# Unsupervised Feature Selection Using Weighted Principal Components

Seoung Bum Kim
Division of Information Management Engineering
Korea University
Anam, Seongbuk, Seoul 136-713
Republic of Korea
sbkim1@korea.ac.kr
Tel. 82-2-3290-3397

Panaya Rattakorn
Department of Industrial and Manufacturing Systems Engineering
University of Texas at Arlington
Arlington, TX 76019
United States of America

# Unsupervised Feature Selection Using Weighted Principal Components


## Abstract

Feature selection has received considerable attention in various areas as a way to select informative features and to simplify the statistical model through dimensional reduction. One of the most widely used methods for dimensional reduction includes principal component analysis (PCA). Despite its popularity, PCA suffers from a lack of interpretability of the original feature because the reduced dimensions are linear combinations of a large number of original features. Traditionally, two or three dimensional loading plots provide information to identify important original features in the first few principal component dimensions. However, the interpretation of what constitutes a loading plot is frequently subjective, particularly when large numbers of features are involved. In this study, we propose an unsupervised feature selection method that combines weighted principal components (PCs) with a thresholding algorithm. The weighted PC is obtained by the weighted sum of the first *k* PCs of interest. Each of the *k* loading values in the weighted PC reflects the contribution of each individual feature. We also propose a thresholding algorithm that identifies the significant features. Our experimental results with both the simulated and real datasets demonstrated the effectiveness of the proposed unsupervised feature selection method.


**Keywords**: Data mining; Feature selection; Principal component analysis; Unsupervised learning

## 1. Introduction

One of the major challenges associated with high-dimensional data is to identify a subset of relevant features of interest. In recent years, feature selection/extraction has received considerable attention in various areas for which datasets with thousands of features are present. The main purpose of feature selection/extraction is to identify a subset of features that are most predictive or informative in a given a dataset. Successful implementation of feature selection/extraction simplifies the entire modeling process and thus reduces computational and analytical efforts.

It is important to distinguish between feature selection and feature extraction, although much of the literature fails to clearly distinguish between them (Jain et al., 2000). Feature selection is a process to select a subset of original features, and feature extraction creates new features through the transformation of the original features (Guyon and Elisseeff, 2003). Widely used feature extraction methods include principal component analysis (PCA) and partial least squares (PLS). PCA is an unsupervised feature extraction method in that the process depends solely upon the input variables, and does not take into account information from the output variable (Jolliffe, 2002). On the other hand, PLS is a supervised feature extraction in that the process takes into account both the input and output variables (Kim, 2008). In general, the first few transformed features obtained from PCA and PLS suffice to provide useful information in the original data. However, because these reduced dimensions from PCA and PLS are linear combinations of a large number of original features, their interpretation cannot be readily made and the extraction of meaningful information is cumbersome.

Interpretation problems posed by the transformation process in PCA and PLS can be overcome by using feature selection methods that simply pick the subset of original features.

Feature selection methods can also be divided into supervised and unsupervised. Supervised feature selection methods use the information of an output variable to identify the best subset of given features in a dataset. Genetic algorithms have been successfully used as an efficient method of supervised feature selection for a high-dimensional spectral dataset (Cho et al., 2008; Davis et al., 2003). Moreover, supervised feature selection problems have been formulated by a multiple hypothesis testing procedure that controls the false discovery rate (Mei et al., 2009; Kim et al., 2008).

Despite extensive research in using the supervised/unsupervised feature extraction and supervised feature selection, relatively few attempts have been made to identify the important features by using unsupervised feature selection methods (Mao, 2005). Unsupervised feature selection methods usually have been divided into three categories — wrapper, filter, and hybrid approaches (Kim and Gao, 2006). The filter approach employs the general characteristics of the data to select a subset of the original data without using any clustering algorithms. In contrast, the wrapper approach necessitates the use of a predetermined clustering algorithm as evaluation criterion. The hybrid approach combines both the filter and wrapper approaches by using different evaluation criteria for each different state (Kim and Gao, 2006).

Dy and Brodley (2000) introduced a wrapper approach that uses an expectation-maximization (EM) clustering algorithm. Hastie et al. (2000) developed a gene-shaving method that used its first principal component to identify the best subsets of those features with large variations. Ding (2003) proposed a two-way ordering approach in which relevant genes were selected based on their similarity information.

Mao (2005) proposed a filter approach that sought to select a subset of original features by using principal components combined with an evaluation based on least square estimation

(LSE). Motivated by Mao's idea, Kim and Gao (2006) developed a two-step hybrid approach. The first step is to subsets of features based on an LSE-based evaluation; the second is to apply a searching algorithm to obtain the best subsets that maximize clustering performance.

Although all of the existing unsupervised feature selection methods performed reasonably well within the limits of the situations for which they were designed, no consensus exists about which of them best satisfies all conditions. Moreover, most of the methods require a high computational load because they involve an extensive search procedure such as the forward selection or the backward elimination. Consequently, the methods based on a search algorithm are not relevant for identifying important features in high-dimensional dataset, often encountered in various applications in these days. In the present study, we propose a new unsupervised feature selection that combines the weighted principal components with a thresholding algorithm. To be specific, the contribution of each feature is represented by a loading value in a weighted principal component, and a thresholding algorithm based on a moving range-based control chart evaluates the significance of its contribution. The proposed method belongs to the filter category and is computationally efficient and easy to implement.

The remainder of this paper is organized as follows. Section 2 presents the proposed unsupervised feature selection method. Section 3 presents the simulation study that examined the performance of the proposed method under various scenarios. Section 4 describes a case study developed to demonstrate the feasibility and effectiveness of the proposed method in real situations. Finally, Section 5 presents our concluding remarks.

## 2. The Proposed Unsupervised Feature Selection Approach

### 2.1. Weighted Principal Components

PCA is one of the most widely used multivariate data analysis techniques and is employed primarily for dimensional reduction and visualization (Jolliffe, 2002). PCA extracts a lower dimensional feature set that can explain most of the variability within the original data. The extracted features, $PC_i$'s ($Y_i$) are each a linear combination of the original features with the loading values ($\alpha_{ij}$, $i, j$=1,2,…,$p$). The $Y_i$'s can be represented as follows:

$$\begin{aligned} Y_1 &= \alpha_{11}X_1 + \alpha_{12}X_2 + \ldots + \alpha_{1p}X_p \\ Y_2 &= \alpha_{21}X_1 + \alpha_{22}X_2 + \ldots + \alpha_{2p}X_p \\ &\vdots \\ Y_p &= \alpha_{p1}X_1 + \alpha_{p2}X_2 + \ldots + \alpha_{pp}X_p \end{aligned} \quad (1)$$

The loading values represent the importance of each feature in the formation of a PC. For example, $\alpha_{ij}$ indicates the degree of importance of the $j$th feature in the $i^{th}$ PC. A two-dimensional loading plot (e.g., PC1 vs PC2 loading plot) may provide a graphical display for identification of important features in the first and second PC domains. However, the interpretation of a two-dimensional loading plot is frequently subjective, particularly in the presence of a large number of features. Moreover, in some situations, consideration of only the first few PCs may be insufficient to account for most of the variability in the data. Determination of the appropriate number of PCs (=$k$) to retain can be subjective. One can use a scree plot that visualizes the proportion of variability of each PC to determine the appropriate number of PCs (Johnson and Wichern, 2002).

If a PCA loading value for the $j$th original feature can be computed from the first $k$ PCs, the importance of the $j$th feature can be represented as follows:

$$\omega_j = \sum_{i=1}^{k} |\alpha_{ij}| \pi_i \text{ , } j=1, 2, \ldots, p, \quad (2)$$

where $k$ is the total number of features of interest and $\pi_i$ represents the weight of $i$th PC. The typical way to determine $\pi_i$ is to compute the proportion of total variance explained by the $i$th PC. $\omega_j$ can be called a weighted PC loading for the feature $j$.

For illustration, Figure 1 displays a plot of $\omega_j$ s, computed from a simulated dataset that contains 1,000 features. A feature with a large value of $\omega_j$ indicates a significant feature. In the next section, we will present a systematic way to obtain a threshold that determines the significance of each feature.

**[Figure 1 about here.]**

2.2. Moving Range-Based Thresholding Algorithm

We propose a moving range-based thresholding algorithm as a way to identify the significant features from the weighted PC loadings discussed in the previous section. The main idea of a moving range-based thresholding algorithm comes from a moving average control chart that has been widely used in quality control (Vermaat et al. 2003). A control chart provides a comprehensive graphical display for monitoring the performance of a process over time so as to keep the process within control limits (Woodall and Montgomery, 2001). A typical control chart comprises monitoring statistics and the control limit. When the monitoring statistics exceed (or fall below) the control limit, an alarm is generated so that proper remedial action can be taken. A moving range control chart is useful when the sample size used for process monitoring is one. Moreover, the average moving range control charts perform reasonably well when the observations deviate moderately from the normal distribution (Vermaat et al. 2003).

In our problem, we can consider the weighted PC loading values as the monitoring statistics. Thus, we plot these loading values on the moving range control chart and identify the significant features when the corresponding weighted PC loading exceeds the control limit (threshold). Given a set of the weighted PC loading values for individual features $(\omega_1, \omega_2, ..., \omega_p)$, the threshold $(\gamma)$ can be calculated as follows (Vermaat et al. 2003):

$$\gamma = \overline{\omega} + \Phi^{-1}(1-\alpha)\frac{\sqrt{\pi}}{2} * \sigma, \tag{3}$$

where $\overline{\omega} = \frac{1}{p}\sum_{i=1}^{p}\omega_i$, $\Phi^{-1}$ is the inverse standard normal cumulative distribution function, and $\alpha$ is the Type I error rate that can be specified by the user. The range of $\alpha$ is between 0 and 1. In typical moving range control charts, $\sigma$ can be estimated by $\overline{MR}$, calculated by the average of the moving ranges of two successive observations.

$$\overline{MR} = \frac{|\omega_1 - \omega_2| + |\omega_2 - \omega_3| + \ldots + |\omega_{p-1} - \omega_p|}{p-1} \tag{4}$$

However, in our feature selection problems, because the weighted PC loading values for individual features $\omega_1, \omega_2, ..., \omega_p$ are not ordered, we cannot simply use (4). To address this issue, we propose a different way of computing the $\overline{MR}$ that can properly handle a set of unordered observations. Given the fact that there is no specific order of observations $\omega_1, \omega_2, ..., \omega_p$, they are randomly reshuffled, and $\overline{MR}s$ are recalculated. Therefore, for $B$=1,000, we obtain a set of $\overline{MR}s$ $\overline{MR}_{(1)}, \overline{MR}_{(2)}, ..., \overline{MR}_{(B)}$. The $\overline{MR}$ for unordered observations is calculated by

$$\overline{MR}^* = \frac{1}{B}\sum_{j=1}^{B}\overline{MR}_{(j)}, \tag{5}$$

Finally, the threshold of the proposed feature selection method can be obtained by the following equation:

$$\gamma = \overline{\omega} + \Phi^{-1}(1-\alpha)\frac{\sqrt{\pi}}{2}\overline{MR}^{*}. \quad (6)$$

A feature is reported as significant if the corresponding weighted PC loading exceeds the threshold $\gamma$.

2.3. Feature Validity Measures

We used sensitivity and specificity as performance measures (Altman and Bland, 1994). Sensitivity and specificity can be expressed as follows:

$$\text{Sensitivity} = \frac{TP}{TP + FN}, \quad (7)$$

$$\text{Specificity} = \frac{TN}{TN + FP}, \quad (8)$$

where *TP* is the number of true positives (number of true significant features identified), *TN* is the number of true negatives (number of true insignificant features identified), *FN* is the number of false negatives, and *FP* is the number of false positives. In short, sensitivity is the proportion of true positives correctly identified by the procedure. Specificity is the proportion of true negatives correctly identified. The range of both sensitivity and specificity is between 0 and 1. The method that produces the largest sensitivity and specificity scores would be considered the better method.

## 3. Simulation Study

### 3.1. Simulated Data

A simulation study evaluated the performance of the proposed method and compared it with other algorithms under various scenarios. Table 1 shows a summary of the simulated data used in this study.

**[Table 1 about here.]**

Each scenario contains the number of observations, the number classes, the number of true significant features, and different degrees of shifts in the mean. Specifically, the simulated data in Scenarios 1 ~ 3 contain two class datasets in which the covariance matrix of each class is the identity matrix ($\Sigma_1 = \Sigma_2 = I$). The mean of Class 1 equals zero, and the mean of Class 2 equals the mean of Class 1 plus the shift in mean as shown in the last column of Table 1. Other scenarios can be explained similarly.

### 3.2. Simulation Results

Table 2 presents the number of identified features, sensitivity, specificity, and computational time (CPU time) in the 10 simulation scenarios. The experiments were conducted on an Intel® Core™2 Duo @ 2.2 GHz computer with 2 GB memory. We compared the proposed weighted PC loading method with the LSE method (Mao, 2005), one of the existing unsupervised feature selection methods. In the LSE method, a subset of significant features was selected based on error reduction after adding additional features. The error reduction function was calculated based on PCs that were obtained from the PCA in the complete data. A sequential forward selection strategy was then used to determine a subset of significant features (Mao, 2005).

The results showed that across all simulation scenarios, the sensitivities and specificities of the proposed method were all one, implying that our method successfully detected all the true significant features. The LSE method also yielded sensitivity and specificity results comparable with the proposed method. However, the LSE method tended to identify more numbers of features than the number of true significant features. More important, the LSE method imposes a high computational load compared with the proposed method. In particular, faced with more than 3,000 features, the LSE method takes a significant amount of time to identify the significant ones.

**[Table 2 about here.]**

## 4. Experiments with Real Data

In addition to the simulation study, we used three real datasets (Wisconsin diagnostic breast cancer, wine, and leukemia microarray) to demonstrate the effectiveness of the proposed weighted PC loading method. These datasets are available on the UCI database (http://archive.ics.uci.edu/ml/), and their summary is shown in Table 3.

**[Table 3 about here.]**

We evaluated the performance of the proposed method and compared it with the Baseline Case and the LSE method. The Baseline Case represents the use of all features for comparison. Table 4 shows feature selection results, classification accuracy derived from a classification algorithm, and CPU time on the real datasets. Classification accuracy is defined as the number of observations correctly classified divided by the total number of observations. To compute classification accuracy, we used a support vector machines (SVM) algorithm, one of the most widely used classification methods (Shawe-Taylor and Cristianini, 2000). The SVM classification accuracy reported here is the average value ± standard deviation from 10-fold cross validation. Note that specificity and sensitivity were not reported here; their omission is because

information about the true clusters is unknown in real data. Moreover, we did not report CPU time for the Baseline Case because this case does not involve any feature selection process but instead uses all of the features.

**[Table 4 about here.]**

In the Wisconsin breast cancer data, our proposed method identified the smallest number of significant features but produced classification accuracy comparable to the Baseline Case and the LSE method. In order to explore more about this outcome, Figure 2 shows the dot plots of two features identified by our proposed method according to the status of patient (malignant, benign). These two features are the mean area of the cell nucleus and the mean of the three largest area values. These features clearly distinguished between benign and malignant samples. Further, we also generated dot plots of two features that the LSE method identified but the proposed method did not (Figure 3). These features are the mean of texture and standard error of perimeter. It can be seen that these feature could not clearly differentiate between benign and malignant samples.

**[Figure 2 about here.]**

**[Figure 3 about here.]**

In the wine data, the proposed weighted PC loading method algorithm identified only one significant (proline) feature out of 13. This one-feature result of SVM classification is not significantly worse in terms of accuracy than the three-feature performance of the LSE method (Table 4). Figure 4 displays the dot plot of the "proline" feature by the type of wine. A clear distinction can be observed between the first and the second and third types. However, this proline may not be a good feature for distinguishing between the second and third types of wine. Figure 5 shows a dot plot of the feature (alkalinity of ash) identified by the LSE method but not

by the weighted PCA loading method. There is an overlapping among the samples, indicating that the feature, alkalinity of ash may not play a significant role in discriminating the type of wine.

**[Figure 4 about here.]**

**[Figure 5 about here.]**

In the microarray leukemia data, our proposed method with $\alpha = 0.05$ identified 457 features as significant out of 7,129 and produced an even better result than the Baseline Case (Table 4). The performance of the LSE method is not reported here because it requires a significant amount of time (more than 48 hours), which of itself is enough to disqualify it as a valid competitor with our proposed method.

## 5. Conclusions

We have presented a new method of unsupervised feature selection for identification of important features in high-dimensional datasets. The proposed method combines PCA techniques and a moving range-based thresholding algorithm. We first obtained the weighted PC, which can be calculated by the weighted sum of the first *k* PCs of interest. Each of the *k* loading values in the weighted PC reflects the contribution of each individual feature. To identify the significant features, we proposed a moving-range thresholding algorithm. Features are considered to be significant if the corresponding weighted PC loadings exceed the threshold obtained by a moving-range thresholding algorithm. Our experimental results with both simulated and real datasets demonstrated that the proposed method could successfully detect the true significant features. Moreover, compared with LSE, which is one of the existing methods of

unsupervised feature selection, the proposed method requires significantly lesser computational loads and thus can efficiently handle high-dimensional datasets.

Our study extends the application scope of both the PCA and control chart techniques. We hope that the procedure discussed here stimulates further investigation into development of better procedures for problems of unsupervised feature selection.

**References**


1. Altman D.G. and Bland J.M. (1994), "Diagnostic tests. 1: sensitivity and specificity," BMJ, Vol. 308, p. 1552.

2. Cadima, J. F. and Jolliffe, I. T. (1995), "Loadings and correlations in the interpretation of principal components," Journal of Applied Statistics, Vol. 22 (2), pp. 203-214.

3. Cadima, J.F. and Jolliffe, I. T. (2001), "Variable selection and the interpretation of principal subspaces," Journal of Agricultural, Biological, and Environmental Statistics, Vol. 6 (1), pp. 62-79.

4. Cho, H.-W., Kim, S.B., Jeong, M., Park, Y., Ziegler, T. R., and Jones, D. P. (2008), "Genetic algorithm-based feature selection in high-resolution NMR spectra," Expert Systems With Applications, Vol. 35, pp. 967-975.

5. Dash, M. and Liu, H. (1997), "Feature selection for classification," Intelligent Data Analysis: An International Journal, Vol. 1 (3), pp. 131-156.

6. Dash, M., Liu, H., and Yao, J. (1997), "Dimensionality reduction of unsupervised data," Proceedings Ninth IEEE International Conference on Tools with AI (ICTAI '97), pp. 532-539.

7. Davis, R.A., Charlton, A.J., Oehlschlager, S. and Wilson, J.C. (2006), "Novel feature selection method for genetic programming using metabolomic H1 NMR data," Chemometrics and Intelligent Laboratory Systems, Vol. 81, pp. 50-59.

8. Ding, C. (2003), "Unsupervised feature selection via two-way ordering in gene expression analysis," Bioinformatics, Vol. 19 (10), pp. 1259-1266.

9. Dy, J. and Brodley, C. (2000), "Feature subset selection and order identification for unsupervised learning," Proceedings of the 17th International Conference on Machine Learning, pp. 247-254.

10. Guyon, I. and Elisseeff, A. (2003), "An introduction to variable and feature selection," Journal of Machine Learning Research, Vol. 3, pp. 1157-1182.

11. Handl, J. and Knowles, J. (2006), "Feature subset selection in unsupervised learning via multiobjective optimization," International Journal on Computational Intelligence Research, Vol. 2 (3), pp. 217-238.

12. Jain, A.K., Duin, R. P.W., and Mao J. (2000), "Statistical pattern recognition: a review," IEEE Transactions on Pattern Analysis & Machine Intelligence, Vol.20, pp. 4-37.

13. Jansen, J. J., Hoefsloot, H. C. J., Boelens, H. F. M., Greef, J. V. D., and Smilde, A. K. (2004), "Analysis of longitudinal metabolomics data," Bioinformatics, Vol. 20, pp. 2438-2446.

14. Johnson, R. A. and Wichern, D. W. (2001), Applied multivariate statistical analysis, 5th Edition, Prentice-Hall, Inc., New Jersey.

15. Jolliffe, I. T. (2002), Principal Component Analysis, Springer-Verlag, New York.

16. Kim, S.B. (2008), Features extraction and selection in high-dimensional spectral data, Encyclopedia of Data Warehousing and Mining (2nd Edition), J. Wang editor, IGI Global, Pennsylvania. pp. 863-869.

17. Kim, S.B., Chen, V. C. P., Park, Y., Ziegler, T. R., and Jones, D. P. (2008), "Controlling the false discovery rate for feature selection in high-resolution NMR spectra," Statistical Analysis and Data Mining, Vol.1, pp. 57-66.

18. Kim, Y. and Gao, J. (2006), "Unsupervised gene selection for high dimensional data," Proceedings of IEEE Symposium of Bioinformatics and Bioengineering (IEEE BIBE), pp. 227-232.

19. Mao, K. Z. (2005), "Identifying critical variables of principal components for unsupervised feature selection," IEEE Transactions on Systems, Man, and Cybernatics - Part B: Cybernatics, Vol. 35 (2), pp. 339-344.

20. Mei, Y., Kim, S.B., and Tsui, K.-L. (2009), "Identification of major metabolite features in high-resolution NMR spectra using linear-mixed effects models," Expert Systems with Applications, Vol. 36, pp. 4703-4708.

21. Morita, M., Sabourin, R., Bortolozzi, F., and Suen, C. Y. (2003), "Unsupervised feature selection using multi-objective genetic algorithms for handwritten word recognition," Proceedings of the Seventh International Conference on Document Analysis and Recognition, Vol. 2, pp. 666-671.

22. Shawe-Taylor, J. and Cristianini, N. (2000), Support Vector Machines and Other Kernel-Based Learning Methods, Cambridge University, New York.

23. Vermaat, M. B., Ion, R. A., Does, R. J. M. M., and Klaassen, C. A. J. (2003), "A comparison of Shewhart individuals control charts based on normal, non-parametric, and extreme-value theory," Quality and Reliability Engineering International, Vol. 19, pp. 337-353.

24. Woodall, W.H. and Montgomery, D.C. (1999), "Research issues and ideas in statistical process control," Journal of Quality Technology, Vol. 31, pp. 376-386.

Table 1. Summary of the simulated data

| Data | Number of features | Number of observations | Number of classes | Number of true significant features | Mean shift |
|---|---|---|---|---|---|
| Scenario 1 | 500 | 200 | 2 | 10 | 5σ |
| Scenario 2 | 100 | 200 | 2 | 10 | 3σ |
| Scenario 3 | 1000 | 200 | 2 | 10 | 1σ |
| Scenario 4 | 1000 | 400 | 4 | 100 | 1σ, 2σ, 3σ |
| Scenario 5 | 1000 | 400 | 4 | 20 | 5σ, 10σ, 20σ |
| Scenario 6 | 1000 | 400 | 4 | 20 | 0.5σ, 1σ, 2σ |
| Scenario 7 | 3000 | 200 | 2 | 100 | 0.5σ |
| Scenario 8 | 3000 | 200 | 2 | 20 | 2σ |
| Scenario 9 | 3000 | 200 | 2 | 300 | 1σ, 2σ, 3σ |
| Scenario 10 | 7000 | 200 | 2 | 300 | 1σ, 2σ, 3σ |

Table 2. Number of identified features, sensitivity (Se), specificity (Sp), and CPU time in 10 scenarios

| Scenario | Method | No. of true significant features | # of identified features | Se | Sp | CPU Time (Sec.) |
|---|---|---|---|---|---|---|
| 1 | LSE | 10 | 11 | 1.000 | 0.998 | 36.02 |
| | WPC + MR ($\alpha = 0.01$) | | 10 | 1.000 | 1.000 | 4.38 |
| | WPC + MR ($\alpha = 0.10$) | | 10 | 1.000 | 1.000 | 4.38 |
| 2 | LSE | 10 | 11 | 1.000 | 0.998 | 7.62 |
| | WPC + MR ($\alpha = 0.01$) | | 10 | 1.000 | 1.000 | 1.49 |
| | WPC + MR ($\alpha = 0.10$) | | 10 | 1.000 | 1.000 | 1.49 |
| 3 | LSE | 10 | 11 | 1.000 | 0.998 | 65.16 |
| | WPC + MR ($\alpha = 0.01$) | | 10 | 1.000 | 1.000 | 10.33 |
| | WPC + MR ($\alpha = 0.10$) | | 19 | 1.000 | 0.991 | 10.33 |
| 4 | LSE | 100 | 94 | 0.940 | 1.000 | 1663 |
| | WPC + MR ($\alpha = 0.01$) | | 100 | 1.000 | 1.000 | 10.73 |
| | WPC + MR ($\alpha = 0.10$) | | 100 | 1.000 | 1.000 | 10.73 |
| 5 | LSE | 20 | 21 | 1.000 | 0.998 | 185.35 |
| | WPC + MR ($\alpha = 0.01$) | | 20 | 1.000 | 1.000 | 10.91 |
| | WPC + MR ($\alpha = 0.10$) | | 20 | 1.000 | 1.000 | 10.91 |
| 6 | LSE | 20 | 21 | 1.000 | 0.998 | 176.89 |
| | WPC + MR ($\alpha = 0.01$) | | 20 | 1.000 | 1.000 | 10.97 |
| | WPC + MR ($\alpha = 0.10$) | | 20 | 1.000 | 1.000 | 10.97 |
| 7 | LSE | 100 | 94 | 0.940 | 1.000 | 3830 |
| | WPC + MR ($\alpha = 0.01$) | | 100 | 1.000 | 1.000 | 64.82 |
| | WPC + MR ($\alpha = 0.10$) | | 100 | 1.000 | 1.000 | 64.82 |
| 8 | LSE | 20 | 20 | 1.000 | 1.000 | 391.23 |
| | WPC + MR ($\alpha = 0.01$) | | 20 | 1.000 | 1.000 | 59.59 |
| | WPC + MR ($\alpha = 0.10$) | | 47 | 1.000 | 0.991 | 59.59 |
| 9 | LSE | 300 | 287 | 0.957 | 1.000 | 47882 |
| | WPC + MR ($\alpha = 0.01$) | | 300 | 1.000 | 1.000 | 60.85 |
| | WPC + MR ($\alpha = 0.10$) | | 300 | 1.000 | 1.000 | 60.85 |
| 10 | LSE | 300 | 378 | 1.000 | 0.988 | 215810 |
| | WPC + MR @ $\alpha = 0.01$ | | 300 | 1.000 | 1.000 | 275.42 |
| | WPC + MR @ $\alpha = 0.10$ | | 300 | 1.000 | 1.000 | 275.42 |

Table 3. Summary of real datasets

| Data | Number of features | Number of observations | Number of classes |
|---|---|---|---|
| Wisconsin diagnostic breast cancer | 30 | 569 | 2 |
| Wine | 13 | 178 | 3 |
| Leukemia | 7129 | 72 | 2 |

Table 4. Comparison of unsupervised feature selection methods on the Wisconsin diagnostic breast cancer, wine, and leukemia datasets

| Data | Method | No. of Identified Features | SVM Classification Accuracy (%) | CPU Time (Second) |
|---|---|---|---|---|
| Wisconsin Diagnostic Breast Cancer | Baseline | 30 | 97.37 ± 2.89 | - |
| | LSE | 12 | 95.59 ± 3.39 | 2.226 |
| | WPC + MR ($\alpha = 0.01$) | 2 | 92.27 ± 1.47 | 1.979 |
| | WPC + MR ($\alpha = 0.05$) | 2 | 92.27 ± 1.47 | 1.979 |
| Wine | Baseline | 13 | 97.19 ± 4.72 | - |
| | LSE | 3 | 98.86 ± 2.41 | 1.363 |
| | WPC + MR ($\alpha = 0.01$) | 1 | 93.79 ± 2.35 | 0.717 |
| | WPC + MR ($\alpha = 0.05$) | 1 | 93.79 ± 2.35 | 0.717 |
| Leukemia | Baseline | 7129 | 88.68 ± 1.59 | - |
| | LSE | - | - | - |
| | WPC + MR ($\alpha = 0.01$) | 384 | 87.03 ± 1.34 | 274.830 |
| | WPC + MR ($\alpha = 0.05$) | 457 | 90.29 ± 1.15 | 274.830 |

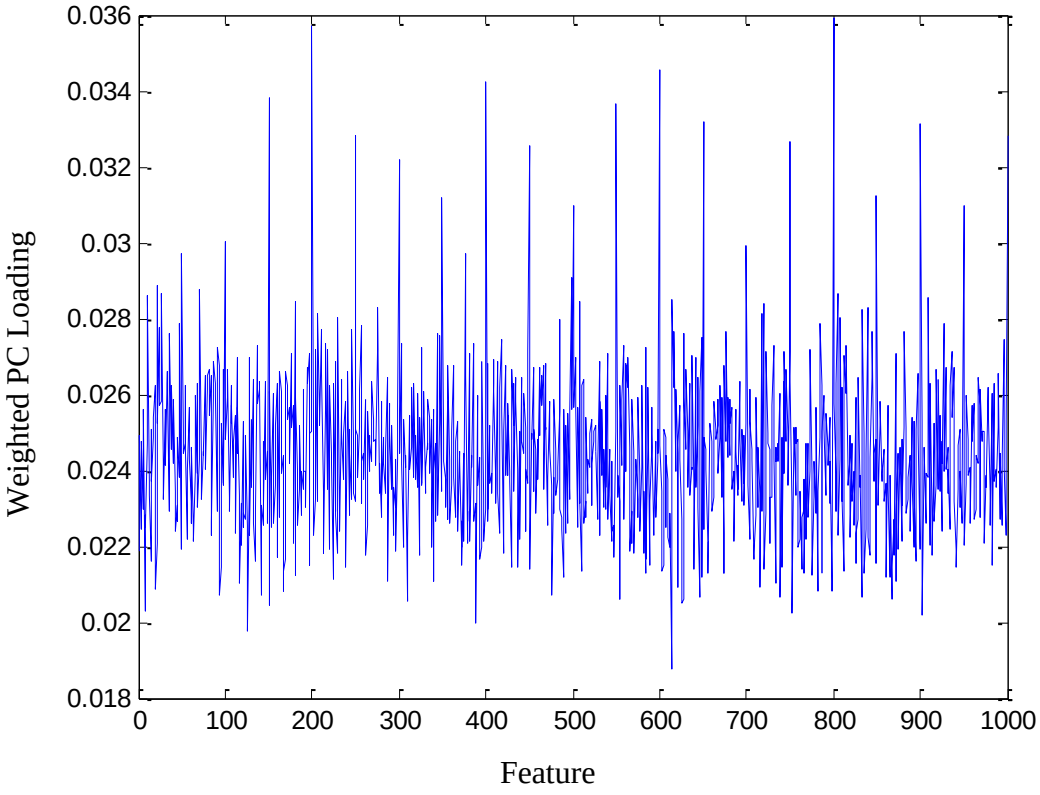



Figure 1. Weighted PC loading values of individual features.

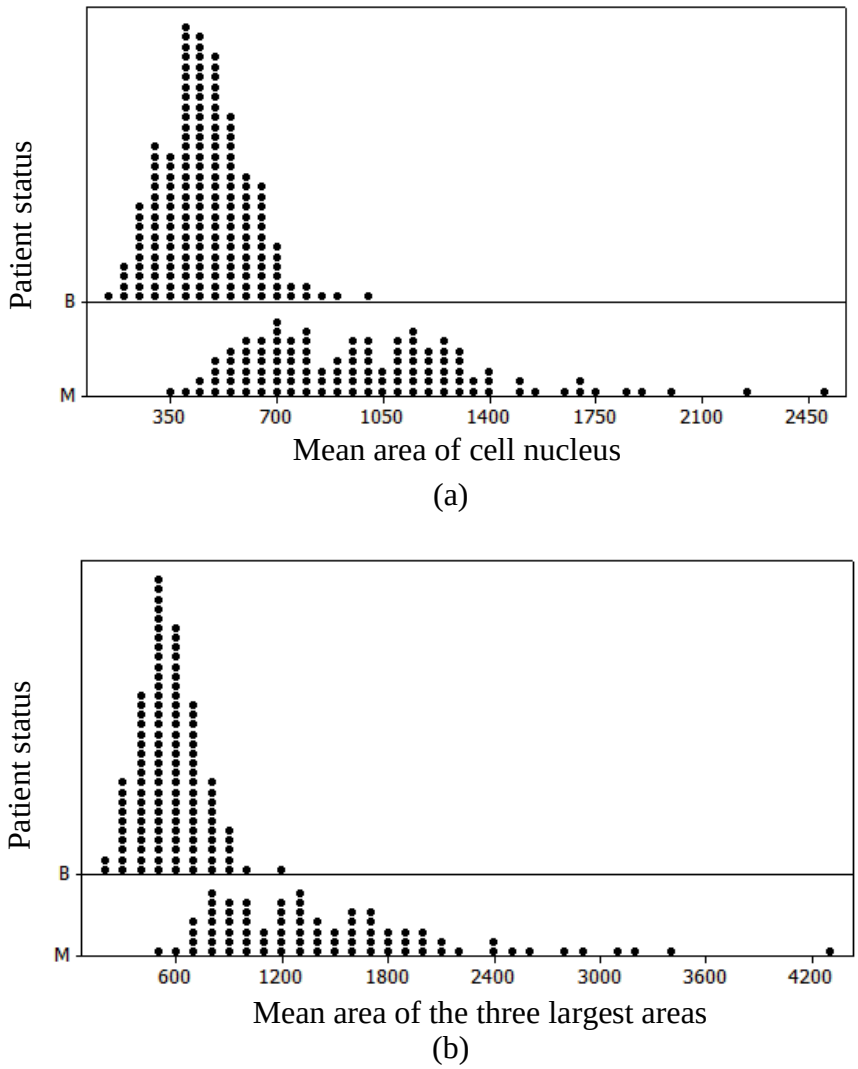
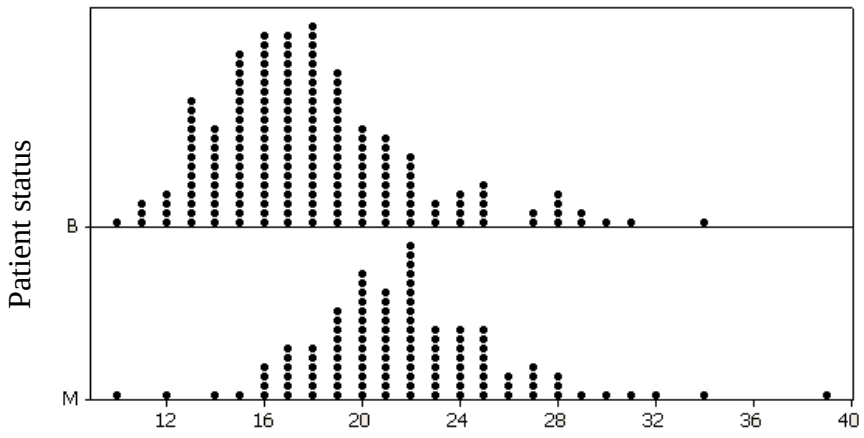
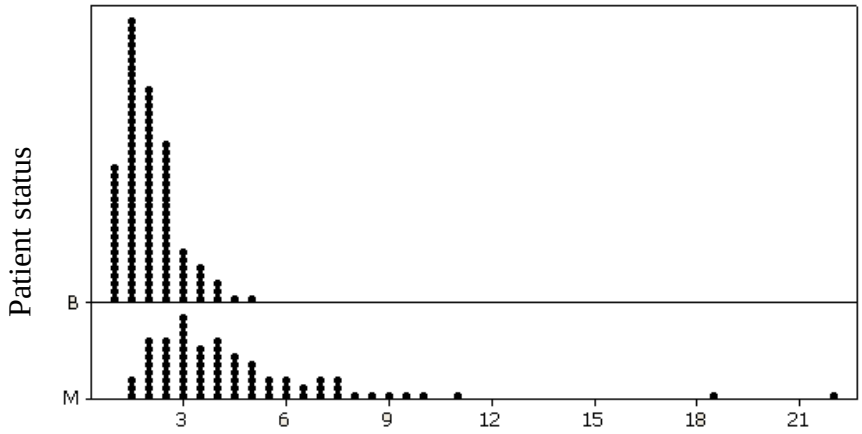



Figure 2. Dot plots of two significant features for Wisconsin diagnostic breast cancer data. (a) mean area of cell nucleus, (b) mean of the three largest areas feature. The features were identified by both the proposed method and the LSE method according to patient types (malignant, benign).




(a)




(b)

Figure 3. Dot plots of two significant features for Wisconsin diagnostic breast cancer data. (a) mean of texture for cell nucleus, (b) standard error of perimeter feature. The features were identified by only the LSE method (not by the proposed method) according to patient types (malignant, benign).

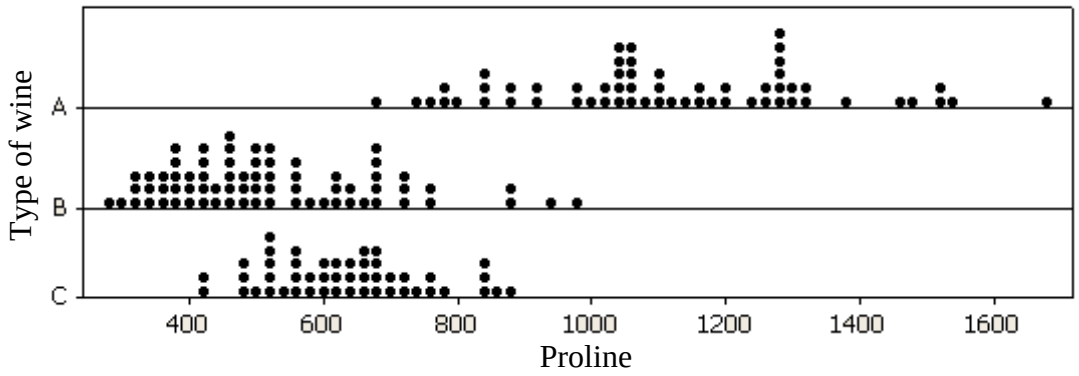



Figure 4. A dot plot of the proline feature by the type of wine. The feature was identified by the proposed method.

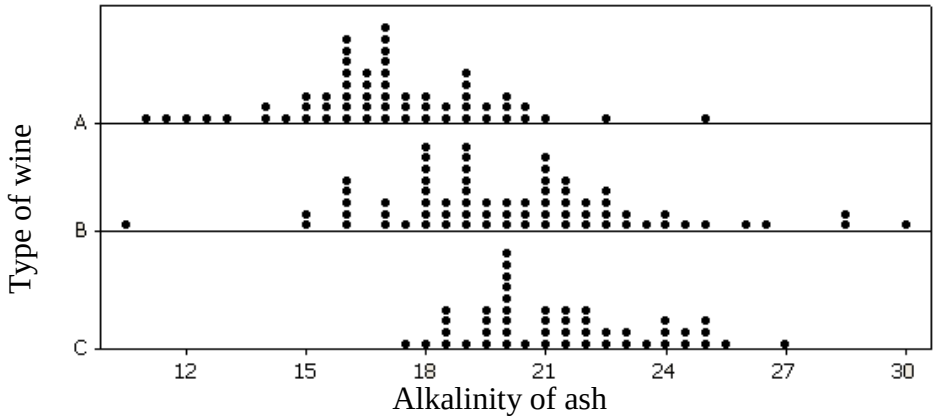



Figure 5. A dot plot of the significant feature (alkalinity of ash) identified by only the LSE method (not by the proposed method) according to wine types.