = 1.095x_{16}(n-1) - 0.4x_{16}(n-2) + 0.7\varepsilon_n + 0.3x_{Lorenz}^2$, where $\varepsilon_n$is Gaussian noise, x_{Lorenz} is the x-component of a Lorenz system: $x' = 10(y - x)$, $y' = (28 - z) - y$, $z' = xy - \frac{8}{3}z$ with time step 0.01. The sample size of each variable is 1000.

Figure 2a shows the dendrogram of hierarchical clustering for the motivating example. Figure 2b shows the clustering results of OPCC. It may be noted that both HC and OPCC cannot identify the cluster structure of variables. This is mainly due to the fact that nonlinear and asymmetric interdependence structures among variables are not considered. Very little work has been done to cluster a large number of variables with complex structures of nonlinear and asymmetric interdependences. In order to tackle these issues and fill the gap, we propose a new strategy that integrates nonlinear coupling analysis with self-organizing networks for variable clustering and predictive modeling.

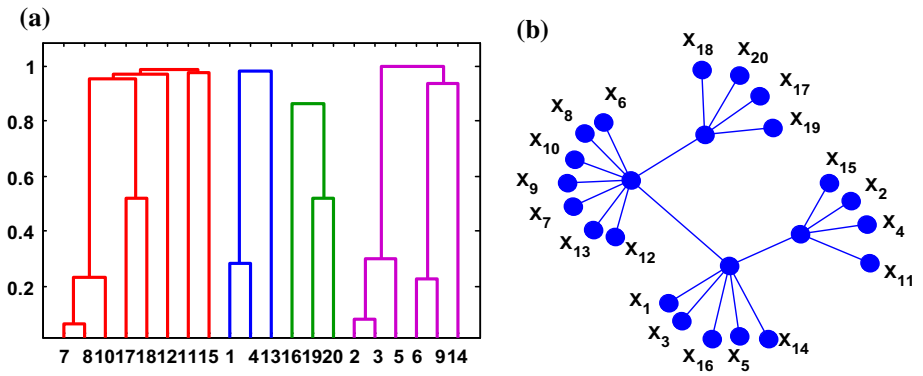

Fig. 2 Clustering of simulated variables using **a** HC and **b** OPCC

3 Research methodology

In this section, we will first introduce nonlinear coupling analysis to characterize and measure nonlinear interdependence structures among variables. Second, we develop a self-organizing network algorithm to cluster variables that involve nonlinear and asymmetric interdependences. Finally, we orthogonalize variables in each of self-organized clusters and then integrate them with group elastic-net model to improve the performance of predictive modeling. We will illustrate the proposed methodology using both simulation and real-world case studies in Sects. 4 and 5.

3.1 Nonlinear coupling analysis

In this investigation, we propose to characterize and quantify nonlinear interrelationships among variables. Traditionally, such interrelationships are estimated with methods such as correlation and mutual information. As aforementioned, correlation is a second-order quantity evaluating merely linear dependency among data. Mutual information quantifies both linear and nonlinear dependency between variables but requires stationarity in the computation. Both of them are limited in the ability to handle nonlinear and asymmetric interdependence structures.

Therefore, we performed nonlinear coupling analysis by exploiting cross recurrences between two variables in the feature space (Kantz and Schreiber 2003; Arnhold et al. 1999). The nonlinear measure is commonly used in neuroscience to study the interrelationship between neurons. Very little has been done to extend nonlinear coupling for variable clustering. Let $x_1(m)$ and $x_2(m)$ be the m-th observation of variables $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. Denote $\Upsilon(x_1(m))$ as the recurrence neighborhood of $x_1(m)$ containing the k closest neighbors of $x_1(m)$. Here, $\Upsilon(x_1(m))$ is the recurrence set $\{x_1(i)\}, i \in \{n_1, \ldots, n_k\}$. If there are some relations between two variables $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$, then the recurrence of $\boldsymbol{x}_1$ will also imply a recurrence of $\boldsymbol{x}_2$, at least with a greater than zero probability. In other words, $x_2(i)$ with the same indices $i \in \{n_1, \ldots, n_k\}$ should also be closer to $x_2(m)$ than the average of randomly chosen observations. Therefore, the quantity of nonlinear interdependence is defined as:

$$\hat{I}_{\boldsymbol{x}_1\boldsymbol{x}_2} = \left\langle \frac{r_m(\boldsymbol{x}_2) - d_m(\boldsymbol{x}_2|\boldsymbol{x}_1)}{r_m(\boldsymbol{x}_2) - d_m(\boldsymbol{x}_2)} \right\rangle_m \tag{6}$$

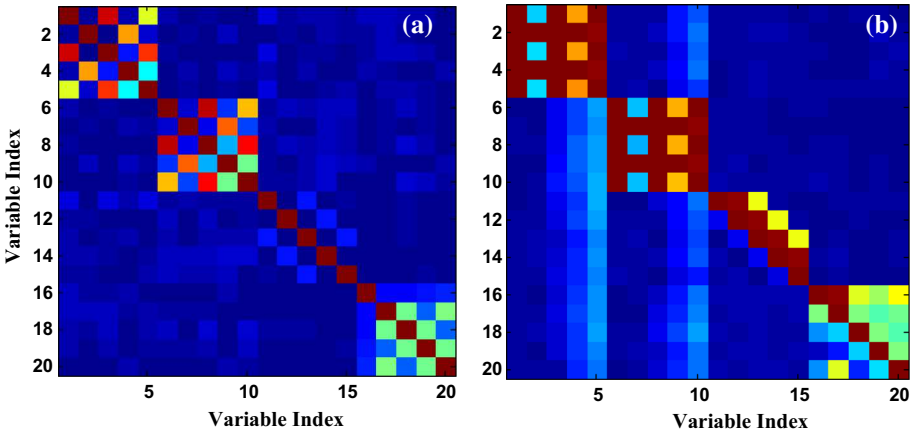

Fig. 3 Matrices of **a** linear correlations and **b** nonlinear interdependences

where $\langle\cdot\rangle_m$ is the ergodic average over the range of variable $\boldsymbol{x}_2$, and $r_m(\boldsymbol{x}_2)$ is the average distance from $x_2(m)$ to k randomly chosen $x_2(i)'$s, i.e., $r_m(\boldsymbol{x}_2) = \frac{1}{k}\sum_{i=1}^{k}(x_2(m) - x_2(i))^2$. Here, $d_m(\boldsymbol{x}_2|\boldsymbol{x}_1)$ is the average conditional distance from $x_2(m)$ to k samples of $\boldsymbol{x}_2$ whose indices $i \in \{n_1, \ldots, n_k\}$ are from the recurrence set $\mathcal{T}(x_1(m))$ of the variable $\boldsymbol{x}_1$:

$$d_m(\boldsymbol{x}_2|\boldsymbol{x}_1) = \frac{1}{k}\sum_{i\in\{n_1,\ldots,n_k\}}(x_2(m) - x_2(i))^2 \tag{7}$$

In addition, $d_m(\boldsymbol{x}_2)$ is the average distance from $x_2(m)$ to k closest neighbors of $x_2(m)$:

$$d_m(\boldsymbol{x}_2) = \frac{1}{k}\sum_{i\in\mathcal{V}(x_2(m))}(x_2(m) - x_2(i))^2 \tag{8}$$

where $\mathcal{V}(x_2(m))$ is the true neighborhood of $x_2(m)$. If $I_{x_1x_2}$ is small (close to zero), then there is no evident interdependence between x_1 and x_2, because true neighbors of $x_2(m)$ are much closer to $x_2(m)$ than those neighbors based on the recurrences in the $\boldsymbol{x}_1$ process. When $I_{x_1x_2}$ is close to unity, there is a strong interdependence between $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$. Figure 3 shows the matrices of both linear correlations and nonlinear interdependences among variables that are computed from the motivating example. The red (grey color along the diagonal in black and white print) color represents a high interdependence, while the blue (black color off diagonal in black and with print) color indicates no interrelationships. Note that nonlinear interdependence in Fig. 3b is significantly different from linear correlation in Fig. 3a. Nonlinear coupling analysis provides a better characterization of complex interdependence structures (i.e., nonlinear and asymmetric) among variables than linear correlations.

3.2 Self-organizing network for variable clustering

Figure 3b shows that nonlinear interdependence is not symmetric. Traditional similarity-based clustering algorithms are not applicable. Latent-variable methods using oblique PCA or factor analysis do not fully consider nonlinear interdependences among variables. To tackle these challenges, we develop a self-organizing network algorithm to cluster variables that involve nonlinear and asymmetric interrelationships. Notably, this present investigation

extends our previous work from self-organizing topology of recurrence networks (Yang and Liu 2013) to self-organizing clustering of highly-redundant variables.

In the literature, very little work has been done to cluster variables with complex nonlinear and asymmetric interdependences. We propose to treat variables as nodes in the network and nonlinear interdependences between variables as the weights of edges, which is varying from 0 to 1. Let $\boldsymbol{G} = \{\boldsymbol{V}, \boldsymbol{E}\}$ be the directed and weighted network, where $\boldsymbol{V}$ is the set of nodes and $\boldsymbol{E}$ is the set of edges. The spring-electrical model assigns two forces, i.e., attractive and repulsive forces between nodes. The repulsive force exists between any pair of nodes while the attractive forces exists between the nodes that have a relation of nonlinear interdependence. The repulsive force is defined as

$$f_r(i, j) = -\frac{1}{||\boldsymbol{s}(i) - \boldsymbol{s}(j)^2||} * \frac{1}{e^{\alpha\left|I_{x_i x_j}\right|}} \tag{9}$$

where α is a system parameter, $\mathbf{s}(i)$ and $\mathbf{s}(j)$ are spatial locations of node i and node j. The repulsive force is inversely proportional to nonlinear interdependence between two nodes (variables), because a bigger repulsive force is expected to separate two nodes when they have a smaller interdependence. The attractive force is defined as

$$f_a(i, j) = ||\boldsymbol{s}(i) - \boldsymbol{s}(j)||^2 * e^{\gamma\left|I_{x_i x_j}\right|}, I_{\boldsymbol{x}_i \boldsymbol{x}_j} \neq 0 \tag{10}$$

where γ is the system parameter. The attractive force is proportional to nonlinear interdependence between two nodes (variables), because a bigger attractive force will pull two nodes closer when they have a higher interdependence. The combined force on a node i is the summation of all repulsive forces and attractive forces on the node:

$$f(i, \boldsymbol{s}, \alpha, \gamma) = \sum_{i \neq j} -\frac{\frac{1}{e^{\alpha\left|I_{x_i x_j}\right|}}}{\boldsymbol{s}(i) - \boldsymbol{s}(j)^3}(\boldsymbol{s}(i) - \boldsymbol{s}(j)) + \sum_{i \leftrightarrow j} e^{\gamma\left|I_{x_i x_j}\right|}\boldsymbol{s}(i) - \boldsymbol{s}(j)(\boldsymbol{s}(i) - \boldsymbol{s}(j)) \tag{11}$$

where $\boldsymbol{s}(i) - \boldsymbol{s}(j)$ is the force-directional vector, which is separated from $f_r(i, j)$ and $f_a(i, j)$ to define the direction of combined force $f(i, \boldsymbol{s}, \alpha, \gamma)$. The attractive and repulsive forces drive the network to self-organize and form a topological structure. The objective of self-organizing process is to identify spatial locations of nodes by minimizing the energy of the network, i.e., the summation of squared combined force on each node:

$$\boldsymbol{s}^* = \arg\min_{\boldsymbol{S}} \left\{ \sum_{i \in \boldsymbol{N}} f^2(i, \boldsymbol{s}, \alpha, \gamma) \right\} \tag{12}$$

The self-organizing process derives a topological structure of the network of variables by minimizing the total energy, thereby clustering homogeneous variables into sub-network communities. Our prior work showed that the variations of system parameters α and γ will not change the structure of the clusters but yield a similar structure only in difference scales (Yang and Liu 2013). Table 2 summarizes the algorithm of self-organizing variable clustering. Note that there are three scenarios for updating the magnitude of move step for network nodes: (i) If the network energy keeps decreasing for 5 iterations, the magnitude will be increased to $\Delta/0.9$; (ii) If the network energy increases in one iteration, the magnitude will be decreased to $0.9 \times \Delta$; (iii) otherwise, the magnitude of move step remains the same.

Figure 4 shows the self-organizing process for clustering 20 variables in the motivating example. In the beginning, 20 variables are randomly distributed in the 3-dimensional space (see Fig. 4a). The topological structures after 200 and 400 iterations are shown in Fig. 4b,

Table 2 The pseudo-agorithm of self-organizing variable clustering

1:	Calculate nonlinear interdependence matrix $\boldsymbol{I}$
2:	Construct G – a directed and weighted network
3:	Initialize the layout of network in the high-dimensional space
4:	Set *flag* as *False* and the magnitude of move step as Δ
5:	**while** *flag* is *False* **do**
6:	//Calculate the energy of network structure EN
7:	Initialize EN as 0
8:	**for** all node i in G **do**
9:	Calculate the combined force on node i
10:	Move node i along the direction of combined force for a small step (magnitude Δ)
11:	Add the square of combined force to energy EN
12:	**end for**
13:	Update the magnitude of move step Δ *//three update scenarios*
14:	**if** the change of network energy is less than the tolerance ε
15:	Set *flag* as True
16:	**end if**
17:	**end while**
18:	**return** spatial structure of the network $\boldsymbol{s}$

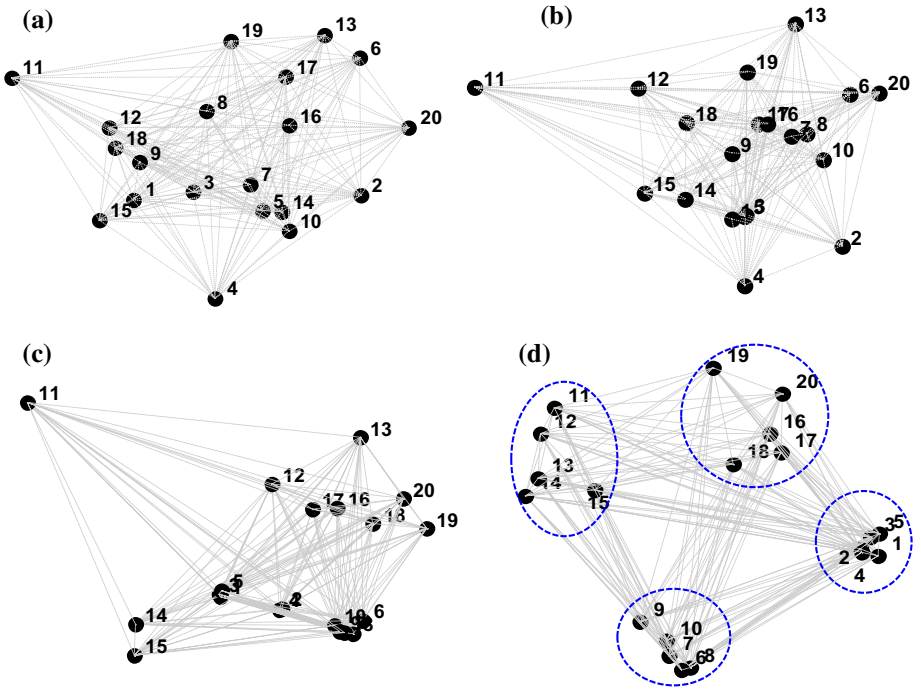

Fig. 4 Self-organizing network for clustering 20 variables: **a** initial topological structure, **b** topological structure after 200 iterations, **c** topological structure after 400 iterations and **d** final topological structure. Note that the computing time for this simulation study is approximately 17 s. The animation video is available at the following link: https://youtu.be/BwgjK8t7Pso

c. The final structure after 600 iterations is shown in Fig. 4d. After 600 iterations, the self-organizing process converges and identifies the underlying cluster structures of 20 variables, which demonstrates the superior performance of self-organizing networks over HC and OPCC (also see Fig. 1).

Our proposed self-organizing network shares some similarities with minimum energy design (Joseph et al. 2014) in the field of design of experiments and self-organizing map (Kohonen 1990; Chen and Yang 2012) in the domain of neural network. However, the spatial location of a design point will not change in the minimum energy design when the experiment has been conducted at this setting. The algorithm will optimize the spatial location of the next design point given that spatial locations of previous design points are fixed. Also, the proposed approach of self-organizing variable clustering is vastly different from the self-organizing map in neural network, which learns self-organizing positions of neurons based on distance measures in the data space. Nonetheless, our proposed research seeks to self-organize the data space of variables.

3.3 Predictive modeling with highly-redundant variables

The self-organizing network drives highly-redundant variables into sub-network clusters. The variables in each cluster bring the redundant information. Note that traditional regularization methods tend to select one variable from the cluster of highly redundant variables, while overlooking other variables. However, extra information does exist to some extent among variables in spite of redundancy. It is necessary to delineate the structure of latent variables hidden in each cluster of homogeneous variables. As such, we propose to minimize the information redundancy within the same cluster before grouped variables are used in predictive models. One idea is to use Gram–Schmidt orthonormalization (GSO) or PCA to maximize the information extraction. Note that GSO and PCA only orthogonalize the variables and make them linearly independent. Such a transformation of data avoids the dropout of variables and improves the extraction of useful information between variables.

Assume we have M clusters and there are K variables, $x_{m1}, x_{m2}, \ldots, x_{mK}$, in the m-th cluster. The GSO minimizes the redundant information by transforming original variables $(\boldsymbol{x}_{m1}, \boldsymbol{x}_{m2}, \ldots, \boldsymbol{x}_{mK})$ into the orthonormal set of new variables $(\boldsymbol{w}_{m1}, \boldsymbol{w}_{m2}, \ldots, \boldsymbol{w}_{mK})$ in each cluster. It begins by normalizing $\boldsymbol{x}_{m1}$,

$$\boldsymbol{v}_{m1} = \boldsymbol{x}_{m1}; \quad \boldsymbol{w}_{m1} = \frac{\boldsymbol{v}_{m1}}{||\boldsymbol{v}_{m1}||} \tag{13}$$

where $\boldsymbol{w}_{m1}$ is the normalized variable of x_{m1}. The second orthogonal vector v_{m2} is obtained as,

$$\boldsymbol{v}_{m2} = \boldsymbol{x}_{m2} - \langle \boldsymbol{x}_{m2}, \boldsymbol{w}_{m1} \rangle \boldsymbol{w}_{m1}; \quad \boldsymbol{w}_{m2} = \frac{\boldsymbol{v}_{m2}}{||\boldsymbol{v}_{m2}||} \tag{14}$$

where w_{m2} is the second orthonormalized vector. The process is repeated to get the k-th orthogonal vector v_{mk},

$$\boldsymbol{v}_{mk} = \boldsymbol{x}_{mk} - \sum_{i=1}^{k-1} \langle \boldsymbol{x}_{mk}, \boldsymbol{w}_{mi} \rangle \boldsymbol{w}_{mi}; \quad \boldsymbol{w}_{mk} = \frac{\boldsymbol{v}_{mk}}{||\boldsymbol{v}_{mk}||} \tag{15}$$

where w_{mk} is the k-th orthonormalized vector.

After variables are orthonormalized in each cluster, information redundancy, at least linear dependency, is minimized to some extent. Further, a group elastic-net model is utilized for variable selection that penalizes in the levels of both individual variables and groups. First,

the logistic regression model is constructed between predictor and response variables as:

$$h_\beta(w) = \frac{1}{1+\exp\left(-\beta^T \boldsymbol{W}\right)}, 0 < h_\beta(w) < 1 \tag{16}$$

where β is model parameter and $h_\beta(w)$ is the estimated probability to get a response when the predictor is $\boldsymbol{W}$. Because the predictors are clustered into M homogeneous groups and each group has K_m variables, we rewrite this model with clustered variables as:

$$h_\beta(w) = \frac{1}{1+\exp\left[-\left(\beta_0 + \left(\sum_{m=1}^{M}\sum_{k=1}^{K_m} w_k \beta_k\right)\right)\right]} \tag{17}$$

If this is a multi-class classification problem, then we will build a separate logistic regression classifier $h_\beta^{(c)}(w)$ for each class c to predict the probability that the response variable $y = c$. When a new input $\boldsymbol{W}$ comes to make a prediction, the class c with the maximal probability $h_\beta^{(c)}(w)$ will be chosen. In this present investigation, we focus on the binary response variable $y = 0$ or 1, i.e., diseased or healthy subjects. The log-likelihood function of logistic regression model is then formulated as:

$$J(\beta) = \sum_{i=1}^{n}\left[y_i \log\left(h_\beta(w,i)\right) + (1-y_i)\log\left(1-h_\beta(w,i)\right)\right]^2 \tag{18}$$

When performing the maximum likelihood estimation, we add the group elastic-net regularization to penalize insignificant cluster of variables and identify key predictors as:

$$\max_{\boldsymbol{\beta}} J(\beta)$$
$$s.t. \sum_{m=1}^{M}\sum_{k}^{K_m}\left(\alpha \boldsymbol{\beta}_k^2 + (1-\alpha)\left|\boldsymbol{\beta}_k\right|\right) \le \lambda \tag{19}$$

where α and λ are regularization parameters. Note that Eq. (19) adds the penalizations in the levels of both individual variables and groups. The integration of orthonormalization with group elastic-net model avoids the leave out of variables in traditional LASSO models and improves the extraction of useful information between variables. In the next two sections, we will evaluate and validate the proposed approach using both simulation and real-world case studies.

4 Experimental design and results

In this section, simulation experiments were designed to evaluate the performance of self-organizing variable clustering (SOC) algorithms. On the basis of 20 variables in the motivating example, we further utilized cubic spline functions to expand each original variable into a basis matrix for representing the family of piecewise polynomials, then these polynomials (or basis functions) corresponding to the same original variable form a natural group. Therefore, a total of 60 variables is generated in 4 groups to evaluate the proposed methodology. The logistic function $h_\beta(x)$ is computed using a sparse set of variables selected from the original 20 variables:

$$h_\beta(x) = \frac{1}{1+\exp\left[-\left(\sum_{i=1}^{20\times p_s} \boldsymbol{x}_i \beta_i + c\varepsilon\right)\right]} \tag{20}$$

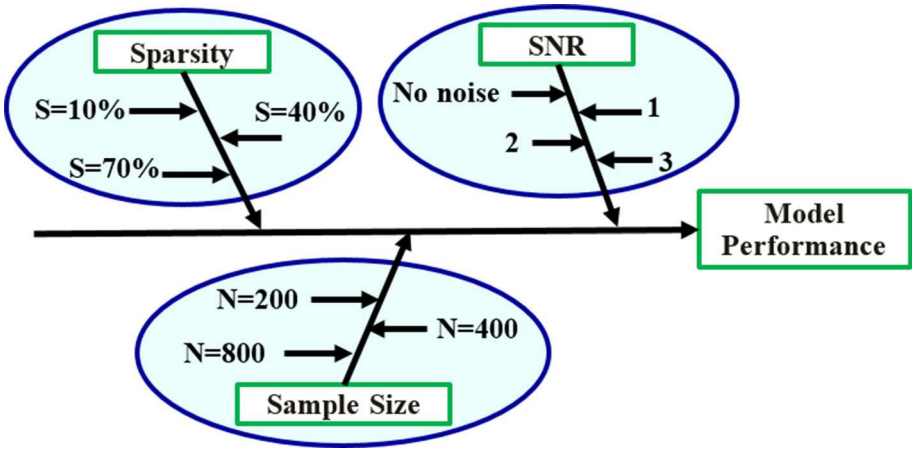

Fig. 5 Cause-and-effect diagram for performance evaluation of the proposed self-organized variable clustering algorithm

where p_s denotes the sparsity level (i.e., the percentage of variables involved to derive the response), c is the amplitude of random noises, and β_j is the model parameter that follows the distribution of $N(\mu_i, 0.1)$, $\mu_i = 1 + (i-1) \times 0.3$, $i = 1, 2, \ldots, 20$. The decision boundary $h_\beta(x) = 0.5$ is used to generate the binary response variable y.

As shown in Fig. 5, a 3-way layout experiment was designed to evaluate the performance of SOC algorithms with three factor groups, i.e., signal-noise-ratio (SNR), sparsity, and sample size. The sample sizes of training set n_1, validation set n_2 and testing set n_3 are varied in three levels (i.e., 50/50/100, 100/100/200, and 200/200/400). The sparsity level p_s is changed from 10, 40 to 70%. The SNR level is varied from no noise, 1, 2, to 3, where SNR is the power ratio between the signal and the background noise, i.e., $\text{var}\left(\sum_{i=1}^{20\times p_s} \boldsymbol{x}_i \beta_i\right) / \text{var}(c\varepsilon)$. As a result, we generated 36 treatment levels of experimental factors and further evaluated and validated the performance of clustering algorithms. In this present investigation, we compared the performance of SOC with no clustering, HC, OPCC. The training dataset is used to train the predictive model, and the validation dataset is to optimally select the penalization parameters in Eq. (17). Model performance is only computed from the testing dataset. Each treatment level is replicated for 100 times.

Table 3 summarizes the averages and standard deviations of prediction errors from 100 replicates with GSO of variables in each cluster. Note that the numbers in the parenthesis are the standard deviations over 100 replicates. As shown in Table 2, all variable clustering methods yield better performance than no clustering of variables. In particular, self-organizing variable clustering outperforms the other two methods and achieves a relatively better performance. No clustering yields the highest prediction errors at all the 36 treatment levels. The comparison results are not surprising because Figs. 2 and 4 show that self-organizing clustering identifies cluster structures better than HC and OPCC. The proposed self-organizing algorithm achieves better performance in 23 out of 36 treatment levels in the experiments. Characterizing nonlinear interdependence structures among variables helps improve the performance of predictive modeling. Also, Table 2 shows that the prediction error decreases as more data samples are available for variable clustering and predictive modeling. Adding noises to data deteriorates the predictive performance of models. Finally, it is worth mentioning that model performance is better when the sparsity level is lower.

Table 3 Averages and standard deviation of prediction errors in the experimental study with 100 replications (GSO)

SNR	Sample size	No clustering			GSO–HC			GSO–OPCC			GSO–SOC		
		S = 0.1	S = 0.4	S = 0.7	S = 0.1	S = 0.4	S = 0.7	S = 0.1	S = 0.4	S = 0.7	S = 0.1	S = 0.4	S = 0.7
1	n1 = 50,	0.361	0.489	0.508	0.278	0.413	0.406	0.259	**0.335**	0.381	**0.252**	0.340	**0.381**
	n2 = 50, n3 = 100	(±0.06)	(±0.05)	(±0.05)	(±0.05)	(±0.09)	(±0.09)	(±0.05)	(±**0.07**)	(±0.09)	(±**0.03**)	(±0.09)	(±**0.06**)
	n1 = 100,	0.351	0.456	0.430	0.276	0.304	0.299	**0.251**	**0.301**	**0.295**	0.278	0.335	0.307
	n2 = 100, n3 = 200	(±0.05)	(±0.04)	(±0.06)	(±0.03)	(±0.04)	(±0.04)	(±**0.03**)	(±**0.04**)	(±**0.03**)	(±0.03)	(±0.06)	(±0.03)
	n1 = 200,	0.283	0.328	0.337	0.270	0.298	0.280	0.264	0.294	**0.255**	**0.259**	**0.266**	0.283
	n2 = 200, n3 = 400	(±0.02)	(±0.02)	(±0.03)	(±0.02)	(±0.02)	(±0.02)	(±0.02)	(±0.02)	(±**0.02**)	(±**0.02**)	(±**0.02**)	(±0.02)
2	n1 = 50,	0.203	0.345	0.438	0.157	0.271	0.263	0.161	0.280	0.255	**0.120**	**0.192**	**0.230**
	n2 = 50, n3 = 100	(±0.05)	(±0.11)	(±0.08)	(±0.03)	(±0.10)	(±0.08)	(±0.03)	(±0.08)	(±0.05)	(±**0.02**)	(±**0.05**)	(±**0.05**)
	n1 = 100,	0.214	0.263	0.284	0.147	0.171	0.188	0.154	**0.165**	0.194	**0.136**	0.199	**0.180**
	n2 = 100, n3 = 200	(±0.02)	(±0.03)	(±0.05)	(±0.02)	(±0.02)	(±0.03)	(±0.02)	(±**0.03**)	(±0.03)	(±**0.02**)	(±0.03)	(±**0.02**)
	n1 = 200,	0.175	0.236	0.227	0.154	0.171	0.174	**0.130**	0.172	**0.167**	0.131	**0.170**	0.185
	n2 = 200, n3 = 400	(±0.01)	(±0.02)	(±0.02)	(±0.01)	(±0.02)	(±0.01)	(±**0.01**)	(±0.02)	(±**0.01**)	(±0.01)	(±**0.01**)	(±0.02)
3	n1 = 50,	0.197	0.355	0.436	0.116	0.169	0.240	0.098	0.210	0.247	**0.097**	**0.165**	**0.236**
	n2 = 50, n3 = 100	(±0.04)	(±0.10)	(±0.08)	(±0.03)	(±0.05)	(±0.07)	(±0.03)	(±0.08)	(±0.05)	(±**0.03**)	(±**0.04**)	(±**0.06**)
	n1 = 100,	0.156	0.262	0.269	0.100	0.154	0.168	0.105	0.137	0.158	**0.099**	**0.135**	**0.157**
	n2 = 100, n3 = 200	(±0.02)	(±0.03)	(±0.04)	(±0.02)	(±0.04)	(±0.02)	(±0.02)	(±0.02)	(±0.02)	(±**0.02**)	(±**0.02**)	(±**0.02**)
	n1 = 200,	0.133	0.194	0.187	0.084	0.138	0.136	0.088	0.150	**0.114**	**0.072**	**0.116**	0.133
	n2 = 200, n3 = 400	(±0.01)	(±0.02)	(±0.02)	(±0.01)	(±0.01)	(±0.02)	(±0.01)	(±0.01)	(±**0.01**)	(±**0.01**)	(±**0.01**)	(±0.01)
No noise	n1 = 50,	0.127	0.304	0.345	0.058	0.166	0.151	**0.051**	0.135	0.146	0.065	**0.109**	**0.145**
	n2 = 50, n3 = 100	(±0.03)	(±0.10)	(±0.09)	(±0.02)	(±0.08)	(±0.05)	(±**0.03**)	(±0.08)	(±0.03)	(±0.02)	(±**0.04**)	(±**0.04**)
	n1 = 100,	0.109	0.199	0.206	0.043	0.063	0.101	**0.043**	**0.056**	**0.095**	0.052	0.062	0.099
	n2 = 100, n3 = 200	(±0.02)	(±0.03)	(±0.03)	(±0.02)	(±0.02)	(±0.02)	(±**0.01**)	(±**0.02**)	(±**0.03**)	(±0.02)	(±0.02)	(±0.03)
	n1 = 200,	0.072	0.145	0.163	0.026	0.054	0.067	0.030	0.056	0.066	**0.023**	**0.044**	**0.066**
	n2 = 200, n3 = 400	(±0.01)	(±0.01)	(±0.02)	(±0.01)	(±0.01)	(±0.02)	(±0.01)	(±0.02)	(±0.02)	(±**0.01**)	(±**0.01**)	(±**0.02**)

Bold values indicate the performances with smallest averages we can achieve at the same settings of SNR, sparsity and sample size

Table 4 shows the averages and standard deviations (i.e., the numbers in the parenthesis) of prediction errors from 100 replicates in each treatment level with PCA of variables in each cluster. Similar to the results in Table 3, the proposed SOC algorithm yields better performance than HC and OPCC algorithms in 23 out of 36 treatment levels in the experiments. The performances of HC and OPCC are close to each other, which is because linear correlation and PCA processing of variables are utilized in both cases. In terms of experimental factors (i.e., sample size, noise and sparsity), Table 4 shows consistent results as in Table 3. The prediction error decreases as the sample size increases. Adding noises to data deteriorates the predictive performance of models. When the sparsity level is lower, the model performance is better.

In addition, if we compared the GSO with PCA-based orthogonalization of variables within each cluster, it may be noted that their performance is similar. For 36 treatment levels, the GSO–SOC approach yields 18 settings that have lower prediction errors than the PCA–SOC approach. The differences between GSO–SOC and PCA–SOC approaches are not statistically significant. However, both Tables 3 and 4 show that the self-organizing network algorithm significantly decreases the errors in predictive models. This demonstrates that variable clustering is critical to improving the performance of predictive models. Further, experimental results show that variable clustering that considers nonlinear correlations among variables yields better results than the one with linear correlations.

5 Model-driven predictive model of space–time vectorcardiogram signals

Furthermore, we evaluate and validate the proposed methodology using a real-world case study that utilizes the extracted parameters from representation models of VCG signals for predictive modeling of myocardial infarctions. Our previous study developed a sparse basis function model to represent 3-lead VCG signals, which minimizes the number of basis functions involved but maintains sufficient explanatory power. As such, large amounts of data are reduced to a parsimonious set of model parameters (i.e., weight, shifting and scaling factors in basis functions) while preserving the information (Liu and Yang 2013). Further, we used model parameters and their derivatives as predictors to build a lasso-penalized logistic regression model for the prediction of cardiac disorders (Liu et al. 2014). However, experimental results show that there are high levels of correlation among parametric features, which lead to sensitive predictive models (i.e., increased variances of estimation). The novelty of this present study is the development of topological network algorithms for handling both redundancy and relevancy among variables and improving the performance of predictive modeling.

As shown in Fig. 6a, cardiac electrical activity is varying across space and time. VCG signals monitor cardiac electrical activity along three orthogonal X, Y, Z planes of the body, namely, frontal, transverse, and sagittal (see Fig. 6b) (Yang et al. 2011, 2012; Yang 2011). Within one cycle, the VCG waveform shows nonlinear variations. Each VCG cycle includes P, QRS and T segments that correspond to different stages of cardiac operations. Between cycles, the VCG waveform is similar to each other but with variations. As shown in Fig. 6b, VCG trajectories of myocardial infarction (red/dashed) yield a different spatial path from the healthy controls (blue/solid) (Yang et al. 2013). In order to reduce large amounts of VCG signals into a sparse set of parameters, we developed the basis function representation of VCG signals (see Fig. 6c). This present paper leverages parametric features for predicting the incidence of myocardial infarction (see Fig. 6d, e). Because the set of parametric features

Table 4 Averages and standard deviation of prediction errors in the experimental study with 100 replications (PCA)

SNR	Sample size	PCA–HC			PCA–OPCC			PCA–SOC		
		S = 0.1	S = 0.4	S = 0.7	S = 0.1	S = 0.4	S = 0.7	S = 0.1	S = 0.4	S = 0.7
1	n1 = 50,	**0.2228**	0.3316	**0.3690**	0.2442	0.3963	0.4271	0.2530	**0.3298**	0.3938
	n2 = 50, n3 = 100	(±**0.0335**)	(±0.0727)	(±**0.0716**)	(±0.0338)	(±0.0804)	(±0.0669)	(±0.0382)	(±**0.0783**)	(±0.0856)
	n1 = 100,	**0.2466**	**0.2750**	0.3359	0.3266	0.3727	0.3070	0.2470	0.3279	**0.2888**
	n2 = 100, n3 = 200	(±**0.0275**)	(±**0.0247**)	(±0.0515)	(±0.0456)	(±0.0635)	(±0.0364)	(±0.0270)	(±0.0499)	(±**0.0327**)
	n1 = 200,	0.2621	0.2927	0.2844	0.2220	**0.2740**	**0.2733**	**0.2151**	0.2957	0.2809
	n2 = 200, n3 = 400	(±0.0181)	(±0.0179)	(±0.0216)	(±0.0164)	(±**0.0194**)	(±**0.0168**)	(±**0.0158**)	(±0.0196)	(±0.0200)
2	n1 = 50,	0.1542	0.2362	0.2623	0.1577	0.2364	**0.2175**	**0.1398**	**0.2269**	0.2659
	n2 = 50, n3 = 100	(±0.0329)	(±0.0743)	(±0.0657)	(±0.0310)	(±0.0563)	(±**0.0479**)	(±**0.0280**)	(±**0.0659**)	(±0.0620)
	n1 = 100,	0.1518	0.2039	0.2007	0.1500	0.1759	**0.1834**	**0.1433**	**0.1600**	0.2082
	n2 = 100, n3 = 200	(±0.0219)	(±0.0260)	(±0.0283)	(±0.0201)	(±0.0255)	(±**0.0230**)	(±**0.0188**)	(±**0.0256**)	(±0.0372)
	n1 = 200,	0.1395	0.1633	0.1698	0.1377	0.1852	0.1671	**0.1327**	**0.1577**	**0.1593**
	n2 = 200, n3 = 400	(±0.0140)	(±0.0151)	(±0.0151)	(±0.0144)	(±0.0179)	(±0.0181)	(±**0.0125**)	(±**0.0142**)	(±**0.0159**)
3	n1 = 50,	0.1064	0.2010	0.2428	0.1253	**0.1518**	0.2409	**0.0960**	0.1653	**0.2330**
	n2 = 50, n3 = 100	(±0.0281)	(±0.0537)	(±0.0547)	(±0.0239)	(±**0.0534**)	(±0.0526)	(±**0.0267**)	(±0.0479)	(±**0.0578**)
	n1 = 100,	0.0966	0.1378	0.1877	**0.0956**	**0.1352**	0.1710	0.1097	0.1419	**0.1691**
	n2 = 100, n3 = 200	(±0.0193)	(±0.0231)	(±0.0236)	(±**0.0176**)	(±**0.0204**)	(±0.0243)	(±0.0184)	(±0.0216)	(±**0.0270**)
	n1 = 200,	0.0838	0.1261	0.1338	**0.0779**	**0.1236**	0.1397	0.0937	0.1317	**0.1255**
	n2 = 200, n3 = 400	(±0.0115)	(±0.0121)	(±0.0151)	(±**0.0114**)	(±**0.0134**)	(±0.0149)	(±0.0121)	(±0.0141)	(±**0.0148**)
No noise	n1 = 50,	0.0576	0.1458	0.1643	0.0605	0.1228	0.1607	**0.0547**	**0.1178**	**0.1470**
	n2=50,n3=100	(±0.0208)	(±0.0345)	(±0.0338)	(±0.0238)	(±0.0600)	(±0.0353)	(±**0.0225**)	(±**0.0617**)	(±**0.0390**)
	n1 = 100,	0.0429	0.0686	0.1125	0.0425	0.0684	0.1052	**0.0388**	**0.0506**	**0.0982**
	n2 = 100, n3 = 200	(±0.0139)	(±0.0209)	(±0.0253)	(±0.0158)	(±0.0212)	(±0.0258)	(±**0.0135**)	(±**0.0195**)	(±**0.0302**)
	n1 = 200,	0.0255	0.0515	0.0647	0.0289	0.0468	0.0612	**0.0205**	**0.0461**	**0.0610**
	n2 = 200, n3 = 400	(±0.0085)	(±0.0149)	(±0.0182)	(±0.0101)	(±0.0132)	(±0.0154)	(±**0.0085**)	(±**0.0119**)	(±**0.0165**)

Bold values indicate the performances with smallest averages we can achieve at the same settings of SNR, sparsity and sample size

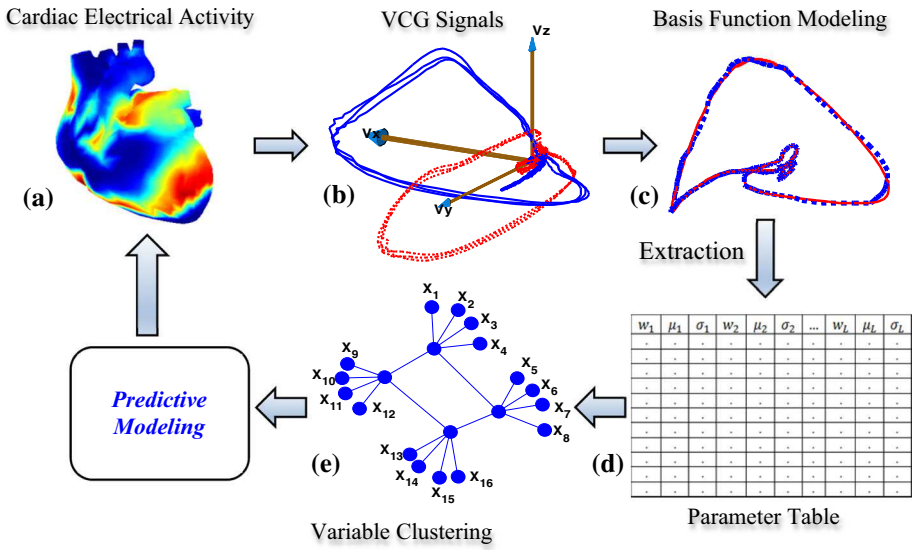

Fig. 6 Flow chart of a real-world case study that extracts model parameters from VCG signals for the identification of myocardial infarctions

contains redundant information that inflates the variance of predictive models, this motivates our further development of the proposed methodology of self-organizing variable clustering.

5.1 Multiscale basis function modeling of VCG signals

Our previous work developed a sparse basis function model to characterize and represent 3-dimensional VCG signals (Liu and Yang 2013). Such a sparse representation reduces large amounts of data to a limited number of model parameters while preserving the same information. This present paper will further develop predictive models of myocardial infarctions using the low-dimensional set of model parameters, as opposed to the original data itself. Figure 6c shows an example of the basis function model of 3D trajectories of VCG signals. In order to capture intrinsic characteristics of cardiac electrical activity, we modeled VCG signals as the superposition of M basis functions:

$$v(t) = w_0 + \sum_{i=1}^{M} w_j \psi_j \left((t - \mu_j)/\sigma_j\right) + \varepsilon \tag{21}$$

where $\psi(t)$ is the basis function, w_j is the weight factor, $\mu_j$is the shifting factor and σ_j is the scaling factor. The objective is to optimize the representation of 3D VCG signals with a sparse basis function model:

$$\text{argmin}\left[\left\| v(t) - w_0 - \sum_{i=1}^{M} w_j \psi_j(t) \right\|^2, \{w, M, \psi(t)\}\right] \tag{22}$$

Compact topological representation calls upon the minimization of the number of basis functions M and the optimal placement of basis function$\psi(t)$. Model parameters w, μ, σ are adaptively estimated by "best matching" projections of VCG signals onto a dictionary of nonlinear basis functions. Our previous work detailed the modeling algorithms to develop a sparse basis function representation of spatiotemporal VCG signals (Liu and Yang 2013).

In addition, our previous experiments show that model goodness-of-fit is greater than 99.9% (R^2) with a parsimonious set of 20 basis functions for a variety of cardiac conditions. In this present study, model parameters, i.e., weight, shifting, scaling factors and residuals, will be further investigated for the identification of myocardial infarctions.

5.2 Predictive modeling of myocardial infarction

This case study focuses on the extraction of parametric features from the sparse basis function model, and their further applications for predictive modeling of myocardial infarctions. If M basis functions are used to represent 3-lead VCG signals, then the set of parameters (i.e., weight, shifting and scaling factors) is $\{w_{3\times M}, \mu_{3\times M}, \sigma_{3\times M}\}$. The total number of parameters will be $3 \times 3 \times M$. Our previous study showed that 20 basis functions yield >99.9% goodness-of-fit in the modeling performance for a variety of cardiac conditions. Hence, we have a total of 180 model parameters that are adaptively estimated from the 3D VCG trajectory. In addition, we added the absolute values of weights, residual sum of squares (RSS) and the RR interval (i.e., heart rate) in this present investigation. The feature matrix is:

$$\boldsymbol{X} = \left\{\boldsymbol{w}_{3\times 20}, \boldsymbol{\mu}_{3\times 20}, \boldsymbol{\sigma}_{3\times 20}, |\boldsymbol{w}|_{3\times 20}, \boldsymbol{RSS}_{3\times 1}, RR_{1\times 1}\right\}$$

where the absolute values of weights $|\boldsymbol{w}|_{3\times 20}$ describes the amplitudes of each basis function, which provide local strengths of a heartbeat. The residual sum of squares $\boldsymbol{RSS}_{3\times 1}$ describes the variations that cannot be adequately explained by the model representation. The RR interval characterizes temporal beat-to-beat variations of cardiac electrical activity. In total, we have 244 parameter-based features that provide effective measures of original VCG signals. Notably, model representation reduces the high-dimensional set of VCG signals into a sparse set of feature matrix.

In this present investigation, we used 3-lead VCG signals from 388 subjects (79 healthy controls and 309 myocardial infarctions), available in the PhysioNet PTB Database (Goldberger et al. 2000). Each recording contains 15 simultaneous heart monitoring signals, i.e., 12-lead ECG and 3-lead VCG. The signals were digitized at 1 kHz sampling rate with a 16-bit resolution over a range of 16.384 mV. The VCG recordings are typically about 2 min duration, and all signals were recorded for at least 30 s. Our previous study showed that most of model-driven parametric features are statistically significant between healthy controls and myocardial infarctions (Liu et al. 2014). Specifically, our experimental results showed that more than 146 features have the Kolmogorov–Smirnov statistic greater than the critical value 0.17, indicating significant differences between control and diseased conditions. It is worth mentioning that weight factors are the most significant group of features among all parametric features. However, a large number of predictors tend to bring the "curse of dimensionality" problem, as well as the overfitting for the predictive modeling. Therefore, our previous study utilized the lasso-penalized logistic regression model to shrink the number of predictors and identify the cases of myocardial infarction.

Nonetheless, our previous study Liu et al. (2014) focused on the relevancy between predictor variables and the response variables, but did not specifically consider complex interdependence structures among predictor variables. Prior research showed that a higher correlation (>0.90) between variables (collinearity) leads to more sensitive estimations of parameters in predictive models (i.e., increased variances of estimation) (Nas and Mevik 2001). This present paper further investigates the nonlinear correlations between variables and then identifies the cluster structures for improving the performance of predictive modeling. Figure 7a shows the plot of nonlinear and asymmetric interdependence structures

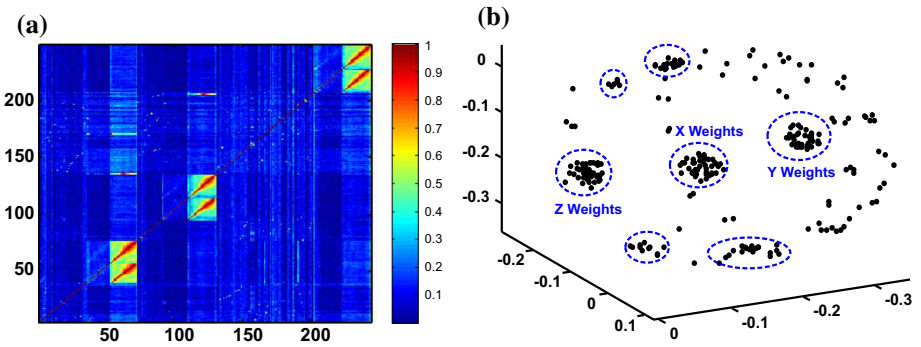

Fig. 7 **a** Nonlinear interdependence matrix; **b** self-organized clustering results of model-based parametric features

among variables. It may be noted that there are three groups of variables with stronger interdependences, and also some groups of variables with weaker interdependence relationships. However, few, if any, previous work has explicitly considered such relationships among variables before predictive modeling. Figure 7b shows sub-network communities in the self-organized clustering of model-based parametric features. Notably, self-organizing algorithms derive a topological structure of the network of variables based on the matrix of nonlinear and asymmetric interdependences in Fig. 7a. As a result, homogeneous variables are clustered into sub-network communities.

Furthermore, we minimized the redundancy of variables within each self-organized cluster through GSO and then integrated this new set of clustered variables with group elastic-net model to improve the performance of predictive modeling. The available data are divided into three parts: a training set (25% of samples), a validation set (25% of samples), and a test set (50% of samples). The training set is to train the predictive model, and the validation set is to cross-validate the model and optimally determine the penalization parameters (also see Eq. 17). Model performance is only computed from the test dataset. The experiments were replicated for 100 times. Figure 8 shows the averages and standard deviation of prediction errors in the real-world case study. "Without clustering" represents the results from the lasso-penalized logistic regression model in our previous study Liu et al. (2014), while "SOC clustering" denotes the results from the present study with self-organizing variable clustering. As shown in Fig. 8, self-organizing variable clustering yields smaller standard deviations of performance metrics (i.e., accuracy, sensitivity, and specificity) than "without clustering". Also, the average accuracy is improved from 89.50 to 94.71%, the average sensitivity is improved from 94.33 to 96.80%, and the average specificity is increased from 84.80 to 92.62%. In addition, our experiments showed that OPCC yields an average accuracy of 93.76% (sensitivity 96.28% and specificity 91.28%), which is better than HC (average accuracy 93.07%, sensitivity 96.05% and specificity 90.18%) but inferior to the proposed SOC method. Experimental results demonstrate that the proposed method outperforms traditional models that do not explicitly consider variable clustering and complex interdependent structures among predictor variables.

The classification of myocardial infarctions with ECG/VCG signals is a general problem in the biomedical domain. See a detailed review of the state of the art in Yang (2011). Our previous investigations focused on the use of nonlinear recurrence quantifiers (Yang 2011) and spatiotemporal features (Yang et al. 2013) for the identification of myocardial infarctions. This present paper approached this problem from a different angle, i.e., leveraging the

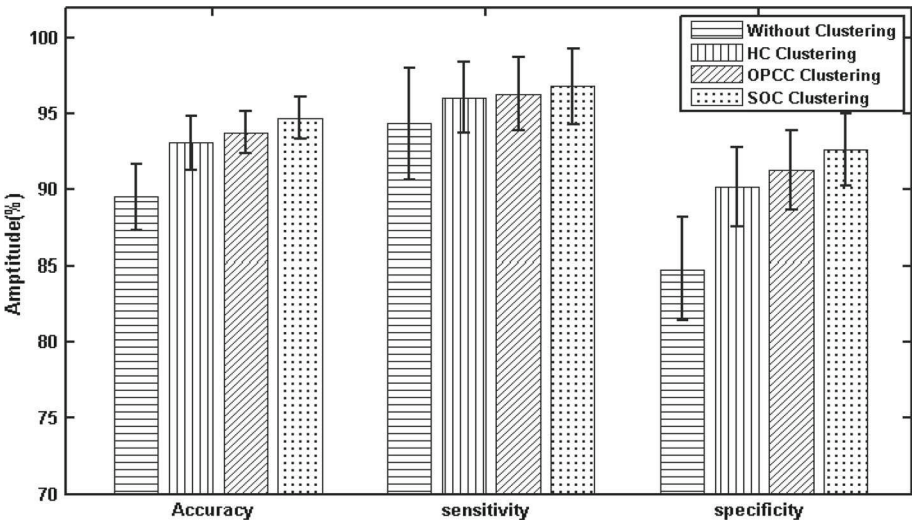

Fig. 8 Averages and standard deviation of prediction errors in the real-world case study that extracts a sparse set of model parameters from VCG signals for the identification of myocardial infarctions

parameters of representation model of ECG/VCG signals to build the classification model. Notably, self-organizing variable clustering yields comparable performance as recurrence quantification analysis (accuracy 92.70%, sensitivity 96.50% and specificity 74.80%) (Yang 2011) and spatiotemporal warping features (accuracy 95.10%, sensitivity 94.90% and specificity 95.70%) (Yang et al. 2013). Our future work will focus on the investigation of ensemble models and feature combination to improve the predictive performance.

6 Discussion and conclusions

Advanced sensing and real-time data acquisition bring "Big Data" which provides an unprecedented opportunity to move forward the new frontier of innovation and knowledge discovery. However, it is common that big data involve large amounts of variables with complex interdependence structures, which pose significant challenges on traditional methodologies in predictive analytics. 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 the response variable but does not explicitly consider the redundancy (i.e., interdependence structures) among variables. The variable clustering, on the other hand, focuses only on the issues of variable redundancy while overlooking the relevancy between variables and the response. In the literature, Kojadinovic proposed to integrate mutual information with agglomerative hierarchical clustering to identify "classes" of continuous variables (Kojadinovic 2004). A homogeneous class refers to the variables in the cluster should be "functional dependent", and a separated class refers to the variables in the cluster that are as "mutually stochastically independent" as possible. Slonim et al. utilized mutual information to cluster nonlinear structures among data samples by designing a tradeoff function between information carried by the cluster identities and average similarity (Slonim et al. 2005). It should be noted that this information-theoretic approach pre-defines the number of clusters to solve the tradeoff function, as well as considers nonlinear correlation structures among data samples instead of continuous variables. Mutual information

shows the advantage to provide an equitable measure of association between two variables but requires stationarity in the computation and is limited in the ability to handle asymmetric interdependence structures. Nonetheless, nonlinear variables (e.g., logistic map and Lorenz variables in the simulation study) show nonstationary properties. Traditional measures that require stationary assumptions tend to yield time-varying dependence structures between variables. In addition, the asymmetric dependence structure between two nonlinear variables is not uncommon, i.e., $Pr(\boldsymbol{x}_1|\boldsymbol{x}_2) \neq Pr(\boldsymbol{x}_2|\boldsymbol{x}_1)$. Indeed, the presence of nonlinear and asymmetric interdependence structures poses a significant challenge for variable clustering.

New methodologies that consider nonlinear interdependence structures among variables and further integrate variable clustering with variable selection to improve the effectiveness of predictive analytics are urgently needed. This paper presents a new strategy that combines the advantages of both variable clustering and variable selection. Complex interdependence structures among variables are characterized and quantified using nonlinear coupling analysis. Then, we developed a self-organizing network algorithm to effectively cluster variables that have nonlinear and asymmetric interdependences. This new method circumvents the limitations of existing methods of variable clustering. For example, the HC algorithm is not applicable when the interdependence structure is asymmetric. The OPCC algorithm cannot adequately handle nonlinear interdependences. Further, the redundant information from related variables in self-organized clusters is minimized. Finally, self-organized clusters are integrated with group elastic net models to improve the performances of predictive models. As such, we handle the relevancy and redundancy among variables simultaneously. Results of simulation experiments demonstrated that the proposed method not only outperforms traditional variable clustering algorithms such as HC and OPCC but also effectively identifies cluster structures among variables, thereby improving the performance of predictive modeling. Also, we evaluated and validated the proposed methodology using a real-world case study that extracts parameters from representation models of VCG signals for the identification of myocardial infarctions. Experimental results showed that the proposed method yields an average sensitivity of 96.80% and an average specificity of 92.62% in the identification of myocardial infarctions using sparse parameters of VCG representation models. The proposed new idea of self-organizing algorithm is generally applicable for variable clustering and predictive modeling in many disciplines that involve a large number of highly-redundant variables.

Acknowledgements The authors would like to thank the National Science Foundation (CMMI-1646660, CMMI-1617148, CMMI-1619648, and IOS-1146882) for the support of this research. The authors also thank Harold and Inge Marcus Career Professorship (HY) for additional financial support. The authors are very grateful to anonymous reviewers for their constructive suggestions that greatly improved the quality of this paper.

References

Arnhold, J., Grassberger, P., Lehnertz, K., & Elger, C. E. (1999). A robust method for detecting interdependence: application to intracranially recorded EEG. *Physica D*, *134*(4), 419–430.

Chen, Y., & Yang, H. (2012). Self-organized neural network for the quality control of 12-lead ECG signals. *Physiological Measurement*, *33*(9), 1399–1418.

Chen, Y., & Yang, H. (2014). Heterogeneous postsurgical data analytics for predictive modeling of mortality risks in intensive care units. In *Engineering in Medicine and Biology Society (EMBC), 2014 36th annual international conference of the IEEE* (pp. 4310–4314). Chicago.

Ding, Y., Elsayed, E. A., Kumara, S., Lu, J., Niu, F., & Shi, J. (2006). Distributed sensing for quality and productivity improvements. *IEEE Transactions on Automation Science and Engineering*, *3*(4), 344–359.

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, *32*(2), 407–499.

Fraser, A. M., & Swinney, H. L. (1986). Independent coordinates for strange attractors from mutual information. *Physical Review A*, *33*(2), 1134–1140.

Friedman, J. (1997). On bias, variance, 0/1-loss, and the curse-of-dimensionality. *Data Mining and Knowledge Discovery*, *1*, 55–77.

Goldberger, A., Amaral, L., Glass, L., Hausdorff, J., Ivanov, P., Mark, R., et al. (2000). PhysioBank, PhysioToolkit, and PhysioNet: Components of a new research resource for complex physiologic signals. *Circulation*, *101*(23), e215–e220.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *American Society for Quality*, *12*(1), 55–67.

Joseph, V. R., Dasgupta, T., Tuo, R., & Jeff Wu, C. F. (2014). Sequential exploration of complex surfaces Using minimum energy designs. *Technometrics*, *57*(1), 64–74.

Kaiser, H. F. (1958). The varimax criterion for analytic rotation in factor analysis. *Psychometrika*, *23*(3), 187–200.

Kantz, H., & Schreiber, T. (2003). Coupling and synchronisation of nonlinear systems. In *Nonlinear time series analysis* (2nd ed., pp. 292–299). Cambridge: Cambridge University Press.

Kohonen, T. (1990). The self-organizing map. *Proceedings of the IEEE*, *78*(9), 1464–1480.

Kojadinovic, I. (2004). Agglomerative hierarchical clustering of continuous variables based on mutual information. *Computational statistics & data analysis*, *46*(2), 269–294.

Lee, T., Duling, D., Liu, S., & Latour, D. (2008). Two-stage variable clustering for large data sets. In *Proceeding of SAS global forum* (pp. 1–14).

Liu, G., Kan, C., Chen, Y., & Yang, H. (2014). Model-driven parametric monitoring of high-dimensional nonlinear functional profiles. In *2014 IEEE international conference on automation science and engineering (CASE)* (pp. 722–727).

Liu, G., & Yang, H. (2013). Multiscale adaptive basis function modeling of spatiotemporal vectorcardiogram signals. *IEEE Journal of Biomedical and Health Informatics*, *17*(2), 484–492.

Narendra, P. M., & Fukunaga, K. (1977). A branch and bound algorithm for feature subset selection. *IEEE Transactions on Computers*, *C–26*(9), 917–922.

Nas, T., & Mevik, B. H. (2001). Understanding the collinearity problem in regression and discriminant analysis. *Journal of Chemometrics*, *15*(4), 413–426.

Neslin, S. A., Gupta, S., Kamakura, W., Lu, J., & Mason, C. H. (2006). Defection detection: Measuring and understanding the predictive accuracy of customer churn models. *Journal of Marketing Research*, *43*(2), 204–211.

Slonim, N., Atwal, G. S., Tkačik, G., & Bialek, W. (2005). Information-based clustering. *Proceedings of the National Academy of Sciences*, *102*, 18297–18302.

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

Wang, H., Zhang, X., Ashok, K., & Huang, Q. (2009). Nonlinear dynamics modeling of correlated functional process variables for condition monitoring in chemical–mechanical planarization. *IEEE Transactions on Semiconductor Manufacturing*, *22*(1), 188–195.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association*, *58*(301), 236–244.

Yang, H. (2011). Multiscale recurrence quantification analysis of spatial cardiac vectorcardiogram (VCG) signals. *IEEE Transactions on Biomedical Egnineering*, *58*(2), 339–347.

Yang, H., Bukkapatnam, S. T., & Komanduri, R. (2012). Spatio-temporal representation of cardiac vectorcardiogram (VCG) signals. *Biomedical Engineering Online*, *11*, 16.

Yang, H., Bukkapatnam, S. T., Le, T., & Komanduri, R. (2011). Identification of myocardial infarction (MI) using spatio-temporal heart dynamics. *Medical Engineering & Physics*, *34*(4), 485–497.

Yang, H., & Chen, Y. (2014). Heterogeneous recurrence monitoring and control of nonlinear stochastic processes. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *24*(1), 013138.

Yang, H., Kan, C., Liu, G., & Chen, Y. (2013). Spatiotemporal differentiation of myocardial infarctions. *IEEE Transactions on Automation Science and Engineering*, *10*(4), 938–947.

Yang, H., & Kundakcioglu, E. (2014). Healthcare intelligence: Turning data into knowledge. *IEEE Intelligent Systems*, *29*(3), 54–68.

Yang, H., & Liu, G. (2013). Self-organized topology of recurrence-based complex networks. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, *23*, 043116.

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