From Genes to Tumors: A Comparison of Dimension Reduction Techniques in Logistic Regression

INFOMDA2 group assignment

F.P. Meijvis, A.E. Maranus, W.B. Eiling

September 16, 2025

Abstract

This study examines the comparative classification performance of (sparse) principle component analysis ((S)PCA) in addition to stepwise logistic regression for the classification of prostate tumor samples versus healthy ones through gene expression data. The three approaches were compared using a 5-fold cross-validation framework in which averaged classification metrics including accuracy, s ensitivity, specificity and area under the curve (AUC) were calculated. Stepwise logistic regression performed poorly, having an AUC of 0.5, indicating that this method performs no better than a random guessing approach. The (S)PCA model with 8 PCs showed distinctly improved predictive performance metrics with AUC of 0.949 and 0.959, respectively. In conclusion, (S)PCA models have a great classification advantage over stepwise logistic regression in gene expression data analysis in which SPCA conceivably preferable for usage in gene expression data analysis due to its model simplicity and reduced overfitting characteristics.

1 Introduction

Prostate cancer arises from the malignant transformation of prostate tissue, driven by various cellular and microenvironmental factors (1). This transformation is marked by sequential alterations in gene expression, which make prostate cells increasingly prone to tumor formation. Genes involved in this process may exhibit upregulation or downregulation. In cancer development, tumor-promoting genes are typically upregulated, while tumor-suppressing genes are often downregulated, interrupting the cells' safety mechanisms.

In this study, we analyze gene expression data from 52 prostate tumor samples and 50 normal tissue samples. **Figure 1a** and **Figure 1b** highlight genes with significantly up- or downregulated expression in tumor samples compared to normal samples. The objective of this investigation is to classify (or predict) a binary outcome—tumor versus normal tissue—based on the expression levels of 6,033 genes. This analysis aims to identify key gene expression patterns indicative of tumorigenesis in prostate cancer.

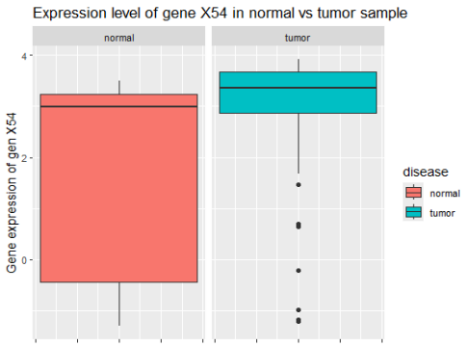
Logistic regression is a commonly used model for binary classification. However, the high dimensionality of this dataset ($p > N$) poses several challenges, including multicollinearity, inflated standard errors (non-significant p-values), non-uniqueness of solutions, and overfitting. To address these issues, a feature selection or dimensionality reduction approach is required.

This study compares the predictive performance of three approaches: (a) stepwise logistic regression, (b) principal component analysis (PCA) with logistic regression, and (c) sparse PCA (SPCA) with logistic regression. In stepwise logistic regression, the high dimensionality is managed by iteratively adding or removing genes based on a statistical criterion (e.g., deviance or AIC) to assess whether the model improves. PCA reduces dimensionality by creating principal components (PCs) as linear combinations of the input variables, capturing maximal variance (3). However, PCA retains contributions from all original features in the transformed components (3; 7)[1], limiting its interpretability and leading to potential theoretical failure in the $p \gg N$ regime (3). SPCA addresses these limitations by enforcing sparsity in the PCs, ensuring that only a subset of variables contributes to each component (3).

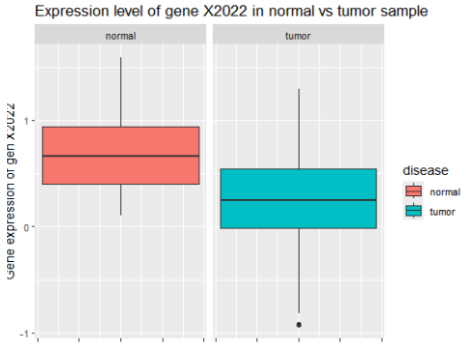
This study addresses the following research question: *What is the comparative classification performance of stepwise logistic regression, PCA with logistic regression, and SPCA with logistic regression?* Although stepwise logistic regression is straightforward to implement, it is expected to yield inferior predictions compared to the PCA-based approaches due to its susceptibility to overfitting and multicollinearity[2]. While SPCA's added sparsity may aid with biological interpretability, its predictive performance might be slightly lower than that of standard PCA, as it could exclude minor variability in the data that are relevant to the outcome.

Figure 1: Gene expression of Two Genes from Prostate Cancer Gen Expression Dataset.

(a) Possible Tumor-Promoting Gene



(b) Possible Tumor Suppressor Gene



[1]This implies that PCA is not a feature-selection method.
[2]Principal components are orthogonal and uncorrelated by design.

2 Methods

2.1 Data Pre-Processing

The gene expression dataset was retrieved from the `spls` package and consisted of 52 prostate tumor samples and 50 normal samples. Each sample contained 6,033 features corresponding to gene expression levels, along with a binary outcome variable indicating sample status (0 = normal, 1 = tumor). The dataset was pre-normalized, log-transformed, and standardized to zero mean and unit variance across genes (8). These preprocessing steps ensured appropriate scaling for subsequent analyses. Further normalization or scaling was unnecessary, as the pre-standardization aligned with the assumptions of PCA and SPCA (7).

2.2 Dimension Reduction Techniques and Model Fitting

To address the high dimensionality of the dataset and mitigate overfitting in logistic regression models, three approaches were applied: (1) hybrid stepwise logistic regression, (2) PCA, and (3) SPCA. All analyses were conducted in `R` version 4.4.2 (4), with the relevant packages specified below.

Hybrid stepwise logistic regression was implemented using the `MASS` package's `stepAIC()` function. This method iteratively selected predictors by minimizing the Akaike Information Criterion (AIC) through a combination of forward selection and backward elimination (2). The algorithm began by fitting an intercept-only model and a full model with all predictors using `glm()`. In each iteration, a forward step added the predictor that produced the largest reduction in AIC (if any), and a backward step removed predictors whose exclusion further reduced AIC. This iterative process continued until no additional improvement in AIC was possible.

PCA was applied using the `stats` package's `prcomp()` function to transform the 6,033 predictors into orthogonal PCs that capture the majority of the dataset's variance. The number of PCs ($k = 1, \ldots, 33$) was optimized using 5-fold cross-validation (CV)[3], selecting the configuration that maximized the average area under the curve (AUC) of logistic regression models fitted using `glm()` across the folds.

SPCA was implemented using the `sparsepca` package's `rspca()` function (6). SPCA introduces a sparsity-inducing LASSO penalty, ensuring that each PC is influenced by only a subset of genes (3)[4]. A grid search was performed over the number of components ($k = 1, \ldots, 33$), the sparsity-controlling parameter ($\alpha = 0.1, 0.01, 0.001, 0.0001$), and the ridge shrinkage parameter ($\beta = 0.1, 0.01, 0.001, 0.0001$). The optimal configuration was selected based on the highest average AUC achieved in 5-fold CV, with logistic regression models fitted to the sparse PCs using `glm()`.

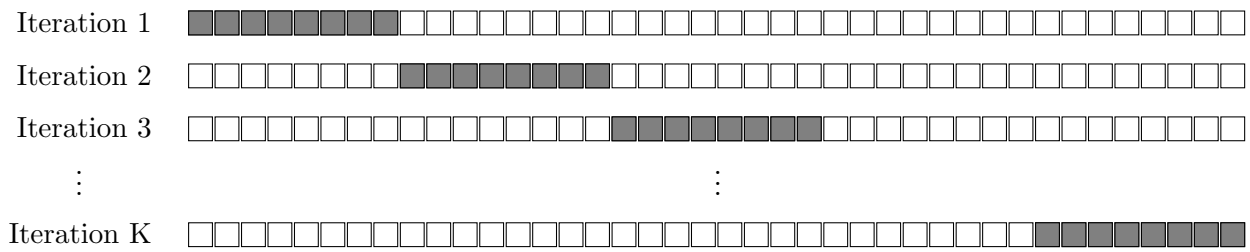
2.3 Comparison and Validation Strategy

The three approaches were compared using a 5-fold cross-validation framework (2), implemented with the `caret` package for data partitioning and metric computation. The dataset was divided into five folds, with models iteratively trained on four folds and tested on the remaining fold (**Figure 2**). This process was repeated for all folds, and average classification metrics—including accuracy, sensitivity, specificity, and AUC—were calculated using the `pROC` package. AUC was the primary metric for tuning the number hyperparameters and for identifying the best-performing approach due to its robustness in imbalanced classification settings. Receiver operating characteristic (ROC) curves were created to visually depict the AUC.

[3]CV selects components based on variance explained in the outcome, while alternative criteria (e.g., scree plot, Kaiser criterion or cumulative variance) focus solely on capturing variance among features, ignoring predictive utility (7).

[4]PCA can be interpreted as a ridge regression problem, while SPCA incorporates elastic-net regularization through the LASSO penalty (5).

Figure 2: 5-Fold Cross-Validation



Note. For a dataset of $T = 40$ observations, the figure represents the elements of the test set as dark-colored squares and the elements of the training set as blank squares. Each iteration consists of T time-points, with K total iterations in K-fold CV.

3 Results

This section outlines the performance of the three approaches, reporting accuracy, sensitivity, specificity, positive predictive value (PPV), negative predictive value (NPV), and AUC, each to three decimal places. First, we discuss the model selection process within each approach, followed by a comparison of the predictive performance of the optimal model of each approach.

3.1 Model Selection Within Approaches

For stepwise logistic regression, the intercept-only model was selected as the best model in all five folds. Since the model contains no predictors, the intercept corresponds to the observed proportion of $Y = 1$ in the training data. Consequently, predictions for the test set are entirely determined by whether this proportion exceeds 0.5, resulting in perfect sensitivity (1) or specificity (1), while the other is 0 (**Table 1**). Due to this structure, either the PPV or NPV is undefined (NaN) for each fold. The AUC is consistently 0.5, indicating no ability to differentiate between positives and negatives.

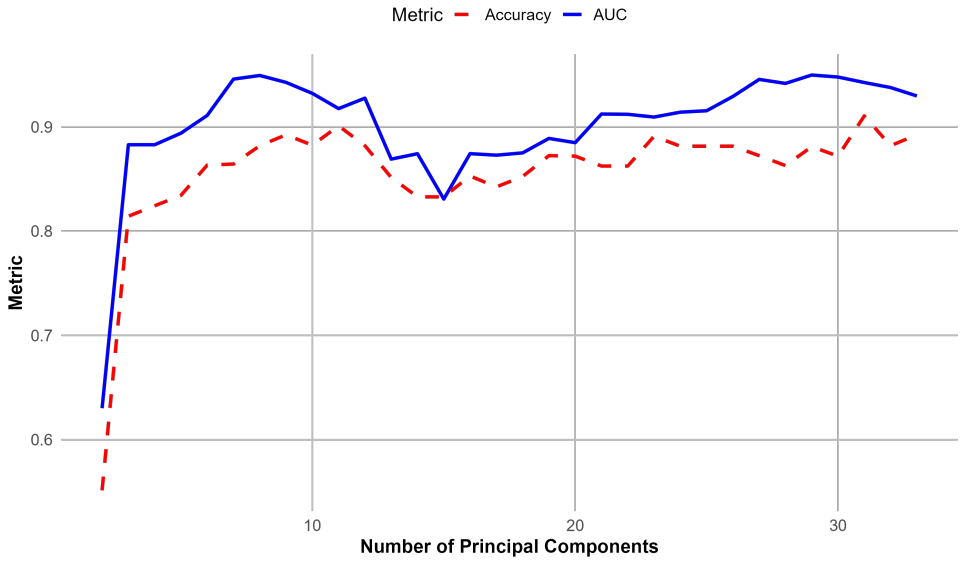
Table 1: Performance Metrics for 5-Fold Stepwise Logistic Regression

logistic regression	accuracy	sensitivity	specificity	PPV	NPV	AUC
fold 1.	0.5	0	1	NaN	0.5	0.5
fold 2	0.429	0	1	NaN	0.429	0.5
fold 3	0.381	0	1	NaN	0.381	0.5
fold 4	0.35	1	0	0.35	NaN	0.5
fold 5	0.4	1	0	0.4	NaN	0.5
Average	0.412	0.4	0.6	NaN	NaN	0.5

Note. NaN values mark divisions by 0.

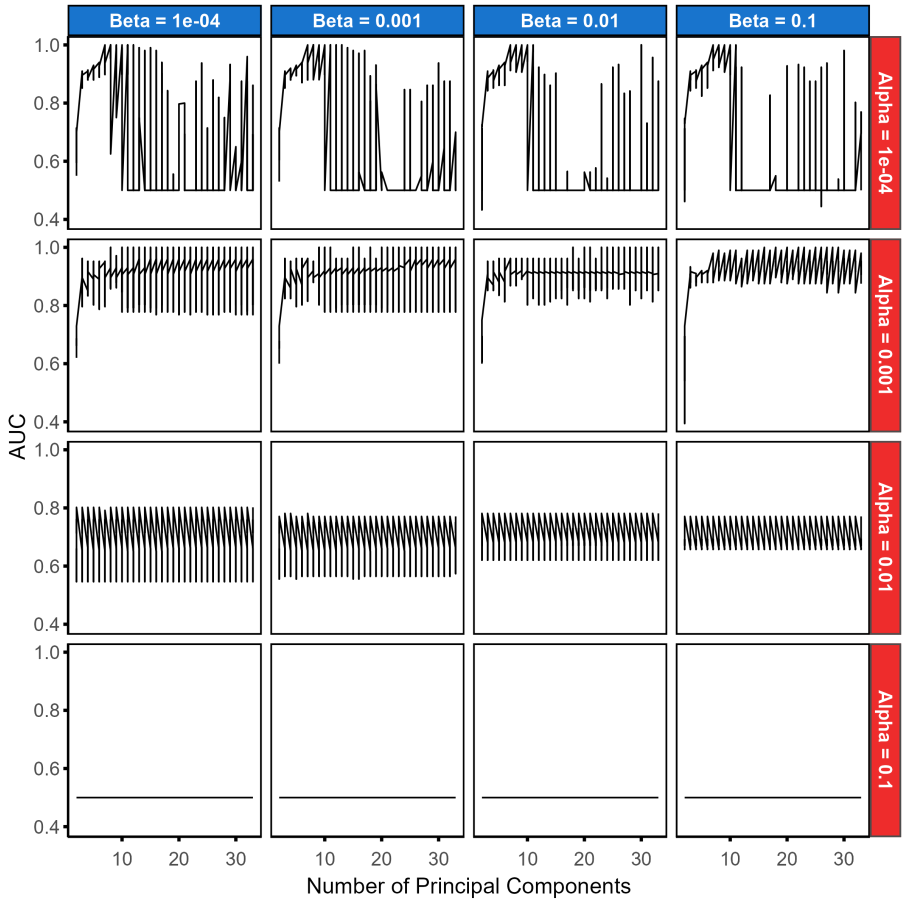
For PCA, the number of PCs was optimized by maximizing the AUC while minimizing model complexity for interpretability. A total of 33 PCs were generated from the gene expression data. **Figure 3** shows the average AUC and accuracy across folds for models with varying numbers of PCs. The model with 29 PCs achieved the highest AUC; however, the model with 8 PCs was selected for parsimony, as its AUC was nearly identical.

Figure 3: Averaged AUC Scores for Each PC in 5-Fold PCA + Logistic Regression



For SPCA, model selection involved optimizing hyperparameters (α and β) and the number of PCs to maximize the AUC. **Figure 4** shows the grid search results, where models with smaller α values tended to favor parsimony while achieving high AUC. The best-performing model used 8 PCs with $\alpha = 0.0001$ and $\beta = 0.1$.

Figure 4: Averaged AUC Scores for Different PC, α and β in 5-Fold SPCA + Logistic Regression



3.2 Comparison of Three Optimal Models

The performance of the three approaches is summarized in **Table 2**, with AUC serving as the primary evaluation metric. Stepwise logistic regression performed the worst, with an AUC of 0.5, equivalent to random guessing. Its poor accuracy (0.412), sensitivity (0.4), and specificity (0.6), along with undefined PPV and NPV, indicate its inability to identify meaningful patterns.

In contrast, PCA combined with logistic regression demonstrated strong performance, achieving an AUC of 0.949 and high accuracy (0.882), sensitivity (0.894), and specificity (0.894). These results suggest that PCA effectively reduced dimensionality while retaining the most predictive features. SPCA outperformed PCA slightly in terms of AUC (0.959) but exhibited lower accuracy (0.754) and specificity (0.698), indicating a higher tendency for false positives. However, its sensitivity (0.858) remained high, reflecting strong detection of true positives.

Overall, PCA demonstrated the most balanced performance across metrics, while SPCA slightly improved AUC at the cost of increased false positives. Stepwise logistic regression, meanwhile, was wholly inadequate for high-dimensional data. These findings underscore the advantages of PCA-based methods in predictive modeling for gene expression datasets.

Table 2: Comparison of Prediction Metrics Between Three Approaches

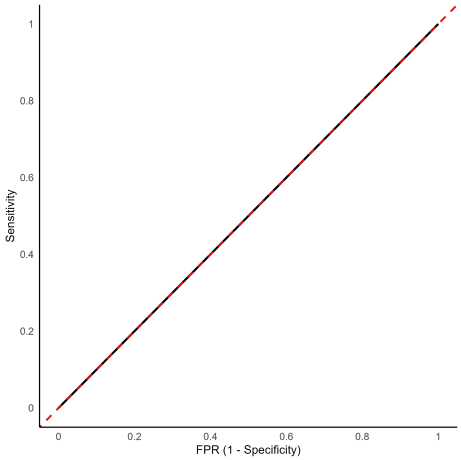
Approach	Accuracy	Sensitivity	Specificity	PPV	NPV	AUC
Stepwise Logistic Regression	0.412	0.4	0.6	NaN	NaN	0.5
PCA + Logistic Regression	0.882	0.894	0.894	0.875	0.916	0.949
SPCA + Logistic Regression	0.754	0.858	0.698	0.743	0.884	0.959

Note. All shown metrics were averaged across the folds. NaN values mark divisions by 0. PPV = positive predicted value, NPV = negative predicted value, AUC = area under the curve.

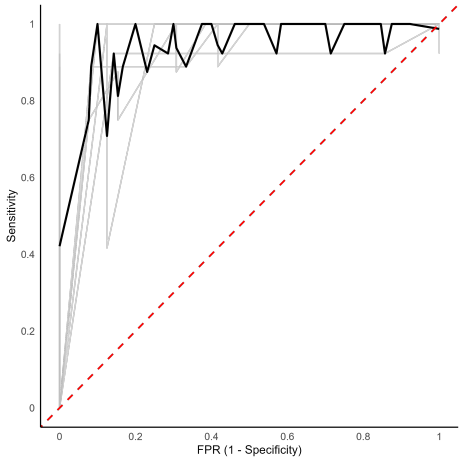
The ROC curves accompanying the AUC values are shown in **Figure 5**. For stepwise logistic regression, the ROC curve is a diagonal line, overlapping the red dashed line of a random classifier, confirming its inability to separate positives from negatives. PCA demonstrates a well-separated ROC curve, albeit with some jaggedness due to the small test set size ($n = 20$), possibly indicating overfitting. The SPCA ROC curve is nearly identical to that of PCA, showing excellent separation and comparable predictive performance.

Figure 5: ROC curves of the different models.

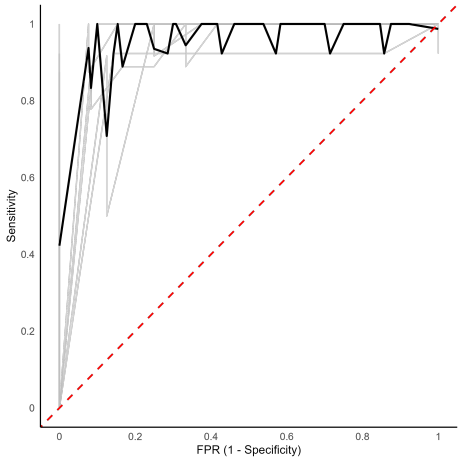
(a) Stepwise Logistic Regression



(b) PCA + Logistic Regression with 8 PCs



(c) SPCA + Logistic Regression with 8 PCs, $\alpha = 0.0001$, $\beta = 0.1$



Note. The dashed red line represents the performance of a random classifier, where there is no separation of positives and negatives; the grey lines represents the ROC-curves of the 5 different folds; and the black line represents the average across the folds. FPR = false positive rate.

4 Conclusion

This study investigated the comparative performance of three approaches—stepwise logistic regression, PCA with logistic regression, and SPCA with logistic regression—for classifying prostate tumor and normal tissue based on gene expression data. The research question sought to evaluate which method offers the best classification performance while maintaining interpretability. The findings align with the hypothesis that stepwise logistic regression would underperform due to its vulnerability to overfitting and multicollinearity. In all five cross-validation folds, the stepwise logistic regression models selected only the intercept, resulting in an AUC of 0.5—indicating an inability to classify healthy samples from tumor samples. This finding emphasizes the method's limitations in high-dimensional datasets, as it fails to account for multicollinearity

or complex interrelationships among predictors, which are typical in gene expression data.

As expected, the PCA-based logistic regression model significantly improved classification performance, achieving an AUC of 0.949 and an accuracy of 0.882. However, the current implementation of the PCA method only detects linear correlations within the data, making it incompetent for the identification of other structures in the gene expression data. Besides that, this method maintain all incorporate all input variables, limiting its interpretability and contributing to a more overfitted model. SPCA introduces sparsity, potentially contributing to enhancing parsimony and reducing overfitting in this prediction model. Although the AUC score of the SPCA model is higher, all other performance metrics are lower compared to the PCA model. This suggests that incorporating multiple performance metrics into the tuning procedure—rather than relying solely on AUC as was done in this investigation—may be necessary to achieve more balanced performance across all evaluation criteria.

Applying dimensionality reduction techniques like (S)PCA to gene expression data can significantly aid in identifying important genes involved in prostate cancer development. Examining the loadings of individual genes that strongly contribute to the principal components (PCs) can help uncover previously unknown genes associated with prostate cancer tumorigenesis. This interpretability is a distinct advantage of SPCA over standard PCA, as SPCA allows for filtering genes with non-zero loadings for each PC, simplifying the identification of relevant genes. Additionally, biplots can be constructed, where genes aligned in similar directions may indicate relatedness, potentially revealing new patterns or gene expression pathways contributing to prostate cancer development. In conclusion, applying PCA or SPCA to gene expression data can facilitate the discovery of previously unidentified genetic patterns indicative of tumorigenesis in prostate cancer, enhancing our understanding of the disease and supporting the development of targeted therapies.

References

[1] Wang G, Zhao D, Spring DJ, DePinho RA. Genetics and biology of prostate cancer. Genes Dev. 2018 Sep 1;32(17-18):1105-1140. https://doi.org/10.1101/gad.315739.118.

[2] Hastie T, Tibshirani R, Friedman JH. The elements of statistical learning: data mining, inference, and prediction. 2nd ed. Springer Series in Statistics. New York: Springer; 2009. ISBN: 9780387848846. Available from: https://books.google.nl/books?id=eBSgoAEACAAJ.

[3] Hastie T, Tibshirani R, Wainwright M. Statistical learning with sparsity: the lasso and generalizations. CRC Press; 2015. ISBN: 9781498712170. Available from: https://books.google.nl/books?id=f-A_CQAAQBAJ.

[4] R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing; 2024. https://www.R-project.org/

[5] Zou, H., Hastie, T., Tibshirani, R. (2006). Sparse Principal Component Analysis. Journal of Computational and Graphical Statistics, 15(2), 265–286. https://doi.org/10.1198/106186006X113430

[6] Erichson, N. B., Zheng, P., Manohar, K., Brunton, S. L., Kutz, J. N., Aravkin, A. Y. (2020). Sparse Principal Component Analysis via Variable Projection. SIAM Journal on Applied Mathematics, 80(2), 977–1002. https://doi.org/10.1137/18M1211350

[7] James, G., Witten, D., Hastie, T., Tibshiran, R. (2023). An introduction to Statistical Learning with Applications in R.

[8] Dettling M, Bühlmann P. Supervised clustering of genes. Genome Biology. 2002;3(12):research0069.1. https://doi.org/10.1186/gb-2002-3-12-research0069