COMMITTEE ON DATA
CODATA
INTERNATIONAL SCIENCE COUNCIL

On the Application of Principal Component Analysis to Classification Problems

RESEARCH PAPER

JIANWEI ZHENG
CYRIL RAKOVSKI

*Author affiliations can be found in the back matter of this article

]u[ubiquity press

ABSTRACT

Principal Component Analysis (PCA) is a commonly used technique that uses the correlation structure of the original variables to reduce the dimensionality of the data. This reduction is achieved by considering only the first few principal components for a subsequent analysis. The usual inclusion criterion is defined by the proportion of the total variance of the principal components exceeding a predetermined threshold. We show that in certain classification problems, even extremely high inclusion threshold can negatively impact the classification accuracy. The omission of small variance principal components can severely diminish the performance of the models. We noticed this phenomenon in classification analyses using high dimension ECG data where the most common classification methods lost between 1 and 6% of accuracy even when using 99% inclusion threshold. However, this issue can even occur in low dimension data with simple correlation structure as our numerical example shows. We conclude that the exclusion of any principal components should be carefully investigated.

CORRESPONDING AUTHOR:
Jianwei Zheng

Chapman University, One University Drive Orange, CA, US

zheng120@mail.chapman.edu

KEYWORDS:
PCA; Dimensionality Reduction; Power; Classification Accuracy

TO CITE THIS ARTICLE:
Zheng, J and Rakovski, C. 2021. On the Application of Principal Component Analysis to Classification Problems. *Data Science Journal*, 20: 26, pp. 1–6. DOI: *https://doi.org/10.5334/dsj-2021-026*

INTRODUCTION

Principal Component Analysis (PCA) (Du et al., 2012; Hsieh et al., 2010; Mehmet Korürek, 2010; Kim et al., 2009) is a popular tool for data dimensionality reduction in the presence of complex correlation structure among a large number of numerical variables. The presence of correlations among the original variables in the data can be used to create new summary variables, principal components (PCs), that are optimal, uncorrelated linear combinations of the original variables. The optimality is represented by the fact that the PCs have the maximum possible variance among all linear combinations of the original variables and thus contain the maximum amount of information. The lack of correlation among the PCs removes the redundancy present in the original variables. The well-known lemma for maximization of quadratic forms for points on the unit sphere shows that the vectors of coefficients that define the PCs are the eigenvectors of the variance matrix. The eigenvalues associated with the eigenvectors equal the variance of the PCs and define an order among all PCs. The ones with the largest variance are considered the main PCs and provide an scheme for dimensionality reduction, and we take the first few PCs that jointly account for more than 80% or 90% of the total variance of the original variance. This approach makes intuitive sense as the PCs associated with the smallest eigenvalues are almost constant and thus have limited classification capability. However, in certain problems dimensionality reduction via PCA with even high cutoff for exclusion is not a good idea. This phenomenon was noticed when we implementing an arrhythmia classification on ECG data, even though some of studies demonstrated the PCA application on same research (Gupta and Mittal, 2019b, 2018b; Gupta et al., 2020; Gupta and Mittal, 2018a, 2016, 2019a). The ECG graph of a normal beat (shown in ***Figure 1***) consists of a sequence of waves, a P-wave presenting the atrial depolarization process, a QRS complex denoting the ventricular depolarization process, and a T-wave representing the ventricular repolarization. Our data consisted of 200 data points per heart beat with complex correlation structure that seemed ideal for preliminary PCA dimensionality reduction step before subsequent classification approach was employed. However, using PCA exclusion cutoffs of 90%, 92%, 95%, 99% for the 200 PCs dramatically improves classification accuracy rate. The PCA application processed a segment of ECG presented one time heartbeat is depicted in ***Figure 2***. This is an example revealing that PCA may not be a good idea for certain types of classification problems. A more detailed results that

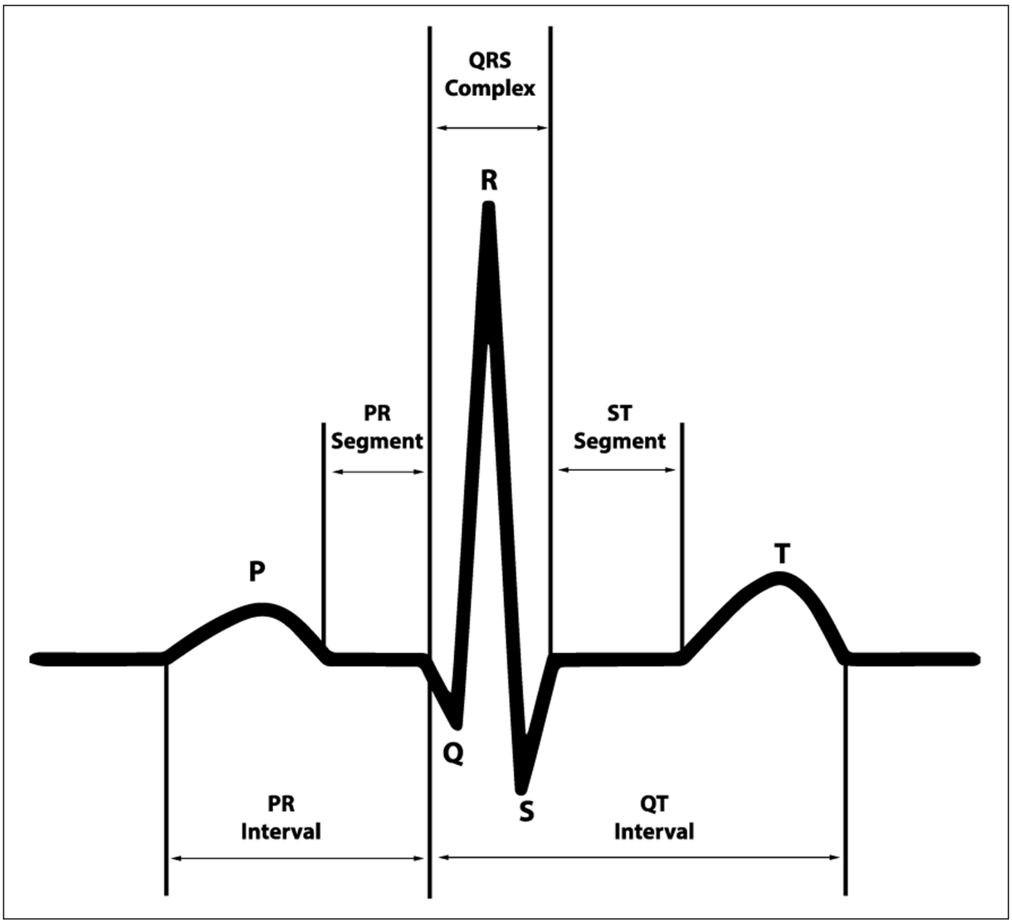

Figure 1 The ECG waveform and segments in lead II that presents a normal cardiac cycle.

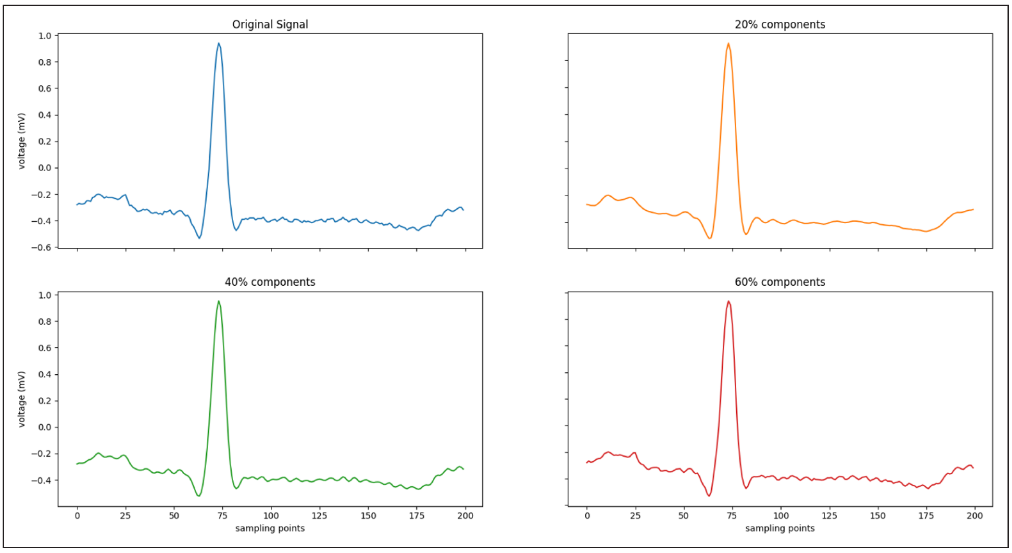

Figure 2 One heartbeat ECG presented by 100%, 20%, 40%, and 60% respectively.

highlight this finding are shown in ***Table 1***. We can see that the loss of classification accuracy using five common classification algorithms (random forest, conditional random forest, naive Bayes, multinomial logistic regression, and quadratic discriminant analysis) using the original ECG data and principal components accounting for 99% of the total variance was between 0.001 and 0.06. In subsequent presentation we show that omission of even the lowest ranked PCs can be disadvantageous to the classification accuracy of the algorithm.

CLASSIFIER NAME	***NON PCA***	***PCA*****	***THE DIFFERENCE***
Random Forest	0.96	0.92	−0.04
Conditional Random Forest	0.96	0.90	−0.06
Naive Bayes	0.92	0.87	−0.05
Multinomial Logistic Regression	0.94	0.94	−0.001
Quadratic Discriminant Analysis	0.93	0.90	−0.02

Table 1 Accuracy* comparison between classification models using original variables and principal components**.

* Accuracy is the average of 10 stratified folds.

** Principal components accounting for 99% of the variance used.

METHODS

Here is a mathematical description of data scenarios where this phenomenon can occur. Let Σ be the covariance matrix of the original variables x_1, x_2, ⋯, x_p and (λ_1, e_1), (λ_2, e_2) ⋯, (λ_p, e_p) be the eigenvalue-eigenvector pairs where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p$. Then, the PCs are $y_1 = e_1^T x_1, y_2 = e_2^T x_2, \ldots, y_p = e_p^T x_p$. The classical approach (Johnson and Wichern, 1988) for dimensionality reduction is to select the first s major PCs that jointly account for at least, say m * 100% of the total variance of the original variables,

$$s = min_{1\leq k\leq p} \frac{\lambda_1+\lambda_2+\ldots+\lambda_k}{\lambda_1+\lambda_2+\ldots+\lambda_p} \geqslant m. \qquad (1)$$

Now assume that we have a classification problem with two groups. Let G_i, i = 1, 2, ⋯, n be dichotomous variables that denote the group classification. Assume that the true underlying model describing the associations between G_i and y_{i1}, y_{i2}, ⋯, y_{ip} are given by the following logistic model,

$$Logit(P(G_i = 1 \mid y_{i1}, y_{i2}, \ldots, y_{ip})) = \beta_0 + \beta_1 y_{is+1} + \beta_2 y_{is+2} + \ldots + \beta_j y_{is+j}, \qquad (2)$$

where β_0, β_1, ⋯, β_j are the true effect sizes and $1 \leq j \leq p - s$. It is clear that under these conditions, the classification will be poor due to the exclusion of the true predictors from the data at the preprocessing step of dimensionality reduction. That omission entails low classification accuracy based on spurious association between the group and noise variables or no detectable classification capability at all.

Therefore, in its classical dimensionality reduction implementation, PCA, might not be useful for certain classification problems. In particular, in classification problems with complex patterns the lower ranked PCs are the ones that carry the information about group differences as the first several main PCs that reflect the correlation structure of the complex mean pattern and do not contain enough information about subtle group differences. Thus, if PCA is employed, we recommend that the PC inclusion thresholds should be carefully considered and based not only on the proportion of explained variance but also on the magnitude of the variance of the excluded PCs and the power to detect effect size of certain magnitude given the sample size (Schoenfeld D. A., 2005; F.Y. Hsieh and Larsen, 1998). In particular, if we consider $y_{s+1} = (y_{1s+1}, y_{2s+1}, \ldots, y_{n\,s+1})$ (with variance λ_{s+1}) for inclusion in subsequent analysis where the first l and subsequent $n - l$ subjects belong to groups 1 and 2 respectively. Let $\pi(\delta)$ denote the power to detect a difference of size δ between the group means subject to the restriction imposed by the fixed variance of the (s+1)-th PC. We will show that $\pi(\delta)$ can be arbitrarily close to 1. It is clear that,

$$\pi(\delta)=\Phi\left(\sqrt{\frac{2l(n-l)\delta}{n(\sigma_1^2+\sigma_2^2)}}-z_{1-\alpha/2}:(n-1)\lambda_{s+1}=\sum_{i=1}^{n}(y_{is+1}-\bar{y}_{s+1})^2\right), \qquad (3)$$

where σ_1^2,σ_2^2 are the variances of two groups, $z_{1-\alpha/2}$ is $(1 - \alpha/2)100 - th$ percentile of the standard normal distribution, $\bar{y}_{s+1}$ is the mean of vector y_{s+1}, and Φ is the cumulative density function of the standard normal distribution.

The ANOVA decomposition of the total sums of squares yields,

$$(n-1)\lambda_{s+1}=l(\bar{y}'_{s+1}-\bar{y}_{s+1})^2+(n-l)(\bar{y}''_{s+1}-\bar{y}_{s+1})^2+\sum_{i=1}^{l}(y_{is+1}-\bar{y}'_{s+1})^2+\sum_{j=l+1}^{n}(y_{js+1}-\bar{y}''_{s+1})^2, \qquad (4)$$

where $\bar{y}'_{s+1}$, $\bar{y}''_{s+1}$, and $\bar{y}_{s+1}$ are the means in the first, second and entire sample respectively.

Letting $\sigma_1^2 \to 0$ and $\sigma_2^2 \to 0$ entails $y_{is+1} \to \bar{y}'_{s+1}$ for all $i = 1, 2, \cdots, l$ and $y_{js+1} \to \bar{y}''_{s+1}$ for all $j = l + 1, l + 2, \cdots, n$. Then,

$$l\left(\bar{y}'_{s+1}-\bar{y}_{s+1}\right)^2+(n-l)\left(\bar{y}''_{s+1}-\bar{y}_{s+1}\right)^2 \to (n-1)\lambda_{s+1}, \qquad (5)$$

Without loss of generality we can assume that the overall mean $\bar{y}_{s+1}$ is zero and that the means of the first group and second groups are d_1 and $-d_2$. Then, from the condition that the overall mean is zero and (5) we deduce that $d_2 = d_1 l/(n - l)$ and $d_1=\sqrt{(n-1)(n-l)\lambda_{s+1}/(nl)}$. From here,

$$\delta=d_1+d_2=\sqrt{\frac{n(n-1)\lambda_{s+1}}{n-l}}, \qquad (6)$$

which is always positive. Clearly, from (3) we get,

$$\pi(\delta)\xrightarrow[\sigma_1^2,\sigma_2^2\to 0]{}1. \qquad (7)$$

This result reveals that any principal component with arbitrarily small variance can have a statistically significant effect with respect to classification which can produce subsequent improvement in the area under the ROC curve and should not be disregarded without further investigation.

RESULTS

We highlight the results through a numerical example. The following positive definite covariance matrix,

$$R=\begin{bmatrix}237 & 134 & 90 & 104\\ 134 & 86 & 68 & 71\\ 90 & 68 & 118 & 39\\ 104 & 71 & 39 & 98\end{bmatrix} \qquad (8)$$

has eigenvalues 419.3, 75.8, 40.8, 3.1 and the first two PCs account for 91.9% of the total variance. The usual dimensionality reduction approach will use the first two PCs for further analysis and disregard the last two. Let the true model for the binary class assignment be given by $Logit(P(G_i = 1)) = 0.5 + \beta_1 y_{i3}$. For effect sizes $\beta_1 = log(2)/4$, $log(2)/2$, $log(2)$, 2 the average areas under the ROC curve (averaged over 10,000 simulated datasets containing 500 subjects) for a logistic regression model that uses PC1 and PC2 were 0.53, 0.54, 0.54, 0.55 and while the corresponding values for a model using PC3 were 0.76, 0.88, 0.95, 0.99. Summary of the results is shown in ***Table 2***.

Table 2 Areas under the ROC curve for both models.

β1	*AUC – PC1, PC2**	*AUC – PC3**
log(2)/8	0.53	0.64
log(2)/4	0.53	0.76
log(2)/2	0.54	0.88
log(2)	0.54	0.95
2	0.55	0.99

* Empirically estimated via 10,000 datasets.

It is clear that even the two smallest effect sizes of $log(2)/8$ and $log(2)/4$ entail dramatic classification accuracy improvement of 0.11 and 0.23 respectively even though the true predictor, PC3, accounts for only 7.5% of the total variance in the data. However, this total variance is 539 and 7.5% of that amount still carries substantial amount of information and subsequent classification power. However, the power to detect effect sizes of $log(2)/8$ and $log(2)/4$ with variable having variance of 40.8 is almost 1 suggesting the inclusion of PC3 in subsequent analyses.

DISCUSSION

In this work we show a potential performance problem of classification algorithms carried out after preliminary dimensionality reduction step via PCA. These scenarios can occur even in simple, low dimensional data cases as our numerical example reveals. However, the issue can regularly arise with higher dimension data that possesses complex patterns and multiple groups. In such cases, the main PCs capture the covariance pattern of combined data while the the lower ranked PCs capture the information about group differences and are therefore vital for classification accuracy. Our results show that PCA with inclusion thresholds based on proportion of total variance explained often decreases classification accuracy even with extremely high inclusion threshold. Thus, we suggest using all PCs in classification problem in order to avoid the omission of PCs with lower ranking that are important classification predictors. In such cases, the benefit of the not using the original variables and switching to PCA might come from the fact that the PCs are uncorrelated and that might be advantagous in certain model building algorithms.

ACKNOWLEDGEMENTS

We are grateful for the support of the Kay Family Foundation.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR AFFILIATIONS

Jianwei Zheng *orcid.org/0000-0002-3930-228X*
Chapman University, One University Drive Orange, CA, US

Cyril Rakovski *orcid.org/0000-0002-9051-1466*
Chapman University, One University Drive Orange, CA, US

REFERENCES

Du, X, Dua, S, Acharya, RU and **Chua, CK.** 2012. Classification of epilepsy using high-order spectra features and principle component analysis. *Journal of Medical Systems*, 36: 1731–1743.

Gupta, V and **Mittal, M.** 2016. Respiratory signal analysis using pca, fft and artfa. 221–225. DOI: *https://doi.org/10.1109/ICEPES.2016.7915934*

Gupta, V and **Mittal, M.** 2018a. Knn and pca classifier with autoregressive modeling during different ecg signal interpretation. *Procedia Computer Science*, 125: 18–24.

Gupta, V and **Mittal, M.** 2018b. R-peak based arrhythmia detection using hilbert transform and principal component analysis. 1–4.

Gupta, V and **Mittal, M.** 2019a. Qrs complex detection using stft, chaos analysis, and pca in standard and real-time ecg databases. *Journal of The Institution of Engineers (India): Series B*, 100.

Gupta, V and **Mittal, M.** 2019b. R-peak detection in ecg signal using yule–walker and principal component analysis. *IETE Journal of Research*, 1–14.

Gupta, V, Mittal, M and **Mittal, V.** 2020. R-peak detection based chaos analysis of ecg signal. *Analog Integrated Circuits and Signal Processing*, 102. DOI: *https://doi.org/10.1007/s10470-019-01556-1*

Hsieh, C-W, Liu, T-C, Jong, T-L and **Tiu, C-M.** 2010. A fuzzy-based growth model with principle component analysis selection for carpal bone-age assessment. *Medical & Biological Engineering & Computing*, 48: 579–588.

Hsieh, FY, Bloch, AB and **Larsen, MD.** 1998. A simple method of sample size calculation for linear and logistic regression. *Statistics in Medicine*, 17: 1623–1634.

Johnson, RA and **Wichern, DW.** (eds.) 1988. *Applied Multivariate Statistical Analysis*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc. DOI: *https://doi.org/10.2307/2531616*

Kim, J, Shin, HS, Shin, K and **Lee, M.** 2009. Robust algorithm for arrhythmia classification in ecg using extreme learning machine. *BioMedical Engineering OnLine*, 8: 31.

Mehmet Korürek, AN. 2010. Clustering mit–bih arrhythmias with ant colony optimization using time domain and pca compressed wavelet coefficients. *Digital Signal Processing*, 20: 1050–1060.

Schoenfeld, DA and **Borenstein, M.** 2005. Calculating the power or sample size for the logistic and proportional hazards models. *Journal of Statistical Computation and Simulation*, 75: 771–785.

TO CITE THIS ARTICLE:
Zheng, J and Rakovski, C. 2021. On the Application of Principal Component Analysis to Classification Problems. *Data Science Journal*, 20: 26, pp. 1–6. DOI: *https://doi.org/10.5334/dsj-2021-026*

Submitted: 13 December 2020
Accepted: 09 August 2021
Published: 18 August 2021

Data Science Journal is a peer-reviewed open access journal published by Ubiquity Press.