Using the Dimension Reduction Method FAMD in the Data Pre-processing Step for Risk Prediction and for Unsupervised Clustering

by

Xinhui Ran

BS, Chongqing University of Post and Telecommunications, China, 2006

MMedSc, Guangzhou University of Chinese Medicine, China, 2010

Submitted to the Graduate Faculty of

the Department of Biostatistics

Graduate School of Public Health in partial fulfillment

of the requirements for the degree of

Master of Science

University of Pittsburgh

2019

UNIVERSITY OF PITTSBURGH

GRADUATE SCHOOL OF PUBLIC HEALTH

This thesis was presented

by

Xinhui Ran

It was defended on

April 12, 2019

and approved by

Thesis Advisor:
Chung-Chou H. Chang, PhD
Professor
Departments of Medicine and Biostatistics
School of Medicine and Graduate School of Public Health
University of Pittsburgh

Committee Members:
Jonathan G. Yabes, PhD
Assistant Professor
Departments of Medicine and Biostatistics
School of Medicine and Graduate School of Public Health
University of Pittsburgh

Florian Mayr, MD
Assistant Professor
Department of Critical Care Medicine
School of Medicine
University of Pittsburgh

Chung-Chou H. Chang, PhD

Using the Dimension Reduction method FAMD in the Data Pre-processing Step for Risk Prediction and for Unsupervised Clustering

Xinhui Ran, MS

University of Pittsburgh, 2019

Abstract

High-dimensional data generated from various resources including the electronic health records (EHRs), Medicare, and Medicaid, are used in multiple research areas such as public health and medical research. However, working with high-dimensional data is a no easy task because of methodological challenges. Dimensionality reduction technique has been used to transform high-dimensional data into a lower dimensional space while preserving meaningful characteristics of the original data. Principal component Analysis (PCA) is the most widely used method for dimension reduction. However, it has its limitation on linearity assumption and is unsuitable for data containing both numeric and categorical types. Factor analysis of mixed data (FAMD) is a dimension reduction method that can be used for data with mixed types of variables. Dimension reduction is often used as a data pre-processing step prior to further analyses. However, this approach should be used with caution as it depends on the purpose of the application. In this thesis, I demonstrate that using the dimension reduction method FAMD in the data pre-processing step for risk prediction can achieve comparable prediction performance as the traditional variable selection procedure; however, when classifying individuals into similar groups using the unsupervising clustering techniques, the clustering results of using principal components generated from FAMD are substantially different from those of using the original variables.

PUBLIC HEALTH SIGNIFICANCE: High-dimensional data often present challenges in building a risk prediction model or in classifying individuals into groups with more homogeneous characteristics. Dimension reduction techniques, such as incorporating dimension reduction tools, can be incorporated in the data pre-processing step for high-dimensional data collected from public health or medical records. The results of the thesis show that using dimension reduction method (e.g., FAMD for mixed variable types) as a data pre-processing step should be used with caution.

Table of Contents

List of Tables

List of Figures

1.0 Introduction

High-dimensional data generated from the electronic health records (EHRs) can be suitable for developing risk prediction models to estimate the intensive care unit (ICU) mortality[1], estimate the hospital readmissions[2], discover causal risk factors of severe acute kidney injury (AKI)[3], develop disease phenotyping algorithms[4], and make the bed-side decision makings.[5]

Working with high-dimensional data may cause several methodological issues. First, if we have thousands of variables in the data set, it is not practical to analyze each one in a microscope level for that will take too much computational time and power. Second, oftentimes there exists intercorrelation between variables which can increase computational complexities tremendously. For example, to include all covariates in a multiple regression model will likely lead to severe multicollinearity issues. Also, if we aim to develop a risk prediction model with a set of intercorrelated variables, it could result in unacceptably large standard errors and inaccurate predictions. Third, it is very difficult to plot or visualize the patterns of data and relationship between data when the feature (or covariate) space is higher than three dimensions. In addition, certain algorithms which rely on distance measures (e.g., clustering algorithms) may struggle to train effective models when there are larger number of covariates than subjects in the dataset, which is referred to the curse of dimensionality phenomenon. Researchers have shown that in the high-dimensional feature space, the distance of each pairs of points are very close for many data distributions.[6] Thus, it may lead some clustering algorithms flawed when applied on high-dimensional data structure.

The aforementioned issues in dealing with high-dimensional data using the traditional multivariate or multivariable methods accentuate the need of a new methodological approaches that can extract the most important information from high-dimensional data in the preprocessing step before proceeding to the subsequent developments.[7]

Dimensionality reduction technique has been used to transform high-dimensional data into a lower dimensional space while preserving meaningful characteristics of the original data.[8] With efficient dimension reduction, not only data visualization is improved, but also a significant amount of computational time can be saved. By excluding the redundant variables in the data set or reconstructing a set of uncorrelated variables using dimension reduction techniques, the multicollinearity problem in the original data set can be resolved. It has been shown that building classification models with data from greatly reduced dimension can result in higher prediction accuracy while utilizing fewer features and training samples.[9] The curse of dimensionality is also expected to get alleviated via dimension reduction.

Principal components analysis (PCA)[10] is one of the most commonly used techniques to perform dimension reduction. For a given high-dimensional data, it produces the best linear combination of original data features. However, as an efficient way to perform data reduction, PCA suffers from some limitations, such as linear relationships between variables are assumed, and its interpretation is based on that the variables are scaled at the numeric levels. When dealing with categorical variables, or variables with mixed measurement types (continuous mixed with categorical), the numeric assumption made by the PCA is violated therefore PCA is inappropriate to be applied. In addition, nonlinear association might also exist even among numeric variables.

Besides PCA, there are some other classical dimension reduction methods, such as multidimensional scaling (MDS)[11] and independent component analysis (ICA)[12]. However, all of these suffer from the drawback of the limitation on linearity as PCA does. Recent techniques, such as kernel PCA[13], locally linear embedding (LLE)[14], Laplacian eigenmaps (LEM)[15], Isomap[16], and semidefinite embedding (SDE)[17] have been developed to overcome the strong assumption of linearity.

When the variables of a given data are all categorical, multiple correspondence analysis (MCA)[18] is often used. MCA represents data as points in a low-dimensional Euclidean space. For a given categorical variable, individuals with the same level are close together and individuals with different levels are far apart. For all the categorical variables under investigation, individuals having higher percent of variables with the same level are closer. The procedure can be considered as a generalization of PCA for categorical data, which allows us to analyze the patterns of relationships of categorical variables that represent the underlying structures.

In practice, it is common that data contain both numeric and categorical variables. Originated from introducing qualitative variables in PCA or introducing quantitative variables in MCA, factor analysis of mixed data (FAMD)[19] acts as PCA to handle numeric variables and as MCA to deal with categorical variables. The Jérôme Pages' multiple FAMD replaces each qualitative variable by a set of dummy variables for each level of the variables.[19] It incorporates the unique scaling of MCA on categorical variables and imposes a refinement to balance dispersions of the numeric variables and dummy variables. Then standard PCA can be applied to analyze the association between all variables using FAMD.

The goal of this research is to perform FAMD on the Recombinant human activated protein worldwide evaluation in severe sepsis (PROWESS) dataset that contains demographic and clinical information for patients admitted to ICUs.[22] After performing dimension reduction, we made two different applications. First, we built risk prediction models of 28-day mortality using data before and after dimension reduction. We then assessed the performance of these two models. Second, we uncovered the hidden covariate patterns using the PAM clustering algorithm for data before and after dimension reduction and compared the clustering findings.

In Section 2, we provide a brief description on PCA, MCA, and FAMD dimension reduction methods. In Section 3, we apply the FAMD method to the PROWESS data. Based on the original and dimension reduction datasets, we build risk prediction models and partition the covariate space using a clustering method. In Section 4, we discuss our findings by focusing on the impact of dimension reduction performed in a data pre-processing step on risk prediction and clustering.

2.0 Methods

2.1 Principal Component Analysis (PCA)

Principal component analysis (PCA) is a commonly applied technique for dimension reduction. The original set of possibly correlated variables in the dataset is transformed into new features that are uncorrelated. The transformed uncorrelated variables are called principal components, which are rank ordered with respect to the variability of the data they explained. Therefore, the first principal component explains the largest variability of the data. By applying the PCA, a high-dimensional feature (or covariate) space spanned by a large set of variables can then be reduced to a lower dimensional space spanned by a smaller number of principal components. The choice of the number of principal components determines the dimension of the reduced space.

The PCA procedure requires that the mean and variance of the data can be used to fully describe the corresponding probability distribution (i.e., mean and variance are the sufficient statistics), which implies that the probability distribution of each variable in the data must be Gaussian distributed.[20] Therefore, proper transformation for variables deviated from normal is needed before applying PCA to avoid poor performance.

We now briefly describe how PCA searches for the new vectors (principal components) in order to retain the maximum variation of the original data.[19] Let matrix **X** represent a data set, with *N* individuals and *K* numeric variables. Therefore, each individual i can be seen as a point

M_i in $\mathbb{R}^K$. In order for computational convenience and for preventing giving more emphasis to variables that having higher variance in developing the principal components[21], we will need to standardize each variable by shifting its mean to 0 and scaling its variance to 1. Let x_k be the standardized variable ($k = 1, \ldots, K$). The squared distance of each individual data point M_i to the origin is:

$$d_i^2 = \sum_k x_{ik}^2.$$

Let each individual is assigned the weight $p_i = 1/N$, where $\sum_i p_i$ =1. The total inertia (variance) N_I then can be expressed as:

$$N_I = \sum_i p_i d_i^2 = \sum_k \sum_i p_i x_{ik}^2 = \sum_k Var[k] = K.$$

On the other hand, each variable can be seen as a point M_K in space $\mathbb{R}^N$, of which each dimension corresponds to an individual. Therefore, the squared distance of each data point M_K to the origin is:

$$d_{M_K}^2 = \sum_i p_i x_{ik}^2 = Var[k] = 1.$$

The total inertia N_K then can be expressed as:

$$N_K = \sum_k d_{M_K}^2 = \sum_k \sum_i p_i x_{ik}^2 = \sum_k Var[k] = K.$$

This inertia value is the same to the one we calculated through N_I as it should be.

The goal of conducting PCA is to find a set of orthogonal axes (principal components) of the maximum inertia through the iterative projection procedure. From this procedure, the amount of inertia explained by the s_th principal component is denoted by λ_s, which is nothing more than the eigenvalue associated to the unit eigenvector v_s.[21] Therefore, given a standardized data set, the sum of all the eigenvalues will be the total inertia, that is, $\sum_s \lambda_s = K$. In PCA analysis, we can examine the eigenvalues to determine the number of principal components to be considered. If we denote a unit vector as μ_s in $\mathbb{R}^K$, project a point M_i onto μ_s , and let $h_S^{M_i}$

denote the projected length of vector $\overrightarrow{OMi}$, then we are searching for the μ_s such that $\sum_i p_i\,(h_S^{M_i})^2$ is maximized, with the constraint of being orthogonal to s-1 directions already found. Or, if we denote a unit vector as v_s in $\mathbb{R}^I$, project a point M_K onto v_s, and let $h_S^{M_K}$ denote the projected length of vector $\overrightarrow{OM_K}$, then we are searching for the v_s such that $\sum_K (h_S^{M_K})^2$ is maximized, with the constraint of being orthogonal to *s* -1 directions already found.

When the data are standardized, an eigenvalue > 1 indicates that the corresponding principal component accounts for more variances than that accounted by one of the original variables. Therefore, it is common to use eigenvalue = 1 as a cutoff point to decide which principal components are remained.

The contribution of individual *i* to compose a principal component can be calculated as:

$$\text{contrib}\ (i, v_s) = \frac{\text{projected inertia of point } M_i \text{ on } v_s}{\text{projected total inertia of all points in } N_I \text{ on } v_s} = \frac{p_i(\,h_S^{M_i})^2}{\sum_i p_i(\,h_S^{M_i})^2} = \frac{p_i(\,h_S^{M_i})^2}{\lambda_s}\,.$$

Similarly, the contribution of variable *k* to compose a principal component can be calculated as:

$$\text{contrib}\ (k, v_s) = \frac{\text{projected inertia of point } M_K \text{ on } v_s}{\text{projected total inertia of all points in } N_K \text{ on } v_s} = \frac{(\,h_S^{M_K})^2}{\sum_K(\,h_S^{M_K})^2} = \frac{\mathrm{r}(\mathrm{k},v_S)^2}{\lambda_s}\,.$$

Note that the contribution of variable *k* to compose a principal component usually is not expressed by the form of proportion calculated above. Instead, it is expressed by the squared correlation coefficient, $\mathrm{r}(\mathrm{k}, v_s)^2$ multiplied by 100. Therefore, for each component, the contributions of all variables sum up to 100; and for each variable, its contributions to all components sum up to 100.

As we described earlier, PCA is only appropriate when the data are comprised of a set of normally distributed variables. When the data contains a categorical variable with *Q* levels, a suggestion is to replace this variable by a set of *Q* – 1 dummy variables and then apply PCA on the reconstructed data. Although the convenience of this strategy, the results could be biased because of the distributions of the numeric variables and the dummy indicator variables are not comparable. Factor Analysis of Mixed Data (FAMD) has been developed to handle data with mixed variable types. Instead of using the 0/1 value for the dummy variables, FAMD incorporates a unique scaling technique in the Multiple Correspondence Analysis (MCA) in order to balance the influence of the two types of variables (numeric and categorical) during the construction of the principal components.[19]

MCA is a dimension reduction method that is used for data only including categorical/qualitative variables. Previously, researchers had suggested a strategy to deal with the data with mixed variable types by discretizing the numeric variables and then use the MCA to analyze all the variables. However, this procedure is not recommended for several reasons. First, it results in a loss of information by discretizing a numeric variable. Second, the choice of the optimal cutoff points are often not obvious, especially when there are only a few number of observations in the data set. In addition, the process could be tedious if there are a considerably large number of numeric variables.

In the following section, we briefly describe the MCA dimension reduction procedure, where PCA and MCA together will be the basis of FAMD.

2.2 Multiple Correspondence Analysis (MCA)

Let matrix **X** represent a data set with *N* individuals and a set of *J* categorical variables. Let K_j denote the number of levels of variable j, and $H = \sum_j K_j$ is the total number of levels for all variables. Let matrix **Y** represent the expanded data with *N* individuals (rows) and *H* total levels (columns), where y_{ik} = 1 if the ith individual possesses the k_jth level of variable j; and y_{ik}

= 0 in all other cases. Let p_k denote the proportion of individuals possessing level *k* so that $p_k = \frac{1}{N}\sum_i y_{ik}$. To standardize the data, we scale each cell in matrix **Y** by $\frac{y_{ik}}{p_k} - 1$, so that the mean of each column (i.e., one level of a variable) is 0. Let matrix **Z** denote this scaled new data, within which the (i,k)th element $z_{ik} = \frac{y_{ik}}{p_k} - 1$.

Using the similar idea in PCA, treating each individual *i* as a data point M_i in space $\mathbb{R}^K$. By assigning the weight of $p_i = \frac{1}{N}$ to each individual and assigning the weight of $m_k = \frac{p_k}{J}$ for each column (one level of a variable), the squared distance of each data point M_i to the origin becomes:

$$d_i^2 = \sum_k \frac{p_k}{J} z_{ik}^2 = \frac{1}{J}\sum_k \left(\frac{y_{ik}}{p_k} - 1\right).$$

Therefore, the total inertia (variance) can be expressed as:

$$N_I = \sum_i p_i d_i^2 = \frac{K}{J} - 1.$$

By treating each level of a variable as a data point M_K in space $\mathbb{R}^N$, of which each dimension corresponds to an individual. The inertia of category *k* can be expressed as:

$$m_k d_{M_K}^2 = \frac{p_k}{J}\sum_i p_i z_{ik}^2 = \frac{1-p_k}{J}.$$

The total inertia then can be expressed as:

$$N_K = \sum_k \frac{1-p_k}{J} = \frac{K}{J} - 1,$$

which is the same to the one that is calculated through N_I as it should be. In addition, a variable *j* is represented by its K_j levels in a vector space. We denote this subspace formed by the K_j levels as E_j, of which the dimension is $K_j - 1$. Because the K_j levels are orthogonal to each other in the subspace E_j, the total inertia of the K_j levels of variable *j* are:

$$\sum_{k \in K_j} \frac{1-p_k}{J} = \frac{K_j - 1}{J}.$$

Just like PCA, the goal of conducting MCA is to find a set of orthogonal axes (principal components) of the maximum inertia. This can be done iteratively by projecting level k of variable j on a centered unit vector ν in $\mathbb{R}^N$. Through this procedure, the projected inertia for each variable can be calculated.

2.3 Factor Analysis of Mixed Data (FAMD)

As we mentioned earlier, when data contains both numeric and categorical variables, FAMD acts as PCA in dealing with numeric variables and acts as MCA in dealing with categorical variables. Suppose there are J numeric variables and Q categorical variables, where k_q denotes the number of categories of the qth qualitative variable. Let p_{k_q} denote the proportion of individuals possessing category k_q. Let H denote the total categories for all the qualitative variables.

When processing the data, the numeric variables and the categorical variables are standardized as those described in PCA and MCA, respectively. The weight for each individual is still $1/N$, but instead of assigning the weight of each level to p_{k_q}/H as that in MCA, the weight of each level of a categorical variable in FAMD is assigned to p_{k_q} in order to balance the contributions of the two types of variables. As a result, in space $\mathbb{R}^K$, each numeric variable has inertia of 1 and is represented by a vector; each categorical variable has total inertia of $k_q - 1$ and is represented by k_q vectors. When projecting the total inertia of $k_q - 1$ on each dimension of the subspace of a categorical variable, the projected inertia is 1. Therefore, when searching for the new axes with maximum inertia, the two types of variables are on the equal step.

When search for a new principal component in FAMD, we maximize the sum of the squared correlation coefficient between numeric variables and the principal component plus the sum of the squared correlation ratio between categorical variables and the principal component. The contribution of individual i (or a variable) to a principal component can be calculated in a similar sense as that in PCA and the quality of representation is defined as the cosine of the angle θ_{kj}, which is the correlation coefficients of variable *k* and variable *j*.

3.0 Application

We demonstrate the use of dimension reduction methods for data collected from a multicenter, randomized, double-blind, placebo-controlled trial (PROWESS trial) of 1,690 sepsis patients. The drug to be tested or placebo was administered intravenously. Details of the trial description and results can be found in the related research publication.[22] Patients' baseline characteristics including demographic, clinical, biomarkers, and laboratory information were assessed within 24 hours before the infusion began. All patients were followed for 28 days after the start of the infusion or until death. For this study, we chose 28 demographic and clinical variables which contain 21 numeric and 7 categorical or qualitative types. Because the distributions of several numeric variables were highly right-screwed, we applied the natural logarithm transformation to these variables before applying FAMD. Figures 9 and 10 depict the density plot of the numeric variables before and after necessary log transformation, respectively. Table 1 summarizes the descriptions of these variables. Mean (standard deviation [SD]) and frequency (percentage) are presented for numeric and categorical/qualitative variables, respectively.

Table 1 Summary statistics for all variables measured at the time of ICU admission

Variable	Mean (SD) or n (%)
Age in year, mean (SD)	60.52 (16.80)
Albumin (g/dL), mean (SD)	2.01 (0.65)
Log transformed Alanine aminotransferase (IU/L), mean (SD)	3.53 (1.14)
Log transformed aspartate aminotransferase (IU/L), mean (SD)	4.01 (1.16)
Log transformed BANDS, mean (SD)	0.04 (1.27)

Table 1 Continued	
Log transformed serum bilirubin, mean (SD)	-0.28 (0.96)
Log transformed BUN, mean (SD)	2.28 (0.67)
Chlorine, mean (SD)	106.21 (7.49)
Log transformed creatinine, mean (SD)	0.44 (0.64)
Log transformed glucose, mean (SD)	5.02 (0.44)
Heart rate, mean (SD)	129.67 (27.87)
Hemoglobin (g/dL), mean (SD)	10.67 (1.97)
Log transformed PaO2 (mmHg), mean (SD)	4.40 (0.49)
Log transformed platelets, mean (SD)	5.05 (0.64)
Respiratory rate, mean (SD)	30.83 (12.36)
Sodium (mEq/L), mean (SD)	139.03 (6.31)
Log transformed systolic blood pressure (mmHg), mean (SD)	4.42 (0.39)
Temperature in Celsius, mean (SD)	38.19 (1.70)
Log transformed white blood cell counts, mean (SD)	2.50 (0.79)
Log transformed prothrombin time in second, mean (SD)	2.98 (0.27)
Log transformed (101% – oxygen saturation), mean (SD)	1.88 (0.82)
Number of comorbidity conditions, n (%)	
1	549 (32.49%)
2	489 (28.93%)
3	392 (23.20%)
4	180 (10.65%)
5 or more	80 (4.73%)
GCS levels, n (%)	
Score of 3-12	562 (33.25%)
Score of 13	90 (5.33%)
Score of 14	211 (12.49%)
Score of 15	827 (48.93%)
Site of infection, n (%)	
Bloodstream	87 (5.15%)
Central nervous system	39 (2.31%)
Genitourinary	179 (10.59%)
Abdominal	337 (19.94%)
Lung	906 (53.61%)
Other	142 (8.40%)
Gram stain of bacterial pathogen, n (%)	
Mixed	610 (36.09%)
Fungus	85 (5.03%)
Purely gram-negative	266 (15.74%)
Purely gram-positive	375 (22.19%)
Organism negative	354 (20.95%)
Male sex, n (%)	964 (57.04%)
Drug resistance, n (%)	499 (29.53%)
Death within 28 days, n (%)	469 (27.75%)

After necessary log transformation of the right-screwed variables, FAMD was applied to the final data set to reduce dimension while keeping the most important information in the data. The associated scree plot (Figure 1) illustrates the eigenvalues of the corresponding principal components (PCs). The variance explained by each PC is depicted in Figure 2 and the cumulative variance explained by the PCs is shown in Figure 3. From Figures 1 to 3, there were 17 eigenvalues that are greater than or equal to 1 (round to the second decimal point), which in total explained 64.12% of the total variances. The first PC explained most of the variation of the data (7.1%) and corresponds to the largest eigenvalue (2.8). The succeeding PC then explains most of the variation of the data among the remaining PCs. In Figures 1 and 2, the slopes of the first 4 PCs were steeper compared to those of the other PCs. This indicates that the first 4 PCs were the most important components that contained the most important information of the data. Thus, we will later focus on interpreting these 4 PCs.

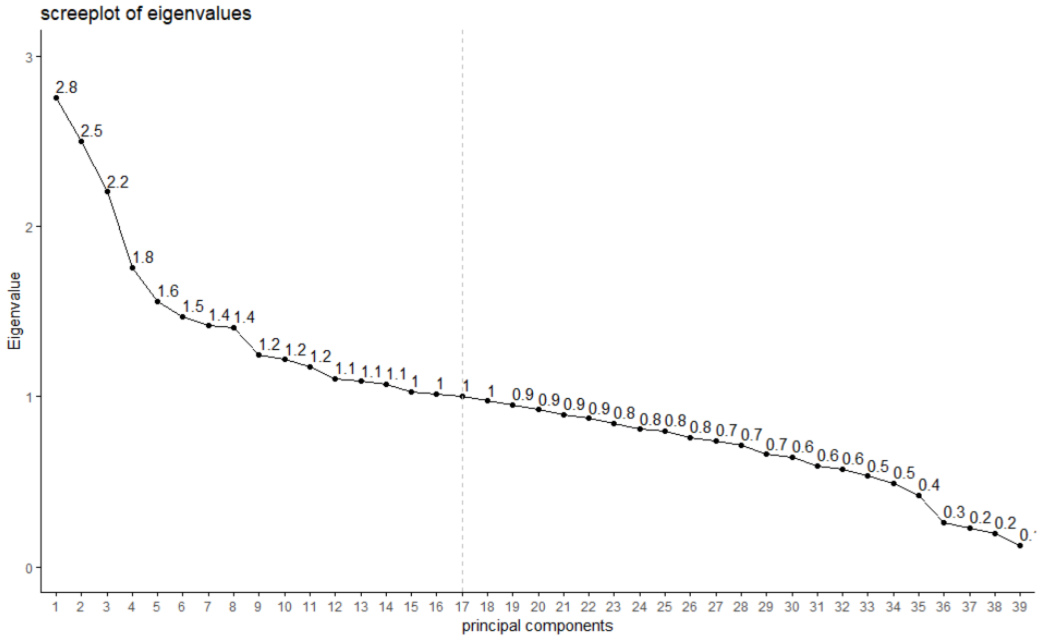

Figure 1 Scree plot of eigenvalues

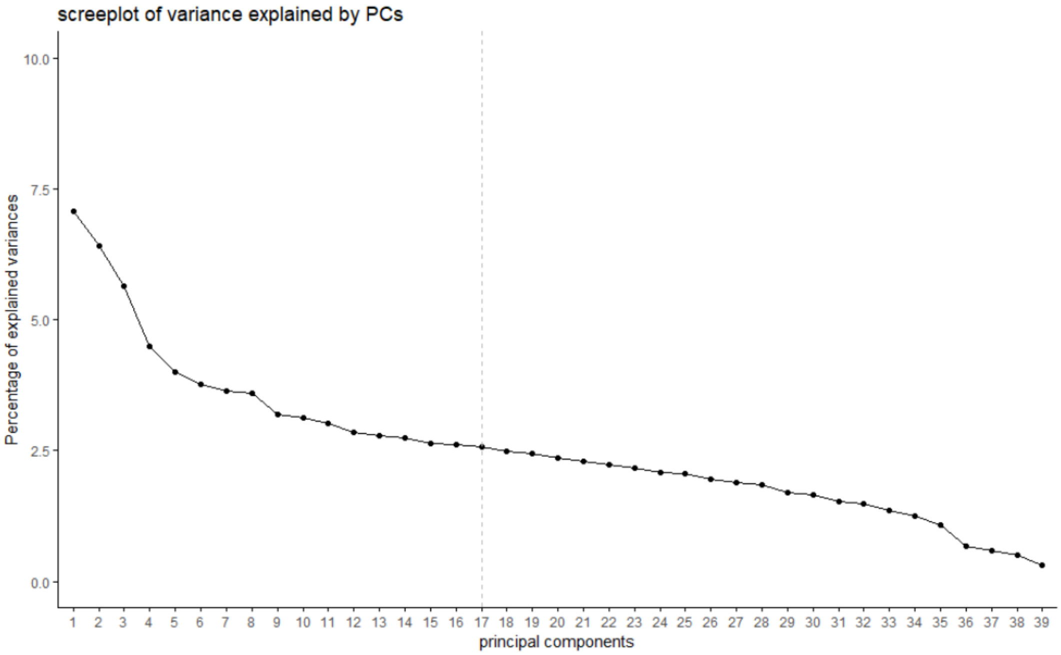

Figure 2 Scree plot of variance explained by principal components

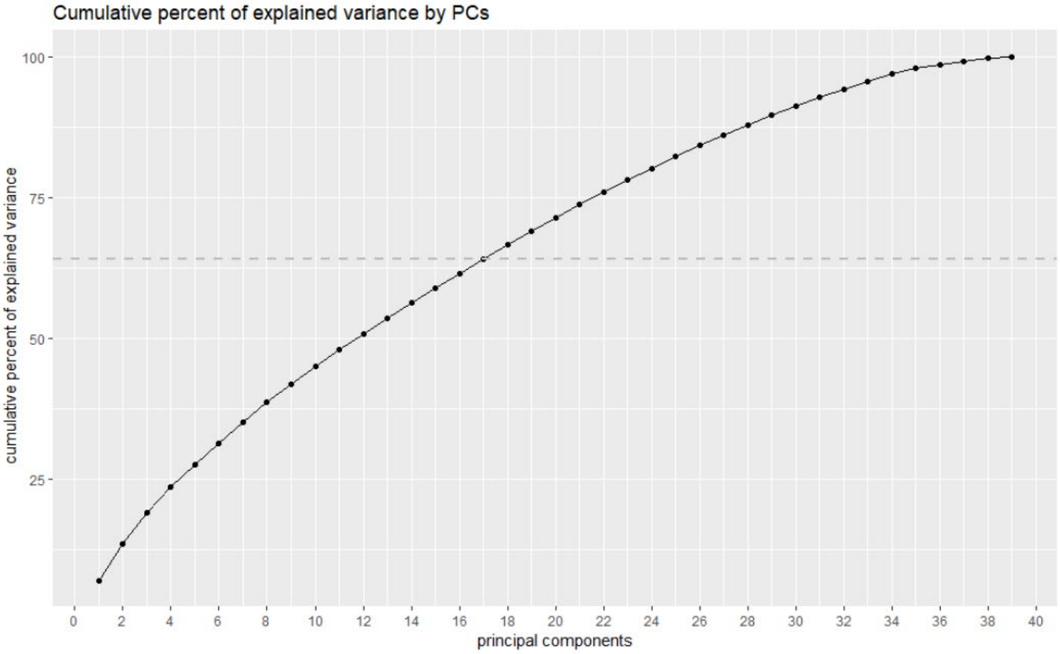

Figure 3 Plot of cumulative percent of explained variance by principal components

Relationship between variables and the principal dimensions can be detected by inspecting the new coordinates of the variables in the new principal dimensions space. Variables that have larger coordinate value on the axis of dimension are more correlated to dimension 1; the same findings also apply to dimension 2 and other dimensions. Figure 4-7 illustrate the relationships between variables with PC1, PC2, PC3, and PC4, which are the first leading and the most important principal components that captures most variability of the data. The coordinates of Figures 4-7 are the variable coordinates in the new principal components space.

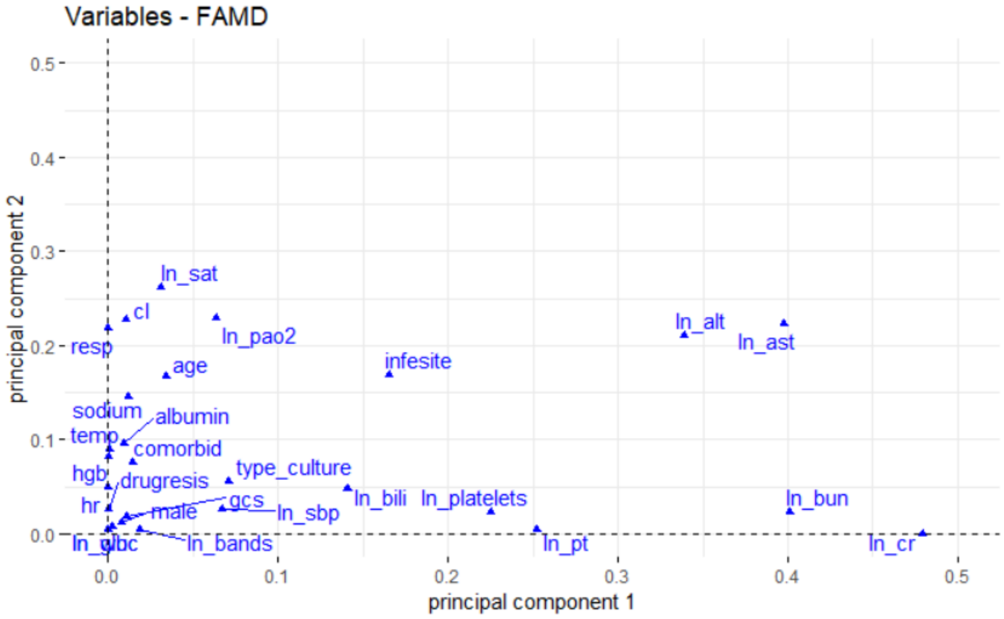

Figure 4 Scatter plot of variables correlations with new PCs (PC1 vs PC2)

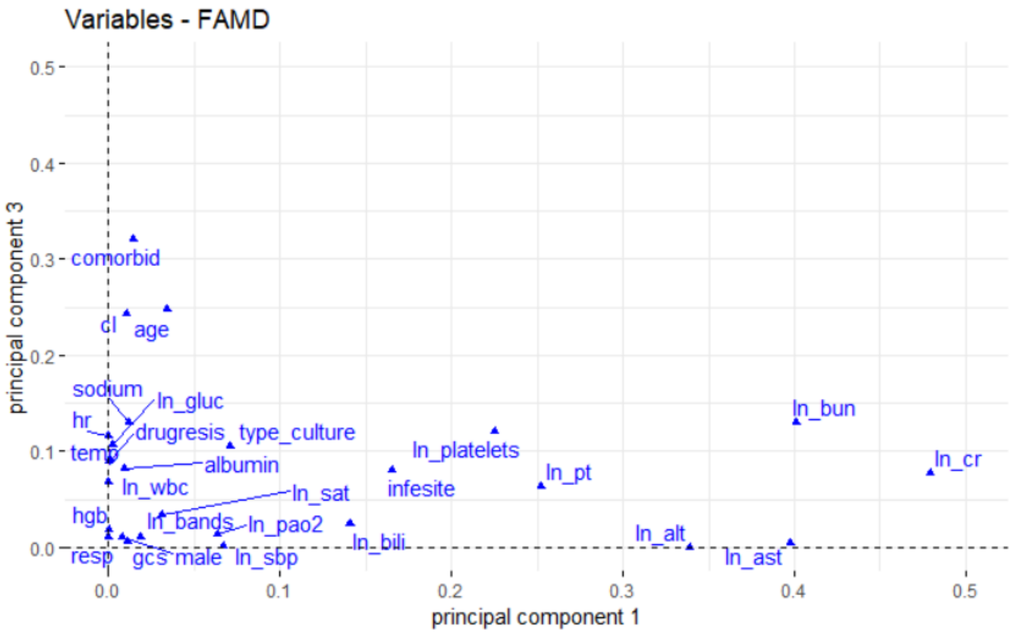

Figure 5 Scatter plot of variables correlations with new PCs (PC1 vs PC3)

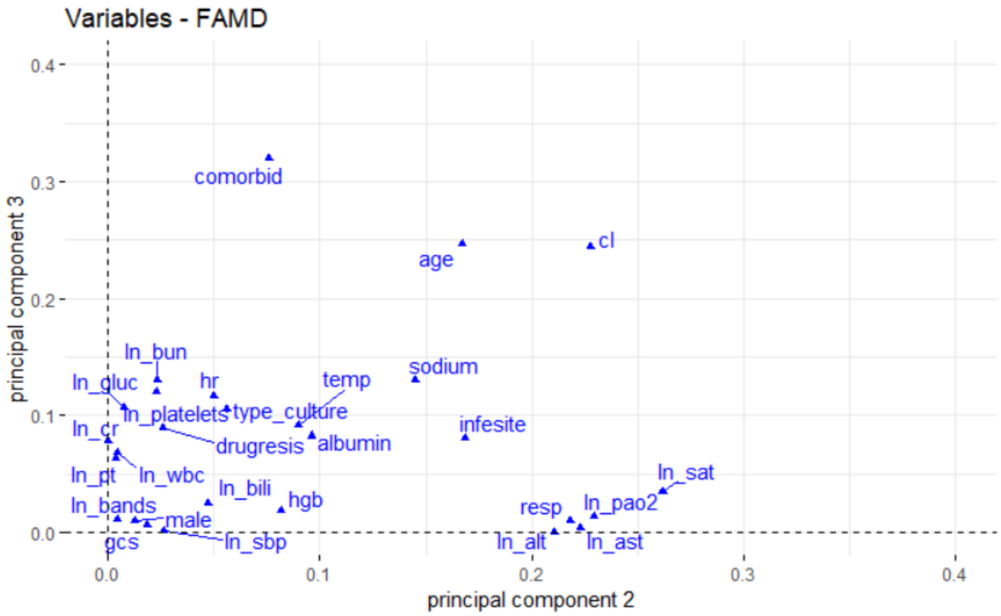

Figure 6 Scatter plot of variables correlations with new PCs (PC2 vs PC3)

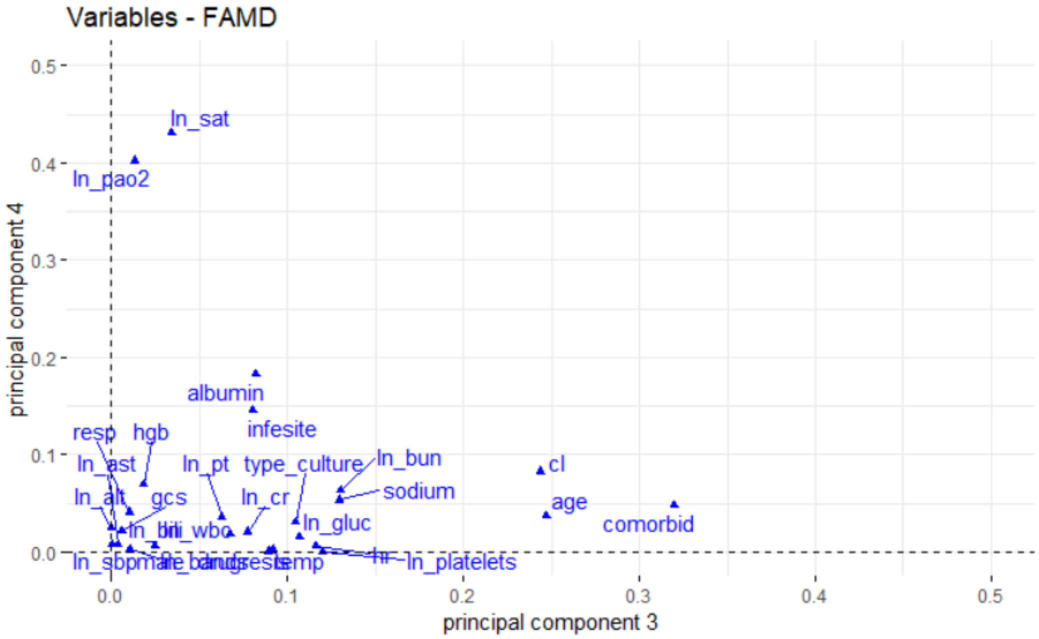

Figure 7 Scatter plot of variables correlations with new PCs (PC3 vs PC4)

Figure 4 to 7 showed the factor loadings for each of the variable on selected two principal components (PCs). Roughly speaking, for a given PC, the percentage of the variance it explains from an original variable can be represented by the squared factor loading. We found that log transformed creatinine level (ln_cr), log transformed blood urea nitrogen (BUN) level (ln_bun), log transformed AST level (ln_ast), log transformed ALT level (ln_alt), log transformed prothrombin time (ln_pt), and log transformed platelets count (ln_platelets) were correlated to PC1 the most. Log transformed SaO2 (ln_sat), respiratory rate (resp), Cl (cl), log transformed partial pressure of oxygen (PaO2) level (ln_pao2), log transformed ALT level (ln_alt), and log transformed AST level (ln_ast) were correlated to PC2 the most. Number of comorbidity conditions (comorbid), age, and Cl (cl) were the most correlated to PC3. Log transformed SaO2 (ln_sat) and log transformed PaO2 level (ln_pao2) were correlated to PC4 the most. We also demonstrated the correlation between any two principal components using a scatter plot. As shown in Figure 6, log transformed AST level (ln_ast) and log transformed ALT level (ln_alt) were correlated to both PC1 and PC2. Cl (cl) was correlated to both PC2 and PC3 (Figure 8). These results indicate that several variables in the original data were intercorrelated; therefore, the application dimension reduction was justified.

Alternatively, we can use cos2 values to discern the relationships between variables and PCs. Cos2 values indicate the quality of representation of variables on the factor map. For a numeric variable, the cos2 value is equal to the squared coordinates on each PC, which is the squared correlation coefficient of that variable and a particular PC. For a categorical variable with *k* levels, the cos2 value is equal to the squared coordinates divided by $k - 1$, which is the squared correlation ratio of that variable and a particular PC. We can plot the cos2 values of a variable for all the principal component dimensions. The heatmap of Figure 8 depicts the cos2 value of each variable on each dimension. A darker color represents a higher cos2 value, which indicates that the variable is more correlated to that corresponding PC.

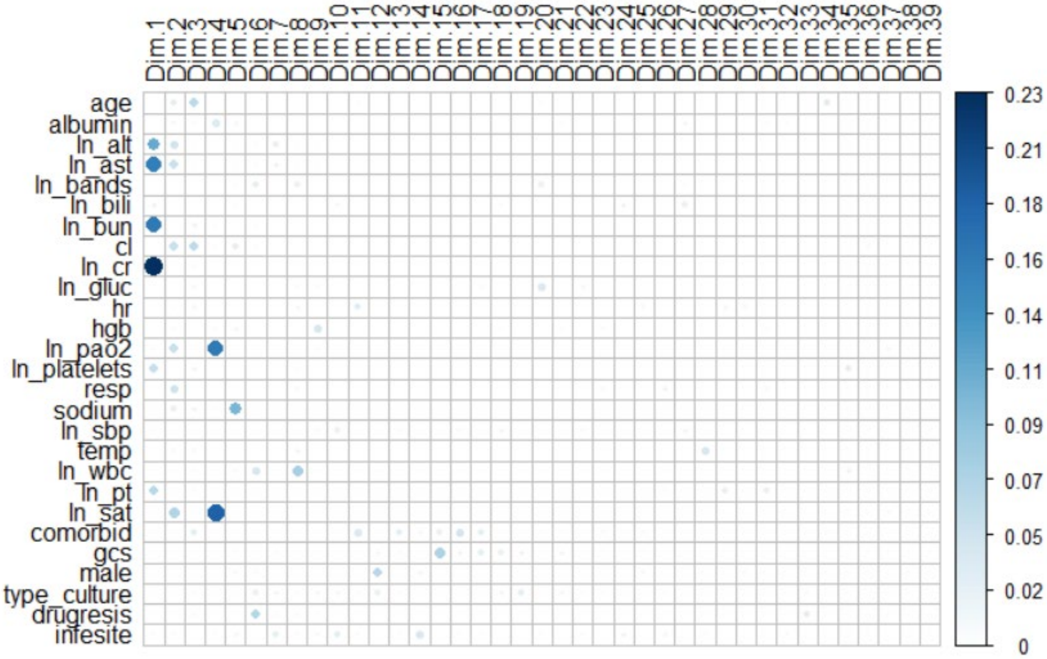

Figure 8 Cos2 values of variables on each PC

From figure 8, again we found that log transformed AST level (ln_ast), log transformed ALT level (ln_alt), log transformed blood urea nitrogen (BUN) level (ln_bun), log transformed creatinine level (ln_cr), log transformed PaO2 level (ln_pao2), and log transformed prothrombin time (ln_pt) were correlated to PC1 the most. Log transformed AST level (ln_ast), log transformed ALT level (ln_alt), Cl (cl), log transformed PaO2 level (ln_pao2), respiratory rate (rep), and log transformed SaO2 (ln_sat) were correlated to PC2 the most. Age, Cl (cl), and number of comorbidity conditions (comorbid) were correlated to PC3 the most. Log transformed PaO2 level (ln_pao2) and log transformed AST level (ln_ast) were correlated to PC4 the mos. Also, we found that after PC16 the correlation of each variable to each dimension is very mild. This suggests that it is reasonable to keep only the first 17 PCs (the ones having corresponding eigenvalue ≥ 1).In fact, by checking the cos2 values, we found that from PC18 to PC39 the maximum cos2 value of a variable to a PC is 0.037, which indicates that the correlations between each variable to any PC is very low. In addition, we found that some of the variables have very low correlation to any of the principal components, such as log transformed bands (ln_bands), log transformed serum total bilirubin level (ln_bili), heart rate (hr), log transformed systolic blood pressure (SBP) level (ln_sbp), and gram stain of bacterial pathogen (type_culture). Those variables can be considered to have little contributions to the total variation compared with other demographic and clinical variables.

3.1 Risk prediction

We used FAMD to generate 39 PCs which contains all variations of the original data. To predict the 28-day mortality of these patients, we built risk prediction models using logistic regression. Accuracy and area under the ROC curve (AUC) were used to evaluate the model performance. A 10-fold cross-validation was conducted for measuring the performance of a given predictive model on new test data sets. We compared the model performance using various sets of leading PCs as predictors. We also compared the model performance using reduced PCs and using the original variables same as above.

The performance of the model based on reduced PCs as predictors is shown in Table 2. The performance of the model based on all original variables and the selected set of variables using stepwise variable selection and AIC as the selection criteria is shown in Table 3. The selected set of variables includes 16 variables: age, albumin, log transformed ALT level (ln_alt), log transformed AST level (ln_ast), log transformed serum total bilirubin level (ln_bili), log transformed blood urea nitrogen (BUN) level (ln_bun), Cl (cl), heart rate (hr), and log transformed platelets count (ln_platelets), log transformed systolic blood pressure (SBP) level (ln_sbp), temperature (temp), log transformed prothrombin time (ln_pt), log transformed SaO2 (ln_sat), number of comorbidity conditions (comorbid), gender (male), site of infection (infisite). The calculation of accuracy is based on 0.5 cut point, meaning that if the predicted probability of having mortality is greater or equal to 0.5, the outcome is labeled as 1 (death) and that if the predicted probability of having mortality is less than 0.5, the outcome is labeled as 0 (alive).

Results in Table 2 show that, in general, by including more PCs one can achieve better prediction in a model (e.g. higher accuracy and higher AUC). When a model included all 39 PCs as predictors, the performance was exactly the same as that included all variables as predictors. Moreover, the predictive model of 16 predictors selected from a forward stepwise procedure produced a similar performance as that produced from the full model (Table 3).

Table 2 Mortality predictive model performance with different number of PCs

Number of PCs	Total variance explained	Accuracy	AUC
1 PC	7.07%	0.73	0.63
2 PCs	13.48%	0.73	0.63
3 PCs	19.12%	0.73	0.65
4 PCs	23.62%	0.74	0.68
5 PCs	27.62%	0.74	0.68
8 PCs	38.62%	0.74	0.70
10 PCs	44.93%	0.74	0.70
12 PCs	50.78%	0.74	0.71
16 PCs	61.56%	0.75	0.72
17 PCs	64.12%	0.75	0.72
29 PCs	89.58%	0.76	0.73
39 PCs (all PCs)	100%	0.75	0.77

Table 3 Mortality predictive model performance with all original variables and selected variables

predictors	Accuracy	AUC
All variables	0.75	0.77
Selected 16 variables	0.76	0.76

3.2 Clustering

We applied the PAM clustering procedure on the lower-dimensional data obtained from FAMD. The result was compared to the same method applied to the original data. The number of clusters was determined from the corresponding k-means consensus matrix. Tables 4 to 6 show the cross tabulations of the clustering results using selected number of PCs vs. using all of the original variables. We found that clustering results were very different using two different sets of input data; even if we used all the PCs, the clustering results still differed substantially from those using all of the original variables.

Table 4 Cross-tabulation of the clustering results of using all 39 PCs and using all original variables

	Clusters (using all 39 PCs)			
Clusters (using all original variables)	1	2	3	4
1	50	88	129	67
2	67	145	168	166
3	44	72	106	82
4	182	106	97	121

Table 5 Cross-tabulation of the clustering results of using first 30 PCs and using all original variables

	Clusters (using first 30 PCs)			
Clusters (using original variables)	1	2	3	4
1	155	64	64	51
2	94	277	99	76
3	94	89	69	52
4	156	81	137	132

Table 6 Cross-tabulation of the clustering results of using first 17 PCs and using all original variables

	Clusters (using first 17 PCs)			
Clusters (using original variables)	1	2	3	4
1	116	90	28	100
2	102	140	131	173
3	47	42	121	94
4	96	144	89	177

4.0 Discussion

We have applied factor analysis of mixed data (FAMD) to achieve dimension reduction on the PROWESS dataset which contains demographic and clinical information measured at the time of ICU admission for sepsis patients. One of the most important advantages of this method is its ability to handle both numeric and categorical variables. In addition, using this method, not only we can analyze the relationships between numeric variables using PCA instead of MCA to analyze the relationships between categories, but also obtain the relationship between numeric and categorical variables. In practice, the number of principal components to be kept was based on the corresponding eigenvalues. We kept the principal components of which the corresponding eigenvalues are greater than or equal to 1. As a result, we reduced the dimensions of the covariate space from 39 to 17 which explained 64.12% of the total variances of data. One can also keep more principal components to maintain a higher proportion of total variances. For example, by keeping 20 principal components in PROWESS data we can explain 71.41% of the total variances.

One potential disadvantage of FAMD is that as it handles numeric variables as PCA does, so the numeric variables are required to be approximately normally distributed in order to achieve good performance. In our application, though some of the numeric variables were log-transformed in order to improve normality, a few of them were still not quite normally distributed or close to symmetric.

Another potential drawback of FAMD is that although it treats each numeric variable as one dimension, it treats a categorical variable with K levels as $K - 1$ dimensions. For example, in our application, we have a total of 27 variables in the raw data set, which can be seen as 27 dimensions. However, while applying the FAMD, although it reduced the dimension to 17 with 64.12% of the total variances remained, internally, it reduced the dimension from 39 to 17. That is saying that it treats each categorical variable as having more than 1 dimension. This suggests that if there are much more categorical variables than numeric variables in the data set and each categorical variable have relatively large number of levels, FAMD may not be appropriate. In addition, as we introduced in the methods section, in FAMD, each numeric variable has inertia of 1 and is represented by a vector; each categorical variable has a total inertia of $k_q - 1$ and is represented by k_q vectors. With this setting, the contributions of the numeric and categorical variables are balanced when searching for the new axes with maximum inertia.

Based on the results from our risk prediction, we concluded that the use of the leading principal components as predictors in the regression model is appropriate. Our analysis demonstrated that using different number of PCs as predictors would achieve comparable model performance to that using the original variables as predictors. When using all PCs, the predictive performance would be exactly the same as using all the original variables. Although, with larger number of PCs in general would achieve better predictive performance, one can find that even with a small number of PCs as predictor (e.g., 1 PC, 2 PCs, or 3 PCs), the predictive performance of the model is still acceptable. Thus, data after dimension reduction can be used in risk prediction. However, one should notice that using the reduced PCs as predictors does not necessarily achieve better performance than using variables obtained from the traditional variable selection method (e.g., stepwise procedure). In our analysis, the 16 variables selected by using the stepwise procedure (using AIC as the selection criterion) performs better than using any reduced PCs. Nevertheless, when there are hundreds or thousands of candidate variables, using principal components from the dimension reduction methods will be a lot more efficient and applicable than the variable selection approach.

When compared between applying dimension reduction as a data preprocessing step before using the PAM clustering algorithm with and directly applying PAM to the original data we found very different results. In fact, even if we used all the principal components for clustering, the results were substantially different from applying clustering algorithm to the original variables. This indicates that principal components do not necessarily keep the underlying clustering structure. Therefore, using dimension reduction as a data pre-processing step for clustering may not be appropriate. Our results are consistent with previous researchers' findings. Previous research showed that applying principal components for dimension reduction before clustering algorithm is not justified in general[23] and that the most important principal components do not necessarily contain any clustering information from the original data.[24] Some research proposed adaptive dimension reduction to deal with this issue.[25-27] Using adaptive dimension reduction, clustering and subspace learning are performed simultaneously to avoid the information loss as in the case when the clustering and subspace learning are separated (e.g., using dimension reduction as a data preprocessing step for clustering). However, in many applications, dimension reduction is still applied as a preprocessing step for clustering algorithms in high-dimensional data structure. Our application demonstrates the inappropriateness of using dimension reduction method to produce principal components as fixed new input for clustering algorithm. In practice, one needs to be cautious of using dimension reduction as a data preprocessing step for clustering application.

In summary, dimension reduction technique is useful in mapping data into a much lower dimensional space to filter out noise while maintaining the most important variabilities in the original high-dimensional data. When applying dimension reduction in a data preprocessing step, it is important to use caution because not all features of the original data will be kept. The loss of a critical data feature during dimension reduction could yield erroneous results in the subsequent analysis.

Appendix A Table Dictionary

Table 7 Data dictionary for all variables

variable	Variable name
age	age
albumin	albumin
ln_alt	Alanine aminotransferase (ALT), log-transformed
ln_ast	aspartate aminotransferase (AST), log-transformed
ln_bands	BANDS, log-transformed
ln_bili	serum total bilirubin level, log-transformed
ln_bun	blood urea nitrogen (BUN) level, log-transformed
cl	Chloride
ln_cr	creatinine, log-transformed
ln_gluc	glucose, log-transformed
hr	heart rate
hgb	Hemoglobin
ln_pao2	PaO2, log-transformed
ln_platelets	platelets count, log-transformed
resp	respiratory rate
sodium	sodium
ln_sbp	systolic blood pressure (SBP) level, log-transformed
temp	temperature
ln_wbc	white blood cell counts, log-transformed
ln_pt	prothrombin time, log-transformed
ln_sat	101 – oxygen saturation, log-transformed
mort28	death status of 28-day
comorbid	number of comorbidity conditions
gcs	GCS levels
male	gender
type_culture	gram stain of bacterial pathogen
drugresis	drug resistance
infesite	site of infection

Appendix B Plots

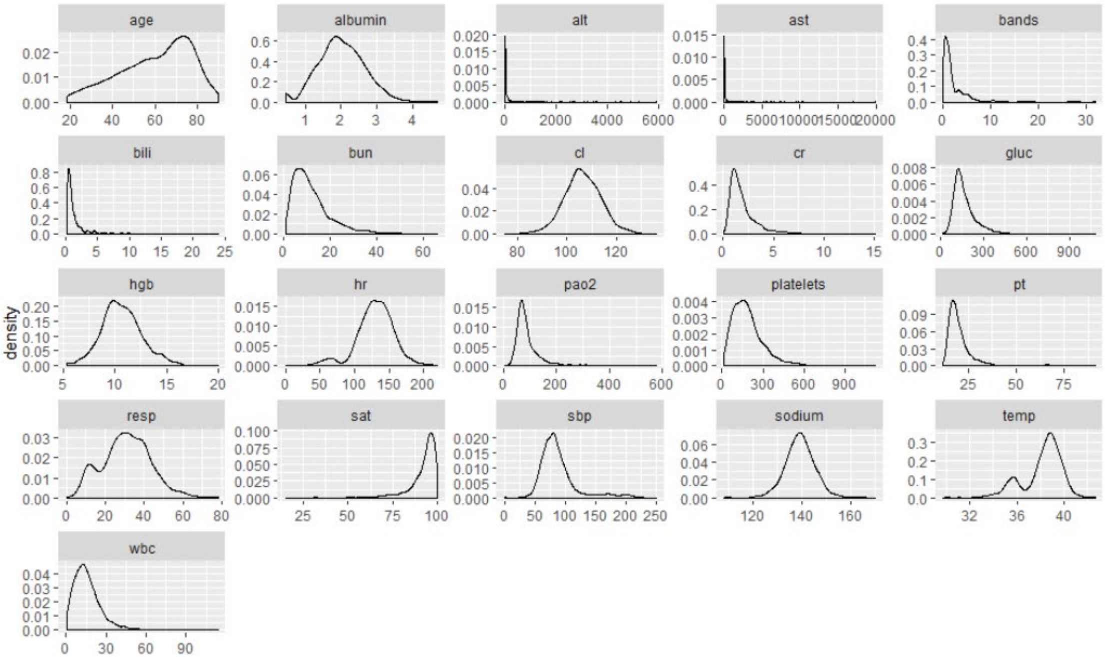

Figure 9 Density plots for raw quantitative variables

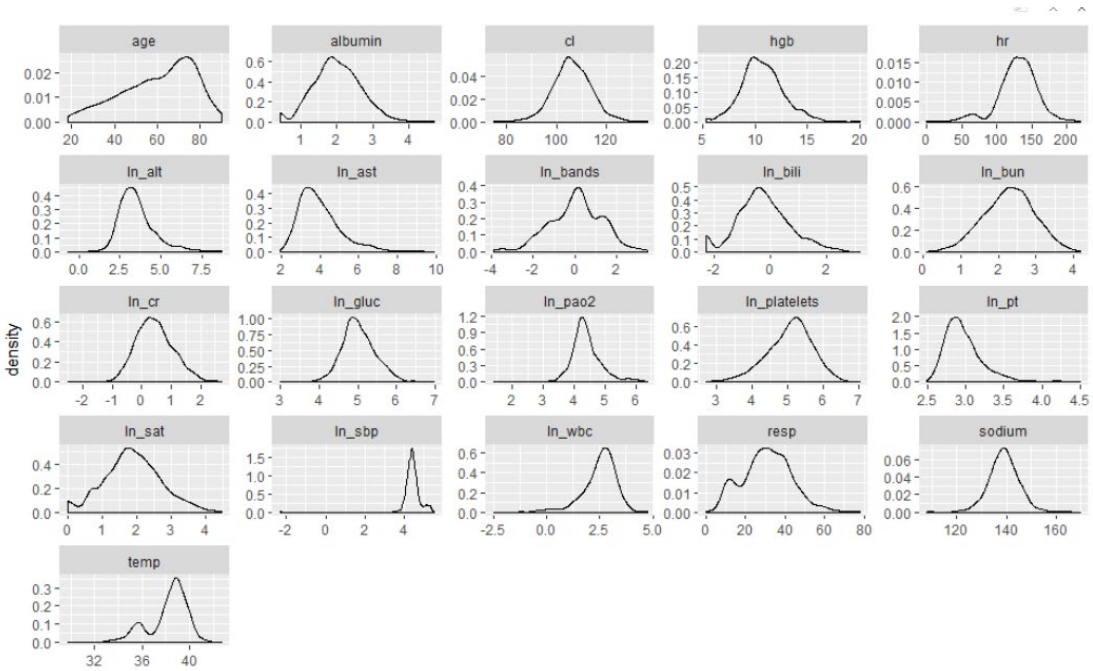

Figure 10 Density plot for quantitative variables with necessary log-transformation

Bibliography

1. Kim S, Kim W, Park RW. A Comparison of Intensive Care Unit Mortality Prediction Models through the Use of Data Mining Techniques. *Healthcare informatics research.* 2011;17(4):232-243.

2. Lodhi MK, Ansari R, Yao Y, Keenan GM, Wilkie D, Khokhar AA. Predicting Hospital Re-Admissions from Nursing Care Data of Hospitalized Patients. Paper presented at: Advances in Data Mining. Applications and Theoretical Aspects; 2017//, 2017; Cham.

3. Chen W, Hu Y, Zhang X, et al. Causal risk factor discovery for severe acute kidney injury using electronic health records. *BMC Medical Informatics and Decision Making.* 2018;18.

4. Koola JD, Davis SE, Al-Nimri O, et al. Development of an automated phenotyping algorithm for hepatorenal syndrome. *Journal of Biomedical Informatics.* 2018;80:87-95.

5. Imhoff M, Fried R, Gather U, Lanius V. Dimension reduction for physiological variables using graphical modeling. *AMIA Annual Symposium proceedings AMIA Symposium.* 2003;2003:313-317.

6. Beyer K, Goldstein J, Ramakrishnan R, Shaft U. When is “nearest neighbor” meaningful? Paper presented at: International conference on database theory1999.

7. Ding C, Xiaofeng H, Hongyuan Z, Simon HD. Adaptive dimension reduction for clustering high dimensional data. Paper presented at: 2002 IEEE International Conference on Data Mining, 2002. Proceedings.; 9-12 Dec. 2002, 2002.

8. Van Der Maaten L, Postma E, Van den Herik JJJMLR. Dimensionality reduction: a comparative. 2009;10:66-71.

9. Wang C-W, You W-H. Boosting-SVM: effective learning with reduced data dimension. *Applied Intelligence.* 2013;39(3):465-474.

10. Jolliffe IT. *Principal component analysis.* 2nd ed. New York: Springer; 2002.

11. Cox TF, Cox MAA. *Multidimensional scaling.* Vol 59;59.;. 1st ed. London: Chapman & Hall; 1994.

12. Hyvarinen AJNcs. Survey on independent component analysis. 1999;2(4):94-128.

13. Mika S, Schölkopf B, Smola AJ, Müller K-R, Scholz M, Rätsch G. Kernel PCA and de-noising in feature spaces. Paper presented at: Advances in neural information processing systems1999.

14. Roweis ST, Saul LK. Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science.* 2000;290(5500):2323-2326.

15. Belkin M, Niyogi P. Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. *Neural Computation.* 2003;15(6):1373-1396.

16. Tenenbaum JB, De Silva V, Langford JCJs. A global geometric framework for nonlinear dimensionality reduction. 2000;290(5500):2319-2323.

17. Weinberger KQ, Saul LKJIjocv. Unsupervised learning of image manifolds by semidefinite programming. 2006;70(1):77-90.

18. Le Roux B, Rouanet H. *Multiple correspondence analysis.* Vol no. 07-163.;163;. Thousand Oaks, Calif: Sage Publications; 2010.

19. Pagès J. *Multiple factor analysis by example using R.* Vol 18. 1 ed. Boca Raton: CRC Press, Taylor & Francis Group; 2015.

20. Shlens JJAowuahwcpepmP-T-Ijp. A Tutorial on Principal Component Analysis: Derivation, Discussion and Singular Value Decomposition. 2003. 2003.

21. Jollife IT, Cadima J. Principal component analysis: A review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences.* 2016;374(2065):20150202.

22. Bernard GR, Vincent J-L, Laterre P-F, et al. Efficacy and Safety of Recombinant Human Activated Protein C for Severe Sepsis. *The New England journal of medicine.* 2001;344(10):699-709.

23. Chang W-C. On Using Principal Components Before Separating a Mixture of Two Multivariate Normal Distributions. *Journal of the Royal Statistical Society Series C (Applied Statistics).* 1983;32(3):267-275.

24. Scrucca L. Dimension reduction for model-based clustering. *Statistics and Computing.* 2010;20(4):471-484.

25. Ding C, He X, Zha H, Simon HD. Adaptive dimension reduction for clustering high dimensional data. 2002.

26. Ding C, Li T. Adaptive dimension reduction using discriminant analysis and K -means clustering. 2007.

27. Wang X-D, Chen R-C, Zeng Z-Q, Hong C-Q, Yan F. Robust Dimension Reduction for Clustering With Local Adaptive Learning. *IEEE Transactions on Neural Networks and Learning Systems*. 2019;30(3):657-669.