Prediction of Stress Increase in Unbonded Tendons using Sparse Principal Component Analysis

Eric Mckinney
Utah State University

Follow this and additional works at: https://digitalcommons.usu.edu/gradreports

Part of the Applied Mathematics Commons, Applied Statistics Commons, and the Statistical Theory Commons

Recommended Citation
https://digitalcommons.usu.edu/gradreports/1034
PREDICTION OF STRESS INCREASE IN UNBONDED TENDONS USING SPARSE PRINCIPAL COMPONENT ANALYSIS

by

Eric McKinney

A project submitted in partial fulfillment of the requirements for the degree of MASTER OF SCIENCE in Statistics

Approved:

________________________________________  ______________________________________
Yan Sun                                         Marc Maguire
Major Professor                                Committee Member

________________________________________  ______________________________________
Adele Cutler                                    Daniel Coster
Committee Member                                Committee Member

UTAH STATE UNIVERSITY
Logan, Utah

2017
ABSTRACT

Prediction of Stress Increase in Unbonded Tendons

using Sparse Principal Component Analysis

by

Eric McKinney, Master of Science

Utah State University, 2017

Major Professor: Dr. Yan Sun
Department: Mathematics and Statistics

While internal and external unbonded tendons are widely utilized in concrete structures, the analytic solution for the increase in unbonded tendon stress, $\Delta f_{ps}$, is challenging due to the lack of bond between strand and concrete. Moreover, most analysis methods do not provide high correlation due to the limited available test data. In this thesis, Principal Component Analysis (PCA), and Sparse Principal Component Analysis (SPCA) are employed on different sets of candidate variables, amongst the material and sectional properties from the database compiled by Maguire et al. [18]. Predictions of $\Delta f_{ps}$ are made via Principal Component Regression models, and the method proposed, a linear model using SPCA on variables with a significant level of correlation with $\Delta f_{ps}$, is shown to improve over current models without increasing complexity.

(67 pages)
PUBLIC ABSTRACT

Prediction of Stress Increase in Unbonded Tendons

using Sparse Principal Component Analysis

by

Eric McKinney, Master of Science
Utah State University, 2017

Major Professor: Dr. Yan Sun
Department: Mathematics and Statistics

While internal and external unbonded tendons are widely utilized in concrete structures, the analytic solution for the increase in unbonded tendon stress, $\Delta f_{ps}$, is challenging due to the lack of bond between strand and concrete. Moreover, most analysis methods do not provide high correlation due to the limited available test data. In this thesis, Principal Component Analysis (PCA), and Sparse Principal Component Analysis (SPCA) are employed on different sets of candidate variables, amongst the material and sectional properties from the database compiled by Maguire et al. [18]. Predictions of $\Delta f_{ps}$ are made via Principal Component Regression models, and the method proposed, a linear model using SPCA on variables with a significant level of correlation with $\Delta f_{ps}$, is shown to improve over current models without increasing complexity.

(67 pages)
ACKNOWLEDGEMENTS

I would like to thank my principle advisor, Dr. Yan Sun, for her countless hours of guidance. I am grateful for how she pushed me to reach deeper levels of understanding in my research. She is a continuing example to me of professional and personal achievement.

I would also like to thank Dr. Marc Maguire for his irreplaceable counsel and direction. His expertise has provided numerous insights into the industrial side of this area of research while also keeping me in touch with the overall goal of the project.

Thanks to the rest of my committee members, namely Dr. Adele Cutler and Dr. Daniel Coster, for their support. I have deeply appreciated their valuable time, interest, and respective expertise.

Lastly, I would like to thank my darling wife, Brianna. She has been a constant source of bright encouragement and unfailing strength to both myself and our children. In many ways, her sacrifice has been greater than mine, and yet she continues to be positive and supportive through it all.

Eric McKinney
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABSTRACT</td>
<td>iii</td>
</tr>
<tr>
<td>PUBLIC ABSTRACT</td>
<td>iv</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>v</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>viii</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>x</td>
</tr>
<tr>
<td>1 INTRODUCTION</td>
<td>1</td>
</tr>
<tr>
<td>1.1 Literature Review</td>
<td>1</td>
</tr>
<tr>
<td>1.2 Approach</td>
<td>2</td>
</tr>
<tr>
<td>2 METHODOLOGY</td>
<td>4</td>
</tr>
<tr>
<td>2.1 Principal Component Analysis</td>
<td>4</td>
</tr>
<tr>
<td>2.2 Lasso and Elastic Net Penalization</td>
<td>7</td>
</tr>
<tr>
<td>2.3 Sparse Principal Component Analysis</td>
<td>8</td>
</tr>
<tr>
<td>3 APPLICATION</td>
<td>11</td>
</tr>
<tr>
<td>3.1 Principal Component Analysis Application</td>
<td>11</td>
</tr>
<tr>
<td>3.2 Sparse Principal Component Analysis Application</td>
<td>25</td>
</tr>
<tr>
<td>4 CONCLUSION</td>
<td>38</td>
</tr>
<tr>
<td>4.1 Discussion Of Results</td>
<td>38</td>
</tr>
</tbody>
</table>
4.2 Summary and Conclusion .................................. 40

BIBLIOGRAPHY ................................................. 42

APPENDIX ...................................................... 45
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1 Projecting two variables onto a new variable. The magenta line segments at the two ends of the data show the direction of maximum variation.</td>
<td>5</td>
</tr>
<tr>
<td>2.2 A visualization of a simple PCA by projecting two variables onto a new variable with the maximum amount of variation. The blue points are the data, and the red points are the scores.</td>
<td>5</td>
</tr>
<tr>
<td>3.1 Cumulative Explained Variation for each Principal Component for Models 1 and 2.</td>
<td>12</td>
</tr>
<tr>
<td>3.2 $\Delta f_{ps}$ vs. $\hat{\Delta f}_{ps}$ for Model 1.</td>
<td>14</td>
</tr>
<tr>
<td>3.3 $\Delta f_{ps}$ vs. $\hat{\Delta f}_{ps}$ for Model 2.</td>
<td>14</td>
</tr>
<tr>
<td>3.4 Cumulative Explained Variation for each Principal Component for Continuous Variables.</td>
<td>15</td>
</tr>
<tr>
<td>3.5 Cumulative Explained Variation for each Principal Component for Categorical Variables.</td>
<td>16</td>
</tr>
<tr>
<td>3.6 $\Delta f_{ps}$ vs. $\hat{\Delta f}_{ps}$ for Model 3.</td>
<td>16</td>
</tr>
<tr>
<td>3.7 $\Delta f_{ps}$ vs. $\hat{\Delta f}_{ps}$ for Model 4.</td>
<td>17</td>
</tr>
<tr>
<td>3.8 Cumulative Explained Variation for each Principal Component for Models 7 and 8.</td>
<td>18</td>
</tr>
<tr>
<td>3.9 $\Delta f_{ps}$ vs. $\hat{\Delta f}_{ps}$ for Model 7.</td>
<td>19</td>
</tr>
<tr>
<td>3.10 $\Delta f_{ps}$ vs. $\hat{\Delta f}_{ps}$ for Model 8.</td>
<td>19</td>
</tr>
<tr>
<td>3.11 Cumulative Explained Variation for each Principal Component for Models 11 and 12.</td>
<td>23</td>
</tr>
<tr>
<td>3.12 $\Delta f_{ps}$ vs. $\hat{\Delta f}_{ps}$ for Model 11.</td>
<td>24</td>
</tr>
<tr>
<td>3.13 $\Delta f_{ps}$ vs. $\hat{\Delta f}_{ps}$ for Model 12.</td>
<td>24</td>
</tr>
</tbody>
</table>
3.14 Number of Loadings for the SPC regression vs. Adjusted $R^2$ Values for Continuous Variables in Models 5 and 6. ......................................................... 26

3.15 Number of Loadings for the SPC regression vs. Adjusted $R^2$ Values for Categorical Variables in Models 5 and 6. ......................................................... 26

3.16 $\Delta f_{ps}$ vs. $\hat{\Delta f}_{ps}$ for Model 5. .......................................................... 29

3.17 $\Delta f_{ps}$ vs. $\hat{\Delta f}_{ps}$ for Model 6. .......................................................... 29

3.18 Number of Loadings for the SPC regression vs. Adjusted $R^2$ Values for Self-Selected Variables in Models 9 and 10. ......................................................... 31

3.19 $\Delta f_{ps}$ vs. $\hat{\Delta f}_{ps}$ for Model 9. .......................................................... 32

3.20 $\Delta f_{ps}$ vs. $\hat{\Delta f}_{ps}$ for Model 10. .......................................................... 33

3.21 Number of Loadings for the SPC regression vs. Adjusted $R^2$ Values for Correlation Cut-off Variables in Models 13 and 14. ......................................................... 34

3.22 $\Delta f_{ps}$ vs. $\hat{\Delta f}_{ps}$ for Model 13. .......................................................... 36

3.23 $\Delta f_{ps}$ vs. $\hat{\Delta f}_{ps}$ for Model 14. .......................................................... 36
## LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Variable Names, Notation, and Type for the Statistical Analysis</td>
<td>11</td>
</tr>
<tr>
<td>3.2</td>
<td>Correlation Cut-off Important Variables for the Internal Data</td>
<td>21</td>
</tr>
<tr>
<td>3.3</td>
<td>Correlation Cut-off Important Variables for the External Data</td>
<td>22</td>
</tr>
<tr>
<td>3.4</td>
<td>SPCA Loadings for Internal Continuous Variables</td>
<td>27</td>
</tr>
<tr>
<td>3.5</td>
<td>SPCA loadings for Internal Categorical Variables</td>
<td>27</td>
</tr>
<tr>
<td>3.6</td>
<td>SPCA loadings for External Continuous Variables</td>
<td>28</td>
</tr>
<tr>
<td>3.7</td>
<td>SPCA loadings for External Categorical Variables</td>
<td>28</td>
</tr>
<tr>
<td>3.8</td>
<td>SPCA loadings for Internal Self-Selected Important Variables</td>
<td>31</td>
</tr>
<tr>
<td>3.9</td>
<td>SPCA loadings for External Self-Selected Important Variables</td>
<td>32</td>
</tr>
<tr>
<td>3.10</td>
<td>SPCA loadings for Internal Correlation Cut-off Important Variables</td>
<td>35</td>
</tr>
<tr>
<td>3.11</td>
<td>SPCA loadings for External Correlation Cut-off Important Variables</td>
<td>35</td>
</tr>
<tr>
<td>4.1</td>
<td>PCA Models’ $R^2$, Adjusted $R^2$, and $\lambda$ Values.</td>
<td>38</td>
</tr>
<tr>
<td>4.2</td>
<td>SPCA Models’ $R^2$, Adjusted $R^2$, and $\lambda$ Values.</td>
<td>38</td>
</tr>
</tbody>
</table>
1.1 Literature Review

The use of unbonded tendons either internal or external, increases cost-efficiency, provides aesthetic satisfaction for users, and achieves fast and efficient construction \[5, 22, 25\]. However, the analysis of structures using unbonded tendons is exceptionally difficult and has been the subject of many international research projects, most of which attempt to simplify the problem considerably. Although numerous studies have been conducted to estimate the tendon stress increases at nominal strength, the analytic solution for the increase in unbonded tendon stress, $\Delta f_{ps}$, is challenging due to the lack of bond between strand and concrete, and most analysis methods do not provide high correlation due to the limited available test data \[18\].

Current design for unbonded tendon reinforced members in the United States uses American Concrete Institute 318 \[9, 10\]:

$$\Delta f_{ps} = 70 + \frac{f'_c}{\mu \rho_p}$$

or American Association of State Highway and Transportation Officials Load and Resistance Factor Design \[1\] guidelines:

$$\Delta f_{ps} = 6200 \left( \frac{d_p - c}{L} \right) \left( 1 + \frac{N}{2} \right)$$

Both of the above methods are relatively easy for implementation in design, however, there are concerns with both. The ACI model is a curve fit to statistical data from only a handful of experimental data prior to 1978 \[20, 21\]. The AASHTO method is not dependent on an experimental curve fit for $\Delta f_{ps}$, but is dependent on an estimation
of the scaled plastic hinge length ($\psi$) from Tam and Pannell [26]. The ACI method especially is well liked by designers due to its simplicity for design.

There are considerably more prediction methods available in the literature as well as international design codes. Maguire et al. [18] performed an in-depth review of various prediction methods based on the common mechanisms and empirical assumptions. The collapse mechanism model uses the relationship between strain, angle of rotation, and applied load. The ASSHTO LRFD method based on Roberts-Wollmann et al. [25], and MagGregeor [17] is considered a collapse mechanism model, which uses the relationship between strain, angle of rotation and applied load. Other collapse mechanism models have been developed by the British Standard Institution [11] and Harajli [7] among others. Another category, called bond-reduction models, calculates a bond-reduction coefficient ($\Omega$) to reduce the strength of a cross section unbonded reinforcement. Probably the most well-known bond reduction model was introduced by Naaman and Alkhairi [23] and at one time was accepted in the 1994 AASHTO LRFD code, but later replaced in the 1998 AASHTO LRFD and also included statistical fitting to some degree. Alternatively, statistical analysis methods have been developed using the available experimental data of their time. The statistically based 1963 ACI code [9] and European design codes, including German [6] and Swiss [24] codes, are widely accepted for design and real world application. The 1963 and current ACI methods purposely under-predict strand stress increase in most cases, and when compared to other methodologies provide closer to a lower bound prediction as opposed to an accurate prediction.

1.2 Approach

Maguire et al. [19] and Maguire et al. [18] indicated considerable phenomenological differences between continuous unbonded tendon reinforced members, which are common, and simply supported members, which are uncommon in design. Interestingly, most methods from the literature compared prediction performance to a majority of simply supported members. In response, Maguire et al. (2017) compiled the largest known international database of 83 continuous members (downloadable at http://www.ascelibrary.org/). In order to consider multiple variables including internal and external tendons, Maguire et al. (2017) also suggested an update to the AASHTO LRFD collapse mechanism model ($\psi = 14$ and $\psi = 18.5$ for internal and external tendons, respectively) based a statistical analysis and found nearly all types of prediction
methods to have very low prediction accuracy with fit statistics $R^2$ of 0.27 and a best measured-to-prediction ratio ($\lambda$) of 1.34, neither of which indicates a good fit.

With the overall lack of data available and targeted research programs to drive better phenomenological models for unbonded tendon reinforced structures, a statistical approach may provide the best prediction for $\Delta f_{ps}$. The advantages of a statistically based model are clear. They can be easily implemented, do not require excessive design time, and they do not burden the engineer with several design iterations (e.g., bond reduction and collapse mechanism models), like the ACI 318 equation. Furthermore, they can be optimized to fit the data and cross validation used to verify their accuracy.

The main objective of this paper is to present a novel approach to predict the increase in tensile strength in unbonded tendons using Principal Component Analysis (PCA), and Sparse Principal Component Analysis (SPCA). PCA is a statistical procedure to select significant variables by converting the variable information into the orthogonal base set [12]. PCA has gained considerable popularity in structural engineering in recent years in combination with machine learning and structural health monitoring [28, 29] vibrations [8, 13, 15, 30, 31] and image based crack detection [2] because it is especially useful for analyzing large dataset with many variables. SPCA uses the Least Absolute Shrinkage and Selection Operator (LASSO) to reduce the contribution of relatively insignificant principal coefficients in the proposed statistical model, which simplifies the model further [4, 33].

In this paper, a focus is made on improving the accuracy of $\Delta f_{ps}$ predictions for the internally reinforced and externally reinforced unbonded tendons separately. Sets of candidate variables, amongst the material and sectional properties from the database compiled by Maguire et al. [18], are considered to analyze the significant factors in the database for prediction of $\Delta f_{ps}$. The analysis results for each set of candidate variables are compared to an initial PCA on all of the variables. The results verify that improvements in predictions can be made with a simplified SPCA regression model.
2.1 Principal Component Analysis

PCA is a widely used statistical technique for dimension reduction. It takes linear combinations of all of the variables to create a reduced number of uncorrelated variables (called principal components, or PC’s) that still express a majority of the information from the original data [16]. The number of principal components selected, which is usually much smaller than the number of original variables, is determined by considering how much information is retained at the cost of simplifying the data. In addition to dimension reduction, another typical scenario where PCA works well is when a level of collinearity exists in the data, i.e., some or all of the predictor variables are correlated. After applying PCA, the resulting principal components are uncorrelated, and hence the replication of information in the original variables is removed.

A visualization of a simplified PCA [3] is exemplified in Figures 2.1 and 2.2. Figure 2.1 demonstrates two variables (horizontal and vertical axes) being projected onto a new axis (the oblique axis) which can be thought of as a new variable. However, PCA is a special type of projection onto the new axis that maximizes the variation in the new projected data points. Hence, Figure 2.2 shows the projection of the two variables onto the new variable or PC. The second minor perpendicular axis shows the choice for the second PC, and demonstrates that any subsequently chosen PC will be orthogonal to the previously chosen PCs [12].

Let \( X = [x_{ij}], i = 1, \cdots, n, j = 1, \cdots, p \), be the \( n \times p \) data matrix of \( n \) observations on the \( p \)-dimensional random vector \( X = [X_1, X_2, \cdots, X_p] \). Define the \( 1 \times p \) mean vector \( \bar{x} \) as

\[
\bar{x} = \left[ \frac{1}{n} \sum_{i=1}^{n} x_{i1}, \cdots, \frac{1}{n} \sum_{i=1}^{n} x_{ip} \right]
\]
Figure 2.1: Projecting two variables onto a new variable. The magenta line segments at the two ends of the data show the direction of maximum variation.

Figure 2.2: A visualization of a simple PCA by projecting two variables onto a new variable with the maximum amount of variation. The blue points are the data, and the red points are the scores.

That is, the $j^{th}$ element of $\bar{x}$ is the sample mean of the $j^{th}$ variable. The $p \times p$ sample covariance matrix $S$ is computed as

$$S = \frac{1}{n-1} (X - 1_n \bar{x})^T (X - 1_n \bar{x}),$$
where \( \mathbf{1}_n \) is an \( n \times 1 \) column vector of ones.

To express a majority of the information from the original data, we want to create new variables with maximum variance from linear combinations of the original variables. This is equivalent to an eigendecomposition of the data matrix [16]. Let \( \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \) be the eigenvalues of \( \mathbf{S} \) in descending order, and let \( \mathbf{u}_1, \mathbf{u}_2, \cdots, \mathbf{u}_p \) be the corresponding eigenvectors. The first principal component \( Y_1 \) is defined as a linear combination of \( X_j \)’s such that it has the largest variance under the constraint that the coefficient vector has unit norm. Hence, we want to find \( \mathbf{u} \) when \( \mathbf{u}^T \mathbf{u} = 1 \) such that the variance of \( Y_k = X \mathbf{u} \) is maximized. Lattin et al. [16] showed that

\[
\text{var}(Y) = \text{var}(X \mathbf{u}) = \mathbf{u}^T \mathbf{S} \mathbf{u}, \quad \mathbf{u}^T \mathbf{u} = 1
\]

can be maximized by forming the Lagrangian, differentiating with respect to \( \mathbf{u} \), setting it equal to zero, and solving, i.e.

\[
L = \mathbf{u}^T \mathbf{S} \mathbf{u} - \lambda (\mathbf{u}^T \mathbf{u} - 1)
\]

\[
\frac{\delta L}{\delta \mathbf{u}} = 2 \mathbf{S} \mathbf{u} - 2 \lambda \mathbf{u} = 0
\]

\[
\mathbf{S} \mathbf{u} = \lambda \mathbf{u} \quad \text{or} \quad (\mathbf{S} - \lambda \mathbf{I}) \mathbf{u} = 0
\]

Hence, the coefficient vector can be estimated by \( \mathbf{u}_1 \), the eigenvector of \( \mathbf{S} \) corresponding to the largest eigenvalue \( \lambda_1 \). The second principal component \( Y_2 \) is the linear combination of \( X_j \)’s with the second largest variance under the unit coefficient vector constraint uncorrelated with \( Y_1 \). The coefficients are estimated by \( \mathbf{u}_2 \). In general, the \( k^{th} \) principal component is estimated by

\[
\overline{Y}_k = X \mathbf{u}_k, \quad k = 1, \cdots, q.
\]

The subsequent analyses will be performed based on these \( q \) uncorrelated principal components (as opposed to the original \( p \) variables), whose observed values are given by the principal component score matrix

\[
\mathbf{Z} = \mathbf{X} \mathbf{U}.
\]
Here \( U = [u_1, u_2, \cdots, u_q] \) is the \( p \times q \) loading matrix containing all of the coefficient vectors. This can also be accomplished by a singular value decomposition of \( X \), so that

\[
X = Z_s D^{\frac{1}{2}} U^T
\]

where \( Z_s \) is the standardized score matrix, \( D^{\frac{1}{2}} \) is a diagonal matrix with \( \lambda_1^2, \lambda_2^2, \cdots, \lambda_p^2 \) diagonal entries, and \( U^T \) simply the transposed eigenvector matrix.

Because covariance is scale sensitive, the result of PCA is significantly affected by scaling. Thus, it is a common practice to standardize the variances before performing a PCA. In such a situation, the sample correlation matrix \( \rho \) is used in replacement of the sample covariance matrix \( S \). It is equivalent to the sample covariance matrix when the variances of all variables are standardized to be 1. Let \( D \) be the diagonal matrix of the diagonal entries of \( S \), i.e.

\[
D = \text{diag} \{ S(1,1), S(2,2), \cdots, S(p,p) \}.
\]

Then, \( \rho \) can be computed as

\[
\rho = \sqrt{D}^{-1} S \sqrt{D}^{-1}.
\]

It is easily seen that the diagonal values of \( \rho \) are uniformly 1, and the off diagonal values are sample correlations of the corresponding rows and columns. Similar as previously, the eigenvectors of \( \rho \) give the estimated coefficients for the principal components, and the input for any subsequent analyses is the corresponding score matrix.

### 2.2 Lasso and Elastic Net Penalization

Before an introduction to sparse principal component analysis, we briefly introduce the \( L_1 \), or lasso penalty, and the elastic net. In linear regression, while the least squares estimates \( \hat{\beta} \) for the coefficients \( \beta \) are unbiased, the model \( \hat{\beta} \)'s can suffer from large variations. Tibshirani \[27\] proposed penalizing the least squares estimates as

\[
\hat{\beta}_{\text{lasso}} = \arg \min_{\beta} \left\| Y - \sum_{j=1}^{p} X_j \beta_j \right\|^2 + \lambda^* \sum_{j=1}^{p} |\beta_j|, \quad \lambda^* \geq 0
\]
where \( \| \cdot \| \) denotes the Euclidean norm.

Here, \( \lambda^* \) allows the lasso estimates to trade off bias for lower variance in order to improve predictive accuracy. The lasso penalty also simplifies the model since large enough choices for \( \lambda^* \) will shrink some of the coefficient estimates to zero. This is a direct result of the \( L_1 \) penalty. \cite{27} Notice that the choice of \( \lambda^* = 0 \) will reduce the problem back to least squares estimates. However, the lasso suffers from the fact that the number of selected variables cannot exceed the number of observations \cite{32}. However, the elastic net improves over the lasso by relaxing this restriction. For \( \lambda_1^* \geq 0 \) and \( \lambda_2^* \geq 0 \) the elastic net estimates \( \hat{\beta}_{en} \) are found as follows:

\[
\hat{\beta}_{en} = (1 + \lambda_2^*) \left\{ \arg \min_{\beta} \left\| Y - \sum_{j=1}^{p} X_j \beta_j \right\|^2 + \lambda_2^* \sum_{j=1}^{p} \beta_j^2 + \lambda_1^* \sum_{j=1}^{p} |\beta_j| \right\}
\]

Within the elastic net are the noticeable ridge and lasso penalties, and by letting \( \lambda_2^* = 0 \) the elastic net reduces to the lasso. A \( \lambda_2^* > 0 \) will allow the elastic net to handle data with \( p > n \). Hence, Zou and Hastie \cite{32} found the elastic net to be a more favorable choice over the lasso in applications such as gene selection when \( p \gg n \).

### 2.3 Sparse Principal Component Analysis

One major drawback of PCA is that each principle component is a linear combination of all of the predictor variables, which often makes the results difficult to interpret. To address this problem, Zou et al. \cite{33} proposed the Sparse Principal Components Analysis (SPCA) as an alternative to reduce some of the linear coefficients to 0 by producing a sparse estimate of the loading matrix via the technique of penalized regression. Intuitively, since each PC is a linear combination of the original \( p \) variables, its loadings can be approximated by regressing the PC on these \( p \) variables. Hence, PCA can be alternatively expressed as a ridge regression problem. Let \( Z_i \) be scores of the \( i^{th} \) principal component, then

\[
\hat{\beta}_{ridge} = \arg \min_{\beta} \left\| Z_i - X \beta \right\|^2 + \lambda^* \| \beta \|^2,
\]

where the estimated loadings are found by

\[
\hat{u} = \frac{\hat{\beta}_{ridge}}{\| \hat{\beta}_{ridge} \|}.
\]
In this case, the ridge penalty is for the reconstruction of the principal components rather than a penalty toward the regression coefficients [33].

Since SPCA requires an ordinary PCA to be performed first, Zou et al. [33] present a more “self-contained” approach. Let $\lambda^* \geq 0$, then

$$\left\{ \hat{A}, \hat{B} \right\} = \arg \min_{A,B} \left\{ \|X - XBA^T\|^2 + \lambda^* \sum_{k=1}^{q} \|\beta_k\|^2 \right\}, \text{ subject to } A^T A = I.$$ 

will produce $\hat{B} \propto U$.

Here, $A = [\alpha_1, \alpha_2, \cdots, \alpha_q]$ and $B = [\beta_1, \beta_2, \cdots, \beta_q]$ are two $p \times q$ coefficient matrices. The normalized vector of $\hat{\beta}_k$ gives the approximation to the loadings of the $k^{th}$ principal component, i.e,

$$\hat{u}_k = \frac{\hat{\beta}_k}{\|\hat{\beta}_k\|}, \quad k = 1, \cdots, q.$$ 

Then, an $L_1$ or Lasso penalty [27] is added to the optimization criterion to induce sparsity, i.e., reduce some of the estimates to 0, via the elastic net [32]. Namely, the problem is formulated as

$$\left\{ \hat{A}, \hat{B} \right\} = \arg \min_{A,B} \left\{ \|X - XBA^T\|^2 + \lambda^* \sum_{k=1}^{q} \|\beta_k\|^2 + \sum_{k=1}^{q} \lambda_{k}^* \|\beta_k\|_1 \right\}, \quad A^T A = I,$$

where $\|\bullet\|_1$ denotes the $L_1$ norm, i.e., summation of the absolute values of the elements.

The resulting estimates $\hat{\beta}_k$ contain 0’s so that the variables associated with 0 coefficients are excluded from the constitution of the PC. The constants $\lambda^*$ and $\lambda_{k}^*$, $k = 1, \cdots, q$ are tuning parameters of the elastic net. Especially, $\lambda_{k}^*$‘s are tuning parameters associated with the lasso penalty, which controls the amount of shrinkage, i.e., how many coefficients are shrunk to 0. Smaller values of $\lambda_{k}^*$ induces more 0’s in $\hat{\beta}_k$.

While imposing a lasso penalty to the PCA is what produces the sparse loadings, the SPCA can be seen as a regression problem with both ridge and lasso penalties, where the PCA is the penalized ridge regression part. Hence, the naive elastic net is present within SPCA by the inclusion of both ridge and lasso penalties. The fitting of SPCA is carried out in the software R using the package elasticnet.
Lastly, although PCs are uncorrelated when using PCA, due to the induced sparsity in SPCA the resulting loadings deviate from being orthogonal, and consequently, the corresponding sparse PCs are not uncorrelated [33]. However, we willingly trade off PCs being uncorrelated for improvements in simplicity and predictive accuracy.
3.1 Principal Component Analysis Application

The unbonded tendon data are split into internally reinforced (internal) and externally reinforced (external) subsets each possessing 15 predictor variables and the response variable, \( \Delta f_{ps} \) based on the variables compiled and observations by Maguire et al. [18]. The internal data has 182 observations, and the external data has 71. The variable names and type that the variable was treated as are found in (Table 3.1). Both data subsets exhibit multicollinearity among predictors in their respective sample covariance matrices suggesting repetition of information. Due to the wide variation in scale of the different variables, the correlation matrix is chosen over the covariance matrix for the PCA.

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Notation</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Loading Type</td>
<td>( l_t )</td>
<td>Categorical</td>
</tr>
<tr>
<td>Span Length</td>
<td>( L )</td>
<td>Continuous</td>
</tr>
<tr>
<td>Beam Depth</td>
<td>( h )</td>
<td>Continuous</td>
</tr>
<tr>
<td>Beam Width</td>
<td>( b )</td>
<td>Continuous</td>
</tr>
<tr>
<td>Depth to P/S</td>
<td>( d_{ps} )</td>
<td>Continuous</td>
</tr>
<tr>
<td>Area of P/S</td>
<td>( A_{ps} )</td>
<td>Continuous</td>
</tr>
<tr>
<td>Ultimate Tendon Strength</td>
<td>( f_{pu} )</td>
<td>Continuous</td>
</tr>
<tr>
<td>Concrete Strength</td>
<td>( f_c )</td>
<td>Continuous</td>
</tr>
<tr>
<td>Area of Tens. Stl.</td>
<td>( A_s )</td>
<td>Continuous</td>
</tr>
<tr>
<td>Yield Strength</td>
<td>( f_y )</td>
<td>Continuous</td>
</tr>
<tr>
<td>Depth to Tens. Stl.</td>
<td>( d_s )</td>
<td>Continuous</td>
</tr>
<tr>
<td>Area of Comp. Stl.</td>
<td>( A_s' )</td>
<td>Continuous</td>
</tr>
<tr>
<td>Depth to Comp. Stl.</td>
<td>( d_{s'} )</td>
<td>Continuous</td>
</tr>
<tr>
<td>Initial Prestress</td>
<td>( f_{pe} )</td>
<td>Continuous</td>
</tr>
<tr>
<td>Elastic Modulus of Prestressing Reinforcement</td>
<td>( E_{ps} )</td>
<td>Continuous</td>
</tr>
<tr>
<td>Stress Increase in Unbonded Tendons</td>
<td>( \Delta f_{ps} )</td>
<td>Continuous</td>
</tr>
</tbody>
</table>

Table 3.1: Variable Names, Notation, and Type for the Statistical Analysis
Multiple approaches were used in selecting important variables for the PCA. The initial approach consisted of merely assuming that all 15 variables were important. An Eigen-decomposition was applied to the correlation matrix calculate the PCs. Figure 3.1 consists of scree-plots showing the proportion of variation and cumulative proportion of variation explained by each principal component for their respective data subset.

![Figure 3.1: Cumulative Explained Variation for each Principal Component for Models 1 and 2.](image)

An ‘elbow’, or change in slope between PCs [12], suggests a good choice for the number of PC’s that express the most information while keeping the model simple [16], e.g. the bend in the plot seen at three PC’s in the external data scree-plot. However, five principal components are selected for both the internal and external data as a means to compare models, and since five PCs capture a majority of proportion of variation in the data, while keeping the models relatively simple. The cumulative proportion of variation for 5 PC’s is 0.81 for the internal tendons, and 0.84 for the external tendons.

From the five selected principal components, linear combinations of the 15 variables can now be expressed as five new uncorrelated variables. By 10-fold cross validation, linear models called Model 1 and Model 2 are then fit to the data using the new five variables. As criterion of how well the models are fitting the data, the coefficient of
determination ($R^2$), adjusted $R^2$, and the average ratio of measured vs. predicted responses, $\lambda$, are calculated for each model and recorded in the first row of Table 4.1 [14].

$R^2$ is the ratio of the explained variation made by the model over the total variation in the data, defined as:

$$R^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

where $\hat{y}_i$ is the $i^{th}$ predicted $\Delta f_{ps}$, $y_i$ is the $i^{th}$ $\Delta f_{ps}$, and $\bar{y}$ is the sample average of $\Delta f_{ps}$.

Adjusted $R^2$ is similar to $R^2$ but it is penalized for more complicated models that involve more predictors. It is calculated as follows:

$$R^2_a = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(n - p - 1)}{\sum_{i=1}^{n}(y_i - \bar{y})^2/(n - 1)}$$

where $p$ is the number of predictor variables used in the model plus one.

Lastly, $\lambda$ is calculated as the mean of all of the ratios of $\Delta f_{ps}$ values and their corresponding linear model predicted values, $\hat{\Delta f_{ps}}$, i.e.

$$\lambda = \frac{1}{n} \sum_{i=1}^{n} \frac{\Delta f_{ps_i}}{\hat{\Delta f_{ps_i}}}$$

A visualization of $\lambda$ for the initial model is seen in Figure 3.2 and Figure 3.3 as plots of the $\Delta f_{ps}$ values against the linear model’s predicted values, $\hat{\Delta f_{ps}}$. 
After calculation, Model 1 and Model 2 have criteria $R^2$ as 0.355, 0.627, adjusted $R^2$ as 0.337, 0.598, and $\lambda$ as 0.997, 1.104 respectively (see Table 4.1).
A second approach was attempted by handling the continuous and categorical variables separately. While all of the variables are continuous except \( l_t \), the variables \( E_{ps} \) and \( d'_s \) behaved as categorical in the data and are treated as such (Table 3.1). A separate PCA was computed for the 12 continuous variables and the three categorical variables within each data set. In order to keep the same number of overall PC’s in the final models, four PC’s are chosen for the continuous variables, and one is chosen for the categorical variables as seen in the scree plots in Figure 3.4 and Figure 3.5. The results were then combined into linear models, called Model 3 and Model 4, and their criteria are \( R^2 = 0.281, 0.625 \), adjusted \( R^2 = 0.260, 0.596 \), and \( \lambda = 1.011, 1.017 \), as shown in Table 4.1. Plots for measured vs. predicted \( \Delta f_{ps} \) are also included as Figure 3.6 and Figure 3.7.

![Figure 3.4: Cumulative Explained Variation for each Principal Component for Continuous Variables.](image-url)
Figure 3.5: Cumulative Explained Variation for each Principal Component for Categorical Variables.

Figure 3.6: $\Delta f_{ps}$ vs. $\hat{\Delta f}_{ps}$ for Model 3.
Figure 3.7: $\Delta f_{ps}$ vs. $\hat{\Delta f}_{ps}$ for Model 4.

Again, linear models 1, 2, 3, and 4 suffer due to the fact that each principal component is a linear combination of all predictor variables, which is also not ideal for structural design. Variable selection restricting only important variables into the PCA would allow for simpler linear models with possibly better predictive power. Two techniques are employed and are also compared to the initial analysis. The first set of selected important variables is decided through professional suggestion. The authors call this set the “self-selected” set. The second set, called the “correlation cutoff” set, is selected by a test of minimum correlation with $\Delta f_{ps}$.

The self-selected important variables are $L$, $h$, $A_{ps}$, $f_c$, $A_s$, $A_{s'}$, $f_{pe}$, & $\Delta f_{ps}$. After a PCA is run on these variables the data is reduced from only seven predictor variables to five. While this is not a gain of much more simplicity to the model, the correlation between our predictors is removed. The scree plots in Figure 6 again show that most of the information is expressed in the first five PC’s chosen.
While there is a noticeable gain in cumulative proportion of variance explained by these 5 PC’s in both data sets (0.896 for the internal data, and 0.977 for the external data), the final models do not make similar gains in modeling the data, as seen by their respective $R^2 = 0.282$, 0.515, adjusted $R^2 = 0.262$, 0.478, and $\lambda = 1.212$, 1.023, values. This is also seen in Figure 3.9 and Figure 3.10.
This process is repeated for the correlation cut-off set as well. However, these variables were selected by first examining their respective correlations with \( \Delta f_{ps} \). While simply selecting predictors with a certain amount of correlation with the response does
not consider collinearity among predictors, the subsequent PCA and SPCA handles this by producing uncorrelated PC’s. A Pearson’s product-moment correlation test is applied with a level of significance set at 0.05. Table 3.2 and Table 3.3 contain the correlations and p-values for both internal and external data, respectively.

Interestingly, Table 3.2 indicates that for internally bonded tendons, the length is not important, which Mojtahedi and Gamble [21] indirectly indicate is important. Concrete strength is not considered important, although it shows up in the Mattock et al. [20] and the current ACI code. The variables $b$, $d$ and $A_{ps}$ are considered important and are also considered in the ACI code as the prestressing reinforcing ratio ($\rho_{ps}$). Interestingly $f_y$ is considered important, although it is not included in any known prediction model, and conversely, $A_s$ is not considered important.

Additionally, Table 3.3 indicates that there are considerable differences in the significance of many variables. There is agreement on several variables, for instance, the loading type, depth of section ($h$ and $d_{ps}$) and $A_{ps}$ are considered important while $d_s$, $d_s'$ and $E_{ps}$ are not considered important in both sets. However, the remaining variables are in contention. For instance, length is considered important in the external dataset as is concrete strength, $f_{pe}$ and $A_s$, but not $f_y$. Interestingly, $A_s'$ is considered important in the external dataset. Furthermore, $h$, $f_{ps}$, $A_s$, and $f_y$ were found to have opposite effect (see difference signs in Table 3.2 and Table 3.3) on the behavior, indicating either very different phenomenological effects or shortcomings in the dataset.

The dataset itself is made of all of the available experimental data, but the dataset is also shaped by the experimental needs. Externally reinforced members tend to be larger bridge girders with higher reinforcing ratios and, often, $A_s'$. The make-up of the externally reinforced dataset reflects this and contains more beam-like members (higher $d_{ps}$, $n$, $A_{ps}$, $A_s'$ etc.), many of them simulating bridge girders. The internally reinforced dataset is made up of many more slab like members that do not contain compression steel, are smaller, and some are scaled. Regardless, one should be aware that the dataset, while the largest available, does contain limited numbers and limited variations for many variables [18]. From this analysis, it is unclear if the difference in variable importance is due to the dataset or phenomenological differences. The analysis does seem to dispute the use of the same equation for internal and external members (like ACI and AASHTO) and indicates predictions that somehow account for the difference may be better [7].
<table>
<thead>
<tr>
<th>Variable</th>
<th>Correlation</th>
<th>p-value</th>
<th>Important</th>
</tr>
</thead>
<tbody>
<tr>
<td>$l_t$</td>
<td>0.514</td>
<td>1.0954e-13</td>
<td>TRUE</td>
</tr>
<tr>
<td>$L$</td>
<td>-0.056</td>
<td>0.45086</td>
<td>FALSE</td>
</tr>
<tr>
<td>$h$</td>
<td>0.28</td>
<td>0.00012805</td>
<td>TRUE</td>
</tr>
<tr>
<td>$b$</td>
<td>-0.171</td>
<td>0.020703</td>
<td>TRUE</td>
</tr>
<tr>
<td>$d_{ps}$</td>
<td>0.289</td>
<td>7.3882e-05</td>
<td>TRUE</td>
</tr>
<tr>
<td>$A_{ps}$</td>
<td>-0.508</td>
<td>2.4338e-13</td>
<td>TRUE</td>
</tr>
<tr>
<td>$f_{pu}$</td>
<td>0.325</td>
<td>7.3457e-06</td>
<td>TRUE</td>
</tr>
<tr>
<td>$f_c$</td>
<td>-0.01</td>
<td>0.89468</td>
<td>FALSE</td>
</tr>
<tr>
<td>$A_s$</td>
<td>0.012</td>
<td>0.87177</td>
<td>FALSE</td>
</tr>
<tr>
<td>$f_y$</td>
<td>0.224</td>
<td>0.002387</td>
<td>TRUE</td>
</tr>
<tr>
<td>$d_s$</td>
<td>-0.039</td>
<td>0.6031</td>
<td>FALSE</td>
</tr>
<tr>
<td>$A'_s$</td>
<td>-0.045</td>
<td>0.54725</td>
<td>FALSE</td>
</tr>
<tr>
<td>$d'_s$</td>
<td>-0.079</td>
<td>0.28672</td>
<td>FALSE</td>
</tr>
<tr>
<td>$f_{pe}$</td>
<td>0.055</td>
<td>0.45837</td>
<td>FALSE</td>
</tr>
<tr>
<td>$E_{ps}$</td>
<td>-0.082</td>
<td>0.26934</td>
<td>FALSE</td>
</tr>
<tr>
<td>$\Delta f_{ps}$</td>
<td>1</td>
<td>0</td>
<td>TRUE</td>
</tr>
</tbody>
</table>

Table 3.2: Correlation Cut-off Important Variables for the Internal Data
<table>
<thead>
<tr>
<th>Variable</th>
<th>Correlation</th>
<th>p-value</th>
<th>Important</th>
</tr>
</thead>
<tbody>
<tr>
<td>$l_t$</td>
<td>0.246</td>
<td>0.038422</td>
<td>TRUE</td>
</tr>
<tr>
<td>$L$</td>
<td>-0.267</td>
<td>0.02437</td>
<td>TRUE</td>
</tr>
<tr>
<td>$h$</td>
<td>-0.567</td>
<td>2.5027e-07</td>
<td>TRUE</td>
</tr>
<tr>
<td>$b$</td>
<td>-0.033</td>
<td>0.78333</td>
<td>FALSE</td>
</tr>
<tr>
<td>$d_{ps}$</td>
<td>0.235</td>
<td>0.048071</td>
<td>TRUE</td>
</tr>
<tr>
<td>$A_{ps}$</td>
<td>-0.488</td>
<td>1.5907e-05</td>
<td>TRUE</td>
</tr>
<tr>
<td>$f_{pu}$</td>
<td>-0.217</td>
<td>0.069402</td>
<td>FALSE</td>
</tr>
<tr>
<td>$f_c$</td>
<td>0.517</td>
<td>3.8469e-06</td>
<td>TRUE</td>
</tr>
<tr>
<td>$A_s$</td>
<td>-0.529</td>
<td>2.1348e-06</td>
<td>TRUE</td>
</tr>
<tr>
<td>$f_y$</td>
<td>-0.229</td>
<td>0.054291</td>
<td>FALSE</td>
</tr>
<tr>
<td>$d_s$</td>
<td>-0.137</td>
<td>0.25533</td>
<td>FALSE</td>
</tr>
<tr>
<td>$A_s'$</td>
<td>-0.349</td>
<td>0.0028228</td>
<td>TRUE</td>
</tr>
<tr>
<td>$d_s'$</td>
<td>-0.14</td>
<td>0.24515</td>
<td>FALSE</td>
</tr>
<tr>
<td>$f_{pe}$</td>
<td>0.346</td>
<td>0.0030873</td>
<td>TRUE</td>
</tr>
<tr>
<td>$E_{ps}$</td>
<td>0.093</td>
<td>0.43923</td>
<td>FALSE</td>
</tr>
<tr>
<td>$\Delta f_{ps}$</td>
<td>1</td>
<td>0</td>
<td>TRUE</td>
</tr>
</tbody>
</table>

Table 3.3: Correlation Cut-off Important Variables for the External Data

If a variable exhibited significant correlation (p-value less than 0.05) with $\Delta f_{ps}$ it was labeled as “important” and kept for subsequent analysis. Model 11 uses the correlation cut-off variables for the internal data are $l_t$, $h$, $b$, $d_{ps}$, $A_{ps}$, $f_{pu}$, $f_y$, & $\Delta f_{ps}$, and Model 12 uses the correlation cut-off variables for the external data are $l_t$, $L$, $h$, $d_{ps}$, $A_{ps}$, $f_{pu}$, $f_c$, $A_s$, $f_y$, $A_s'$, $f_{pe}$, & $\Delta f_{ps}$. The scree plots in Figure 3.9 show a cumulative proportion of variation for the internal data is 0.974, and 0.904 for the external data.
Figure 3.11: Cumulative Explained Variation for each Principal Component for Models 11 and 12.

By using Pearson’s product-moment correlation test to remove variables that exhibit low correlations with $\Delta f_{ps}$, and then applying a PCA on the remaining predictors, the linear models tend to model the data better as seen in their respective $R^2 = 0.489, 0.640$, adjusted $R^2 = 0.475, 0.612$, and $\lambda = 1.010, 1.029$, values (see Table 4.1).
Figure 3.12: $\Delta f_{ps}$ vs. $\hat{\Delta f}_{ps}$ for Model 11.

Figure 3.13: $\Delta f_{ps}$ vs. $\hat{\Delta f}_{ps}$ for Model 12.
3.2 Sparse Principal Component Analysis Application

SPCA was applied to both internal and external data sets on the Continuous and Categorical, Self-Selected, and Minimum Correlation Cut-off sets of variables producing six additional linear models. Models 5 and 6 handle the Continuous and Categorical variables separately as before. Models 9 and 10 use the Self-Selected variables, and Models 13 and 14 use the Correlation Cut-off variables. In all of these cases, a decision must be made about how much sparsity is desirable. Again, sparsity in the Principal Components is the reduction of some of the coefficients, or loadings, for the linear combinations of the predictor variables to zero.

Since Models 5 and 6 handle the continuous and categorical variables separately, Figure 3.14 and Figure 3.15 reveal optimal choices for the number of sparse coefficients per PC by maximization of adjusted $R^2$. The variation in the external subset is being explained significantly better by the data than the internal subset as seen by the consistently higher adjusted $R^2$ in Figure 3.14. On the other hand, Figure 3.15 shows little variation in data, which is explained by the variables that were treated as categorical variables. The maximum adjusted $R^2$ for the internal continuous variables is achieved by reducing all but five loadings to zero. The external continuous variables reach their highest adjusted $R^2$ by using only one non-zero loading per PC. Both maximum adjusted $R^2$ values, along with their corresponding $R^2$ and $\lambda$ values are recorded in Table 4.1.
Additionally, Table 3.4 and Table 3.5 show which continuous and categorical variable loadings were not reduced to zero. In effect, the reduction of these loadings to zero
acts as another type of variable selection.

<table>
<thead>
<tr>
<th>Variable</th>
<th>PC1</th>
<th>PC2</th>
<th>PC3</th>
<th>PC4</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$h$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$b$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$d_{ps}$</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$A_{ps}$</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>$f_{pu}$</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$f_c$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>$A_s$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$f_y$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$d_s$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$A'_s$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>$f_{pe}$</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 3.4: SPCA Loadings for Internal Continuous Variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>PC1</th>
</tr>
</thead>
<tbody>
<tr>
<td>$l_t$</td>
<td>0.00</td>
</tr>
<tr>
<td>$d'_s$</td>
<td>0.71</td>
</tr>
<tr>
<td>$E_{ps}$</td>
<td>0.70</td>
</tr>
</tbody>
</table>

Table 3.5: SPCA Loadings for Internal Categorical Variables
<table>
<thead>
<tr>
<th>Variable</th>
<th>PC1</th>
<th>PC2</th>
<th>PC3</th>
<th>PC4</th>
</tr>
</thead>
<tbody>
<tr>
<td>$L$</td>
<td>0.45</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>$h$</td>
<td>0.64</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>$b$</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.83</td>
</tr>
<tr>
<td>$d_{ps}$</td>
<td>0.00</td>
<td>-0.47</td>
<td>0.00</td>
<td>0.001</td>
</tr>
<tr>
<td>$A_{ps}$</td>
<td>0.40</td>
<td>0.00</td>
<td>0.00</td>
<td>0.01</td>
</tr>
<tr>
<td>$f_{pu}$</td>
<td>0.10</td>
<td>0.00</td>
<td>-0.53</td>
<td>-0.04</td>
</tr>
<tr>
<td>$f_{c} $</td>
<td>0.00</td>
<td>-0.67</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>$A_{s}$</td>
<td>0.47</td>
<td>0.00</td>
<td>-0.01</td>
<td>0.00</td>
</tr>
<tr>
<td>$f_{y} $</td>
<td>0.00</td>
<td>0.25</td>
<td>-0.43</td>
<td>0.00</td>
</tr>
<tr>
<td>$d_{s}$</td>
<td>0.00</td>
<td>0.03</td>
<td>0.73</td>
<td>0.00</td>
</tr>
<tr>
<td>$A'_{s}$</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.56</td>
</tr>
<tr>
<td>$f_{pc}$</td>
<td>0.00</td>
<td>-0.51</td>
<td>-0.09</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Table 3.6: SPCA loadings for External Continuous Variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>PC1</th>
</tr>
</thead>
<tbody>
<tr>
<td>$l_{t}$</td>
<td>0</td>
</tr>
<tr>
<td>$d'_{s}$</td>
<td>-1</td>
</tr>
<tr>
<td>$E_{ps}$</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 3.7: SPCA loadings for External Categorical Variables

Again, Figure 3.16 and Figure 3.17 compare the measured vs. predicted stress increase in the unbounded tendons.
Figure 3.16: $\Delta f_{ps}$ vs. $\Delta \hat{f}_{ps}$ for Model 5.

Figure 3.17: $\Delta f_{ps}$ vs. $\Delta \hat{f}_{ps}$ for Model 6.

The following equations are the SPCA linear models with their respective PC’s.
Prediction Equation for Internal Continuous and Categorical SPCA (Model 5)

\[ \hat{\Delta f_{ps}} = 42.82 - 6.17PC_1 + 6.41PC_2 - 11.09PC_3 - 1.74PC_4 + 0.94PC_5 \]

- \( PC_1 = -1d_{ps} \)
- \( PC_2 = 1f_{pu} \)
- \( PC_3 = 1A_{ps} \)
- \( PC_4 = 1f_c \)
- \( PC_5 = 0.71d_s' + 0.70E_{ps} \)

Prediction Equation for External Continuous and Categorical SPCA (Model 6)

\[ \hat{\Delta f_{ps}} = 68.24 - 14.93PC_1 - 19.51PC_2 - 3.33PC_3 + 1.66PC_4 - 8.37PC_5 \]

- \( PC_1 = 0.45L + 0.64h + 0.40A_{ps} + 0.10f_{pu} + 0.47A_s \)
- \( PC_2 = -0.47d_{ps} - 0.67f_c + 0.25f_y + 0.03d_s - 0.51f_{pe} \)
- \( PC_3 = -0.53f_{pu} - 0.01A_s - 0.43f_y + 0.73d_s - 0.09f_{pe} \)
- \( PC_4 = 0.83b + 0.001d_{ps} + 0.01A_{ps} - 0.04f_{pu} + 0.56A_s' \)
- \( PC_5 = -1d_s' \)

Next, SPCA was applied to the Self-Selected variables. As with the Continuous and Categorical variables, Figure 3.18 shows the consistently higher adjusted \( R^2 \) for the external data. Surprisingly, only one loading for each PC in the internal and external model is calculated to maximize the adjusted \( R^2 \). This suggests that a simple linear model is sufficient in modeling the variation in the stress increase \( \Delta f_{ps} \). Again, these adjusted \( R^2 \) values, along with their corresponding \( R^2 \) and \( \lambda \) values are recorded in Table 4.1.
Table 3.8 shows that the five variables needed for the linear model are \( L, h, A_{ps}, f_c, \) and \( A_s \). However, Table 3.9 suggests variables \( h, f_c, f_{pc}, A_s, \) and \( A'_s \). While two variables differ between models, the loadings are the same for the variables in common, \( h, f_c, \) and \( A_s \).

<table>
<thead>
<tr>
<th>Variable</th>
<th>PC1</th>
<th>PC2</th>
<th>PC3</th>
<th>PC4</th>
<th>PC5</th>
</tr>
</thead>
<tbody>
<tr>
<td>( L )</td>
<td>0</td>
<td>-1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>( h )</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>( A_{ps} )</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>( f_c )</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>-1</td>
</tr>
<tr>
<td>( f_{pc} )</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>( A_s )</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>( A'_s )</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>

Table 3.8: SPCA loadings for Internal Self-Selected Important Variables
Again, Figure 3.19 and Figure 3.20 show the measured vs. predicted stress increase.

![Graph showing measured vs. predicted stress increase](image)

**Figure 3.19**: $\Delta f_{ps}$ vs. $\hat{\Delta f}_{ps}$ for Model 9.
The linear equations for Models 9 and 10 are much simpler when compared to their corresponding PCA models, namely Models 7 and 8, since each PC in the PCA models have seven variables whereas each PC in the SPCA have only one. This gain in simplicity is paired with gains in $R^2$, and adjusted $R^2$, along with $\lambda$ values closer to one (Table 4.1).

**Prediction Equation for Internal Self-Selected SPCA (Model 9)**

$$\hat{\Delta f_{ps}} = 42.82 - 1.92 PC_1 - 0.36 PC_2 - 10.64 PC_3 + 6.12 PC_4 + 1.78 PC_5$$

$PC_1 = 1A_s$

$PC_2 = -1L$

$PC_3 = 1A_{ps}$

$PC_4 = 1h$

$PC_5 = -1f_c$
**Prediction Equation for External Self-Selected SPCA (Model 10)**

\[ \Delta f_{ps} = 68.24 - 26.81PC_1 - 8.98PC_2 + 3.80PC_3 - 13.03PC_4 + 0.69PC_5 \]

\[ PC_1 = 1h \]
\[ PC_2 = -1f_c \]
\[ PC_3 = 1A'_s \]
\[ PC_4 = -1f_{pe} \]
\[ PC_5 = 1A_s \]

Lastly, SPCA is applied to the Correlation Cut-off variables. Recall that while eleven variables were retained for the external data, only seven were kept for the internal data. Hence, the number of non-zero loadings for each SPC for the internal data only extends to seven in Figure 3.21. Table 3.10 and Table 3.11 show the specific choices for the sparse loadings for each PC.

![Graph](image-url)

**Figure 3.21:** Number of Loadings for the SPC regression vs. Adjusted $R^2$ Values for Correlation Cut-off Variables in Models 13 and 14.
Table 3.10: SPCA loadings for Internal Correlation Cut-off Important Variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>PC1</th>
<th>PC2</th>
<th>PC3</th>
<th>PC4</th>
<th>PC5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$l_t$</td>
<td>0.002</td>
<td>-0.68</td>
<td>0.23</td>
<td>0.07</td>
<td>0.00</td>
</tr>
<tr>
<td>$h$</td>
<td>0.73</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>$b$</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>-0.99</td>
</tr>
<tr>
<td>$d_{ps}$</td>
<td>0.68</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>$A_{ps}$</td>
<td>0.00</td>
<td>0.00</td>
<td>-0.95</td>
<td>0.004</td>
<td>-0.0002</td>
</tr>
<tr>
<td>$f_{pu}$</td>
<td>0.00</td>
<td>-0.74</td>
<td>-0.21</td>
<td>0.00</td>
<td>-0.0004</td>
</tr>
<tr>
<td>$f_y$</td>
<td>0.00</td>
<td>-0.05</td>
<td>0.00</td>
<td>-0.998</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Table 3.11: SPCA loadings for External Correlation Cut-off Important Variables

<table>
<thead>
<tr>
<th>Variable</th>
<th>PC1</th>
<th>PC2</th>
<th>PC3</th>
<th>PC4</th>
<th>PC5</th>
</tr>
</thead>
<tbody>
<tr>
<td>$l_t$</td>
<td>0.00</td>
<td>0.00</td>
<td>0.97</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>$L$</td>
<td>0.34</td>
<td>0.07</td>
<td>0.00</td>
<td>0.00</td>
<td>0.58</td>
</tr>
<tr>
<td>$h$</td>
<td>0.57</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>$d_{ps}$</td>
<td>0.00</td>
<td>0.51</td>
<td>0.18</td>
<td>0.03</td>
<td>0.00</td>
</tr>
<tr>
<td>$A_{ps}$</td>
<td>0.51</td>
<td>0.00</td>
<td>0.00</td>
<td>0.06</td>
<td>0.00</td>
</tr>
<tr>
<td>$f_{pu}$</td>
<td>0.00</td>
<td>0.00</td>
<td>0.01</td>
<td>-0.88</td>
<td>0.16</td>
</tr>
<tr>
<td>$f_c$</td>
<td>-0.03</td>
<td>0.52</td>
<td>-0.18</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>$A_s$</td>
<td>0.40</td>
<td>-0.01</td>
<td>0.00</td>
<td>-0.11</td>
<td>0.03</td>
</tr>
<tr>
<td>$f_y$</td>
<td>0.00</td>
<td>-0.19</td>
<td>-0.03</td>
<td>-0.43</td>
<td>-0.37</td>
</tr>
<tr>
<td>$A_{s}'$</td>
<td>0.39</td>
<td>0.00</td>
<td>0.03</td>
<td>0.00</td>
<td>-0.68</td>
</tr>
<tr>
<td>$f_{pe}$</td>
<td>0.00</td>
<td>0.65</td>
<td>0.00</td>
<td>-0.16</td>
<td>-0.17</td>
</tr>
</tbody>
</table>

It should be noted that the predicted stress increase, $\hat{\Delta f_{ps}}$ is consistently under predicting for higher measured values of $\Delta f_{ps}$ in the internal data (Figure 3.22). Some of this is also exhibited in Figure 3.23 though not as strongly for the external data.
Figure 3.22: $\Delta f_{ps}$ vs. $\hat{\Delta f_{ps}}$ for Model 13.

Figure 3.23: $\Delta f_{ps}$ vs. $\hat{\Delta f_{ps}}$ for Model 14.

While we include Models 13 and 14 here with their PCs explicitly listed, with some algebraic manipulation simpler versions of Models 13 and 14 are presented in the following discussion section of this paper.
Prediction Equation for Internal Correlation Cut-off SPCA (Model 13)

$$\Delta f_{ps} = 42.82 + 5.70PC_1 - 6.21PC_2 + 11.23PC_3 + 4.56PC_4 - 3.35PC_5$$

- $PC_1 = 0.002l_t + 0.73h + 0.68d_{ps}$
- $PC_2 = -0.68l_t - 0.74f_{pu} - 0.05f_y$
- $PC_3 = 0.23l_t - 0.95A_{ps} - 0.21f_{pu}$
- $PC_4 = 0.07l_t + 0.004A_{ps} - 0.998f_y$
- $PC_5 = -0.99b - 0.0002A_{ps} - 0.0004f_{pu}$

Prediction Equation for External Correlation Cutoff SPCA (Model 14)

$$\Delta f_{ps} = 68.24 - 14.34PC_1 + 17.66PC_2 + 9.31PC_3 - 3.44PC_4 - 1.09PC_5$$

- $PC_1 = 0.34L + 0.57h + 0.51A_{ps} - 0.03f_c + 0.40A_s + 0.39A'_s$
- $PC_2 = 0.07L + 0.51d_{ps} + 0.52f_c - 0.01A_s - 0.19f_y + 0.65f_{pe}$
- $PC_3 = 0.97l_t + 0.18d_{ps} + 0.01f_{pu} - 0.18f_c - 0.03f_y + 0.03A'_s$
- $PC_4 = 0.03d_{ps} + 0.06A_{ps} - 0.88f_{pu} - 0.11A_s - 0.43f_y - 0.16f_{pe}$
- $PC_5 = 0.58L + 0.16f_{pu} + 0.03A_s - 0.37f_y - 0.68A'_s - 0.17f_{pe}$
CHAPTER 4
CONCLUSION

4.1 Discussion Of Results

From Table 4.1 the $R^2$, adjusted $R^2$, and $\lambda$ values for the initial models involving all 15 variables (Models 1 and 2) are 0.355, 0.337, 0.997, for the internal data, and 0.627, 0.598, 1.104, for the external data. Comparatively, these initial PCA linear models improve significantly over previous methods [18], where $\lambda = 1.85$ and $R^2 = 0.16$ for the AASHTO, being the most accurate and precise of the available American codified methods, as well as $\lambda = 1.34$ and $R^2 = 0.27$ for the previously proposed method modification to the AASHTO prediction [18].

<table>
<thead>
<tr>
<th>Variables</th>
<th>Internal Data</th>
<th></th>
<th>External Data</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$R^2$</td>
<td>Adj. $R^2$</td>
<td>$\lambda$</td>
<td>$R^2$</td>
</tr>
<tr>
<td>All Variables</td>
<td>0.355</td>
<td>0.337</td>
<td>0.997</td>
<td>0.627</td>
</tr>
<tr>
<td>Cont. &amp; Cate.</td>
<td>0.281</td>
<td>0.260</td>
<td>1.011</td>
<td>0.625</td>
</tr>
<tr>
<td>Self-Selected</td>
<td>0.282</td>
<td>0.262</td>
<td>1.212</td>
<td>0.515</td>
</tr>
<tr>
<td>Corr. Cut-off</td>
<td>0.489</td>
<td>0.475</td>
<td>1.010</td>
<td>0.635</td>
</tr>
</tbody>
</table>

Table 4.1: PCA Models’ $R^2$, Adjusted $R^2$, and $\lambda$ Values.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Internal Data</th>
<th></th>
<th>External Data</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$R^2$</td>
<td>Adj. $R^2$</td>
<td>$\lambda$</td>
<td>$R^2$</td>
</tr>
<tr>
<td>All Variables</td>
<td>0.378</td>
<td>0.360</td>
<td>0.996</td>
<td>0.677</td>
</tr>
<tr>
<td>Cont. &amp; Cate.</td>
<td>0.440</td>
<td>0.424</td>
<td>1.018</td>
<td>0.634</td>
</tr>
<tr>
<td>Self-Selected</td>
<td>0.332</td>
<td>0.313</td>
<td>0.984</td>
<td>0.555</td>
</tr>
<tr>
<td>Corr. Cut-off</td>
<td>0.492</td>
<td>0.478</td>
<td>1.013</td>
<td>0.654</td>
</tr>
</tbody>
</table>

Table 4.2: SPCA Models’ $R^2$, Adjusted $R^2$, and $\lambda$ Values.

The linear equations for the initial SPCA models are much simpler when compared to their corresponding PCA models since each PC is required to have 15 loadings,
whereas each SPC have only six. This gain in simplicity is paired with gains in $R^2$, and adjusted $R^2$, along with $\lambda$ values closer to one (compare the first rows in Table 4.1 and Table 4.2).

The PCA models handling the continuous and discrete variables separately (Models 3 and 4) did not perform better than the initial models involving all 15 variables. This may be due to the unaccounted covariances between the continuous and discrete variables. However, the SPCA on the continuous and categorical variables did improve over the initial model. A possible contributing reason for this increase is the ability of the SPCA to create new variables which are not required to be linear combinations of all of the continuous and categorical variables, i.e. a type of variable selection within the SPCA.

While the PCA and SPCA models for the self-selected variables (Models 7 - 10) did improve over the AASHTO and proposed modified AASHTO predictions [18], they performed poorer than the initial PCA on all of the variables. This suggests that variables engineers commonly associate with $\Delta f_{ps}$ may not be as impactful as thought, underscoring the necessity for further phenomenological study.

Notice only one loading for each PC in the internal and external model for the self-selected subset and internal continuous subset is suggested to maximize the adjusted $R^2$. This suggests that a simple linear model is sufficient in modeling the variation in the tendon stress increase for these cases. Additionally, Table 3.10 and Table 3.11 show that only the five variables are needed for the self-selected linear models, namely $L$, $h$, $A_{ps}$, $f_c$, and $A_s$ are the internal variables, whereas the external variables are $h$, $f_c$, $f_{pe}$, $A_s$, and $A_s'$. While two variables differ between models, the loadings are the same for the variables in common: $h$, $f_c$, and $A_s$.

It should be noted that the predicted stress increase, $\hat{\Delta f_{ps}}$, is consistently under predicting for higher measured values of $\Delta f_{ps}$ in the internal data (see Figures 3.2, 3.6, 3.9, 3.12, 3.16, 3.19, and 3.22). Some of this is also exhibited in the external data though not as strongly (see Figures 3.3, 3.7, 3.10, 3.13, 3.17, 3.20, and 3.23). This suggests that an underlying non-linear relationship may be present in the data, and suggests further analysis possibly involving more advanced models.
Additionally, a property of the coefficient of determination is $R^2$ must increase for every additional variable added to a specific model. (This is part of the motivation for also using adjusted $R^2$, which penalizes $R^2$ for every additional variable.) However, there is an increase in $R^2$ seen in the PCA model using all of the variables and the PCA model using only the correlation cut-off variables. (Compare $R^2$ in the first and last lines in Table 4.1, and the internal data in Table 4.2). While seemingly contradictory, the results and the principle are not at odds. While the correlation cut-off models have PCs using fewer of the original variables, the number of PCs used in each model is unchanged. Furthermore, the calculated loadings are different for their respective variables between the two models. Hence, a comparison of this type is not applicable. It is possible to compute increases or decreases in $R^2$ between these models as the analysis has shown.

Most notably, the $R^2$, adjusted $R^2$, and $\lambda$ values for the correlation cutoff SPCA models (Models 13 and 14) are 0.492, 0.478, 1.013 for the internal data, and 0.654, 0.627, 1.018 for the external data. Notice that while the difference in increased $R^2$ and adjusted $R^2$ for the internal model is 0.137 and 0.141, a substantial amount, the external model does not improve over the initial SPCA for all external variables. However, the small decrease in $R^2$ for the external model is sacrificed in exchange for the simpler equations for these models. Therefore, the SPCA correlation cutoff model equations are recommended over the initial equations.

*Simplified Prediction Equation for Internal Correlation Cutoff SPCA*

$$\hat{\Delta f_{ps}} = 42.82 + 7.05l_t + 4.17h + 3.35b + 3.88d_{ps} - 10.67A_{ps} + 2.23f_{pu} - 4.25f_y$$

*Simplified Prediction Equation for External Correlation Cutoff SPCA*

$$\hat{\Delta f_{ps}} = 68.24 + 8.99l_t - 4.29L - 8.11h + 10.61d_{ps} - 7.51A_{ps} + 2.96f_{pu}$$

$$\quad + 7.90f_c - 5.53A_s - 1.84f_y - 4.57A'_s + 12.27f_{pe}$$

### 4.2 Summary and Conclusion

The data in Maguire, et al. [18] were separated into two data sets determined by internal or external tendons. Stochastic linear models based on PCA and SPCA were
constructed as prediction equations for $\Delta f_{ps}$. Eight linear models involved all of the available explanatory variables, of which four handled the Continuous and Discrete variables separately. The remaining eight models used only subsets of important variables, which were the Self-Selected, or Correlation Cutoff important variable subsets. Upon comparison, the simplified linear models using SPCA on the Correlation Cutoff variables performed notably with $R^2$, adjusted $R^2$, and $\lambda$ values of 0.492, 0.478, 1.013, and 0.654, 0.627, 1.018 for the internal and external data respectively.

The following conclusions can be made from the above work

1. External and internal members show vastly different importance for different variables within the dataset. The reason for this is unclear, but is likely due to the differences in data contained in the dataset and, to a lesser extent, phenomenologically differences between the two structural systems.

2. Based on the above conclusion, there is a significant need for more data in order to obtain better understanding, statistically and phenomenologically, of unbonded tendon reinforced members. This is ideally accomplished through additional testing, as the available database is relatively small.

3. The PCA analysis on all of the variables produced $R^2$, adjusted $R^2$, and $\lambda$ values are 0.355, 0.337, 0.997, for the internal data, and 0.627, 0.598, 1.104, for the external data respectively.

4. The SPCA analysis on the correlation cut-off variables produced $R^2$, adjusted $R^2$, and $\lambda$ values of 0.492, 0.478, 1.013, for the internal data, and 0.654, 0.627, 1.018 for the external data respectively.

5. The PCA and SPCA analysis predicted significantly better than codified methods ($R^2 = 0.16$ and 0.08, $\lambda = 1.85$ and 2.01 for AASHTO and ACI respectively) and the optimized semi-empirical model presented by Maguire et al. [18] ($R^2 = 0.27$ and $\lambda = 1.34$).


APPENDIX
R Code

```r
library(knitr)
library(elasticnet)

# Reading in the data
ContData = read.csv("data_continuous.csv", header = FALSE, col.names = c("tendonType", "loadingPattern", "loadingType", "spanLength", "totalSpanLength", "beamDepth", "beamWidth", "depthToPS", "areaOfPS", "ultimateTendonStrength", "concreteStrength", "areaOfTensStl.", "yieldStrength", "depthToTensStl.", "areaOfComp.Stl.", "depthToComp.Stl.", "initialPrestress", "e_ps", "epsilon_cu", "beta_1", "deltaF")

SimpSupData = read.csv("data_simplysupported.csv", header = FALSE, col.names = c("tendonType", "loadingType", "spanLength", "beamDepth", "beamWidth", "depthToPS", "areaOfPS", "ultimateTendonStrength", "concreteStrength", "areaOfTensStl.", "yieldStrength", "depthToTensStl.", "areaOfComp.Stl.", "depthToComp.Stl.", "initialPrestress", "e_ps", "epsilon_cu", "beta_1", "deltaF")

ContData$epsilon_cu = NULL # Constant for all observations
SimpSupData$epsilon_cu = NULL

ContData$beta_1 = NULL # Advised to remove by engineering professional
SimpSupData$beta_1 = NULL

tenData = rbind(subset(ContData, select = c(tendonType, loadingType, spanLength, beamDepth, beamWidth, depthToPS, areaOfPS, ultimateTendonStrength, concreteStrength, areaOfTensStl., yieldStrength, depthToTensStl., areaOfComp.Stl., depthToComp.Stl., initialPrestress, e_ps, deltaF)), SimpSupData)

tenData = dplyr::select(tenData, tendonType, loadingType, spanLength, beamDepth, beamWidth, depthToPS, areaOfPS, ultimateTendonStrength, concreteStrength, areaOfTensStl., yieldStrength, depthToTensStl., areaOfComp.Stl., depthToComp.Stl., initialPrestress, e_ps, deltaF))

intData = tenData[tenData$tendonType == "1", -1]
row.names(intData) = 1:dim(intData)[1]

# Principal components using prcomp function.
```
intPC = prcomp(subset(intData, select=-deltaF), center = TRUE, scale. = TRUE, retx = TRUE)
kable(intPC$rotation[, 1:5])
screepplot(intPC,npcs = length(intPC$rotation[, 1]), type = "l", main = "Scree Plot for PCA on Internal Tendon Data")
somey(intPC)

# Building a linear model with PCA
intlm = lm(intData$deltaF ~ intPC$x[, 1] + intPC$x[, 2] + intPC$x[, 3] + intPC$x[, 4] + intPC$x[, 5])
somey(intlm)

plot(intData$deltaF, intlm$fitted.values, xlim = c(0, 100), ylim = c(0, 100),
    xlab = "Measured Delta Fps", ylab = "Predicted Delta Fps", main = "Measured Delta Fps vs. Predicted Delta Fps\non Internal Tendon Data", asp = 1)
abline(a = 0, b = 1)

RsqrdInt = summary(intlm)$r.squared
RsqrdInt

adjRsqrdInt = summary(intlm)$adj.r.squared
adjRsqrdInt

lambdaInt = mean(intData$deltaF/intlm$fitted.values)
lambdaInt

# Using Marc's suggested important variables
names(subset(intData, select=c(spanLength, beamDepth, areaOfPS,
    concreteStrength, initialPrestress, areaOfTensStl., areaOfComp.Stl., deltaF )))

# Removing unimportant variables
intMImp = subset(intData, select=c(spanLength, beamDepth, areaOfPS,
    concreteStrength, initialPrestress, areaOfTensStl., areaOfComp.Stl., deltaF ))

# Principal components using prcomp function.
intMImpPC = prcomp(subset(intMImp, select=-deltaF), center = TRUE, scale. = TRUE, retx = TRUE)
kable(intMImpPC$rotation[, 1:5])
screepplot(intMImpPC, npcs = length(intMImpPC$rotation[, 1]), type = "l", main = "Scree plot for PC on internal Tendon Data")
summary(intMImpPC)

# Building a linear model with PCA
intMImplm = lm(intMImp$deltaF ~ intMImpPC$x[, 1] + intMImpPC$x[, 2] + intMImpPC$x[, 3] + intMImpPC$x[, 4] + intMImpPC$x[, 5])
summary(intMImplm)

plot(intData$deltaF, intMImplm$fitted.values, xlim = c(0, 100), ylim = c(0, 100), xlab = "Measured delta F", ylab = "Predicted delta F", main = "Measured vs. Predicted")
abline(a = 0, b = 1)

RsqrdintMImp = summary(intMImplm)$r.squared
RsqrdintMImp

adjRsqrdintMImp = summary(intMImplm)$adj.r.squared
adjRsqrdintMImp

lambdaintMImp = mean(intMImp$deltaF/intMImplm$fitted.values)
lambdaintMImp

# SPCA on internal important variables

# Defining new function to determine number of non-zero loadings that produce max Adj R^2 for SPCA
intMImpSPCA <- spca(subset(intMImp, select=-deltaF), K = 5, type = "predictor", sparse = "varnum", para = c(rep(dim(subset(intMImp, select=-deltaF))[2], 5)), use.corr = TRUE) # Same as normal PCA.
intMImpSPCA$loadings

maxAdjR2SPCA = function(dataframe) {
    # dataframe is a data frame containing deltaF
}
for (i in 1:dim(subset(dataframe, select=-deltaF))[2]) {
  dataframeSPCA <- spca(subset(dataframe, select=-deltaF), K = 5, type = "predictor", sparse = "varnum", para = c(rep(i, 5)), use.corr = TRUE)
  if (i == 1) {
    standardize = function(x) {(x - mean(x))/sd(x)}
    dataX = apply(subset(dataframe, select=-deltaF), MARGIN=2, FUN=standardize)
    dataScores = NULL
    rsqrd = NULL
    adjrsqrd = NULL
    print("Max i is:")
    print(dim(subset(dataframe, select=-deltaF))[2])
  }
  dataScores = dataX %*% dataframeSPCA$loadings
  dataframeSPCAlm = lm(dataframe$deltaF ~ dataScores[, 1] + dataScores[, 2] +
                        dataScores[, 3] + dataScores[, 4] + dataScores[, 5])
  rsqrd = c(rsqrd, summary(dataframeSPCAlm)$r.squared)
  adjrsqrd = c(adjrsqrd, summary(dataframeSPCAlm)$adj.r.squared)
  print(i)
}
print("Number of non-zero loadings")
m1 = cbind(1:dim(subset(dataframe, select=-deltaF))[2], rsqrd, adjrsqrd)
print(m1)
if (which.max(m1[, 2]) == which.max(m1[, 3])) {
  print("Number of non-zero loadings suggested:")
  print(which.max(m1[, 3]))
} else {
  print("Number of non-zero loadings suggested:")
  print(which.max(m1[, 2]))
  print("or")
  print(which.max(m1[, 3]))
}
print("Max R^2")
print(max(rsqrd))
print("Max Adj R^2")
print(max(adjrsqrd))
plot(1:dim(subset(dataframe, select=-deltaF))[2], adjrsqrd, xlab = "Non-zero Loadings for each SPC", ylab = "Adj R^2", type = "b")
abline(v = which.max(m1[, 3]), col = "red")
legend(x = "right", legend = c("Max Adj. R^2"), cex = 0.8, lty=c(1), lwd = c (1.5), col = c("red"), inset = 0.02)

maxAdjR2SPCA(intMImp)

intMImpSPCA <- spca(subset(intMImp, select=-deltaF), K = 5, type = "predictor",
  sparse = "varnum", para = c(rep(1, 5)), use.corr = TRUE)
intMImpSPCA$loadings
intMImpSPCA$pev
sum(intMImpSPCA$pev)

# transforming data to zero mean and unit variance
standardize = function(x) {(x - mean(x))/sd(x)}
intMX = apply(subset(intMImp, select=-deltaF), MARGIN=2, FUN=standardize)
intMScores = intMX *%*% intMImpSPCA$loadings # Matrix multiplication to rotate
  the standardized data.

# Linear Model using SPCA
intMImpSPCAlm = lm(intMImp$deltaF ~ intMScores[, 1] + intMScores[, 2] +
  intMScores[, 3] + intMScores[, 4] + intMScores[, 5])
summary(intMImpSPCAlm)

plot(intMImp$deltaF, intMImpSPCAlm$fitted.values, xlim = c(0, 100), ylim = c(0, 100), xlab = "Measured delta F", ylab = "Predicted delta F", main = "Measured vs. Predicted")
abline(a = 0, b = 1)

RsqrdintMImpSPCA = summary(intMImpSPCAlm)$r.squared
RsqrdintMImpSPCA

adjRsqrdintMImpSPCA = summary(intMImpSPCAlm)$adj.r.squared
adjRsqrdintMImpSPCA

lambdaintMImpSPCA = mean(intMImp$deltaF/intMImpSPCAlm$fitted.values)
lambdaintMImpSPCA
# Finding unimportant variables

```r
impvars = function(data) {
  for (i in 1:dim(data)[2]) {
    if (i == 1) {
      est = NULL
      pvlue = NULL
      includ = NULL
    }
    cort = cor.test(data[, i], data$deltaF)
    est[i] = cort$estimate
    pvlue[i] = cort$p.value
    includ[i] = (cort$p.value < 0.05)
    if (i == dim(data)[2]) {
      tab = as.table(cbind(names(data), round(est, 3), signif(pvlue, 5), includ))
      colnames(tab) = c("Variable", "Correlation", "p-value", "Important")
      rownames(tab) = NULL
      intImpTab = tab
      print(kable(intImpTab, digits = 4))
    }
  }
}
```

`impvars(intData)`

# Same as Corr > 0.15

```r
intCorDeltaF[abs(intCorDeltaF) > 0.15]
```

# Removing unimportant variables

```r
names(intCorDeltaF[abs(intCorDeltaF) < 0.15]) # Unimportant variables
names(intCorDeltaF[abs(intCorDeltaF) >= 0.15]) # Important variables
```

```r
intImp = subset(intData, select = names(intCorDeltaF[abs(intCorDeltaF) >= 0.15]))
```

# PCA with lm on Correlation cut-off important internal variables.
intImpPC = prcomp(subset(intImp, select=-deltaF), center = TRUE, scale. = TRUE, retx = TRUE)
kable(intImpPC$rotation[, 1:5])
screepplot(intImpPC, npcs = length(intImpPC$rotation[, 1]), type = "l", main = "Scree plot for PC on internal Tendon Data")
summary(intImpPC)

# Building a linear model with PCA
intImplm = lm(intImp$deltaF ~ intImpPC$x[, 1] + intImpPC$x[, 2] + intImpPC$x[, 3] + intImpPC$x[, 4] + intImpPC$x[, 5])
summary(intImplm)

plot(intImp$deltaF, intImplm$fitted.values, xlim = c(0, 100), ylim = c(0, 100), xlab = "Measured delta F", ylab = "Predicted delta F", main = "Measured vs Predicted")
abline(a = 0, b = 1)

RsqrdintImp = summary(intImplm)$r.squared
RsqrdintImp

adjRsqrdintImp = summary(intImplm)$adj.r.squared
adjRsqrdintImp

lambdaintImp = mean(intImp$deltaF/intImplm$fitted.values)
lambdaintImp

# SPCA on internal important variables

# Deciding on how many non-zero loadings to keep per sparse PC.
intImpSPCA <- spca(subset(intImp, select=-deltaF), K = 5, type = "predictor", sparse = "varnum", para = c(rep(dim(subset(intImp, select=-deltaF))[2], 5)), use.corr = TRUE) # Same as normal PCA.
intImpSPCA

maxAdjR2SPCA(intImp)

intImpSPCA <- spca(subset(intImp, select=-deltaF), K = 5, type = "predictor", sparse = "varnum", para = c(rep(3, 5)), use.corr = TRUE)
intImpSPCA$loadings
intImpSPCA$pev
sum(intImpSPCA$pev)

# transforming data to zero mean and unit variance
intX = apply(subset(intImp, select=-deltaF), MARGIN=2, FUN=standardize)
intScores = intX %*% intImpSPCA$loadings # Matrix multiplication to rotate the
standardized data.

# Linear Model using SPCA
intImpSPCAlm = lm(intImp$deltaF ~ intScores[, 1] + intScores[, 2] + intScores[, 3] + intScores[, 4] + intScores[, 5])
summary(intImpSPCAlm)

plot(intImp$deltaF, intImpSPCAlm$fitted.values, xlim = c(0, 105), ylim = c(0, 105), xlab = "Measured delta F", ylab = "Predicted delta F", main = "Measured vs. Predicted")
abline(a = 0, b = 1)

RsqrdfintImpSPCA = summary(intImpSPCAlm)$r.squared
RsqrdfintImpSPCA

adjRsqrdfintImpSPCA = summary(intImpSPCAlm)$adj.r.squared
adjRsqrdfintImpSPCA

lambdadintImpSPCA = mean(intImp$deltaF/intImpSPCAlm$fitted.values)
lambdadintImpSPCA

# Splitting internal data into Continuous and Categorical subsets.
intCont = subset(intData, select=-c(loadingType, depthToComp.Stl., e_ps))
intCate = subset(intData, select=c(loadingType, depthToComp.Stl., e_ps, deltaF))

# PCA on Internal Continuous Variables
intContPC = prcomp(subset(intCont, select=-deltaF), center = TRUE, scale. = TRUE, retx = TRUE)
kable(intContPC$rotation[, 1:5])
screeplot(intContPC, npcs = length(intContPC$rotation[, 1]), type = "l", main = "Scree plot for PC on intCont Tendon Data")
summary(intContPC)

# PCA on Internal Categorical Variables
intCatePC = prcomp(subset(intCate, select=-deltaF), center = TRUE, scale. = TRUE, retx = TRUE)
kable(intCatePC$rotation[, 1:3])
screeplot(intCatePC, npcs = length(intCatePC$rotation[, 1]), type = "l", main = "Scree plot for PC on intCate Tendon Data")
summary(intCatePC)

# Combined Cont and Cate PCA
intContCatelm = lm(intCont$deltaF ~ intContPC$x[, 1] + intContPC$x[, 2] + intContPC$x[, 3] + intContPC$x[, 4] + intCatePC$x[, 1])
summary(intContCatelm)
plot(intCont$deltaF, intContCatelm$fitted.values, xlim = c(0, 105), ylim = c(0, 105), xlab = "Measured delta F", ylab = "Predicted delta F", main = "Measured vs. Predicted")
abline(a = 0, b = 1)
RsqrdintContCate = summary(intContCatelm)$r.squared
RsqrdintContCate
adjRsqrdintContCate = summary(intContCatelm)$adj.r.squared
adjRsqrdintContCate
lambdaIntContCate = mean(intData$deltaF/intContCatelm$fitted.values)
lambdaIntContCate

# SPCA on Internal Continuous Variables

# Deciding on how many non-zero loadings to keep per sparse PC.
intContSPCA <- spca(subset(intCont, select=-deltaF), K = 4, type = "predictor", sparse = "varnum", para = c(rep(dim(subset(intCont, select=-deltaF))[2], 4)), use.corr = TRUE) # Same as normal PCA.
intContSPCA
maxAdjR2ContSPCA = function(dataframe) {
  # dataframe is a data frame containing deltaF
  for (i in 1:dim(subset(dataframe, select=-deltaF))[2]) {
    dataframeSPCA <- spca(subset(dataframe, select=-deltaF), K = 4, type = "predictor", sparse = "varnum", para = c(rep(i, 4)), use.corr = TRUE)
    if (i == 1) {
      standardize = function(x) {(x - mean(x))/sd(x)}
      dataX = apply(subset(dataframe, select=-deltaF), MARGIN=2, FUN=standardize)
      dataScores = NULL
      rsqrd = NULL
      adjrsqrd = NULL
      print("Max i is:")
      print(dim(subset(dataframe, select=-deltaF))[2])
    }
    dataScores = dataX %*% dataframeSPCA$loadings
    dataframeSPCAlm = lm(dataframe$deltaF ~ dataScores[, 1] + dataScores[, 2] + dataScores[, 3] + dataScores[, 4])
    rsqrd = c(rsqrd, summary(dataframeSPCAlm)$r.squared)
    adjrsqrd = c(adjrsqrd, summary(dataframeSPCAlm)$adj.r.squared)
    print(i)
  }
  print("Number of non-zero loadings")
  m1 = cbind(1:dim(subset(dataframe, select=-deltaF))[2], rsqrd, adjrsqrd)
  print(m1)
  if (which.max(m1[, 2]) == which.max(m1[, 3])) {
    print("Number of non-zero loadings suggested:"
    print(which.max(m1[, 3]))
  } else {
    print("Number of non-zero loadings suggested:"
    print(which.max(m1[, 2]))
    print("or")
    print(which.max(m1[, 3]))
  }
  print("Max R^2")
  print(max(rsqrd))
  print("Max Adj R^2")
}
print(max(adjrsqrd))
plot(1:dim(subset(dataframe, select=-deltaF))[2], adjrsqrd, xlab = "Non-zero Loadings for each SPC", ylab = "Adj R^2", type = "b")
abline(v = which.max(m1[, 3]), col = "red")
legend(x = "right", legend = c("Max Adj. R^2"), cex = 0.8, lty=c(1), lwd = c(1.5), col = c("red"), inset = 0.02)
}
maxAdjR2ContSPCA(intCont)

intContSPCA <- spca(subset(intCont, select=-deltaF), K = 4, type = "predictor", sparse = "varnum", para = c(rep(1, 4)), use.corr = TRUE)
intContSPCA$loadings
intContSPCA$pev
sum(intContSPCA$pev)

# transforming data to zero mean and unit variance
intContX = apply(subset(intCont, select=-deltaF), MARGIN=2, FUN=standardize)
intContScores = intContX %*% intContSPCA$loadings # Matrix multiplication to rotate the standardized data.

# SPCA on Internal Categorical Variables
intCateSPCA <- spca(subset(intCate, select=-deltaF), K = 3, type = "predictor", sparse = "varnum", para = c(rep(dim(subset(intCate, select=-deltaF))[2], 3)), use.corr = TRUE) # Same as normal PCA.
intCateSPCA

maxAdjR2CateSPCA = function(dataframe) {
# dataframe is a data frame containing deltaF
for (i in 1:dim(subset(dataframe, select=-deltaF))[2]) {
  dataframeSPCA <- spca(subset(dataframe, select=-deltaF), K = 1, type = "predictor", sparse = "varnum", para = c(rep(i, 1)), use.corr = TRUE)
  if (i == 1) {
    standardize = function(x) {(x - mean(x))/sd(x)}
    dataX = apply(subset(dataframe, select=-deltaF), MARGIN=2, FUN=standardize)
    dataScores = NULL
  }
}
rsqrd = NULL
adjrsqrd = NULL
print("Max i is:")
print(dim(subset(dataframe, select=-deltaF))[2])
}
dataScores = dataX %*% dataframeSPCA$loadings
dataframeSPCAlm = lm(dataframe$deltaF ~ dataScores[, 1])
rsqrd = c(rsqrd, summary(dataframeSPCAlm)$r.squared)
adjrsqrd = c(adjrsqrd, summary(dataframeSPCAlm)$adj.r.squared)
print(i)
}
print("Number of non-zero loadings")
m1 = cbind(1:dim(subset(dataframe, select=-deltaF))[2], rsqrd, adjrsqrd)
print(m1)
if (which.max(m1[, 2]) == which.max(m1[, 3])) {
  print("Number of non-zero loadings suggested:")
  print(which.max(m1[, 3]))
} else {
  print("Number of non-zero loadings suggested:")
  print(which.max(m1[, 2]))
  print("or")
  print(which.max(m1[, 3]))
}
print("Max R^2")
print(max(rsqrd))
print("Max Adj R^2")
print(max(adjrsqrd))
plot(1:dim(subset(dataframe, select=-deltaF))[2], adjrsqrd, xlab = "Non-zero Loadings for each SPC", ylab = "Adj R^2", type = "b")
abline(v = which.max(m1[, 3]), col = "red")
legend(x = "right", legend = c("Max Adj. R^2"), cex = 0.8, lty=c(1), lwd = c(1.5), col = c("red"), inset = 0.02)

maxAdjR2CateSPCA(intCate)

intCateSPCA <- spca(subset(intCate, select=-deltaF), K = 1, type = "predictor", sparse = "varnum", para = c(rep(2, 1)), use.corr = TRUE)
intCateSPCA$loadings
intCateSPCA$pev
sum(intCateSPCA$pev)

# transforming data to zero mean and unit variance
intCateX = apply(subset(intCate, select=-deltaF), MARGIN=2, FUN=standardize)
intCateScores = intCateX %*% intCateSPCA$loadings  # Matrix multiplication to rotate the standardized data.

# Linear Model using SPCA on Internal Continuous and Categorical Variables
intContCateSPCAlm = lm(intData$deltaF ~ intContScores[, 1] + intContScores[, 2] + intContScores[, 3] + intContScores[, 4] + intCateScores[, 1])
summary(intContCateSPCAlm)

plot(intData$deltaF, intContCateSPCAlm$fitted.values, xlim = c(0, 105), ylim = c(0, 105), xlab = "Measured delta F", ylab = "Predicted delta F", main = "Measured vs. Predicted")
abline(a = 0, b = 1)

RsqrdintContCateSPCA = summary(intContCateSPCAlm)$r.squared
RsqrdintContCateSPCA

adjRsqrdintContCateSPCA = summary(intContCateSPCAlm)$adj.r.squared
adjRsqrdintContCateSPCA

lambdaintContCateSPCA = mean(intData$deltaF/intContCateSPCAlm$fitted.values)
lambdaintContCateSPCA

# External unbonded tendon data analysis
extData = tenData[tenData$tendonType == "2", -1]
row.names(extData) = 1:dim(extData)[1]

# Checking for correlated variables using covariance and correlation tables.
cor(subset(extData, select=-deltaF))

# Many pairs of variables have colinearity, and data is "highly" dimensional.
# Principal components using prcomp function.

```r
extPC = prcomp(subset(extData, select=-deltaF), center = TRUE, scale. = TRUE, retx = TRUE)
kable(extPC$rotation[, 1:5])
screeplot(extPC, npcs = length(extPC$rotation[, 1]), type = "l", main = "Scree Plot for PCA on External Tendon Data")
summary(extPC)
```

# Building a linear model with PCA

```r
extlm = lm(extData$deltaF ~ extPC$x[, 1] + extPC$x[, 2] + extPC$x[, 3] + extPC$x[, 4] + extPC$x[, 5])
summary(extlm)
```

```r
plot(extData$deltaF, extlm$fitted.values, xlim = c(0, 200), ylim = c(0, 200),
     xlab = "Measured Delta Fps", ylab = "Predicted Delta Fps", main = "Measured Delta Fps vs. Predicted Delta Fps\non External Tendon Data", asp = 1)
abline(a = 0, b = 1)
```

```
RsqrdsExt = summary(extlm)$r.squared
RsqrdsExt

adjRsqrdsExt = summary(extlm)$adj.r.squared
adjRsqrdsExt

lambdaExt = mean(extData$deltaF/extlm$fitted.values)
lambdaExt
```

# Using self selected important variables

```r
extMImp = subset(extData, select=c(spanLength, beamDepth, areaOfPS, concreteStrength, initialPrestress, areaOfTensStl., areaOfComp.Stl., deltaF ))
```

# PCA with lm on important external variables.

```r
extMImpPC = prcomp(subset(extMImp, select=-deltaF), center = TRUE, scale. = TRUE, retx = TRUE)
kable(extMImpPC$rotation[, 1:5])
screeplot(extMImpPC, npcs = length(extMImpPC$rotation[, 1]), type = "l", main = "Scree plot for PC on external Tendon Data")
```
# Building a linear model with PCA

extMImplm = lm(extMImp$deltaF ~ extMImpPC$x[, 1] + extMImpPC$x[, 2] + extMImpPC$x[, 3] + extMImpPC$x[, 4] + extMImpPC$x[, 5])
summary(extMImplm)

plot(extMImp$deltaF, extMImplm$fitted.values, xlim = c(0, 200), ylim = c(0, 200), xlab = "Measured delta f", ylab = "Predicted delta f", main = "Measured vs. Predicted")
abline(a = 0, b = 1)

RsqrdextMImp = summary(extMImplm)$r.squared
RsqrdextMImp
adjRsqrdextMImp = summary(extMImplm)$adj.r.squared
adjRsqrdextMImp

lambdaextMImp = mean(extMImp$deltaF/extMImplm$fitted.values)
lambdaextMImp

# SPCA on self selected external important variables

extMImpSPCA <- spca(subset(extMImp, select=-deltaF), K = 5, type = "predictor", sparse = "varnum", para = c(rep(dim(subset(extMImp, select=-deltaF))[[2]], 5)), use.corr = TRUE)
extMImpSPCA

maxAdjR2SPCA(extMImp)

extMImpSPCA <- spca(subset(extMImp, select=-deltaF), K = 5, type = "predictor", sparse = "varnum", para = c(rep(1, 5)), use.corr = TRUE)
extMImpSPCA$loadings
extMImpSPCA$pev
sum(extMImpSPCA$pev)

# transforming data to zero mean and unit variance
extMX = apply(subset(extMImp, select=-deltaF), MARGIN=2, FUN=standardize)
extMScores = extM %*% extMImpSPCA$loadings # Matrix multiplication to rotate
the standardized data.

# Linear Model using SPCA
extMImpSPCA lm = lm(extMImp$deltaF ~ extMScores[, 1] + extMScores[, 2] +
extMScores[, 3] + extMScores[, 4] + extMScores[, 5])
summary(extMImpSPCA lm)

plot(extMImp$deltaF, extMImpSPCA lm$fitted.values, xlim = c(0, 200), ylim = c(0,
200), xlab = "Measured delta f", ylab = "Predicted delta f", main = "
Measured vs. Predicted")
abline(a = 0, b = 1)

RsqrdextMImpSPCA = summary(extMImpSPCA lm)$r.squared
RsqrdextMImpSPCA

adjRsqrdextMImpSPCA = summary(extMImpSPCA lm)$adj.r.squared
adjRsqrdextMImpSPCA

lambdaextMImpSPCA = mean(extMImp$deltaF/extMImpSPCA lm$fitted.values)
lambdaextMImpSPCA

# Finding correlation cut-off important variables
extCorDeltaF = cor(extData[, 16]

# p-value > 0.05
impvars(extData)

# Same as Corr < 0.23
extCorDeltaF[abs(extCorDeltaF) < 0.23]

# Removing unimportant variables
names(extCorDeltaF[abs(extCorDeltaF) < 0.23])
names(extCorDeltaF[abs(extCorDeltaF) >= 0.23]) # Important variables

extImp = subset(extData, select = names(extCorDeltaF[abs(extCorDeltaF) >=
0.23]))
# PCA with lm on important external variables.

```r
extImpPC = prcomp(subset(extImp, select=-deltaF), center = TRUE, scale. = TRUE, retx = TRUE)
kable(extImpPC$rotation[, 1:5])
screeplot(extImpPC,npcs = length(extImpPC$rotation[, 1]), type = "l", main = "Scree plot for PC on external Tendon Data")
summary(extImpPC)
```

# Building a linear model with PCA

```r
ext Implm = lm(extImp$deltaF ~ extImpPC$x[, 1] + extImpPC$x[, 2] + extImpPC$x[, 3] + extImpPC$x[, 4] + extImpPC$x[, 5])
summary(extImplm)
```

```r
plot(extImp$deltaF, extImplm$fitted.values, xlim = c(0, 200), ylim = c(0, 200),
     xlab = "Measured delta f", ylab = "Predicted delta f", main = "Measured vs Predicted")
abline(a = 0, b = 1)
```

```r
RsqrdextImp = summary(extImplm)$r.squared
RsqrdextImp
```

```r
adjRsqrdextImp = summary(extImplm)$adj.r.squared
adjRsqrdextImp
```

```r
lambdaextImp = mean(extImp$deltaF/extImplm$fitted.values)
lambdaextImp
```

# SPCA on external important variables

```r
extImpSPCA <- spca(subset(extImp, select=-deltaF), K = 5, type = "predictor",
                    sparse = "varnum", para = c(rep(dim(subset(extImp, select=-deltaF))[2], 5))
                    , use.corr = TRUE)
extImpSPCA
```

```r
maxAdjR2SPCA(extImp)
```

```r
extImpSPCA <- spca(subset(extImp, select=-deltaF), K = 5, type = "predictor",
                    sparse = "varnum", para = c(rep(2, 5)), use.corr = TRUE)
```
extImpSPCA$loadings
extImpSPCA$pev
sum(extImpSPCA$pev)

# transforming data to zero mean and unit variance
extX = apply(subset(extImp, select=-deltaF), MARGIN=2, FUN=standardize)
extScores = extX %*% extImpSPCA$loadings # Matrix multiplication to rotate the
standardized data.

# Linear Model using SPCA
extImpSPCAlm = lm(extImp$deltaF ~ extScores[, 1] + extScores[, 2] + extScores[,
3] + extScores[, 4] + extScores[, 5])
summary(extImpSPCAlm)

plot(extImp$deltaF, extImpSPCAlm$fitted.values, xlim = c(0, 200), ylim = c(0,
200), xlab = "Measured delta f", ylab = "Predicted delta f", main = "
Measured vs. Predicted")
abline(a = 0, b = 1)

RsqrdextImpSPCA = summary(extImpSPCAlm)$r.squared
RsqrdextImpSPCA
adjRsqrdextImpSPCA = summary(extImpSPCAlm)$adj.r.squared
adjRsqrdextImpSPCA

lambdaextImpSPCA = mean(extImp$deltaF/extImpSPCAlm$fitted.values)
lambdaextImpSPCA

# Splitting external data into Continuous and Categorical subsets.
extCont = subset(extData, select=-c(loadingType, depthToComp.Stl., e_ps))
extCate = subset(extData, select=c(loadingType, depthToComp.Stl., e_ps, deltaF)
)

# PCA on External Continuous Variables
extContPC = prcomp(subset(extCont, select=-deltaF), center = TRUE, scale. =
TRUE, retx = TRUE)
kable(extContPC$rotation[, 1:4])
screeplot(extContPC, npcs = length(extContPC$rotation[, 1]), type = "l", main = "Scree plot for PC on extCont Tendon Data")

summary(extContPC)

# PCA on External Categorical Variables

extCatePC = prcomp(subset(extCate, select=-deltaF), center = TRUE, scale. = TRUE, retx = TRUE)
kable(extCatePC$rotation[, 1:3])

screeplot(extCatePC, npcs = length(extCatePC$rotation[, 1]), type = "l", main = "Scree plot for PC on extCate Tendon Data")
summary(extCatePC)

# PCA on Cont and Cate combined

extContCatelm = lm(extCont$deltaF ~ extContPC$x[, 1] + extContPC$x[, 2] + extContPC$x[, 3] + extContPC$x[, 4] + extCatePC$x[, 1])

summary(extContCatelm)

plot(extCont$deltaF, extContCatelm$fitted.values, xlim = c(0, 200), ylim = c(0, 200), xlab = "Measured delta f", ylab = "Predicted delta f", main = "Measured vs. Predicted")
abline(a = 0, b = 1)

RsqrdextContCate = summary(extContCatelm)$r.squared
RsqrdextContCate

adjRsqrdextContCate = summary(extContCatelm)$adj.r.squared
adjRsqrdextContCate

lambdaextContCate = mean(extData$deltaF/extContCatelm$fitted.values)
lambdaextContCate

# SPCA on External Continuous Variables

# Deciding on how many non-zero loadings to keep per sparse PC.

extContSPCA <- spca(subset(extCont, select=-deltaF), K = 4, type = "predictor",
            sparse = "varnum", para = c(rep(dim(subset(extCont, select=-deltaF))[2], 4)), use.corr = TRUE)

extContSPCA$loadings
extContSPCA$pev
sum(extContSPCA$pev)

maxAdjR2ContSPCA(extCont)

extContSPCA <- spca(subset(extCont, select=-deltaF), K = 4, type = "predictor",
             sparse = "varnum", para = c(rep(5, 4)), use.corr = TRUE)

# transforming data to zero mean and unit variance
extContX = apply(subset(extCont, select=-deltaF), MARGIN=2, FUN=standardize)
extContScores = extContX %*% extContSPCA$loadings # Matrix multiplication to
# rotate the standardized data.

# SPCA on External Categorical Variables
extCateSPCA <- spca(subset(extCate, select=-deltaF), K = 1, type = "predictor",
             sparse = "varnum", para = c(rep(dim(subset(extCate, select=-deltaF))[2],
             1)), use.corr = TRUE) # Same as normal PCA.
extCateSPCA

maxAdjR2CateSPCA(extCate)

extCateSPCA <- spca(subset(extCate, select=-deltaF), K = 1, type = "predictor",
             sparse = "varnum", para = c(rep(1, 1)), use.corr = TRUE)
extCateSPCA$loadings
extCateSPCA$pev
sum(extCateSPCA$pev)

# transforming data to zero mean and unit variance
extCateX = apply(subset(extCate, select=-deltaF), MARGIN=2, FUN=standardize)
extCateScores = extCateX %*% extCateSPCA$loadings # Matrix multiplication to
# rotate the standardized data.

# SPCA on External Continuous and Categorical Variables
extContCateSPCAlm = lm(extData$deltaF ~ extContScores[, 1] + extContScores[, 2]
+ extContScores[, 3] + extContScores[, 4] + extCateScores[, 1])
summary(extContCateSPCAlm)
plot(extData$deltaF, extContCateSPCA$predicted.values, xlim = c(0, 200), ylim = c(0, 200), xlab = "Measured delta f", ylab = "Predicted delta f", main = "Measured vs. Predicted")
abline(a = 0, b = 1)

RsqrdContCateSPCA = summary(extContCateSPCA)$r.squared
RsqrdContCateSPCA

adjRsqrdContCateSPCA = summary(extContCateSPCA)$adj.r.squared
adjRsqrdContCateSPCA

lambdaContCateSPCA = mean(extData$deltaF/extContCateSPCA$predicted.values)
lambdaContCateSPCA

# Analysis Results Table
RsqrdTabl = matrix(c(RsqrdInt, adjRsqrdInt, RsqrdExt, adjRsqrdExt,
                     RsqrdIntContCate, adjRsqrdIntContCate, RsqrdExtContCate,
                     adjRsqrdExtContCate,
                     RsqrdIntContCateSPCA, adjRsqrdIntContCateSPCA, RsqrdExtContCateSPCA,
                     adjRsqrdExtContCateSPCA,
                     RsqrdIntMImp, adjRsqrdIntMImp, RsqrdExtMImp,
                     adjRsqrdExtMImp,
                     RsqrdIntMImpSPCA, adjRsqrdIntMImpSPCA, RsqrdExtMImpSPCA,
                     adjRsqrdExtMImpSPCA,
                     RsqrdIntImp, adjRsqrdIntImp, RsqrdExtImp, adjRsqrdExtImp,
                     RsqrdIntImpSPCA, adjRsqrdIntImpSPCA, RsqrdExtImpSPCA,
                     adjRsqrdExtImpSPCA),
                   ncol = 4, byrow = TRUE)
RsqrdTabl = signif(RsqrdTabl, 4)
RsqrdTabl = cbind(c("All", "Continuous and Categorical", "", "Self-Selected Important Variables", "", "Correlation Cutoff Important Variables", ""), c ("PCA", "PCA", "SPCA", "PCA", "SPCA", "PCA", "SPCA"), RsqrdTabl)
RsqrdTabl <- as.data.frame(RsqrdTabl)
colnames(RsqrdTabl) <- c("Variables", "Method", "Internal R^2", "Adj R^2", "External R^2", "Adj R^2")
kable(RsqrdTabl, align = c("l", "l", rep("c", 4)), caption = "$R^2$ & Adjusted $R^2$ Values")
$$\lambda$$ Values