JOURNAL OF STATISTICAL THEORY AND PRACTICE
2018, VOL. 12, NO. 2, 450–461
https://doi.org/10.1080/15598608.2017.1413456

Dimension reduction of gene expression data

Jaylen Lee[a], Shannon Ciccarello[b], Mithun Acharjee[c] and Kumer Das[c]

[a]Department of Mathematics and Statistics, James Madison University, Harrisonburg, Virginia, USA; [b]Department of Mathematics and Statistics, Hollins University, Roanoke, Virginia, USA; [c]Deparment of Mathematics, Lamar University, Beaumont, Texas, USA

ABSTRACT

DNA methylation of specific dinucleotides has been shown to be strongly linked with tissue age. The goal of this research is to explore different analysis techniques for microarray data in order to create a more effective predictor of age from DNA methylation level. Specifically, this study compares elastic net regression models to principal component regression, supervised principal component regression, *Y*-aware principal component regression, and partial least squares regression models and their ability to predict tissue age based on DNA methylation levels. It has been found that the elastic net model performs better than latent variable models when considering less than ten principal components for each method, but *Y*-aware principal component regression predicts more accurately (with a reasonably low testing RMSE) and captures more of the desired structure when the number of principal components increases to 20. Coding limitations inhibited forming conclusive results about the performance of supervised principal component regression as the number of components increases.

ARTICLE HISTORY
Received 8 September 2016
Revised 12 September 2017
Accepted 1 December 2017

KEYWORDS
Principal component analysis; DNA methylation; elastic net regression; *Y*-aware PCR; supervised PCR; PLS regression

AMS SUBJECT CLASSIFICATION
62H25; 62J99; 62N86

1. Introduction

Big data and microarray data are new and vital components of modern statistics. Because the analysis of such information can be lengthy and time-consuming, methods of analysis are needed in the field that are proven to be efficient and accurate. Although there are many categories of big data, of greatest interest in this study are human genomic data: age predicted by methylation levels at specific gene sights. One of the challenges in working with microarray data is its high dimensionality coupled with a relatively small sample size. While there is a plethora of information that can be obtained from such a data set, the high dimensionality can often cover or distort the information of interest. Ideally, when one works with data sets with a large amount of variables, one will use subject expertise to filter out useless information. However, when the variables number in the thousands, this quickly becomes a nearly impossible task. To solve this issue we look toward dimension reduction techniques such as principal component analysis (PCA) and singular value decomposition (SVD).

Finding a model that accurately predicts age based on the methylation levels, or uncovering a commonality between the strongest predictors, will help in processing

CONTACT Kumer Das kumer.das@lamar.edu Department of Mathematics, Lamar University, PO Box 10047, 4400 MLK Parkway,Beaumont, TX 77710, USA.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/UJSP.

large amounts of gene data so that further study can be made of the connection of methylation levels and genetic diseases (Phillips 2008). The current model and first of its kind was developed by Horvath et al. (2012), whose DNAm age calculator is highly accurate and based on a combination of 80 data sets from 27k and 480k Illumina platforms. This model employed elastic net regression resulting in 353 predictor genes with a median error of 3.6 years. A subsequent study conducted by the German Cancer Research Center used genes on the 480k Illumina platform from 965 whole blood samples to investigate age prediction. This research found 65 age-related CpGs and used 17 of those to predict age with a margin of error of 2.6 years (Florath et al. 2014).

The purpose of this study is to analyze the effectiveness of latent variable regression models with regard to predicting age through CpG (shorthand for 5'-C-phospate-G-3') methylation. We do so by regressing age upon the principal components provided by four different methods: standard principal component analysis, partial least squares, *Y*-aware principal component analysis, and supervised principal component analysis. The hope is that, through predicting with these latent variables and exploring their connection to age, there can be an understanding of some overarching structure or mechanism that links DNAm age to tissue age. Additionally, by comparing the different methods, one may hope to create a more accurate age calculator than the one provided by Horvath et al. (2012).

The rest of this article is organized into six sections: Section 2 describes the mathematical ideas behind the dimension reduction techniques. Section 3 describes the data set and steps taken to treat the data. Section 4 delivers the results of analysis while section 5 describes and compares the findings, and section 6 provides concluding remarks.

2. Statistical methods

2.1. *Principal components regression*

The first attempt at dimension reduction and analysis used a combination of singular value and multiple linear regression, known as principal components regression (Jolliffe 1982). The basic SVD was performed on our centered data matrix X such that

$$X_{n\times p} = U_{n\times k}\theta_{k\times k}V^T_{k\times p} \tag{1}$$

where $k = \min(n, p)$, the columns of U are the eigenvectors of XX^T, θ is a diagonal matrix of the singular values in descending order, and the columns of V are the eigenvectors of X^TX. A significant property of the matrix U is that the first column vector, $\mathbf{u_1}$, is in the direction of the most variation of X. Additionally, $\mathbf{u_2}$ is orthogonal to $\mathbf{u_1}$ and in the direction of the second most variation of X. The pattern continues as we might expect for each succeeding column vector of U (Wall, Rechtsteiner, and Rocha. 2003). This orthogonality is a highly desirable trait as it eliminates collinearity when used in a regression model (Jolliffe 1982).

Our standard regression model is

$$Y = X\beta + \varepsilon, \tag{2}$$

where Y is the response vector, β is a vector of least squares coefficients, and ε is an error vector. Replacing X by the SVD approximation, the regression model takes the form

$$Y = U\theta V^T\beta + \varepsilon. \tag{3}$$

Letting $\alpha = \theta V^T\beta$ our equation simplifies to

$$Y = U\alpha + \varepsilon \tag{4}$$

where U is our matrix of latent variables $\mathbf{u_1}, \mathbf{u_2}, \cdots, \mathbf{u_k}$. It can be shown that ordinary least square solution to Eq. (3) is $\hat{\alpha} = U^T Y = \theta V^T\hat{\beta}$ (Jolliffe 1982), which reduces our model to

$$\hat{Y} = U\hat{\alpha}. \tag{5}$$

This method not only reduces the dimension of the data set, but also allows regression upon uncorrelated predictors (Shlens 2014). It can be shown that as we regress upon less $\mathbf{u_i}$, we accomplish a reduction in variance but introduce bias into our estimate. Such a phenomenon is commonly known as the bias-variance trade-off (Jolliffe 1982).

2.2. *Elastic net regression*

Rather than combining our observed information in an attempt to construct latent variables, we can solve the issue of high dimensionality through automated variable selection techniques. One such method is elastic net regression, the same analysis technique that Horvath et al. (2012) used to create their model. A combination of LASSO and Ridge regression, elastic net regression seeks to solve the minimization problem

$$\min_{\beta} \| y - X\beta\|^2 + \lambda_2 \| \beta\|^2 + \lambda_1 \| \beta\|_1 \quad s.t \quad \alpha \| \beta\|^2 + (1-\alpha) \| \beta\|_1 \leq t \tag{6}$$

for $\alpha = \frac{\lambda_2}{\lambda_2+\lambda_1}$ and $t > 0$.

This method performs the desirable variable selection of LASSO while not disregarding variables that vary similarly. As we increase α we lean more toward the stricter L1 regularization of LASSO. On the other hand, decreasing α leads toward the less harsh variable shrinkage of Ridge regression (Zou and Hastie 2005). Horvath used $\alpha = 0.5$ when using this method for variable selection purposes (Horvath et al. 2012).

2.3. *Supervised principal component regression*

Usually during principal component regression we regress our latent variables $\mathbf{u_1}, \mathbf{u_2}, \cdots, \mathbf{u_k}$ upon our dependent variable. Often, to reduce the complexity of the model, we simply get rid of the latter principal components (the ones that account for less of the variation in the data). This is one of the most common mistakes made in using this process. It is quite possible that unimportant variables load more heavily onto the first few principal components while our signal variables load more heavily onto our latter principal components. In getting rid of those latter components we might lose much of the explanatory power of our model! Bair et al. (2006) seek to rectify this issue by performing a variable selection step prior to forming our principal components. This method is known as supervised principal component regression (SPCR).

They do so by computing the p univariate regression coefficients for each feature: $\beta_1, \beta_2, \cdots, \beta_p$.Next, a reduced data matrix, X_θ, is formed where the columns are the x_j such that $\beta_j > \theta$ for some threshold θ. This θ is determined by cross-validation of the

log-likelihood ratio statistic. Afterward, we perform principal component regression as previously demonstrated on X_θ.

The idea is that by filtering out presumed unimportant variables, we can ignore the effect of ignoring latter occurring principal components. In fact, Bair et al. (2006) suggest using the first principal components exclusively.

2.4. Y-aware principal component regression

Similar to SPCR, Y-aware principal component regression attempts to solve the issue of unimportant variables loading more heavily onto the first few principal components. This method also solves this issue by utilizing the univariate regression coefficient for each feature. The difference lies in that, instead of using the β_j as a means to filter out unimportant variables, we use it as a means to scale our matrix. Our transformed X is thus defined as

$$x_j^* = \beta_j x_j - \frac{\sum_{i=1}^{n} \beta_j x_{ij}}{n} \tag{7}$$

where X^* is our transformed data matrix (Zumel 2016). To the extent that we can, each column of our matrix is scaled by their effect on Y. As such, our presumed unimportant variables are now transformed so that they vary less than our signal variables. Because our first principal components are in the direction of the most variation, our signal variables will now load more heavily onto our first principal components. As a result, the intuitive step of pruning the later occurring principal components prior to regressing should be much more reasonable.

2.5. Partial least squares regression

Partial least squares (PLS) regression uses least squares regression on a set of uncorrelated predictor components. This methods works regardless of highly collinear variables, or when there are more predictors than observations (Rosipal 2011). PLS regression produces weight matrices W and C in order to create a linear combination of the columns of X and Y so that their covariance,

$$[cov(t,u)]^2 = [cov(Xw, Yc)]^2 = max_{|r|=|s|=1}[cov(Xr, Ys)]^2,$$

is maximized (Rosipal 2011). The data matrices X and Y are column centered and normalized. A u vector with random values is generated. At this point, there are two algorithms that can be used in PLS regression. We used the nonlinear iterative partial least squares (NIPALS) algorithm. This algorithm repeats until t converges (Abdi 2003).

(1) $w = X^T u$ (estimate X weights). Then normalize w. Note that this is similar to alternating least squares, as described in Liu et al. (2003). The elements of the first right vector are just the simple, centered regression coefficients.
(2) $t = Xw$ (estimate X factor scores).
(3) $c = Y^T t$ (estimate Y weights). Then normalize c.
(4) $u = Yc$ (estimate Y scores).

Once t has converged, then X is regressed on t and Y on u. The effect of t is subtracted from X and Y (tp^T and tc^T, respectively) to deflate them, resulting in the following decompositions of X and Y:

$$X = TP^T + E,$$

$$Y = UQ^T + F.$$

Ordinary least squares procedures for the regression of Y on T are then performed to determine Q, the weights for Y. With Q produced, the completed prediction model, where $B = WQ$, is given as

$$Y = XB + E. \tag{8}$$

3. Data

The data used in our analysis are a small subset of the original data set used by Horvath et al.(2012). The full data set can be obtained from the Gene Expression Omnibus database, series code GSE41037. It consists of DNA methylation levels from 719 whole blood samples measured on the 27k Illumina platform. Although the set was originally used in schizophrenia research, the disease status of the patients was shown to have no impact on their methylation levels (Abdi 2003). The data was split into two groups, training and testing, for cross-validation purposes. All 719 subjects were sorted into five partitions based on the percentile in which their age fell in. Using the Pareto principle, also known as the 80/20 principle (Kurucz, Benczr, and Csalogny 2007; Kiremire 2011), within each partition, 80% of the subjects were assigned to the training set, with the remaining 20% being in the testing set.

The response variable and the predictor variables were log transformed prior to running regression to satisfy assumptions of multiple linear regression. Afterward, each column of the data matrix was centered to have mean 0.

There were 7,307 missing values dispersed throughout the data matrix. Given that some packages required that every entry be filled by a numeric value, missing value imputation was a necessity. Since there are more than 19 million entries in our data matrix, imputing values for 0.03% of those entries should have a negligible effect upon any results. The missing values were first imputed by using the mean of the CpGs, as it was shown that the difference between using the minimum possible value, 0, and the maximum possible value, 1, was negligible in a set of 27,555 genes. However, a more valid method specific to SVD, as detailed by Hastie et al. (1999), was investigated. All statistical analysis has been performed with the statistical programming R (R Core Team 2013).

3.1. *Missing values*

This expectation maximization method begins by assigning column averages to the missing values, but follows with an algorithm for imputing better estimates (Li et al. 2016). After generating a decomposed matrix from the average-modified original matrix, the formerly missing values are imputed by regressing their corresponding gene against the columns of V^T, or the eigengenes. A linear combination of the resulting coefficients is

used to determine a better estimate for the missing value. This algorithm is continued, each time using the previously generated matrix, until 100 iterations is reached. The convergence tolerance, a measure of total change, must be below 0.01. Running this algorithm on our gene data set resulted in a convergence tolerance of 0.001 and allowed for further decomposition of the completed matrix (Hastie et al. 1999).

While the expectation that the SVD method produced a more accurate missing value estimation than column means, it has been shown to perform better on time-series data and is less accurate for microarray data than a newer method, known as *K*-nearest neighbors (Hastie et al. 1999; Troyanskaya et al. 2001; Kurucz, Benczr, and Csalogny 2007). This method begins by locating a missing value in gene A for a certain subject B. Next, k subjects with similar gene expression patterns are identified whose value for gene A is given in the data set. The gene A values of the k subjects are weighted according to their corresponding subject's similarity with the gene expression of subject B and averaged.

4. Results

Each method was implemented, regressing each of their respective variables upon $log(Age)$. There was no stepwise variable removal after regressing. In order to calculate the number of CpGs shared with the Horvath et al.(2012) list for the principal component methods, each component was sorted in decreasing order by the absolute value of the CpG loadings. The top 100 CpGs for each sorted component were taken to be compared to Horvath's list of 353 CpGs. We then summed the number of CpGs that were found to be in the list. Lastly, each model was tested by predicting ages for the testing data set.

4.1 *Elastic net regression*

Elastic net regression was implemented using $\alpha = 0.05$. Using 10-fold cross-validation of the mean squared error, $\lambda = 5.06 \times 10^{-3}$was chosen as the fit parameter. With these parameters, 295 CpGs were selected to regress upon $log(Age)$. Of these 295 CpGs, 28 were found to be in Horvath's 353 CpGs. This small number makes sense since we are using a very small subset of the Horvath et al.(2012) original data set.

4.2. *Supervised principal component regression*

The data matrix was decomposed into three principal components (the maximum allowed under the superpc package). Tenfold cross-validation of the likelihood ratio test statistic was performed. The results are summarized in Figure 1, where through 10 randomly selected cross-validation folds, $\theta = 8$ was chosen as the best selection threshold. A feature threshold of $\theta = 8$ was chosen as a result. A regression was performed upon these three principal components. The overall model showed significant explanatory power, generating an F-statistic of $848.4 > F_{3,573,0.05}$. Each principal component was deemed significant $(P < 0.05)$ with the first principal component having the largest $\hat{\beta}$.

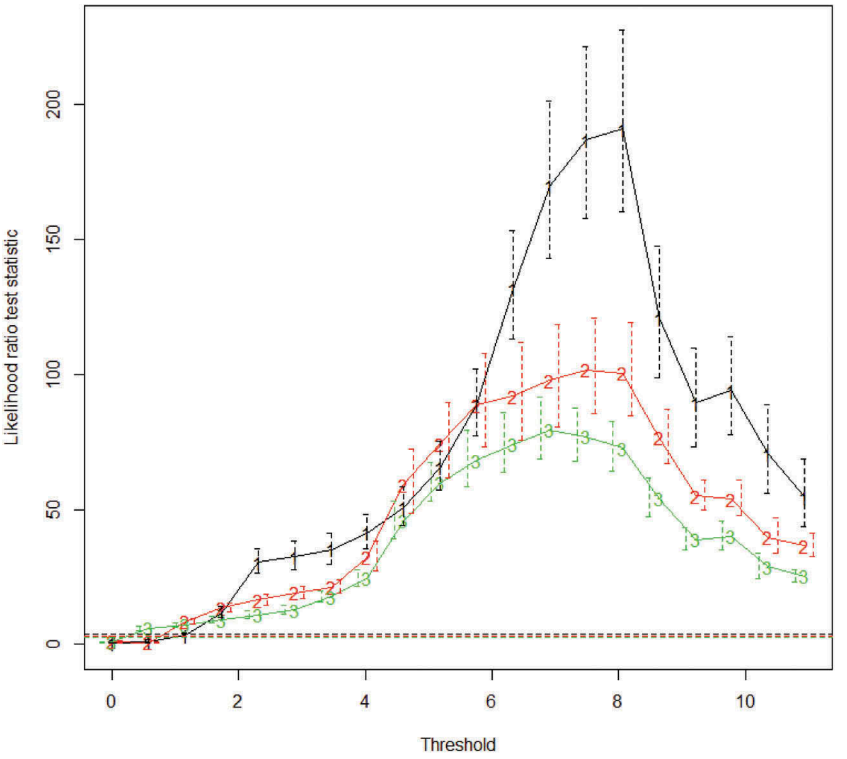

Figure 1. Cross validation of likelihood ratio statistic for SPCR.

4.3. *PCR, Y-aware PCR, and PLS regression (three principal components)*

Our data matrix was decomposed into three principal components for each respective method. This was done to allow these models to be fairly judged against the SPCR model. The cumulative proportion of the variation of X represented by these 3 components is given in Table 1, where regression results for each method have been displayed. # of Prin Comp refers to the number of principal components used for each method; Prop of X Var refers to the cumulative proportion of the variation in X represented by three principal components; # Clock CpGs refers to the number of Horvath's 353 clock CpGs captured by the method; Adj. R^2refers to the coefficient of determination for the training model; Testing Sum of Squared Error (SSE) $= \sum_{i=1}^{142} (y_i - \hat{y}_i)^2$; and Testing Root Mean Squared Error (RMSE) is the predictive error using the testing data set. A low RMSE indicates that the predictive error for the testing data is low. The residual plots for the latent variable methods have been displayed in Figure 2. PLS regression and Y-aware PCR have pretty similar residuals. SPCR has a noticeably slimmer spread, indicating a better fit model. There is a rather strange cluster of subjects that seem to have been underestimated on the right side of the PCR residual plot. For PCR, only the second principal component was significant. For Y-aware PCR, the first component and second component were deemed significant. For PLS regression, every single component was deemed significant($P < 0.05$).

Table 1. Regression results using three principal components.

Method	# of Prin Comp	Prop of *X* Var	# Clock CpGs	Adj. R^2	Testing SSE	Testing RMSE
Elastic net	N/A	N/A	28	0.93	2.201	0.125
Y-aware PCR	3	0.45	9	0.40	14.76	0.325
PCR	3	0.49	0	0.02	23.22	0.407
SPCR	3	0.37	13	0.77	4.37	0.177
PLS regression	3	0.25	0	0.66	9.10	0.255

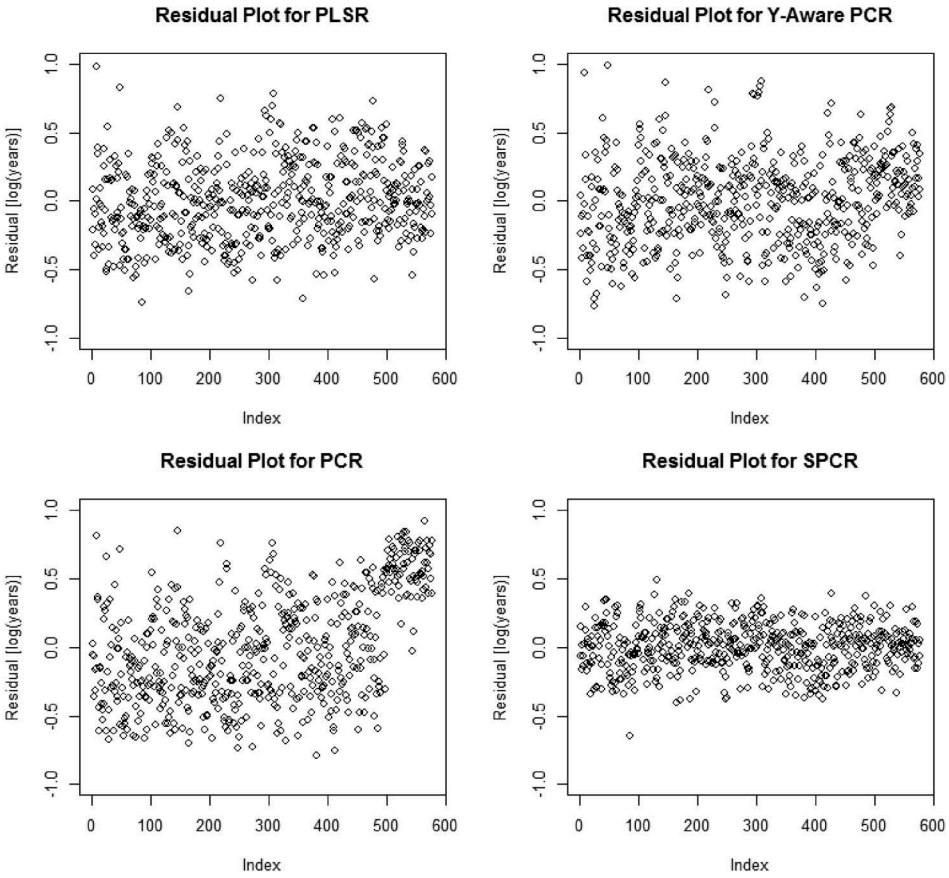

Figure 2. Residual plots for latent variable models.

4.4. *PCR, Y-aware PCR, and PLS regression (twenty principal components)*

A process similar to the one discussed in section 4.3 was executed, this time using 20 principal components. The cumulative proportion of the variation of X represented by these 20 components is given in Table 2. Prediction results are summarized in Table 3, where four subjects of ages 20, 40, 60, and 80 were chosen from the training data set. Prediction intervals were created based on their gene expression profile. For latent variable models, 20 principal components were used, with the exception of SPCR, which utilized 3, due to limitations of the superpc package only allowing up to three principal components to be analyzed in the SPCR method.

Table 2. Regression results using twenty principal components.

Method	# of Prin Comp	Prop of *X* Var	# Clock CpGs	Adj. R^2	Testing SSE	Testing RMSE
Y-aware PCR	20	0.62	60	0.89	1.90	0.124
PCR	20	0.67	27	0.80	5.14	0.204
PLS regression	20	0.67	12	0.99	9.10	0.272

Table 3. Prediction intervals for four subjects.

Age	Bound	*Y*-aware PCR	SPCR	PLS regression	PCR	Elastic net
	Lower	16.5	16.6	19.8	19.5	17.4
20	Estimated	21.5	24.2	19.9	28.0	20.1
	Upper	28.0	35.3	20.1	40.1	23.3
	Lower	34.1	22.1	39.7	28.6	35.1
40	Estimated	44.4	32.1	40.0	41.0	41.0
	Upper	57.8	46.5	40.2	58.7	47.9
	Lower	47.1	46.5	59.9	40.4	53.5
60	Estimated	61.3	67.7	60.4	57.9	61.4
	Upper	79.7	98.7	60.8	82.9	70.4
	Lower	51.1	49.9	79.6	52.7	62.9
80	Estimated	66.8	72.7	80.2	75.8	72.7
	Upper	87.2	105.9	80.7	109.2	84.0

5. Discussion

One of the first things that sticks out from Table 1 is the noteworthy R^2 of the SPCR model compared to the other latent variable models. The SPCR model having a R^2 larger than the PLS regression model was most surprising since the partial least squares components are formed with the express purpose of maximizing the covariance between X and Y. One would expect that a model made with such component would construct a model with much more explanatory power. With an exception of the elastic net regression model, SPCR model has the lowest testing RMSE in Table 1, which indicates that the predictive error is the least using this model.

Both PCR and PLS regression failed to load any of Horvath's 353 clock CpGs in their first three principal components. It seems to be the case that unimportant variables, CpGs that had very little to do with age, varied more highly than the clock CpGs. This indicates a need for methods that take into account our y variable, like SPCR and *Y*-aware PCR. Both methods managed to extract a small subset of the clock CpGs in their first three principal components.

Another significance is the discrepancy in performance between SPCR and *Y*-aware PCR. Given their similar methodology, it should be the case that they would give similar results. However, this is not so as SPCR is shown to be better across all testing criteria (Figure 3), with a higher R^2 indicating a better fit of the testing data, a higher number of clock CpGs represented, and lower prediction error for the testing dataset. Since *Y*-aware PCR merely dampens the effect of unimportant variables while SPCR gets rid of them entirely, they may still be able to worm their way into loading heavily onto the first principal components, resulting in lower predictive power.

The elastic net model out-performs the latent variables models by a wide margin when keeping the number of principal components low. This model demonstrated the greatest R^2 value, collected the largest amount of clock CpGs, and demonstrated the lowest predictive error for the testing data set. This clear outperformance is mostly due to the

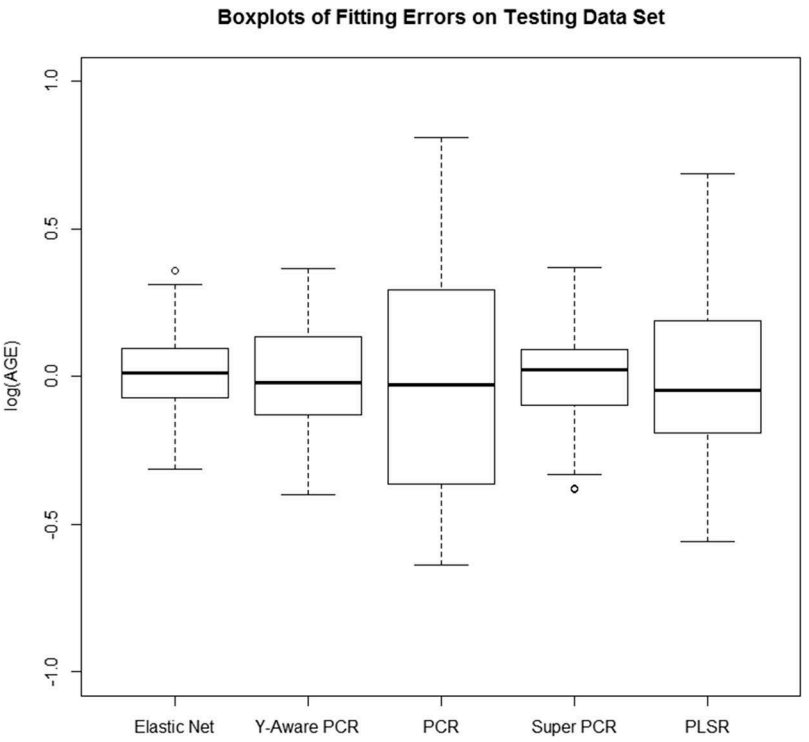

Figure 3. Box plots of fitting errors for testing data set.

limited number of principal components used in the latent variable models. For each of the latent variable methods, the three principal components explain less than half of the variation in methylation levels. As the number of principal components considered increases we find that the performance of the latent variable models also increases. This makes sense because we are using more information about the variation in X (Figure 4). As also shown by Figure 4, when the number of principal components used increases, both Y-aware PCR and PLS regression experience a tremendous increase in their predictive accuracy. This suggests that if there are latent variables that allow prediction of age, then there are certainly more than three important ones. The results in Table 2 showcasing this more readily; in particular, the testing RMSE for Y-aware PCR is significantly lower (0.124) when the number of principal components is 20 compared to the testing RMSE (0.325) for Y-aware PCR when the number of principal components is 3.

Most notable among these results are the R^2 value of PLS regression and the overall performance of Y-aware PCR. Elastic net clearly had a better fit to the training data than PLS regression with three components, but utilizing more latent variables allowed PLS regression to fit the training data nearly perfectly. This proved to be a case of overfitting, however, as the predicting error did not improve at all. Y-aware PCR managed to perform very well by increasing the number of principal components. Both its coefficient of determination and its predicting error experienced significant improvement. It is suspected that SPCR would experience the same improvement. However, due to limitations of the superpc package only allowing up to three principal components to be analyzed, this could not be fully investigated.

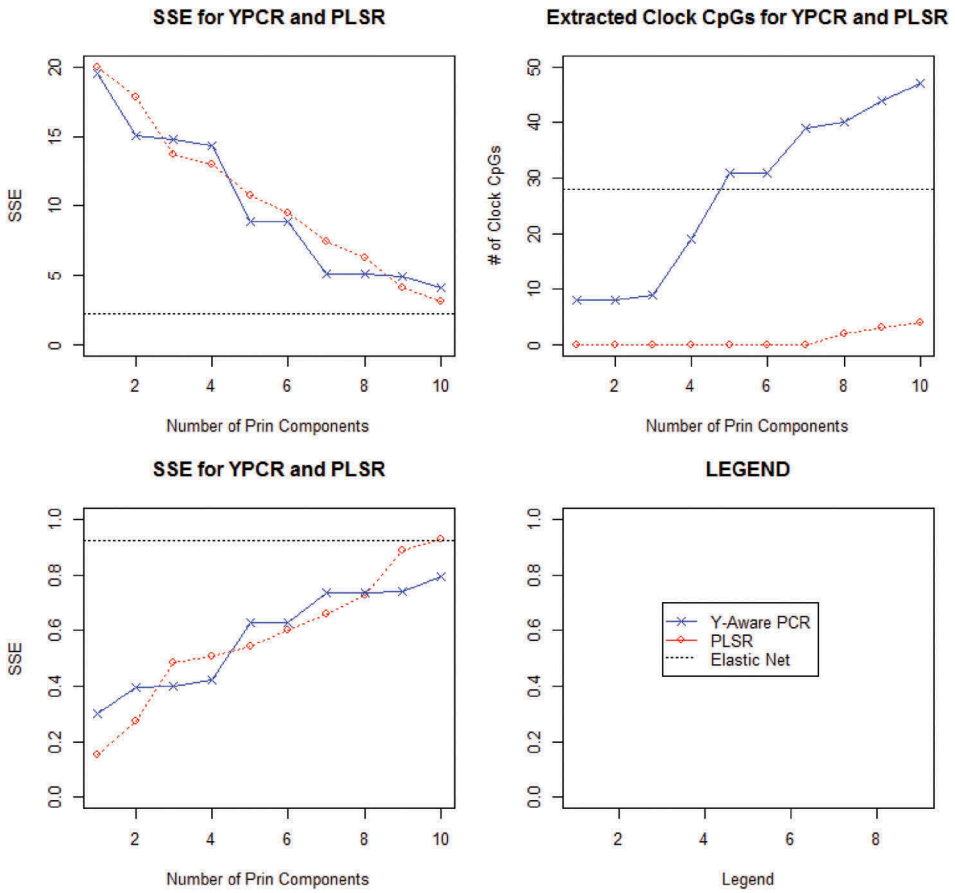

Figure 4. Comparison of latent variable models and elastic net.

6. Conclusion

Microarray data sets are often overwhelmingly large, so understanding any sort of structure goes a long way toward deciphering the genetic mechanisms that influence aging. For the purposes of prediction, elastic net serves as a fine method. However, if there are suspicions of latent mechanisms involved, *Y*-aware PCR should be considered. This method managed to show satisfactory predictive abilities while maintaining a structure to our data matrix.

There are many limitations to our models. They did not take into consideration factors such as gender or tissue type, which were shown to be significant in the Horvath et al. original publication (Horvath et al. 2012). Additionally, our study used data measured on the 27k Illumina platform only, not 450k Illumina. However, it is believed the principles shown in this article will still apply.

Considerations for further research include comparisons to a SPCR model with more than three components, and comparisons on data with multiple units. Additionally, it is hypothesized that the CpGs loaded heavily onto the first few components of the SPCR model may have similar "purposes," such as silencing the transcription of analogous

proteins. This requires using a data set containing multiple tissue types performed by researchers with much more subject knowledge.

Funding

This work was supported by the National Science Foundation (1560332).

References

Abdi, H. 2003. Partial least squares (PLS) regression. In *Encyclopedia of social sciences research methods*, ed. M. Lewis-Beck, A. Bryman, and T. Futing, 792–95. Thousand Oaks (CA): Sage.

Bair, E., T. Hastie, D. Paul, and R. Tibshirani. 2006. Prediction by supervised principal components. *Journal of the American Statistical Association* 101 (473):119–37. doi:10.1198/016214505000000628.

Florath, I., K. Butterbach, H. Muller, M. Bewerunge-Hudler, and H. Brenner. 2014. Cross-sectional and longitudinal changes in DNA methylation with age. *Human Molecular Genetics* 23 (5):1186–201. doi:10.1093/hmg/ddt531.

Hastie, T., R. Tibshirani, G. Sherlock, E. Michael, P. Brown, and D. Botstein, 1999. Imputing Missing Data for Gene Expression Arrays (Technical Report). Division of Biostatistics, Stanford University, Stanford, CA.

Horvath, S., Z. Yafeng, P. Langfelder, R. S. Kahn, M. P. M. Boks, K. V. Eijk, L. H. Berg, and R. A. Ophoff. 2012. Aging effects on DNA methylation modules in human brain and blood tissue. *Genomic Biology* 13 (10):R97.

Jolliffe, I. T. 1982. A note on the use of principal components in regression. *Journal of the Royal Statistical Society, Series C* 31 (3):300–03.

Kiremire, A. R. 2011. The application of the Pareto principal in software engineering (Consulted). Ruston (LA): Louisiana Tech University; http://tinyurl.com/Ankunda-termpaper (accessed July 2016).

Kurucz, M., A. A. Benczr, and K. Csalogny. 2007. Methods for large scale SVD with missing values. *Proceedings of KDD Cup and Workshop* 12:31–38.

Li, H., H. Bangzheng, M. Lublin, and Y. Perez. 2016. *Distributed algorithms and optimization*. Stanford, CA: Stanford University.

Liu, L., D. M. Hawkins, S. Ghosh, and S. S. Young. 2003. Robust singular value decomposition analysis of microarray data. *Proceedings of the National Academy of Sciences of the United States of America* 100 (23):13167–72. doi:10.1073/pnas.1733249100.

Phillips, T. 2008. The role of methylation in gene expression. *Nature Education* 1 (1):116.

R Core Team. 2013. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org (accessed October 2016).

Rosipal, R. 2011. Nonlinear partial least squares: An overview. In *Chemoinformatics and advanced machine learning perspectives: Complex computational methods and collaborative techniques*, ed. H. Lodhi, and Y. Yamanishi, 169–89. ACCM, IGI Global. http://aiolos.um.savba.sk/~roman/Papers/npls_book11.pdf (accessed May 2016).

Shlens, J. 2014. *A tutorial on principal component analysis*. Cornell University Library. https://arxiv.org/pdf/1404.1100.pdf (accessed April 2016).

Troyanskaya, O., M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein, and R. B. Altman. 2001. Missing value estimation methods for DNA microarrays. *Bioinformatics* 17 (6):520525. doi:10.1093/bioinformatics/17.6.520.

Wall, M., M. Rechtsteiner, and L. M. Rocha. 2003. Singular value decomposition and principal component analysis. In *A practical approach to microarray data analysis*, ed. D. P. Berrar, W. Dubitzky, and M. Granzow, 91–109. Los Alamos National Laboratory LA-UR-02-4001.

Zou, H., and T. Hastie. 2005. Regularization and variable selection via the elastic net. *Journal of Royal Statistical Society: Series B* 67 (Part 2):301–20. doi:10.1111/j.1467-9868.2005.00503.x.

Zumel, N. (2016). Principal components regression, Pt. 2: Y-aware methods [Web log comment]. http://www.win-vector.com/blog/2016/05/pcr_part2_yaware (accessed July 2016).