Predictive Modeling

Department of Mathematical Sciences
Qiuying Sha, Professor

Michigan Tech

Chapter 6. Linear Regression and Its Cousins

We will discuss several models

- Ordinary linear regression
- Partial Least squares (PLS)
- Penalized models such as ridge regression, lasso, and elastic net.

Each of them can directly or indirectly be written in the form

$$y_i = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_P x_{iP} + \varepsilon_i$$

Advantages of these models:

- Highly interpretable
- Can compute standard errors of the coefficients, can assess the statistical significance of each predictor in the model.

Limitations:

- Models are appropriate when the relationship between the predictors and response is linear.
  - If there is a curvilinear relationship between the predictors and response (e.g., such as quadratic, cubic, or interactions among predictors), then linear regression models can be augmented with additional predictors.
  - However, nonlinear relationships between predictors and the response may not be adequately captured with these models.

6.1 Case Study: Quantitative Structure-Activity Relationship Modeling

Chemicals, including drugs, can be represented by **chemical formulas**.

Quantitative measurements can be derived, such as the molecular weight, electrical charge, or surface area. These are called ***chemical descriptors***.

Some characteristics of molecules **cannot be analytically determined** from the chemical structure. For example, **biological activity** of a compound.

The relationship between **the chemical structure** and **its activity** can be complex. As such, the relationship is **usually determined empirically using experiments**.

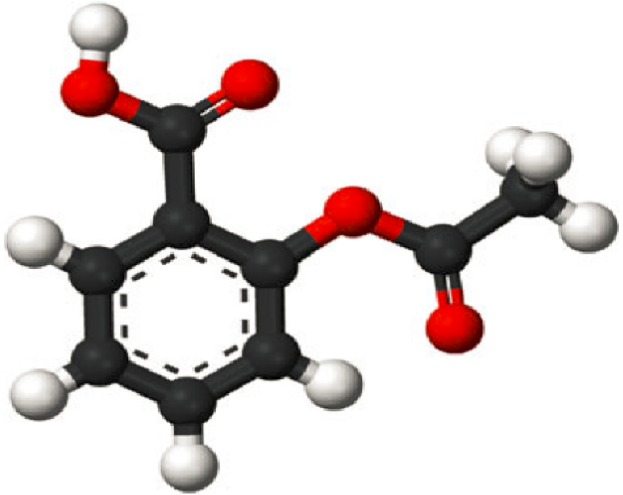

Fig. 6.1: A representation of aspirin, which contains carbon atoms (shown as *black balls*) and hydrogen (*white*) and oxygen atoms (*red*). The chemical formula for this molecule is `O=C(Oc1ccccc1C(=O)O)C`, from which molecular descriptors can be determined, such as a molecular weight of 180.2 g/mol

Physical qualities, such as the **solubility**, are evaluated as well as other properties, such as toxicity.

- A compound's solubility is very important if it is to be given orally or by injection.

We will demonstrate various regression modeling techniques by **predicting solubility using chemical structures.**

Tetko et al. (2001) and Huuskonen (2000) investigated a set of compounds with corresponding experimental solubility values using complex sets of descriptors.

Data:

- There are **1,267 compounds** and a set of more understandable **descriptors** that fall into one of three groups:
  - 208 binary “fingerprints” that indicate the presence or absence of a particular chemical substructure.
  - 16 count descriptors, such as the number of bonds or the number of bromine atoms.
  - 4 continuous descriptors, such as molecular weight or surface area.

- There are many pairs that show **strong positive correlations**; 47 pairs have correlations greater than 0.90.
- The count-based descriptors show a significant **right skewness.**
- The outcome data were measured on the log10 scale and ranged from −11.6 to 1.6 with an average log solubility value of −2.7.

- The data were split using random sampling into a training set (*n* = 951) and test set (*n* = 316).
  - ➢ The training set will be used to tune and estimate models, as well as to determine initial estimates of performance using repeated 10-fold cross-validation.
  - ➢ The test set will be used for a final characterization of the models of interest.

It is useful to explore the training set to understand the characteristics of the data prior to modeling.

Evaluate the continuous predictors for **skewness:**

- The average skewness statistic was 1.6 (with a minimum of 0.7 and a maximum of 3.8).
- To correct for this skewness, a Box–Cox transformation was applied to all predictors.

Figure 6.3 shows scatter plots of the predictors against the outcome along with a regression line.

- This figure indicates that there are some linear relationships between the predictors and the outcome (e.g., molecular weight) and some nonlinear relationships (e.g., the number of origins or chlorines).
- Because of this, we might consider augmenting the predictor set with quadratic terms for some variables.

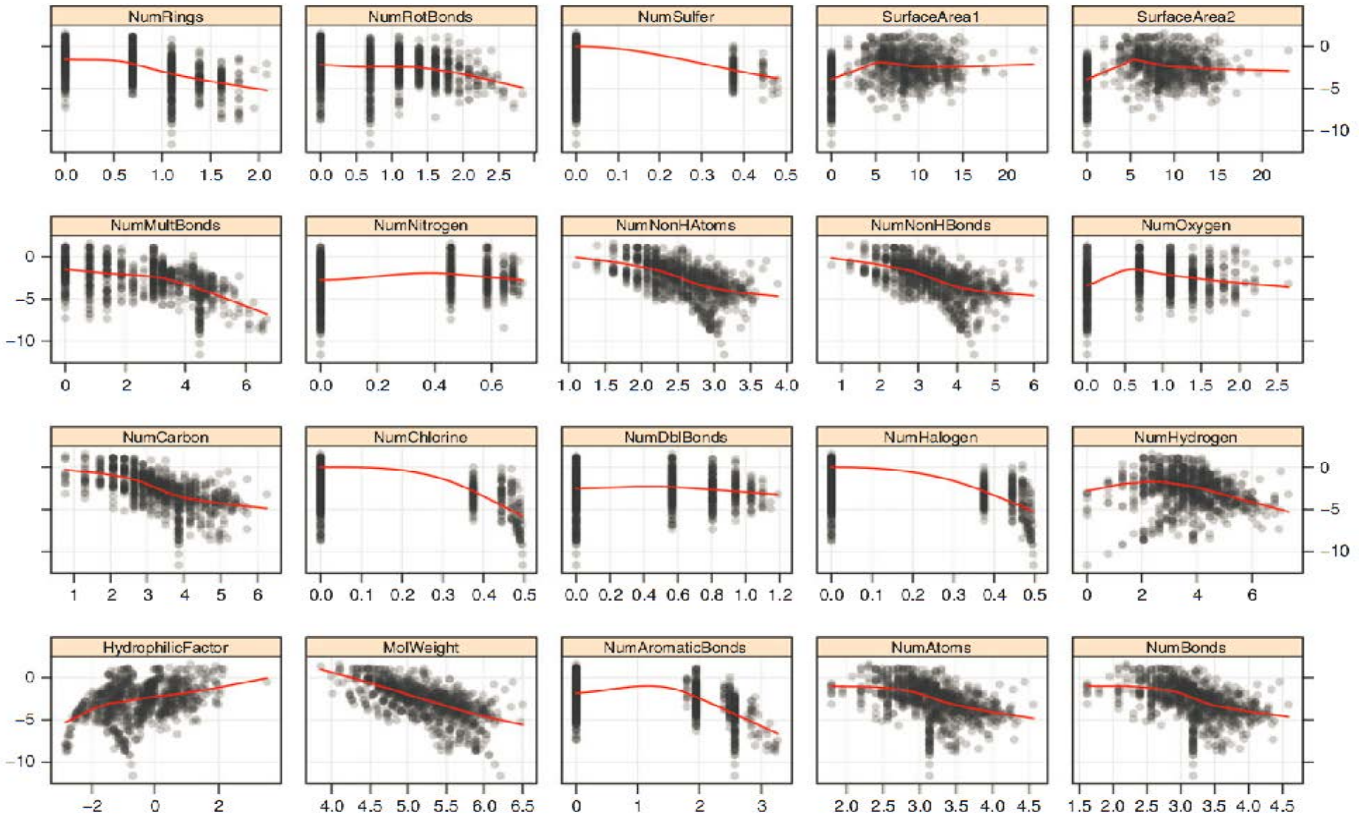

Fig. 6.3: Scatter plots of the transformed continuous predictors in the solubility data set. The *red line* is a scatt smoother

Are there significant **between-predictor correlations**?

- Principal component analysis (PCA) was used on the full set of transformed predictors.
- Figure 6.4 is a scree plot and displays the variability accounted for by each component.

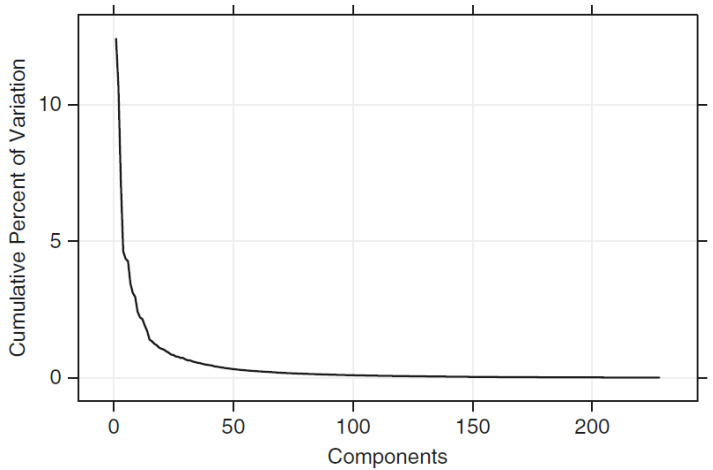

Fig. 6.4: A scree plot from a PCA analysis of the solubility predictors

- The amount of variability summarized by component drops sharply, with no one component accounting for more than 13% of the variance.

- This indicates that the structure of the data is contained in a much smaller number of dimensions than the number of dimensions of the original space; this is often due to a large number of collinearities among the predictors

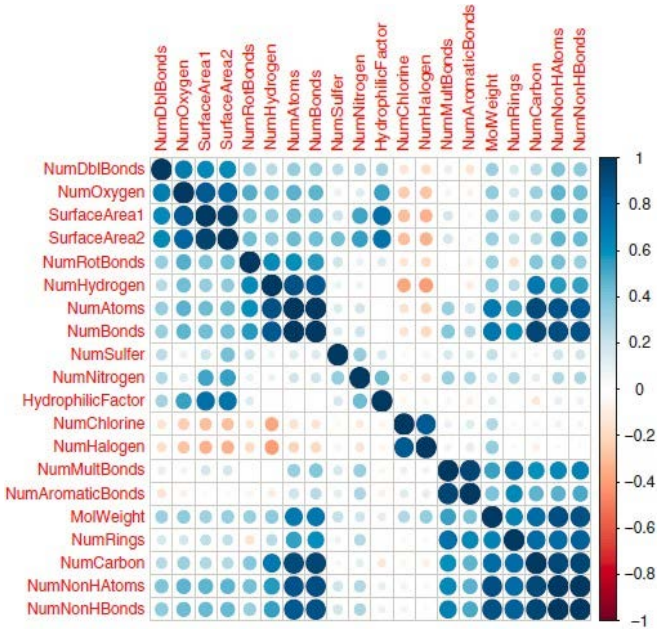

Fig. 6.5: Between-predictor correlations of the transformed continuous solubility predictors

Figure 6.5 shows the correlation structure of the transformed continuous predictors;

- There are many strong positive correlations.
- This could create problems in developing some models. Pre-processing steps will need to be taken to account for this problem.

6.2 Linear Regression

Ordinary least squares linear regression:

- Estimate parameters that minimize the sum-of-squared errors (SSE):

$$\mathrm{SSE} = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2,$$

- The parameter estimates

$$\hat{\beta} = (X^T X)^{-1} X^T y$$

  - The parameter estimates have **the least bias** of all possible parameter estimates.
  - These estimates minimize the bias component of the bias-variance trade-off.

- **A unique inverse** of this matrix exists when

  (1) No predictor can be determined from a combination of one or more of the other predictors **and**

  (2) The number of samples is greater than the number of predictors.

  - If the data fall under either of these conditions, then a unique set of regression coefficients does not exist.
  - However, a unique set of predicted values can still be obtained for data that fall under condition (1) by either replacing with a **conditional inverse** (Graybill 1976) or by **removing predictors** that are collinear (supplement).

See supplemental materials for conditional inverse.

- When condition (2) is true for a data set, the practitioner can take several steps to attempt to build a regression model
  - ✓ Using pre-processing techniques presented in Sect. 3.3 to remove pairwise correlated predictors.
  - ✓ After pre-processing the data, if the number of predictors is still greater than the number of observations, then we will need to take other measures to reduce the dimension of the predictor space.
    - PCA pre-processing (Sect. 3.3)
    - PLS or
    - Employing methods that shrink parameter estimates such as ridge regression, the lasso, or the elastic net.

- Another drawback of multiple linear regression is that its solution is linear in the parameters.

- A third notable problem with multiple linear regression is that it is prone to chasing observations that are away from the overall trend of the majority of the data (minimize SSE).
- To illustrate the problem of correlated predictors, linear models were fit with combinations of descriptors related to the number of non-hydrogen atoms and the number of hydrogen bonds.

  ➢ In the training set, these predictors are highly correlated (correlation: 0.994).

Table 6.1: Regression coefficients for two highly correlated predictors across four separate models

Model	NumNonHAtoms	NumNonHBonds
NumNonHAtoms only	−1.2 (0.1)	
NumNonHBonds only		−1.2 (0.1)
Both	−0.3 (0.5)	−0.9 (0.5)
All predictors	8.2 (1.4)	−9.1 (1.6)

- Standard errors are shown in parentheses.
- This reflects the instability in the regression linear caused by the between - predictor relationships.

- In practice, such highly correlated predictors might be managed manually by removing one of the offending predictors.
- If the number of predictors is large, this may be difficult. Also, on many occasions, relationships among predictors can be complex and involve many predictors. In these cases, manual removal of specific predictors may not be possible and models that can tolerate collinearity may be more useful.

Linear Regression for Solubility Data

Procedures to analyze this data:

- Split the solubility data into training and test sets
- Applied a Box–Cox transformation to the continuous predictors.
- Remove predictors with pairwise correlations greater than 0.9.

  ➢ 38 predictors were identified and removed.

- Fit a linear model to the training data.
  - The linear model was resampled using 10-fold cross-validation.
  - The estimated root mean squared error (RMSE) was 0.71.
  - R^2 value was 0.88.

- Use this model to predict the value in the testing set.
  - The R^2 value between the observed and predicted values was 0.87.
  - The regression diagnostic plots are displayed in Fig. 6.7.

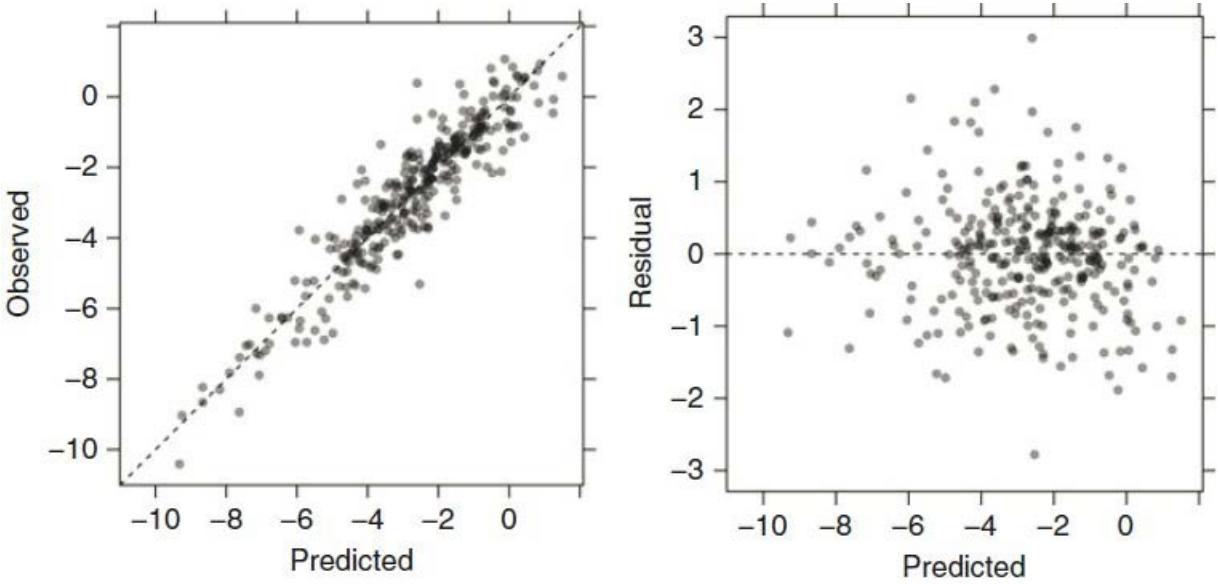

Fig. 6.7: *Left*: Observed versus predicted values for the solubility test set. *Right*: Residuals versus the predicted values. The residuals appear to be randomly scattered about 0 with respect to the predicted values

6.3 Partial Least Squares (PLS)

- If the correlation among predictors is high, then the ordinary least squares solution for multiple linear regression will have high variability and will become unstable.

- If the number of predictors is greater than the number of observations, ordinary least squares will be unable to find a unique set of regression coefficients.

- Pre-processing predictors via PCA prior to performing regression is known as principal component regression (PCR).
  - PCA does not consider any aspects of the response when it selects its components.

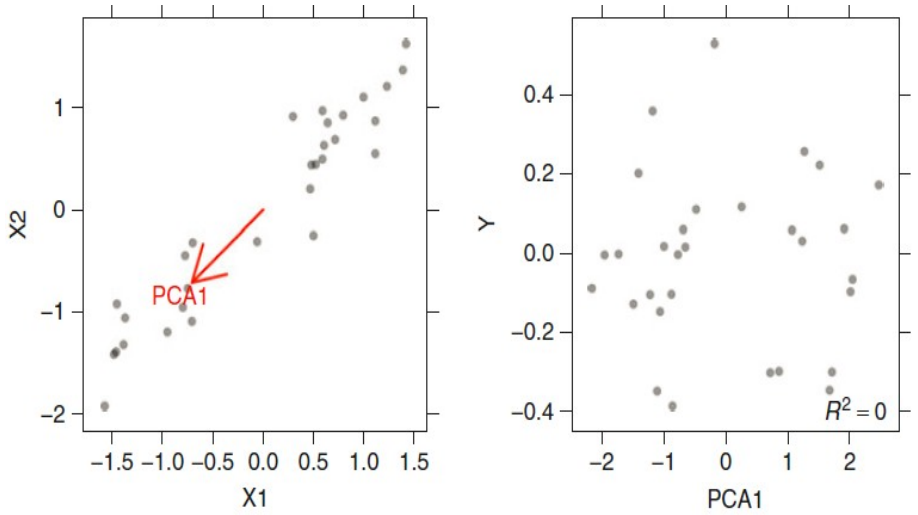

Fig. 6.8: An example of principal component regression for a simple data set with two predictors and one response. *Left*: A scatter plot of the two predictors shows the direction of the first principal component. *Right*: The first PCA direction contains no predictive information for the response

- Authors recommend using PLS when there are correlated predictors and a linear regression-type solution is desired (supplement, PLS-1).

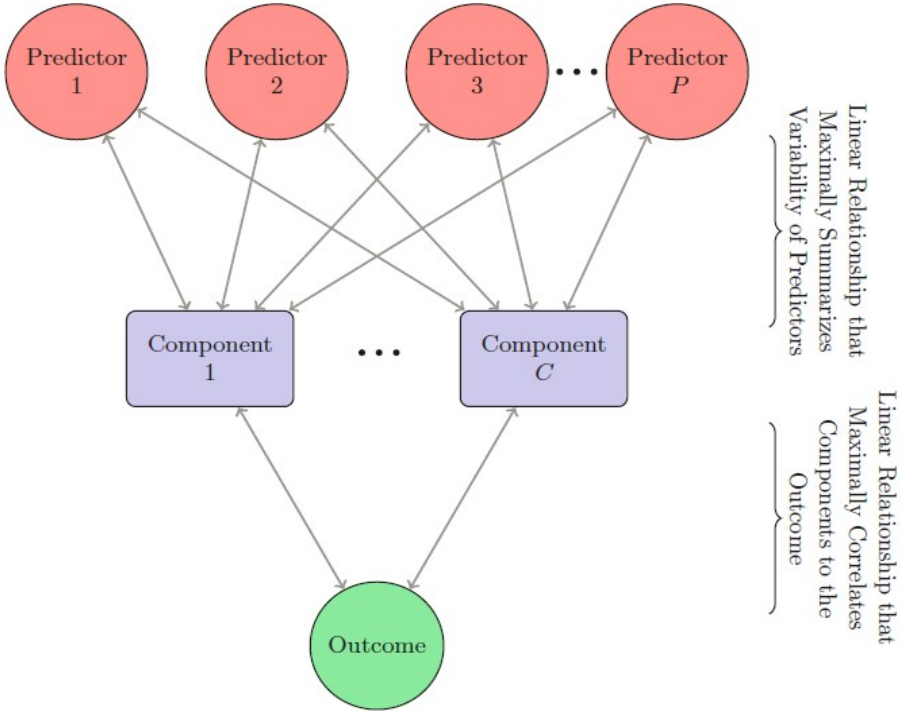

Fig. 6.9: A diagram depicting the structure of a PLS model. PLS finds components that simultaneously summarize variation of the predictors while being optimally correlated with the outcome

- PCA and PLS:
  - Both find linear combinations of the predictors. These linear combinations are commonly called *components* or latent variables.
  - The PCA linear combinations are chosen to maximally summarize predictor space variability.
  - PLS finds components that **maximally summarize the variation of the predictors** while simultaneously requiring these components to have **maximum correlation with the response.**
  - PLS can be viewed as a *supervised* dimension reduction procedure; PCR is an *unsupervised* procedure.
- Prior to performing PLS, the predictors should be **centered and scaled**.
- PLS has **one tuning parameter**: the number of components to retain.
  - Resampling techniques can be used to determine the optimal number of components

PCR and PLSR for Solubility Data

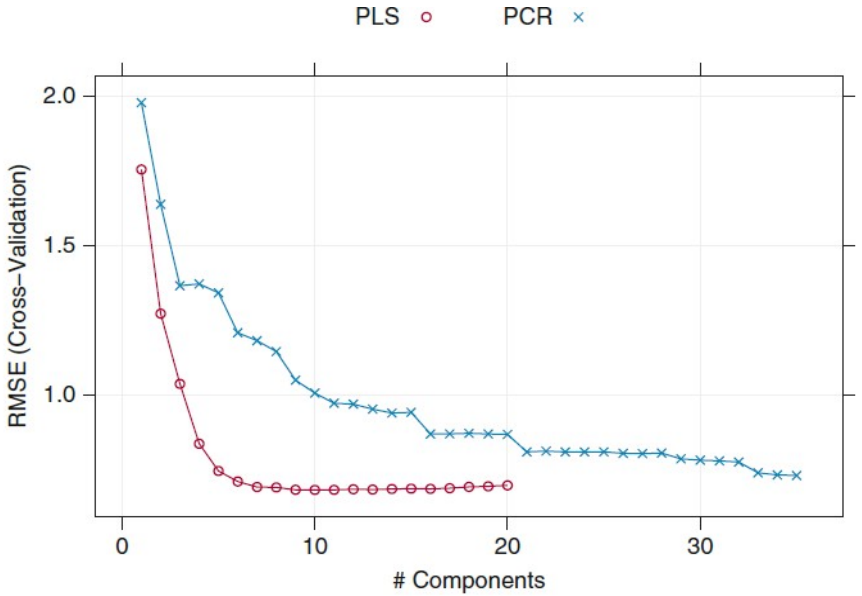

Fig. 6.11: Cross-validated RMSE by component for PLS and PCR. RMSE is minimized with ten PLS components and 35 PCR components

- There are 228 predictors.
- Many predictors are highly correlated and that the overall information within the predictor space is contained in a smaller number of dimensions.
- These predictor conditions are very favorable for applying PLS.

Michigan Tech

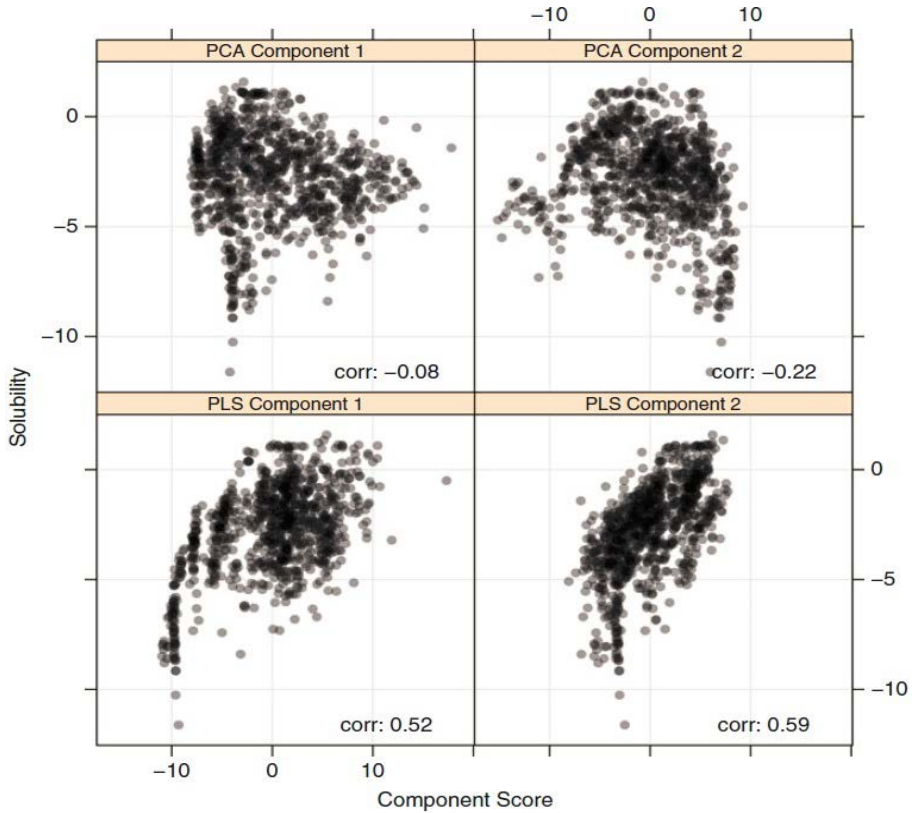

Fig. 6.12: A contrast of the relationship between each of the first two PCR and PLS components with the solubility response. Because the dimension reduction offered by PLS is supervised by the response, it is more quickly steered towards the underlying relationship between the predictors and the response

Figure 6.11: the supervised dimension reduction finds a minimum RMSE with significantly fewer components than unsupervised dimension reduction.

Michigan Tech

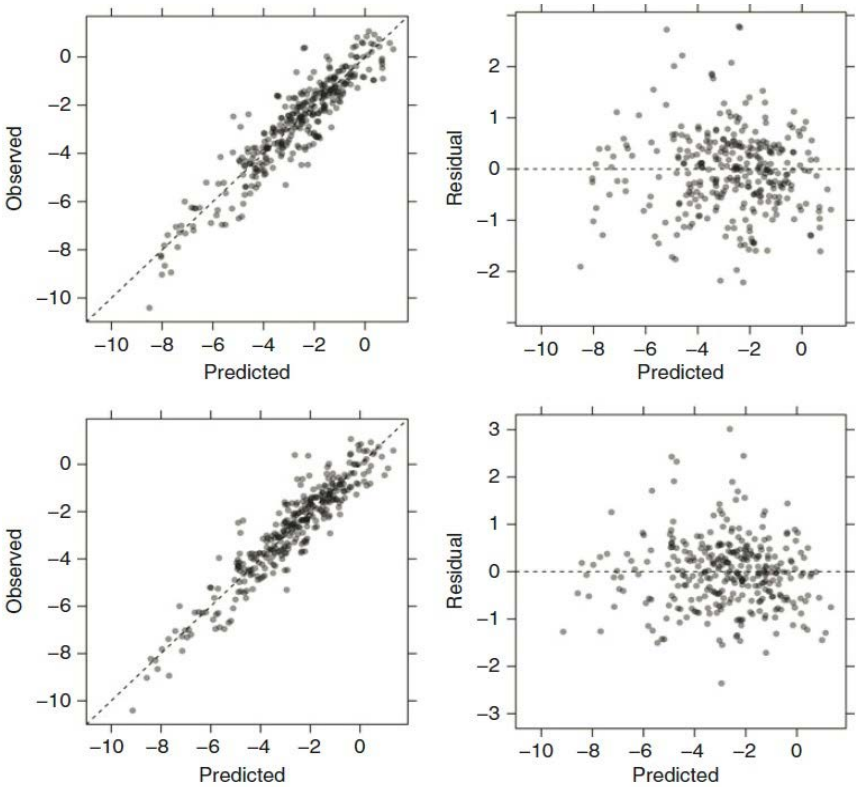

Fig. 6.13: *Left side*: Observed versus predicted values for the solubility test set for PCR (*upper*) and PLS (*lower*). *Right side*: Residuals versus the predicted values for PCR and PLS. The residuals appear to be randomly scattered about 0 with respect to the predicted values. Both methods have similar predictive ability, but PLS does so with far fewer components

Figure 6.13: prediction of the test set using the optimal PCR and PLS models can be seen.

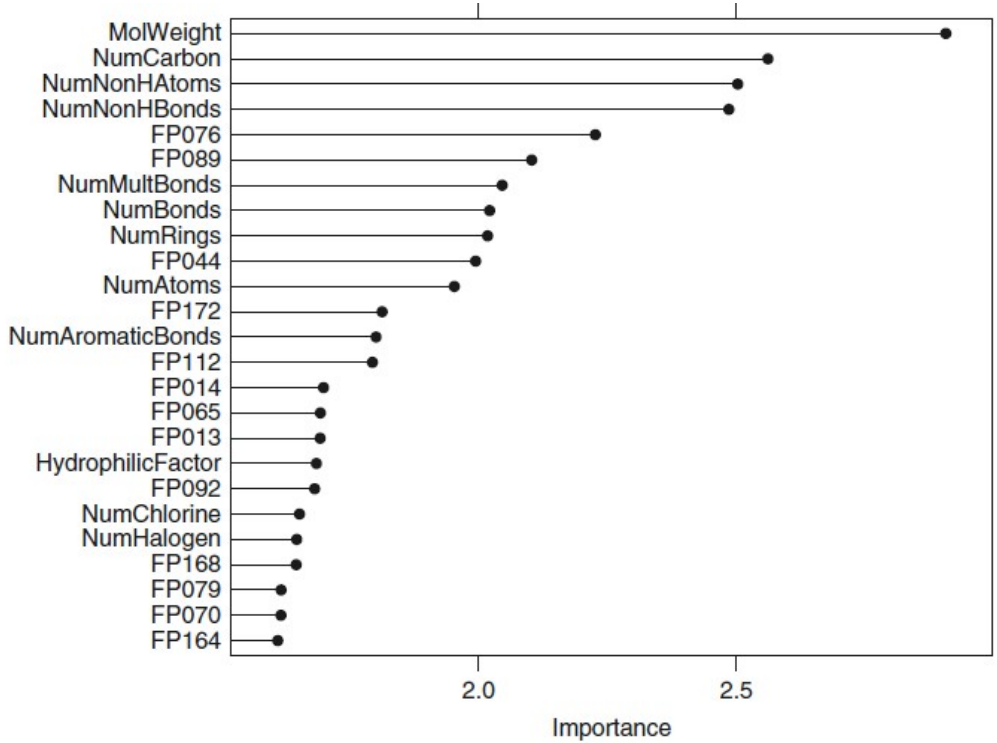

Fig. 6.14: Partial least squares variable importance scores for the solubility data

For the solubility data, the top 25 most important predictors are shown in Fig. 6.14.

A rule-of-thumb: VIP values exceeding 1 are considered to contain predictive information for the response.

6.4 Penalized Models

- The coefficients produced by ordinary least squares regression are **unbiased** and, of all unbiased linear techniques, this model also has the **lowest variance**.

- The MSE is a combination of variance and bias. It is very possible to produce models with **smaller MSEs** by allowing the parameter estimates to be biased.

- One method of creating biased regression models is to **add a penalty to SSE**.

Ridge regression (Hoerl 1970)

- Add a second-order penalty to SSE

$$SSE_{L_2} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{P} \beta_j^2$$

$$= \sum_{i=1}^{n}(y_i - \sum_{j=1}^{P} x_{ij}\beta_j)^2 + \lambda \sum_{j=1}^{P} \beta_j^2$$

which is equivalent to minimization of $\sum_{i=1}^{n}(y_i - \sum_{j=1}^{P} x_{ij}\beta_j)^2$ subject to, for some $c > 0$, $\sum_{j=1}^{P} \beta_j^2 < c$, i.e. constraining the sum of the squared coefficients.

- *Ridge regression shrinks* the estimates towards 0 as the *λ* becomes large (these techniques are sometimes called "shrinkage methods").

Fig. 6.15 shows the *path* of the regression coefficients for the solubility data over different values of λ.

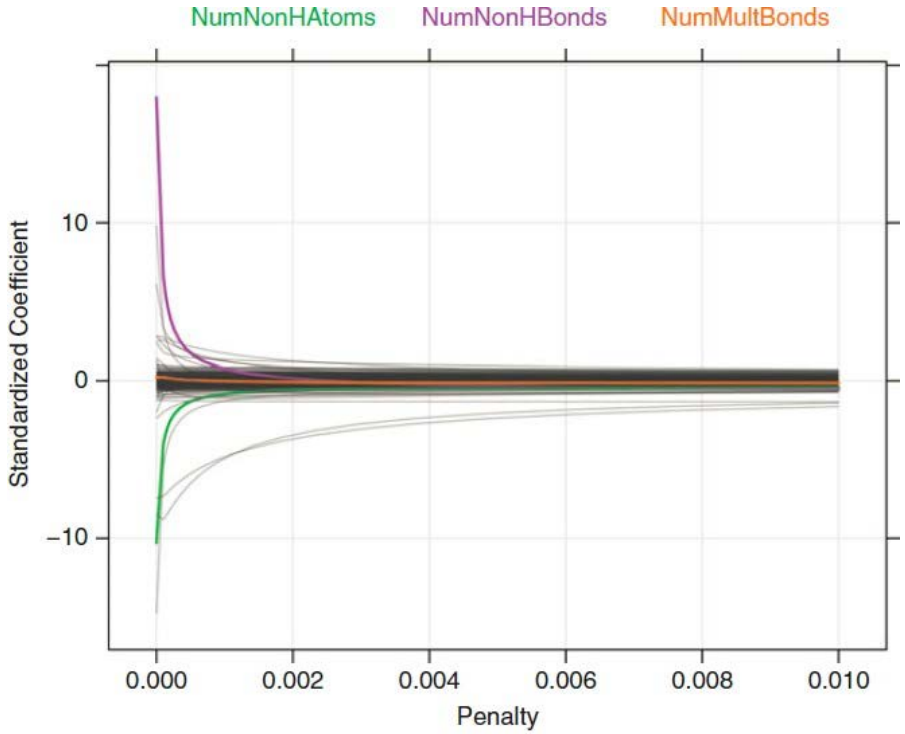

Fig. 6.15: The ridge-regression coefficient path

Parameter estimates for the number of non-hydrogen atoms (in green) and the number of non-hydrogen bonds (purple) are abnormally large.

These large values are indicative of collinearity issues.

As the penalty is increased, the parameter estimates move closer to 0 at different rates.

Using cross-validation, the penalty value was optimized.

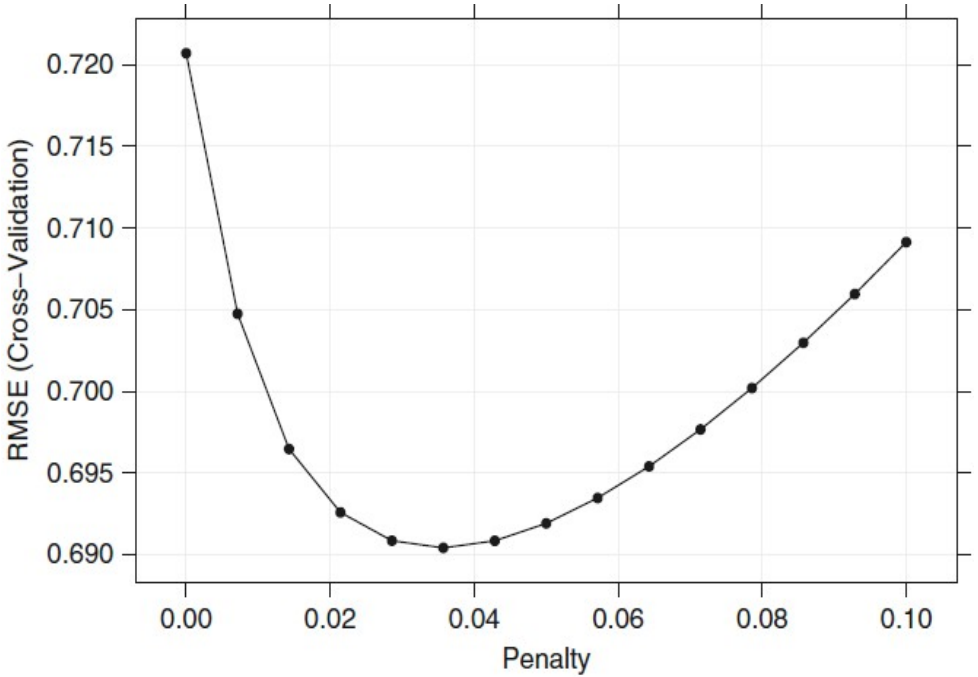

Fig. 6.16: The cross-validation profiles for a ridge regression model

Fig. 6.16 shows how the RMSE changes with λ

- Ridge regression shrinks the parameter estimates towards 0
- Even though some parameter estimates become negligibly small, this model does not conduct *feature selection*.

***Lasso: Least Absolute Shrinkage and Selection Operator* model (Tibshirani 1996)**

- Add a first-order penalty to SSE

$$\text{SSE}_{L_1} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{P} |\beta_j|.$$

- The lasso yields models that simultaneously use regularization to improve the model and to conduct **feature selection**.

Friedman et al. (2010) compared the two types of penalties, stated

- Ridge regression is **known to shrink the coefficients of correlated predictors towards each other**, allowing them to **borrow strength from each other**.

- In the extreme case of *k* identical predictors, they each get identical coefficients with 1/*k*th the size that any single one would get if fit alone.

- Lasso is somewhat indifferent to very correlated predictors, and will tend to **pick one and ignore the rest.**

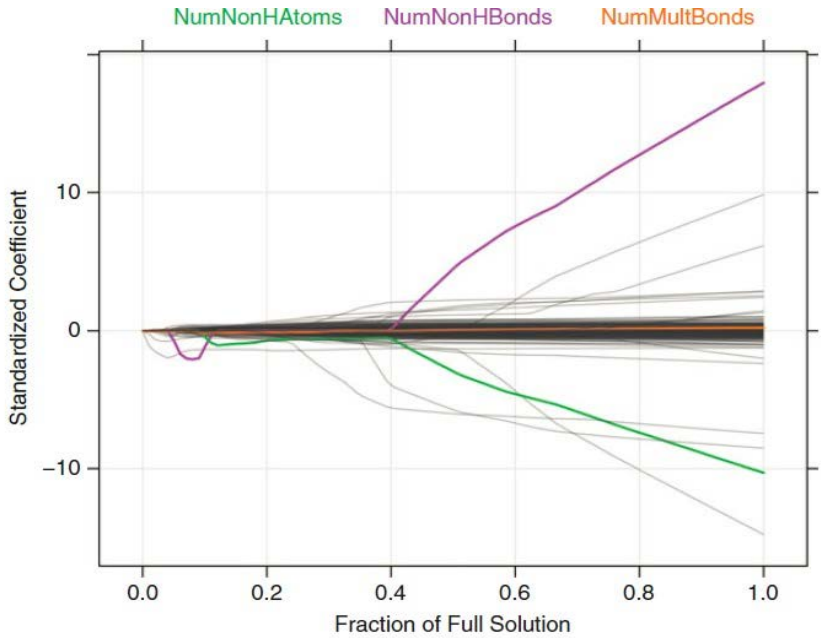

Fig. 6.17: The lasso coefficient path for the solubility data. The x-axis is the fraction of the full least squares solution. As the fraction increases, the lasso penalty (λ) decreases

Smaller values on the *x*-axis indicate that a large penalty has been used.

When the penalty is large, many of the regression coefficients are set to 0.

As the penalty is reduced, many have nonzero coefficients.

Table 6.2: Regression coefficients for two highly correlated predictors for PLS, ridge regression, the elastic net and other models

Model	NumNonHAtoms	NumNonHBonds
NumNonHAtoms only	−1.2 (0.1)	
NumNonHBonds only		−1.2 (0.1)
Both	−0.3 (0.5)	−0.9 (0.5)
All predictors	8.2 (1.4)	−9.1 (1.6)
PLS, all predictors	−0.4	−0.8
Ridge, all predictors	−0.3	−0.3
lasso/elastic net	0.0	−0.8

Penalty for ridge regression was 0.036; for lasso was 0.15. The PLS model used 10 components.

Cross-validation error for

Lasso: 0.67

PLS: 0.68

Ridge reg.: 0.69

Lasso has been extended to many other techniques, such as

- Linear discriminant analysis (Clemmensen et al. 2011; Witten and Tibshirani 2011),

- Least angle regression (LARS, Efron et al. 2004) is a broad framework that encompasses the lasso and similar models.

  The LARS model can be used to fit lasso models more efficiently, especially in high-dimensional problems.

A generalization of the lasso model is the *elastic net* (Zou and Hastie 2005).

- This model combines the two types of penalties:

$$\text{SSE}_{\text{Enet}} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda_1 \sum_{j=1}^{P} \beta_j^2 + \lambda_2 \sum_{j=1}^{P} |\beta_j|.$$

- There are two tuning parameters – two penalty parameters.

- Zou and Hastie (2005) suggest that this model will more effectively deal with groups of high correlated predictors.

Figure 6.18 shows the performance profiles across three values of the ridge penalty and 20 values of the lasso penalty. The pure lasso model is with $\lambda_1 = 0$.

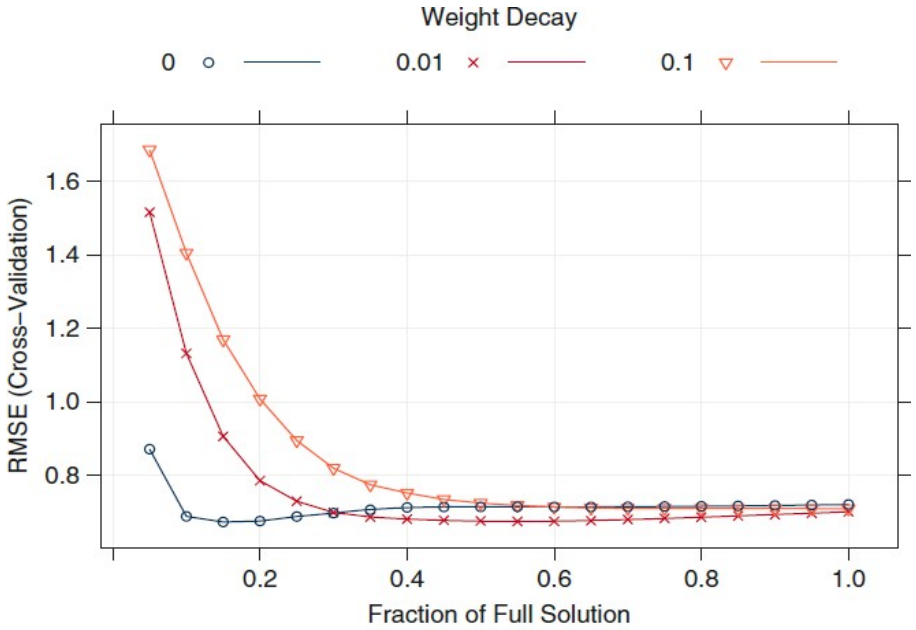

Fig. 6.18: The cross-validation profiles for an elastic net model

The optimal performance was associated with the lasso model with a fraction of 0.15, corresponding to 130 predictors out of a possible 228.