CHAPTER 6
LINEAR REGRESSION AND ITS COUSINS

APPLIED PREDICTIVE MODELING BY KUHN & JOHNSON
COLLATED BY PROF. CHING-SHIH (VINCE) TSOU (PH.D.)
CENTER FOR APPLICATIONS OF DATA SCIENCE (CADS)
GRADUATE INSTITUTE OF INFORMATION AND DECISION SCIENCES (GIIDS)
NATIONAL TAIPEI UNIVERSITY OF BUSINESS (NTUB)
CHINESE ACADEMY OF R SOFTWARE (CARS)
DATA SCIENCE & BUSINESS APPLICATIONS (DSBA) ASSOCIATION OF TAIWAN

AGENDA

- Introduction
- Case Study: Quantitative Structure-Activity Relationship Modeling
- Linear Regression
- Partial Least Squares
- Penalized Methods

INTRODUCTION – MATHEMATICAL MODEL

- In this chapter we will discuss several models, all of which are akin to linear regression in that each can directly or indirectly be written in the form

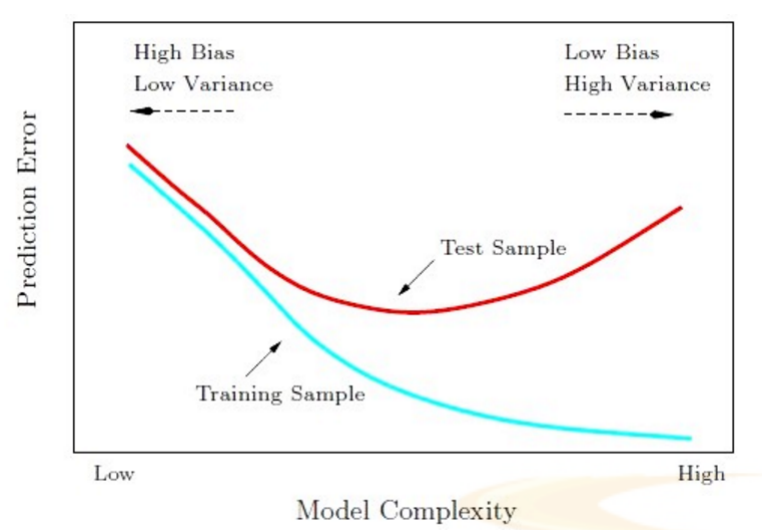
$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + \cdots + b_P x_{iP} + e_i$$

- where y_i represents the numeric response for the ith sample, b0 represents the estimated intercept, b_j represents the estimated coefficient for the jth predictor, x_{ij} represents the value of the jth predictor for the ith sample, and e_i represents random error that cannot be explained by the model.

OBJECTIVE

The objective of the methods presented in this chapter find parameter estimates that fall along the spectrum of the bias-variance trade-off.

BIAS-VARIANCE TRADE-OFF



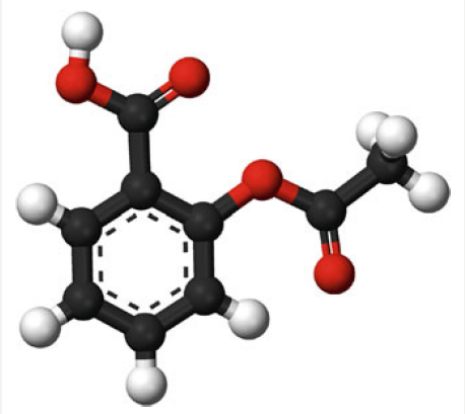
INTRODUCTION – SPECTRUM OF BIAS-VARIANCE

- The objectives of the methods presented in this chapter find parameter estimates that fall along the spectrum (i.e. whole range) of the bias-variance trade-off.
- Ordinary linear regression, at one extreme, finds parameter estimates that have minimum bias, whereas ridge regression, the lasso (Least Absolute Shrinkage and Selective Operator), and the elastic net find estimates that have lower variance.
- The impact of this trade-off on the predictive ability of these models will be illustrated in the sections to follow.
- A distinct advantage of models that follow the form of the equation in the previous slide is that they are highly interpretable.
- Another advantage of these kinds of models is that their mathematical nature enables us to compute standard errors of the coefficients, provided that we make certain assumptions about the distributions of the model residuals. (i.e. can make more inferences, but it needs assumptions.)

INTRODUCTION

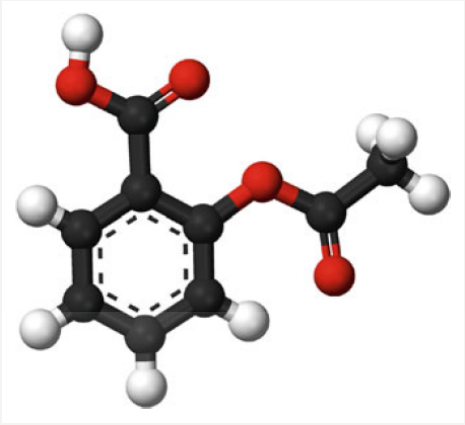
- While linear regression-type models are highly interpretable, they can be limited in their usefulness *(too simple to model the reality!)*. (trade-off again!)
- First, these models are appropriate when the relationship between the predictors and response falls along a hyperplane.
- With more predictors, the relationship would need to fall close to a flat hyperplane.
- If there is a curvilinear relationship between the predictors and response, then linear regression models can be augmented with additional predictors that are functions of the original predictors in an attempt to capture these relationships.

CASE STUDY: QUANTITATIVE STRUCTURE-ACTIVITY RELATIONSHIP (QSAR) MODELING



- Chemicals, including drugs, can be represented by chemical formulas.
- This figure shows the structure of aspirin, which contains nine carbon (in black), eight hydrogen (in white), and four oxygen (in red) atoms.
- From this configuration, quantitative measurements can be derived, such as the molecular weight, electrical charge, or surface area. (Can be determined from the chemical structure.)

CASE STUDY: QUANTITATIVE STRUCTURE-ACTIVITY RELATIONSHIP (QSAR) MODELING



- These quantities are referred to as *chemical descriptors*, and there are myriad types of descriptors that can be derived from a chemical equation.
- Some are simplistic, such as the number of carbon atoms, while others could be described as arcane (e.g., the coefficient sum of the last eigenvector from Barysz matrix weighted by the van der Waals volume).

CASE STUDY: QUANTITATIVE STRUCTURE-ACTIVITY RELATIONSHIP (QSAR) MODELING

- Some characteristics of molecules cannot be analytically determined from the chemical structure. (Still something can nor be determined analytically from the chemical structure, that's the point !)
- The relationship between the chemical structure and its activity can be complex. (How to model such complex relationship ! Big Data and adequate modeling.)
- As such, the relationship is usually determined empirically using experiments.
- One way to do this is to create a biological assay for the target of interest (i.e., the protein).

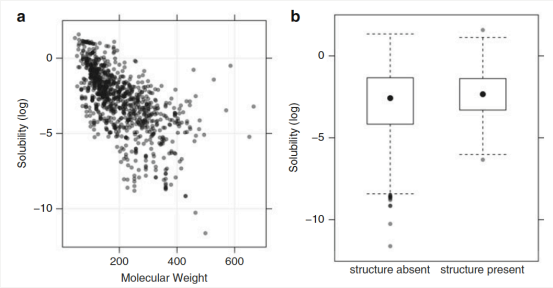
CASE STUDY: QUANTITATIVE STRUCTURE-ACTIVITY RELATIONSHIP (QSAR) MODELING

- A set of compounds can then be placed into the assay and their activity, or inhibition, is measured.
- This activity information generates data which can be used as the training set for predictive modeling so that compounds, which may not yet exist, can be screened for activity.
- While activity is important, other characteristics need to be assessed to determine if a compound is "drug-like" (Lipinski et al. 1997).
- Physical qualities, such as the solubility or lipophilicity (i.e., "greasiness"), are evaluated as well as other properties, such as toxicity.

CASE STUDY: QUANTITATIVE STRUCTURE-ACTIVITY RELATIONSHIP (QSAR) MODELING

- Tetko et al. (2001) and Huuskonen (2000) investigated a set of compounds with corresponding experimental solubility values using complex sets of descriptors.
- They used linear regression and neural network models to estimate the relationship between chemical structure and solubility.
- For our analyses, we will use 1,267 compounds and a set of more understandable descriptors that fall into one of three groups:
  - Two hundred and eight binary "fingerprints" that indicate the presence or absence of a particular chemical substructure. (Lots of binary predictors.)
  - Sixteen count descriptors, such as the number of bonds or the number of bromine atoms.
  - Four continuous descriptors, such as molecular weight or surface area.

CASE STUDY: QUANTITATIVE STRUCTURE-ACTIVITY RELATIONSHIP (QSAR) MODELING



- Left: As molecular weight of a molecule increases, the solubility generally decreases.
- Right: For a particular fingerprint descriptor, there is slightly higher solubility when the substructure of interest is absent from the molecule.
- The outcome data were measured on the log_{10} scale and ranged from − 11. 6 to 1.6 with an average log solubility value of − 2. 7.
- This figure shows the relationship between the experimentally derived solubility values and two types of descriptors in the example data.

CASE STUDY: QUANTITATIVE STRUCTURE-ACTIVITY RELATIONSHIP (QSAR) MODELING

- In the solubility data, for example, the surface area of a compound is calculated for regions associated with certain atoms (e.g., nitrogen or oxygen).
- One descriptor in these data measures the surface area associated with two specific elements while another uses the same elements plus two more.
- The small differences between surface area predictors may contain some important information for prediction, but the modeler should realize that there are implications of redundancy on the model.
- Another relevant quality of the solubility predictors is that the count-based descriptors show a significant right skewness, which may have an impact on some models (see Chap.3 for a discussion of these issues).

CASE STUDY: QUANTITATIVE STRUCTURE-ACTIVITY RELATIONSHIP (QSAR) MODELING

- It is useful to explore the training set to understand the characteristics of the data prior to modeling.
- Recall that 208 of the predictors are binary fingerprints.
- Since there are only two values of these variables, there is very little that pre-processing will accomplish.
- The average skewness statistic was 1.6 (with a minimum of 0.7 and a maximum of 3.8), indicating that these predictors have a propensity to be right skewed.
- To correct for this skewness, a Box–Cox transformation was applied to all predictors.

CASE STUDY: QUANTITATIVE STRUCTURE-ACTIVITY RELATIONSHIP (QSAR) MODELING

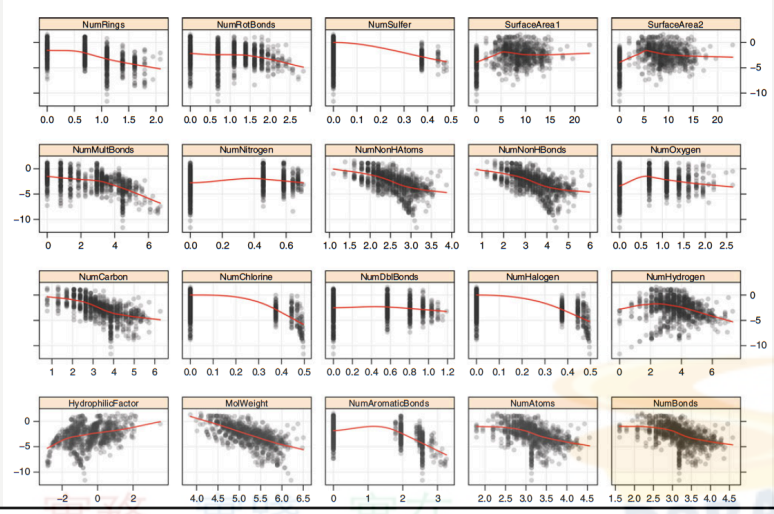
Using these transformed predictors, is it safe to assume that the relationship between the predictors and the outcome is linear? (An important question !)

- The smoothed regression lines indicate 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.

Are there significant between-predictor correlations?

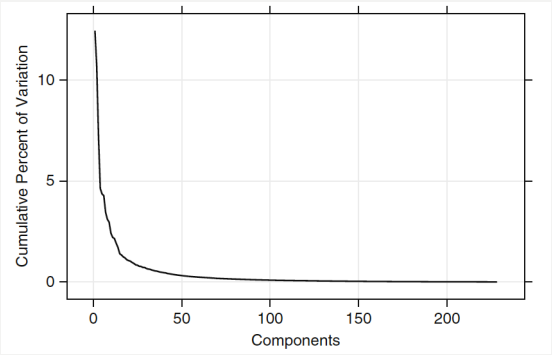
- Principal component analysis (PCA) was used on the full set of transformed predictors, and the percent of variance accounted for by each component is determined.

CASE STUDY: QUANTITATIVE STRUCTURE-ACTIVITY RELATIONSHIP (QSAR) MODELING



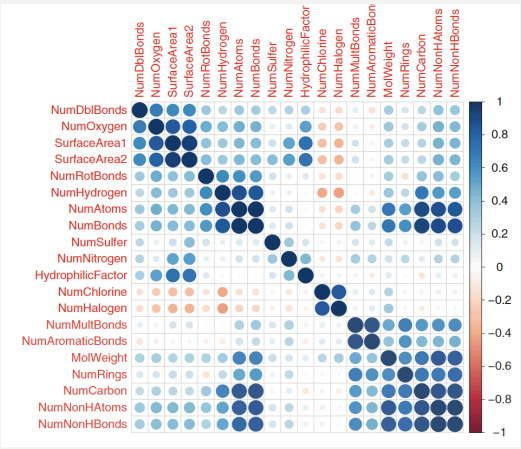
- Scatter plots of the transformed continuous predictors in the solubility data set.
- The red line is a scatter plot smoother called loess.

CASE STUDY: QUANTITATIVE STRUCTURE-ACTIVITY RELATIONSHIP (QSAR) MODELING



- This figure is commonly known as a scree plot and displays a profile of the variability accounted for by each component.
- The amount of variability summarized by component drops sharply, with no one component accounting for more than 13% of the variance.
- This is often due to a large number of collinearities among the predictors.

CASE STUDY: QUANTITATIVE STRUCTURE-ACTIVITY RELATIONSHIP (QSAR) MODELING



- This figure shows the correlation structure of the transformed continuous predictors (Box-Cox transformation); there are many strong positive correlations (indicated by the large, dark blue circles).
- This could create problems in developing some models (such as linear regression), and appropriate pre-processing steps will need to be taken to account for this problem.

LINEAR REGRESSION – OBJECTIVE FUNCTIONS AND COMPUTATION

- The objective of ordinary least squares linear regression is to find the plane that minimizes the sum-of-squared errors (SSE) between the observed and predicted response:

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

- Where y_i is the outcome and y_i *head* is the model prediction of that sample's outcome.
- Mathematically, the optimal plane can be shown to be (a (p + 1)*1 vector)

$$(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T y$$

- Where X is the matrix of predictors and y is the response vector.
- This quantity (the second equation) is easy to compute, and the coefficients are directly interpretable.
- Making some minimal assumptions about the distribution of the residuals, it is straightforward to show that the parameter estimates that minimize SSE are the ones that have the least bias of all possible parameter estimates (Graybill 1976).

LINEAR REGRESSION – COMPUTATION FAILURES

- The interpretability of coefficients makes it very attractive as a modeling tool.
- At the same time, the characteristics that make it interpretable also make it prone to potentially fatal flaws. (Trade-off again !)
- A unique inverse of trans(X)*X matrix exists when
  1. No predictor can be determined from a combination of one or more of the other predictors. (i.e. linearly independent)
  2. The number of samples is greater than the number of predictors. (n > p)
- If the data fall under either of these conditions, then a unique set of regression coefficients does not exist.

LINEAR REGRESSION – RECOVERY MECHANISM

- By default, when fitting a linear model with R and collinearity exists among predictors, ". . .R fits the largest identifiable model by removing variables in the reverse order of appearance in the model formula" (Faraway 2005). (lm(Y ~ X1 + X2 + X3 + X4, data=…), remove which one first?)
- The upshot (i.e. result) of these facts is that linear regression can still be used for prediction when collinearity exists within the data.
- But since the regression coefficients to determine these predictions are not unique, we lose our ability to meaningfully interpret the coefficients.

LINEAR REGRESSION – RECOVERY MECHANISM IN R

- When condition (2) (The number of samples is greater than the number of predictors) is true for a data set, the practitioner can take several steps to attempt to build a regression model.
- As a first step we suggest using pre-processing techniques to remove pairwise correlated predictors, which will reduce the number of overall predictors.
- This pre-processing step may not completely eliminate collinearity, since one or more of the predictors may be functions of *two* or more of the other predictors. (To diagnose multicollinearity, please try variance inflation factor!)
- After pre-processing the data, if the number of predictors still outnumbers the number of observations, then we will need to take other measures to reduce the dimension of the predictor space. (PCA->Regression = PCR, PLS, ridge, lasso, elastic net, and what is the difference!)

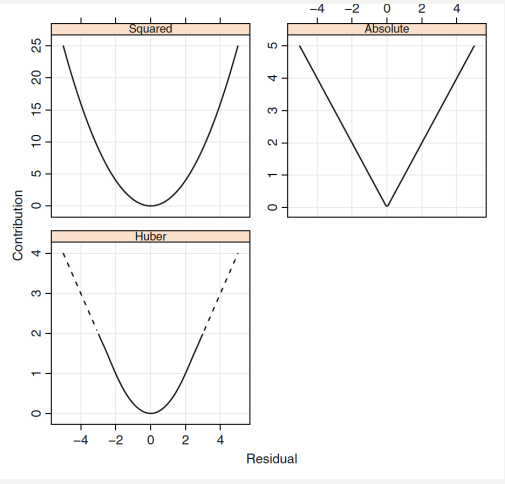
LINEAR REGRESSION – ADDING NONLINEAR TERMS

- Another drawback of multiple linear regression is that its solution is linear in the parameters.
- This means that the solution we obtain is a flat hyperplane.
- If the data have curvature or nonlinear structure, then regression will not be able to identify these characteristics.
- Quadratic, cubic, or interactions between predictors can be accommodated in regression by adding quadratic, cubic, and interactions of the original predictors.
- But the *larger* the number of *original* predictors, the *less practical* including *some or all of these terms* becomes.
- Taking this approach can cause the data matrix to have more predictors than observations (n < p), and we then again cannot invert the matrix.

LINEAR REGRESSION – OUTLIERS ISSUES

- It is prone to chasing observations that are away from the overall trend of the majority of the data.
- Recall that linear regression seeks to find the parameter estimates that minimize SSE; hence, observations that are far from the trend of the majority of the data will have exponentially large residuals.
- In order to minimize SSE, linear regression will adjust the parameter estimates to better accommodate these unusual observations.

LINEAR REGRESSION - OUTLIERS ISSUES



- One common approach is to use an alternative metric to SSE that is less sensitive to large outliers.
- For example, finding parameter estimates that minimize the sum of the absolute errors is more resistant to outliers.
- The Huber function uses the squared residuals when they are "small" and the simple different between the observed and predicted values when the residuals are above a threshold.
- This approach can effectively minimize the influence of observations that fall away from the overall trend in the data.

LINEAR REGRESSION - ALERTS ON RESAMPLING METHODS

- When using resampling techniques such as bootstrapping or cross-validation, the practitioner must still be conscious of the problems described above.
  - Consider, for example, a data set where there are 100 samples and 75 predictors. If we use a resampling scheme that uses two-thirds of the data for training (100*2/3 ~= 67 < 75), then we will be unable to find a unique set of regression coefficients, since the number of predictors in the training set will be larger than the number of samples.
- For the individual models, the regression coefficients are almost identical as are their standard errors.
  - When fitting a model with both terms, the results differ; the slope related to the number of non-hydrogen atoms is greatly decreased.

LINEAR REGRESSION - DEMO ON COLLINEARITY

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)

- The standard errors are increased fivefold when compared to the individual models.
- This reflects the instability in the regression linear caused by the between-predictor relationships and this instability is propagated directly to the model predictions.
- This table also shows the coefficients for these two descriptors when all of the predictors are put into the model.

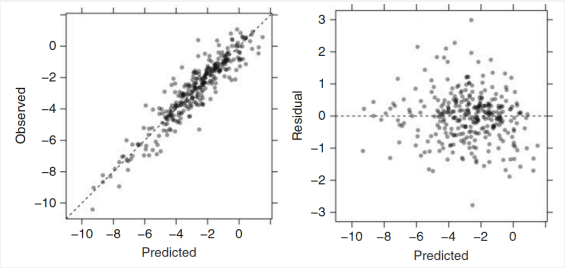
LINEAR REGRESSION - MANUALLY REMOVE CORRELATED PREDICTORS

- 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 - PREPROCESSING STEPS

- Recall that in the Case Study we split the solubility data into training and test sets and that we applied a Box–Cox transformation to the continuous predictors in order to remove skewness.
- The next step in the model building process for linear regression is to identify predictors that have high pairwise correlations and to remove predictors so that no absolute pairwise correlation is greater than some pre-specified level.
- Upon removing these predictors, a linear model was fit to the training data.

LINEAR REGRESSION FOR SOLUBILITY DATA - CHECK THE FIT



- The linear model was resampled using 10-fold cross-validation and the estimated root mean squared error (RMSE) was 0.71 with a corresponding R2 value of 0.88.
- The R2 value between the observed and predicted values was 0.87, and the basic regression diagnostic plots are displayed in this figure.

PARTIAL LEAST SQUARES – COLLINEARITY ISSUE AGAIN

- For many real-life data sets, predictors can be correlated and contain similar predictive information like illustrated with the solubility data.
- 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.
- For other data sets, the number of predictors may be greater than the number of observations.

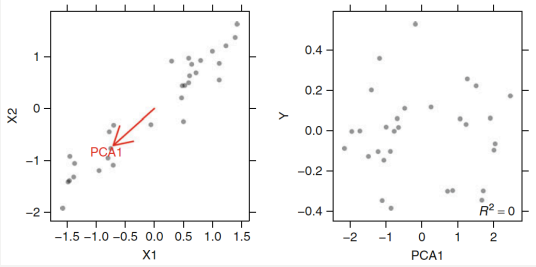
PARTIAL LEAST SQUARES – REMOVAL VS PCA

- A couple of common solutions to the regression problem under these conditions include pre-processing the predictors by either
  1. Removal of the highly correlated predictors using techniques.
  2. Conducting PCA on the predictors.
- The first process does not necessarily ensure that linear combinations of predictors are uncorrelated with other predictors.
- If this is the case, then the ordinary least squares solution will still be unstable.
- Using PCA for pre-processing guarantees that the resulting predictors, or combinations thereof, will be uncorrelated.
- The trade-off in using PCA is that the new predictors are linear combinations of the original predictors, and thus, the practical understanding of the new predictors can become murky.

PARTIAL LEAST SQUARES – PCR AND ITS ISSUE

- Pre-processing predictors via PCA prior to performing regression is known as principal component regression (PCR) (Massy 1965).
- This technique has been widely applied in the context of problems with inherently highly correlated predictors or problems with more predictors than observations.
- While this two-step regression approach (dimension reduction, then regression) has been successfully used to develop predictive models under these conditions, it can easily be misled.
- Specifically, dimension reduction via PCA does *not necessarily produce new predictors that explain the response.*

PARTIAL LEAST SQUARES



- The two predictors are correlated, and PCA summarizes this relationship using the direction of maximal variability.
- The right-hand plot of this figure, however, illustrates that the first PCA direction contains no predictive information about the response.

PARTIAL LEAST SQUARES

- As this simple example illustrates, PCA does not consider any aspects of the response when it selects its components.
- Instead, it *simply chases the variability present throughout the predictor space.*
  - If that variability happens to be related to the response variability, then PCR has a good chance to identify a predictive relationship.
  - If, however, the variability in the predictor space is not related to the variability of the response, then PCR can have difficulty identifying a predictive relationship when one might actually exist.
- Because of this inherent problem with PCR, we recommend using PLS when there are correlated predictors and a linear regression-type solution is desired.

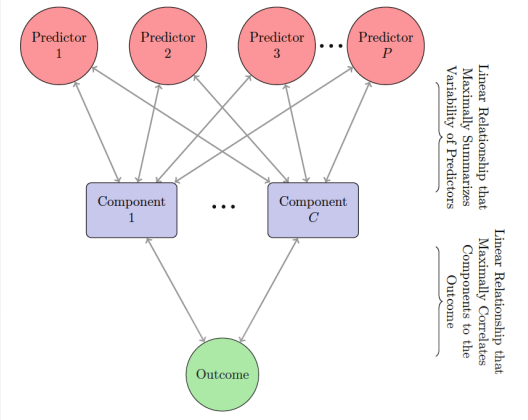
PARTIAL LEAST SQUARES – NIPALS ALGORITHM

- PLS originated with Herman Wold's *Nonlinear Iterative Partial Least Squares* (NIPALS) algorithm (Wold 1966 , 1982) which linearized models that were nonlinear in the parameters.
- Subsequently, Wold et al. (1983) adapted the NIPALS method for the regression setting with correlated predictors and called this adaptation "PLS."Briefly, the NIPALS algorithm iteratively seeks to find underlying, or latent, relationships among the predictors which are highly correlated with the response.

PARTIAL LEAST SQUARES

- For a univariate response, each iteration of the algorithm assesses the relationship between the predictors (X) and response (y) and numerically summarizes this relationship with a vector of weights (w).
- The predictor data are then orthogonally projected onto the direction to generate scores (t).
- The scores are then used to generate loadings (p), which measure the correlation of the score vector to the original predictors.
- At the end of each iteration, the predictors and the response are "deflated" by subtracting the current estimate of the predictor and response structure, respectively.

PARTIAL LEAST SQUARES – SCHEME OF PLS

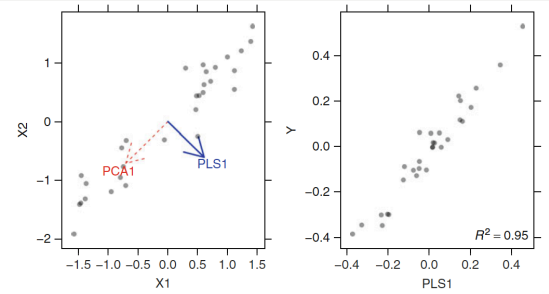


- The new deflated predictor and response information are then used to generate the next set of weights, scores, and loadings.
- These quantities are sequentially stored in matrices W , T , and P , respectively, and are used for predicting new samples and computing predictor importance.
- A schematic of the PLS relationship between predictors and the response can be seen in this figure, and a thorough explanation of the algorithm can be found in Geladi and Kowalski (1986).

PARTIAL LEAST SQUARES – LATENT VARIABLES IN BIG DATA

- These linear combinations are commonly called *components* or latent variables.
- While the PCA linear combinations are chosen to maximally summarize predictor space variability, the PLS linear combinations of predictors are chosen to maximally summarize covariance with the response.
- This means that PLS finds components that maximally summarize the variation of the predictors while simultaneously requiring these components to have maximum correlation with the response.

PARTIAL LEAST SQUARES

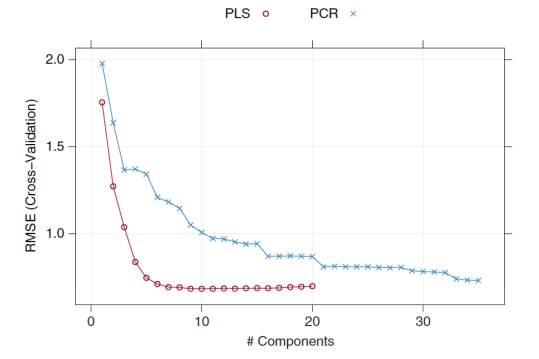


- This time we seek the first PLS component.
- The left-hand scatter plot in this figure contrasts the first PLS direction with the first PCA direction.
- For this illustration the two directions are nearly orthogonal, indicating that the optimal dimension reduction direction was not related to maximal variation in the predictor space.

PARTIAL LEAST SQUARES

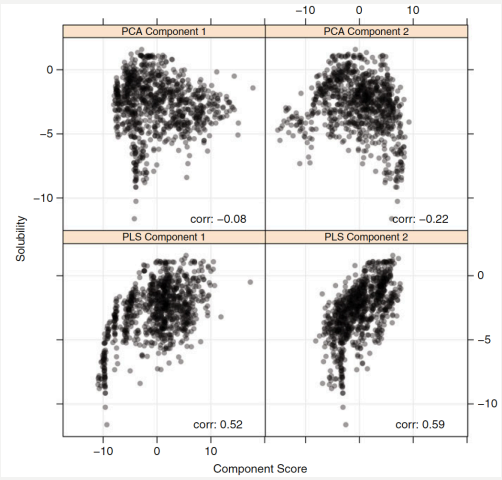
- Prior to performing PLS, the *predictors should be centered and scaled*, especially if the predictors are on scales of differing magnitude.
- PLS will seek directions of maximum variation while simultaneously considering correlation with the response.
- Once the predictors have been preprocessed, the practitioner can model the response with PLS. PLS has one tuning parameter: the number of components to retain.

PCR AND PLS FOR SOLUBILITY DATA – COMPARISONS BETWEEN PCR AND PLS



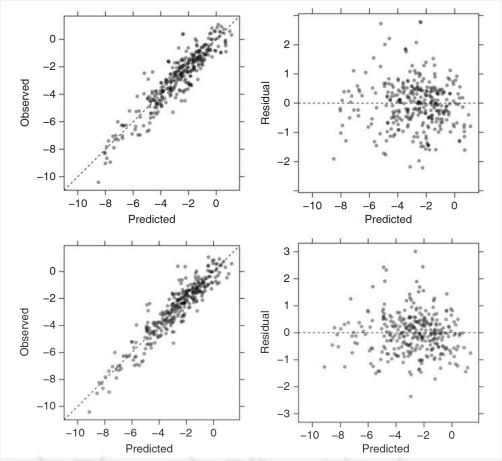
- Cross-validation was used to determine the optimal number of PLS components to retain that minimize RMSE.
- At the same time, PCR was performed using the same cross-validation sets to compare its performance to PLS.
- This figure contains the results, where PLS found a minimum RMSE (0.682) with ten components and PCR found a minimum RMSE (0.731) with 35 components.
- We see with these data that the supervised dimension reduction finds a minimum RMSE with significantly fewer components than unsupervised dimension reduction.

PCR AND PLSR FOR SOLUBILITY DATA



- Because the RMSE is lower for each of the first two PLS components as compared to the first two PCR components, it is no surprise that the correlation between these components and the response is greater for PLS than PCR.
- This figure illustrates that PLS is more quickly being steered towards the underlying relationship with the response.

PCR AND PLSR FOR SOLUBILITY DATA

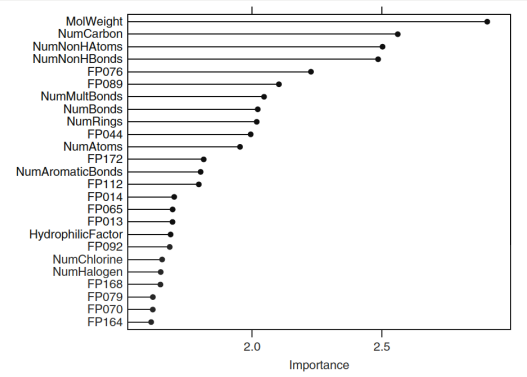


- The predictive ability of each method is good, and the residuals appear to be randomly scattered about zero.
- Although the predictive ability of these models is close, PLS finds a simpler model that uses far fewer components than PCR.

PCR AND PLSR FOR SOLUBILITY DATA

- Because the latent variables from PLS are constructed using linear combinations of the original predictors, it is more difficult to quantify the relative contribution of each predictor to the model.
- The importance of the *j*th predictor is then proportional to the value of the normalized weight vector, w , corresponding to the *j*th predictor.
- When the relationship between predictors and the response requires more than one component, the variable importance calculation becomes more involved.

PCR AND PLSR FOR SOLUBILITY DATA – VARIABLES IMPORTANCE



- For the solubility data, the top 25 most important predictors are shown in this figure.
- The larger the VIP value, the more important the predictor is in relating the latent predictor structure to the response.
- By its construction, the squared VIP values sum to the total number of predictors.

ALGORITHMIC VARIATIONS OF PLS

- The NIPALS algorithm works fairly efficiently for data sets of small-to-moderate size (e.g., < 2,500 samples and < 30 predictors) (Alin 2009).
- But when the number of samples (n) and predictors (P) climbs, the algorithm becomes inefficient.
- This inefficiency is due to the way the matrix operations on the predictors and the response are performed.

ALGORITHMIC VARIATIONS OF PLS

- Specifically, both the predictor matrix and the response must be deflated for each latent variable.
- Lindgren et al. (1993) showed that the constructs of NIPALS could be obtained by working with a "kernel" matrix of dimension P×P .
- The covariance matrix of the predictors (also of dimension P×P), and the covariance matrix of the predictors and response (of dimension P×1).
- This adjustment improved the speed of the algorithm, especially as the number of observations became much larger than the number of predictors.

ALGORITHMIC VARIATIONS OF PLS

- At nearly the same time as the kernel approach was developed, de Jong (1993) improved upon the NIPALS algorithm by viewing the underlying problem as finding latent orthogonal variables in the predictor space that maximize the covariance with the response.
- de Jong (1993) showed that the SIMPLS latent variables are identical to those from NIPALS when there is only one response.

ALGORITHMIC VARIATIONS OF PLS

- Dayal and MacGregor (1997) developed two efficient modifications, especially when n >> P, and, similar to SIMPLS, only require a deflation of the covariance matrix between the predictors and the response at each step of the iterative process.
  - In their first alteration to the inner workings of the algorithm, the original predictor matrix is used in the computations (without deflation).
  - In the second alteration, the covariance matrix of the predictors is used in the computations (also without deflation).

ALGORITHMIC VARIATIONS OF PLS

- Alin (2009) provided a comprehensive computational efficiency comparison of NIPALS to other algorithmic modifications.
- In nearly every scenario, the second kernel algorithm of Dayal and MacGregor was more computationally efficient than all other approaches and provided superior performance when n > 2, 500 and P > 30.
- And in the cases where the second algorithm did not provide the most computational efficiency, the first algorithm did.

ALGORITHMIC VARIATIONS OF PLS

- Rannar et al. (1994) constructed a kernel based on the predictor matrix and response that had dimension n×n .
- A usual PLS analysis can then be performed using this kernel, the outer products of the predictors, and the outer products of the response (each with dimension n×n).
- This algorithm is computationally more efficient when there are more predictors than samples.

ALGORITHMIC VARIATIONS OF PLS

- While many methods exist, the most easily adaptable approaches using the algorithms explained above are provided by Berglund and Wold (1997) and Berglund et al. (2001).
- In Berglund and Wold (1997), the authors show that adding squared predictors (and cubic, if necessary) can be included with the original predictors.
- PLS is then applied to the augmented data set.
- The authors also show that there is no need to add cross-product terms, thus greatly reducing the number of new predictors added to the original data.

ALGORITHMIC VARIATIONS OF PLS

- The original predictors that were binned are then excluded from the data set that includes the binned versions of the predictors. PLS is then applied to the new predictor set in usual way.
- Both of these approaches have successfully found nonlinear relationships between the predictors and the response.
- But there can be a considerable amount of effort required in constructing the data sets for input to PLS, especially as the number of predictors becomes large.
- If a more intricate relationship between predictors and response exists, then we suggest employing one of the other techniques rather than trying to improve the performance of PLS through this type of augmentation.

PENALIZED MODELS – OBJECTIVE FUNCTION

- It is common that a small increase in bias can produce a substantial drop in the variance and thus a smaller MSE than ordinary least squares regression coefficients.
- One method of creating biased regression models is to add a penalty to the sum of the squared errors. Recall that original least squares regression found parameter estimates to minimize the sum of the squared errors:

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

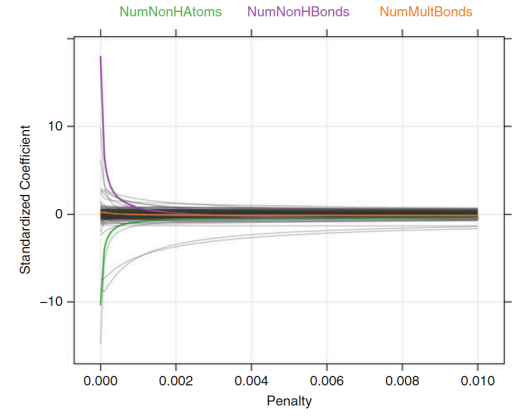
PENALIZED MODELS

- We may want to control the magnitude of these estimates to reduce the SSE. Controlling (or regularizing) the parameter estimates can be accomplished by adding a penalty to the SSE if the estimates become large.
- Ridge regression (Hoerl 1970) adds a penalty on the sum of the squared regression parameters:

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

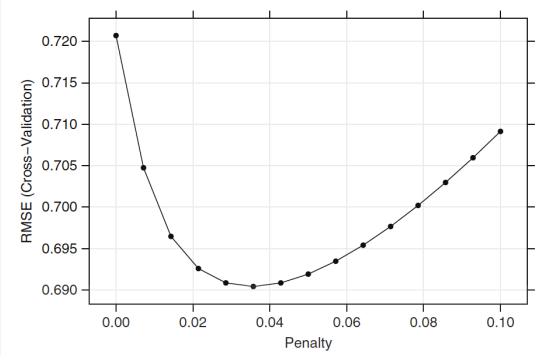
- The "L2 " signifies that a second-order penalty is being used on the parameter estimates.
- The effect of this penalty is that the parameter estimates are only allowed to become large if there is a proportional reduction in SSE.
- In effect, this method shrinks the estimates towards 0 as the λ penalty becomes large (these techniques are sometimes called "shrinkage methods").

PENALIZED MODELS - SHRINKAGE PATHS



- This figure shows the path of the regression coefficients for the solubility data over different values of λ .
- Each line corresponds to a model parameter and the predictors were centered and scaled prior to this analysis so that their units are the same.
- Some parameter estimates are abnormally large, such as the number of non-hydrogen atoms (in green) and the number of non-hydrogen bonds (purple).

PENALIZED MODELS - OPTIMAL PENALTY



- Using cross-validation, the penalty value was optimized.
- When the penalty is increased, the error drops from 0.72 to 0.69 (1st pt. to 2nd pt.). As the penalty increases beyond 0.036 (lowest pt. to its right), the bias becomes to large and the model starts to under-fit, resulting in an increase in MSE.

PENALIZED MODELS

- Even though some parameter estimates become negligibly small, this model does not conduct feature selection. (as compared to lasso)
- A popular alternative to ridge regression is the *least absolute shrinkage and selection operator* model, frequently called the *lasso* (Tibshirani 1996).
- This model uses a similar penalty to ridge regression:

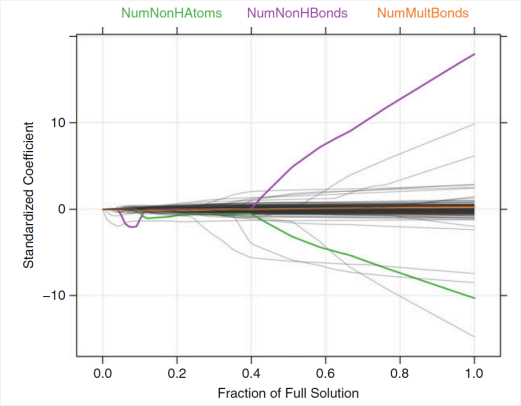
$$\text{SSE}_{L_1} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda \sum_{j=1}^{P}|\beta_j|$$

PENALIZED MODELS

- Thus the lasso yields models that simultaneously use regularization to improve the model and to conduct feature selection.
- In comparing, the two types of penalties, Friedman et al. (2010) 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/kth the size that any single one would get if fit alone.[. . .] lasso, on the other hand, is somewhat indifferent to very correlated predictors, and will tend to pick one and ignore the rest."

PENALIZED MODELS - REGULARIZATION PATHS



- The x -axis is the fraction of the full solution.
- 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.
- When the fraction is around 0.4, this predictor is entered back into the model with a nonzero coefficient that consistently increases

PENALIZED 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

- The ridge-regression penalty used in this table is 0.036 and the lasso penalty was 0.15.
- The ridge-regression model shrinks the coefficients for the non-hydrogen atom and non-hydrogen bond predictors significantly towards 0 in comparison to the ordinary least squares models while the lasso model shrinks the non-hydrogen atom predictor out of the model.

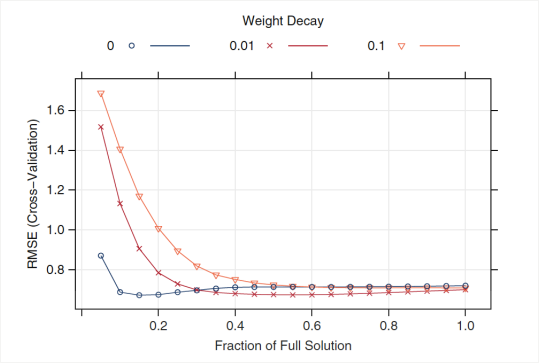
PENALIZED MODELS – TOWARDS ELASTIC NETS MODEL

- A generalization of the lasso model is the *elastic net* (Zou and Hastie 2005).
- This model combines the two types of penalties:

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

- The advantage of this model is that it enables effective regularization via the ridge-type penalty with the feature selection quality of the lasso penalty.
- The Zou and Hastie (2005) suggest that this model will more effectively deal with groups of high correlated predictors.

PENALIZED MODELS – PARAMS FOR LAMBDA1 AND LAMBDA2



- Both the penalties require tuning to achieve optimal performance.
- Again, using resampling, this model was tuned for the solubility data.
- This figure shows the performance profiles across *three values of the ridge penalty* and *20 values of the lasso penalty*. The pure lasso model (with $\lambda 1 = 0$) has an initial drop in the error and then an increase when the fraction is greater than 0.2.

PENALIZED MODELS – OPTIMAL PERFORMANCE

- The two models with nonzero values of the ridge penalty have minimum errors with a larger model.
- In the end, the optimal performance was associated with the lasso model with a fraction of 0.15, corresponding to *130 predictors out of a possible 228*