Regularization and Dimension Reduction on Rocky Mountain River Data

Cason Wight

Brigham Young University

Abstract. In many analyses, excess response variables make it difficult to explain the underlying patterns of a response. Overfit can be a problem for data like these, if a linear model fit is even possible to mathematically obtain to begin with. This report applies two data dimension/regularization methods on river flow data. These data include several covariates for river flow such as temperatures, precipitation levels, and many others. To best predict and make inference on these variables, an elastic net and a lasso regression on principal components were used. These two methods give information on which factors are most impactful on flow. The included factors of these methods will be assessed for fit and prediction ability.

1 River Flow Factors

Rivers are a vital aspect to a geography's infrastructure. For wildlife, plants, and farmers, the flow of rivers can have a large effect on their lives. Irrigation or habitats could be ruined if water is not flowing as it should. In the Rocky Mountains, the factors that influence the flow of rivers are not entirely understood. Data collected on these rivers include many potential factors that could impact the flow of rivers in the Rocky Mountains.

These data include 97 measurements such as monthly temperature, precipitation, human population density, and others. In these data, each of the 102 observations also has a unitless metric on river flow. The goal of this report is determine which of these environmental factors have the most impact on riverflow and see how well they fit and predict on the data. Figure 1 shows scatterplots of a few of the explanatory variables against each other and with the riverflow metric.

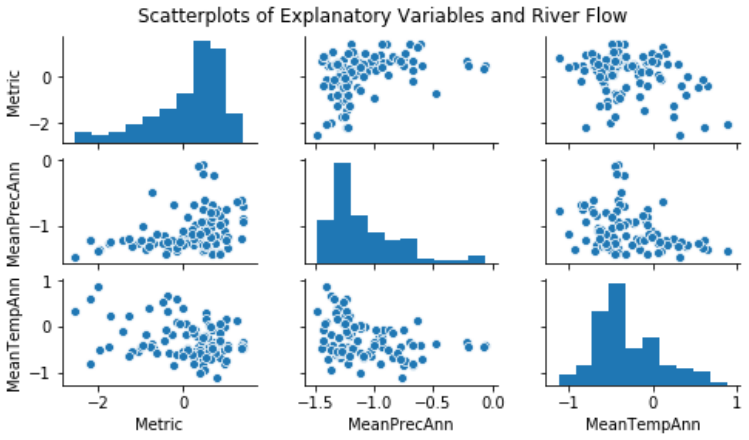

Fig. 1: Explanatory Variables and River Flow

While summary statistics on all of the variables are not realistic, the following are a few notable summary statistics. River flow has a mean of 0.125 and a standard deviation of 0.876. The mean temperature for each month ranges between $-.590°C$ in Decmeber and $0.583°C$ in July. The average mean human population densities (per kilometer) range from $-.134$ in 2000 to $-.131$ in 2015. These are only a few of the variables,

but it is clear that the data have been transformed or standardized in some way (mean density cannot be negative, for example). Thus all interpretations in this report are on the standardized/transformed variables.

One problem with having so may variables is that it is unlikely that each of the 97 variables have a significant impact on river flow. Including all of these variables as separate effects could lead to highly variable estimated effects due to overfit. This means that adding or taking away a single data point could have too large an impact on the estimated effects, which is not desirable. Also, including variable that do not really impact river flow could lead to bad prediction. The best statistical models have little variance from new data points and make accurate predictions, as measured by mean squared error (MSE).

The following are the goals of this report: This report will reduce the 97 possible effects to a more maneagable number of effects through an elastic net and through a lasso on a principal component analysis (PCA). This reduction will give insight to the biggest climate/river network/human factors that affect overall river flow, looking specifically into the five most impactful factors from each method. These factors will be analyzed to see how well they explain overall river flow. Using these estimated effects, the prediction ability will be assessed. A brief comparison will discuss which method, an elastic net or a lasso regression on the principal components of the explanatory variables, is better in this specific example.

2 Elastic Nets and Principal Components Analysis

Multiple methods exist for selecting important variables. Lasso regression and ridge regression take a least squares approach and give an additional penalty proportional to the size of the estimated effects. An elastic net also takes an ordinary least squares and adds a penalty, but the penalty is a weighted average of the two penalties given by lasso and ridge regression.

PCA reduces the vectors of the explanatory variables down to a given number of vectors that represent the variability in different directions. These vectors can then be used in standard regression models to predict the response. PCA is a more challenging concept that will be explained subsequently; the main idea is that the bulk of the variability in the many inputs is captured by just a few representative vectors. This report will now define lasso regression, ridge regression, elastic nets, and PCA in more detail.

As stated above, lasso regression is a least squares approach to effect estimation, with an additional penalty term for the size of the estimates. A lasso regression will have a penalized least squares model with an additional term for the sum of the absolute size of the effects to get the estimates $\hat{\boldsymbol{\beta}}$:

$$\min_{\hat{\boldsymbol{\beta}}} \sum_{i=1}^{n} (y_i - \mathbf{x}_i \hat{\boldsymbol{\beta}})^2 + \lambda \sum_{p=1}^{P} |\hat{\beta}_p|$$

In the above equation, y_i is a single response from observation i and $\mathbf{x}_i$ is a single row of explanatory variables from observation i. $\hat{\boldsymbol{\beta}}$ is a vector of the additive effects of each element of the $\mathbf{x}_i$ row vector on the response y_i. In the penalty term, λ is a specific penalty coefficient and $\hat{\beta}_p$ is the estimated additive effect of the pth element of each $\mathbf{x}_i$ row on each response y_i.

Lasso gets the $\hat{\boldsymbol{\beta}}$ that minimizes the squared residuals under the constraint that the sum of the absolute size of the effects is equal to some constant $s(\lambda)$. The idea here is that if the sum of the total absolute effects is limited to a certain number, only those effects that are most important for predicting the response will be included and others will get an estimated effect of 0. The optimal constraint value, $s(\lambda)$, is selected by picking the value λ that minimizes MSE after k-means cross-validation at many different values of λ.

Ridge regression is extremely similar, except that the penalty takes the sum of the squared estimated effects, instead of the sum of the absolute effects. λ is selected by minimizing MSE, as in lasso regression. The estimation $\hat{\boldsymbol{\beta}}$ in ridge regression is calculated as follows:

$$\min_{\hat{\boldsymbol{\beta}}} \sum_{i=1}^{n} (y_i - \mathbf{x}_i \hat{\boldsymbol{\beta}})^2 + \lambda \sum_{p=1}^{P} \hat{\beta}_p^2$$

An elastic net has the same approach of minimizing the squared residuals, with an additional penalty term, but simply takes a weighted average of the penalties of lasso and ridge regression, giving the equation

$$\min_{\hat{\boldsymbol{\beta}}} \sum_{i=1}^{n} (y_i - \mathbf{x}_i\hat{\boldsymbol{\beta}})^2 + \lambda \sum_{p=1}^{P} \Big[\alpha|\hat{\beta}_p| + (1-\alpha)\hat{\beta}_p^2\Big],$$

where α is a weight between 0 and 1 for the lasso penalty. In elastic nets, both values α and λ are determined through gridsearch, minimizing the MSE using k-means cross-validation. All of these techniques require standardizing the $\mathbf{X}$ matrix before fitting so that the impact of a variable is proportional to its own variability.

PCA is unlike lasso, ridge, or elastic net; it is simply a method of dimension reduction for the $\mathbf{X}$ matrix, which is a $n \times p$ matrix including all the n observation rows $\mathbf{x}_i$. $\boldsymbol{\Psi}$ is a $p \times m$ transformation matrix on $\mathbf{X}$, converting it to a $n \times m$ matrix $\mathbf{Z}$ that contains the variability of $\mathbf{X}$ in the m most variable orthogonal directions such that $\mathbf{Z} = \mathbf{X}\boldsymbol{\Psi}$. The transformation matrix $\boldsymbol{\Psi}$ is calculated by getting the first m eigenvectors of the variance-covariance matrix of $\mathbf{X}$. m is the number of eigenvalues that sum to nearly the total sum of all of the eigenvalues. In this report, m is the number of eigenvalues that make up 95% of the sum of all of the eigenvalues. The newly defined $\mathbf{Z}$ matrix can be used in regression methods to estimate the effects (defined by the vector $\hat{\boldsymbol{\theta}}$) of each element of the $\mathbf{z}_i$ row vectors. Thus regression will be fitting the model,

$$\mathbf{y} = \beta_0 + \mathbf{Z}\boldsymbol{\theta} + \boldsymbol{\epsilon},$$

where $\mathbf{y}$ is a vector of responses and $\boldsymbol{\epsilon}$ is a vector of errors, which is typically assumed to be distributed as a multivariate normal distribution with a mean of $\mathbf{0}$ and variance-covariance matrix $\sigma^2\mathbf{I}$. From the $\hat{\theta}_m$s, the $\hat{\beta}_p$s can be approximated by the transformation $\hat{\boldsymbol{\beta}} \approx \boldsymbol{\Psi}\hat{\boldsymbol{\theta}}$ because $y = \beta_0 + \mathbf{Z}\boldsymbol{\theta} = \beta_0 + \mathbf{X}(\boldsymbol{\Psi}\boldsymbol{\theta})$. This report will analyze the river data through an elastic net and through a lasso regression on the PCA of $\mathbf{X}$.

These two methods have various strengths and weaknesses to them. Elastic nets, ridge regression models, and lasso regression models are easy to interpret. Similar to a simple linear regression model, elastic nets (along with ridge or lasso models) explain the additive effects of explanatory variables on a response variable. Another benefit to elastic nets, ridge regression, and lasso regression is that they naturally avoid unimportant variables, making them useful in avoiding overfit. Lasso regression is especially good at "zeroing" out unimportant variables one at a time. Lasso performs better when most variables actually have no effect. Ridge is preferable with lots of small coefficients present. An elastic net balances the two methods. A weakness of these models is that they still require a linear relationship between the explanatory variables and the response. Another factor of these tools is that no distributional assumptions come with them. This could be a benefit, because the error need not be normally distributed, making it appropriate for more sets of data. This could also be a weakness because inference like confidence intervals cannot be calculated without a distribution, as in a Gaussian model. These tools assume linearity and independence.

PCA also has several strengths and a few weaknesses. One strength is that an $\mathbf{X}$ with too many columns can be reduced to only a few, without losing much of the original information from the $\mathbf{X}$ matrix. The estimated linear effects of the PCA can be easily transformed to approximately fit the original $\mathbf{X}$ matrix, which is another benefit. PCA is not a modeling method itself, but linear models and others can still be applied to a PCA. One downside to PCA is that it is slightly more difficult to understand and there is some variability lost in the transformation, although the amount of lost variability can be predetermined. Another weakness is that the $\mathbf{X}$ matrix is transformed based on its own variability and not based on its predictability of the response. Thus if $\mathbf{X}$ is highly variable in many "directions" that have no impact on the response, they are still not too helpful as predictors on the response. A linear model fit on PCA still assumes linearity, independence, normality, and equal variance as usual in a linear model, but on the $\mathbf{Z}$ matrix instead of the $\mathbf{X}$ matrix.

3 Justification

In elastic nets, none of the explanatory variables are directly "excluded", but some coefficients may turn out to be 0, which has the same effect as excluding it. In the elastic net that was fit on these data, 64 of the 97 original variables resulted in an estimated effect on response of 0. The estimated effects of each factor will be discussed later. The 64 variables with coefficients of 0 were not relatively important in the model, so the available $s(\lambda)$ coefficient allotment was given to other variables. Based on cross validation on MSE, the optimal value of λ and α are .1559 and .1177, respectively.

In the lasso regression on PCA approach, none of the variables are outright excluded, because each of them are transformed into the reduced $\mathbf{Z}$ matrix. However, not all directions of variability were included. The first 19 eigenvectors accounted for just over 95% of the original variability, as shown in figure 2. Thus the dimensions of $\mathbf{Z}$ matrix were 102×19. From the PCA, most simple linear regression (SLR) effects were statistically insignificant ($> .05$), which is why lasso regression on the PCA is preferable to SLR on the PCA.

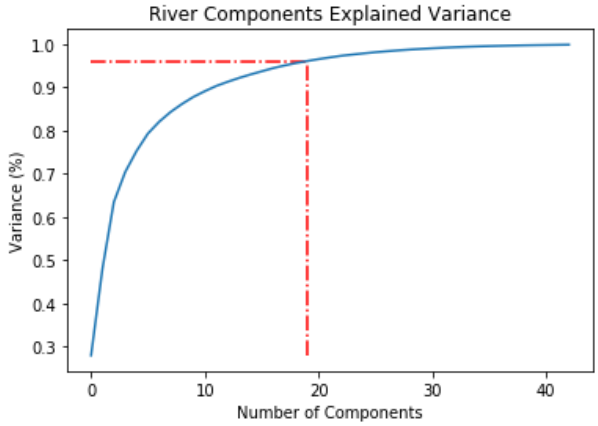

Fig. 2: Number of Components in PCA

Lasso did estimate effects of 0 for some of the eigenvectors. This means that some of the directions of variability in $\mathbf{X}$ were relatively unimportant in predicting river flow. Of the 19 eigenvectors, 3 had estimated effects of 0. The effects of these "directions of variability" themselves are not interpretable, but backtransforming these estimates back to the effects of the original variables can help us gain inference on their linear effects on river flow. The estimated effects of the original variables are discussed in the results. Through the lasso on the PCA, two of the original variables ended up with estimated coefficients of exactly 0. These two were the average percent of soil that had very poor drainage, and average percent of soil that was well drained. The latter was also one of the excluded variables in the elastic net approach. The optimal λ of the lasso was .0160.

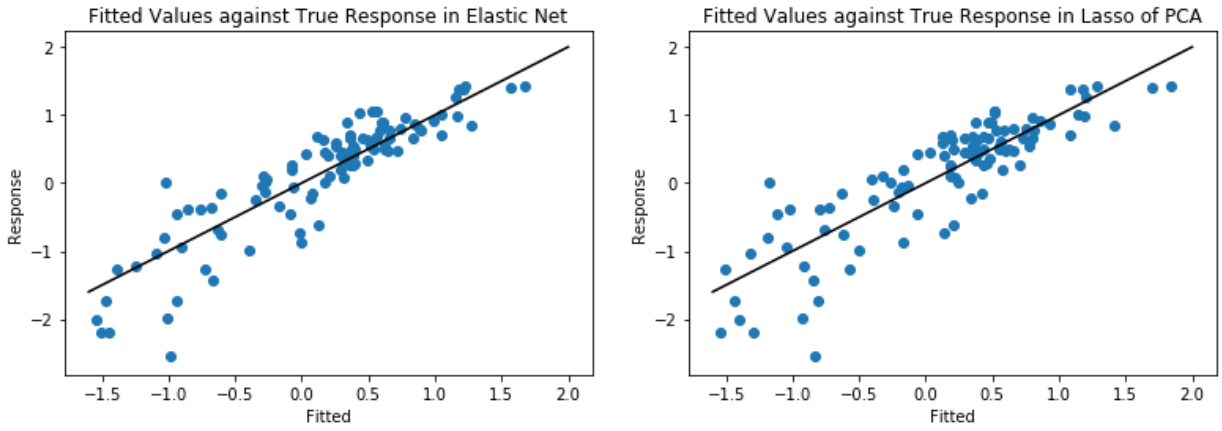

Fig. 3: Evidence for Linearity Assumption

The assumptions of these two methods are extremely similar. For both, the assumptions of normality and equal variance do not apply because distributional assumptions are not included in lasso regression or

in elastic nets. Thus only independence and linearity need to be confirmed. Independence is more difficult to check. Spatial correlation is clearly a factor here, but is beyond the scope of this report. A fitted values vs actual response plot may show violations of independence and/or linearity. For these two assumptions, this scatterplot should have a linear slope. Any curvature would be an indication that something problematic is violating our model assumptions. The fitted values vs actual response plots for the two methods are shown in figure 3.

From figure 3, there is no concerning curvature in either plot. With the linearity assumption confirmed, the methods need no other adjustments. An analysis of fit and predictive ability reveals that the two methods are extremely similar, but an elastic net fits the data better and makes superior predictions. A discussion of the fit and prediction performance of these two methods is included in the following section.

4 Performance and Results

These two approaches came to different conclusions for which variables have the largest impact. The effects of the variables in the elastic net method ranged from −.13 to .17 for each covariate. 64 of the variables had an estimated effect of 0. The variables with the top five absolute effects on response are shown in table 1, along with their 95% bootstrapped credible intervals. The other 28 non-zero variables and their estimates are not reported. As an example of interpretation, as global stream order (number of total branches merging to current point) increases by 1, the river flow is estimated to increase by .1678 (95% CI: .0528 to .2240).

Table 1: Elastic Net, Top 5 Effects on River Flow

Variable	Model Coefficient	95% CI Lower	95% CI Upper
Global Stream Order	.1678	.0528	.2240
Evergreen Needle Trees (%)	.1402	.0349	.1921
Longitude	−.1284	−.1817	−.0076
Somewhat Excessive Drainage (%)	.1184	.0289	.2056
Evergreen Broadleaf (%)	.1176	.0363	.1861

After converting the estimated effects of the PCA to the estimated effects of the original variables, only two of the variables had an estimated effect of 0. Of the remaining 95 non-zero effects, estimates ranged from about −.08 to .08. The original variables with the top five absolute effects on response are shown in table 2, along with their 95% bootstrapped credible intervals. The most important variable from the elastic net approach, global stream order, is the fifth largest effect in this approach. As global stream order (number of total branches merging to current point) increases by 1, the river flow is estimated to increase by .0654 (95% CI: .0397 to .1043). It makes sense that this estimate is smaller, because this method includes many more and smaller effects of the variables (because PCA is a transformation on all of the variables instead of a form of variable selection).

Table 2: Lasso on PCA, Top 5 Effects on River Flow

Variable	Model Coefficient	95% CI Lower	95% CI Upper
Stream Order (Threshold of 100)	.0832	.0501	.1245
Mean Monthly Temperature Range	−.0801	−.1208	−.0429
Stream Drop (Threshold of 100)	−.0712	−.1680	.0042
Poor Excessive Drainage (%)	−.0661	−.1336	−.0040
Global Stream Order	.0654	.0397	.1043

As stated, the elastic net fit the data much better and made more accurate predictions than the lasso on PCA approach. The R^2 was .8080 for the elastic net method and .7522 for the lasso on PCA method. This

means that of the variability in the data, these two methods can account for 75-81% of it, which is decent fit. There is still some variability of river flow which has not been explained by these two appraches.

The standard deviation of the original response values (river flow metrics) was 0.8759. Through 6-means cross validation, the RMSE for the elastic net approach was found to be 0.4962 and for the lasso on PCA was .5744. Performing SLR on the PCA gave an RMSE of .6871, which shows the predictive benefit of using lasso, instead of SLR, on the PCA. This means that without the model, the average deviation of the river flow was roughly .3015-.3800 higher than the average deviation of the predicted values from the true river flow using these two methods.

5 Conclusion

The goals of this report are summarized by the following:

1. Reduce the 97 possible effects to a more maneagable number using an elastic net and a lasso on PCA
2. Identify the variables with the largest impact on river flow based on the two methods
3. Interpret these effects and assess fit/prediction performance

This report has fulfilled the three goals by using an elastic net and using lasso regression on a PCA. The elastic net reduced the 97 original variables to a subset of 33 effects. The PCA reduced the dimensions of the variables from 97 variables to 19 directions of variability, 3 of which had estimated effects of 0 through lasso. Elastic net identified global stream order, percent of evergreen needle trees, longitude, percent of land with excessive drainage, and percent evergreen broadleaf as the five most important factors. The lasso on PCA identified stream order (with a threshold of 100), mean monthly temperature range, stream drop (with a threshold of 100), percent of land with excessive drainage, and global stream order as the five most important factors. Using a PCA meant that the estimated effects are spread across more variables, so they are generally smaller than those few effects identified by the elastic net.

The original standard deviation in river flow was 0.8759, as compared to the RMSE for elastic net of 0.4962 and for lasso on PCA of .5744. The R^2 of these two approaches were .8080 and .7522, respectively. This means that for this example, elastic net is preferable for both model fit and mean squared error of predictions.

The two methods of this report also come with some shortcomings. An obvious shortcoming is that neither approach accounts for spatial correlation, which would intuitively have a big impact on the results. Another shortcoming of these methods is that they have an increased computational cost over a SLR model. Because the optimal weight, α, and the optimal penalty term, λ, have to be solved by grid search, these techniques are much slower.

Next steps for this analysis would first include some approach to fitting the spatial correlation. Another step could also be to dive deeper into each of the estimated effects and look at the possibility of non-linear relationships or interaction effects, which are beyond the scope of this report.