# Model building

Dimension reduction – principal components regression (PCR) and partial least squares (PLS)

# Dimension reduction methods

- All of these methods are defined using the original $p$ predictors, $X_1, \ldots, X_p$.
- We explore a class of approaches that *transform* the predictors and then fit a least squares model using the transformed variables.
- We will refer to these techniques as *dimension reduction* methods.

# Dimension reduction methods

- Let $Z_1, \ldots, Z_M$ represent $M < p$ linear combinations of our original $p$ predictors $X_1, \ldots, X_p$, such that

$$Z_m = \sum_{j=1}^{p} \varphi_{jm} X_j, \quad m = 1, 2, \ldots, M,$$

- We can then fit the linear regression model

$$y_i = \theta_0 + \sum_{i=1}^{M} \theta_m Z_{im}, \quad i = 1, \ldots, n,$$

using ordinary least squares.

- The dimension of the problem has been reduced from $p+1$ to $M+1$, where $M < p$.
- With appropriate choices of $\varphi_{jm}$, the *dimension reduction* procedure can often outperform OLS regression

# Principal components regression (PCR)

- We apply principal components analysis (PCA) (Chapter 3) to define the linear combinations of the predictors, for use in our regression.
- The first principal component (PC) is the (normalized) linear combination of the variables with the largest variance.
- The second PC has the second largest variance and is uncorrelated with the first.
- And so on.
- Hence with many correlated original variables, we replace them with a small set of uncorrelated PCs that capture their joint variation.

# Application to PCR

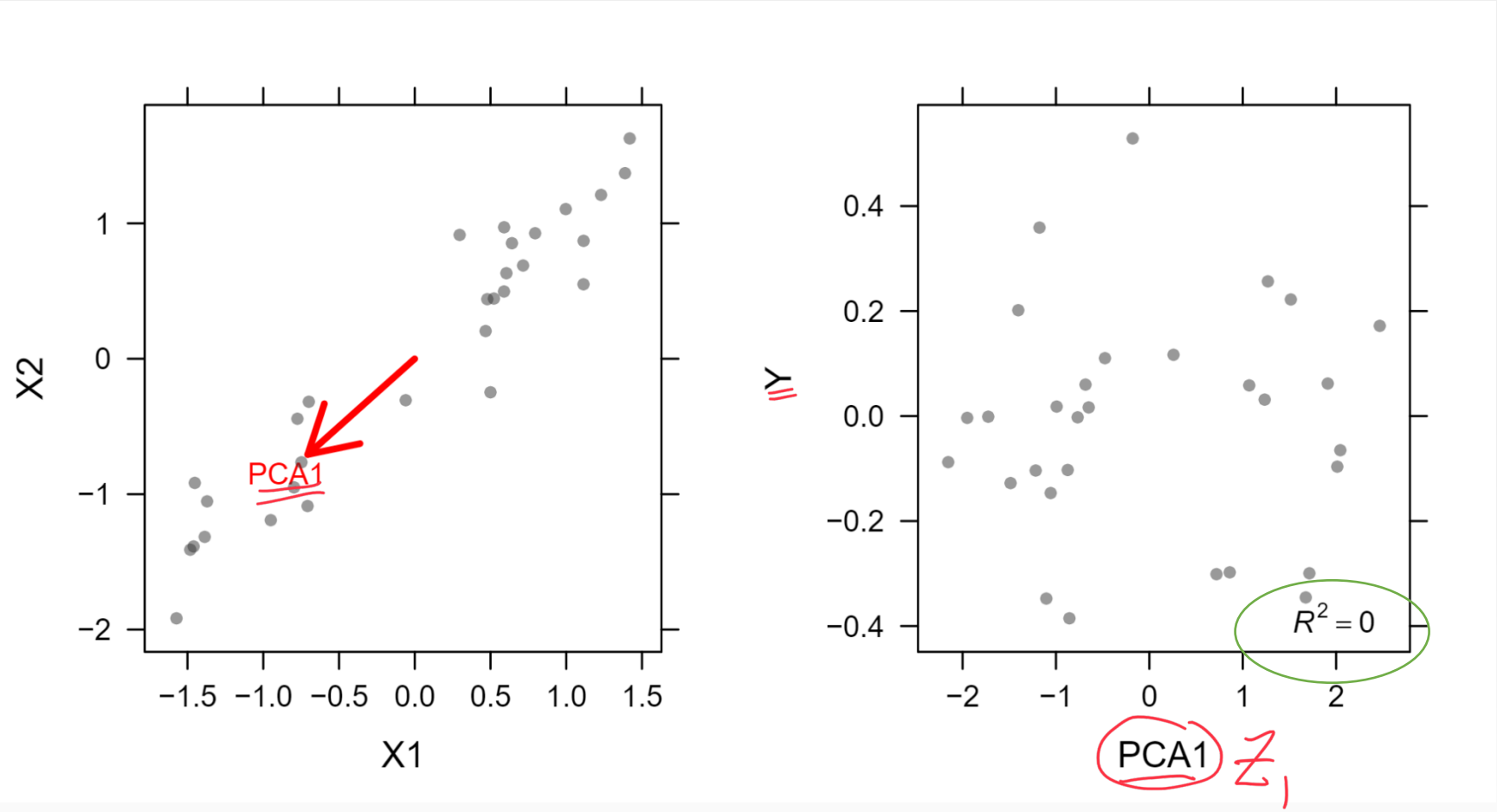
- The PCR approach involves constructing the first $M < p$ PCs, $Z_1, \ldots, Z_M$, and then using these components as the predictors in a linear regression model that is fit using least squares.
- If the assumption underlying PCR holds, by estimating only M <p coefficients we can mitigate overfitting.
- When performing PCR, we generally recommend standardizing each predictor prior to generating the PCs. This standardization ensures that all variables are on the same scale.
- Remark: we choose M by using cross-validation

# Potential problems of PCR

based on PCA ← unsupervised learning

- PCR is an *unsupervised* procedure for dealing with problems with inherently highly correlated predictors or problems with more predictors than observations (i.e., n < p) ← $M < n < P$
- New predictors produced by PCA may not explain the response, since PCA only summarizes this relationship of the predictors using the direction of maximal variability of the predictors.
- In other words, PCA does not consider any aspects of the response when it selects its components for new predictors.

# Example

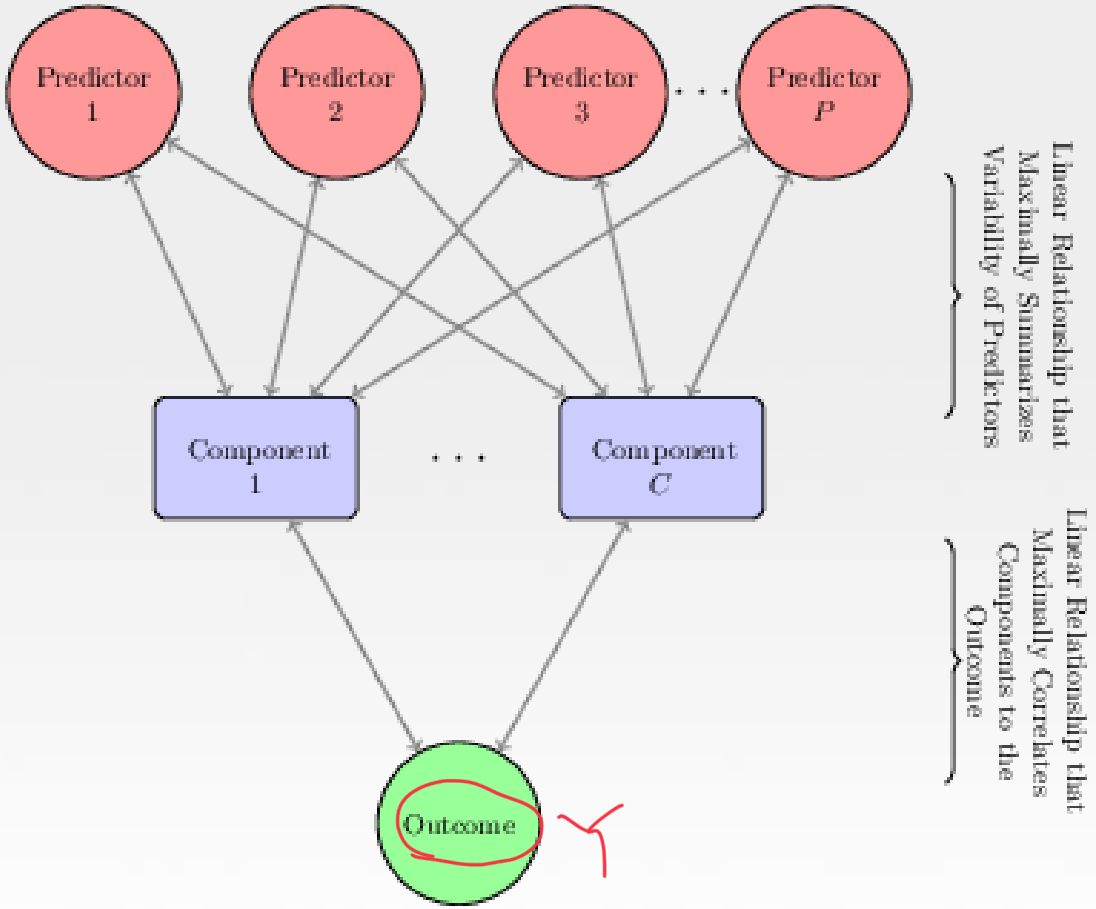



**Left**: A scatter plot of the two predictors shows the direction of the first principal component.
**Right**: The first PCA direction contains no predictive information for the response.

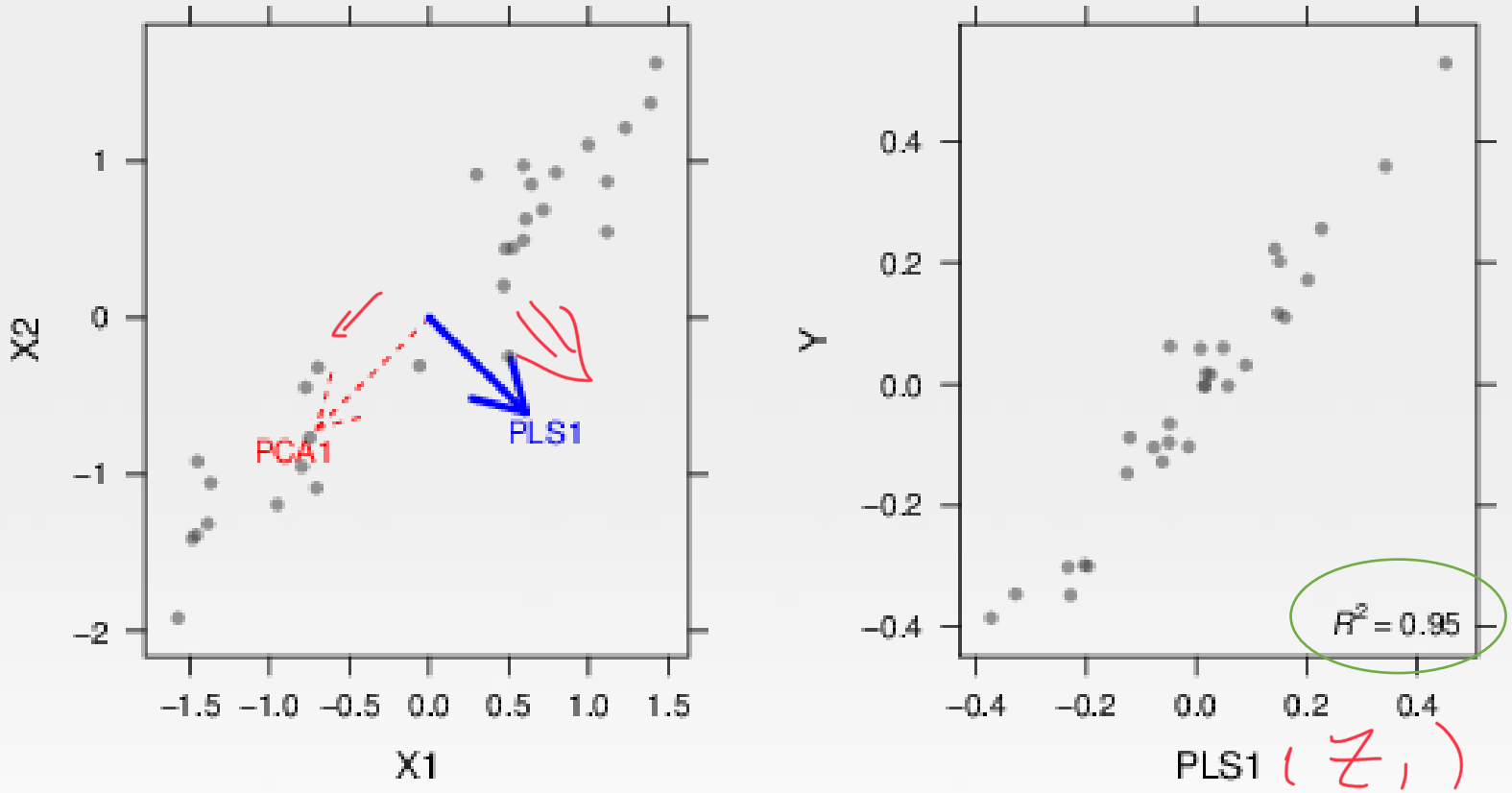
# Partial least squares (PLS)

- PCR identifies PCs in an *unsupervised* way, since the response *Y* is not used to help determine the PC directions.
- PCR suffers from a drawback: there is no guarantee that the directions that best explain the predictors will also be the best directions to use for predicting the response.
- PLS is a supervised alternative to PCR. It makes use of *Y* in order to identify new features that not only approximate the old features well, but also that are related to the response.
- Roughly speaking, the PLS approach attempts to find the new features $Z_1$, ..., $Z_M$ that help explain both *Y* and *X*.

# A diagram depicting the structure of a PLS model



# Example (cont's)




**Left**: The first PLS direction is nearly orthogonal to the first PCA direction. **Right**: Unlike PCA, the PLS direction contains highly predictive information for the response.

# PCR vs. PLS

- As with PCR, M is typically chosen by cross-validation.
- We generally standardize the predictors and response before performing PLS.
- PLS is popular in the field of chemometrics, where many variables arise from digitized spectrometry signals.
- In practice it often performs no better than PCR.

# Model building

Shrinkage-Penalized models

# Ridge regression

- Given the training data, OLS estimate $\beta_0, \beta_1, \ldots, \beta_p$ by minimizing

$$\text{RSS} = \sum_{i=1}^{n}\left(Y_i - \beta_0 - \sum_{j=1}^{p}\beta_j X_{ij}\right)^2$$

- Ridge regression minimizes a slightly different equation given by

$$\left\{\sum_{i=1}^{n}\left(Y_i - \beta_0 - \sum_{j=1}^{p}\beta_j X_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\beta_j{}^2 = \text{RSS} + \lambda\sum_{j=1}^{p}\beta_j{}^2\right\},$$

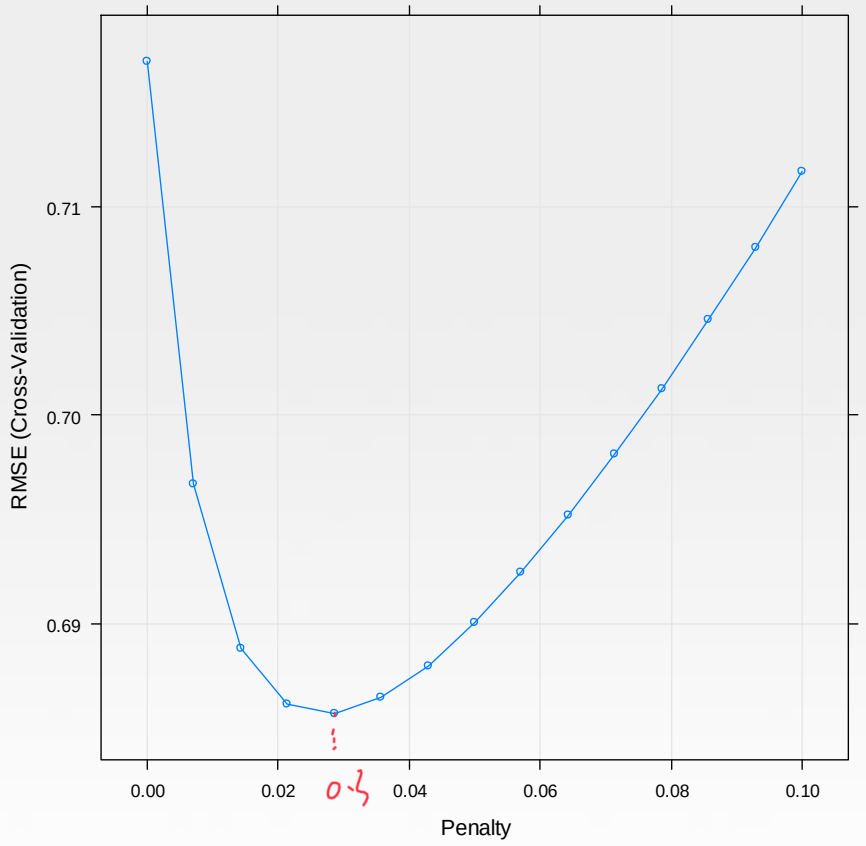
Where λ is a *tuning* parameter to be determined separately.

# Ridge regression

- The second term $\lambda \sum_{j=1}^{p} \beta_j^2$ is called a shrinkage penalty.
- This has the effect of shrinking the estimates of $\beta_j$ towards zero. The tuning parameter $\lambda$ serves to control the relative impact of these two terms on the regression coefficient estimates.
- Ridge regression will produce a different set of coefficient estimates, $\hat{\beta}_\lambda^R = (X^T X + \lambda I_p)^{-1} X^T Y$, for each value of $\lambda$, where $I_p$ is the p-dimensional identity matrix.
- When $\lambda = 0$, we get the OLS. It is critical to select a good value for $\lambda$.

# Cross-validation for λ

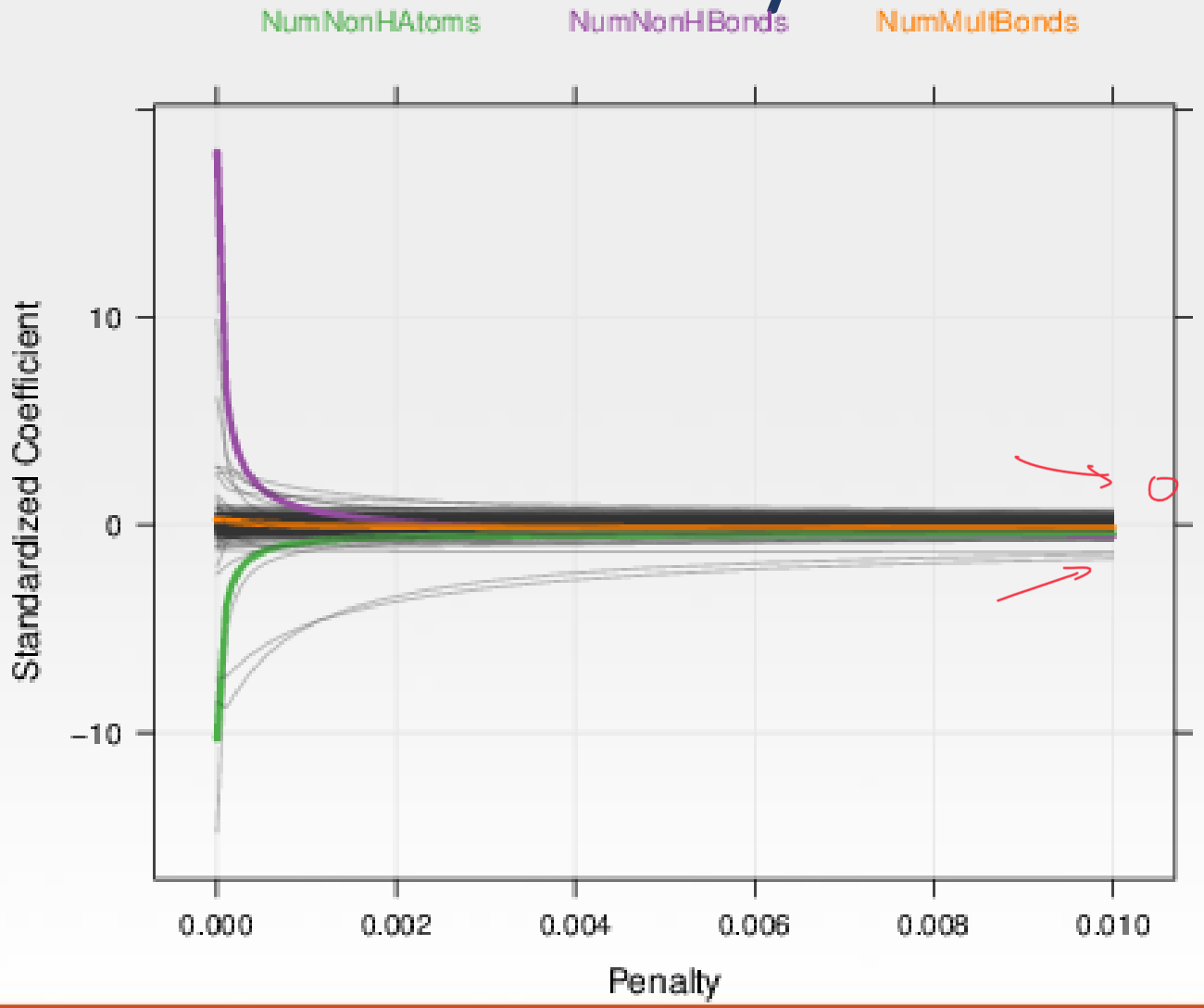
λ is determined by CV




λ = 0.3

λ

# Ridge regression for solubility data




$\lambda \to 0$

Ridge $\approx$ OLS

$\lambda = \infty$

Never to be zero exactly

# Ridge regression for solubility data

- The standardized ridge regression coefficients are displayed.
- The tuning parameter λ is determined by cross-validation.
- **Caution**: Standardize the predictors before ridge regression.
- As $\lambda$ increases, the standardized coefficients shrinks towards zero.
- The notation $||\beta||_2 = \sqrt{\sum_{j=1}^{p} \beta_j^2}$ is the $l_2\ norm$ of β.

# Ridge regression vs. OLS

prediction OLS is less preferred comparing with ridge,

- In general, the ridge regression estimates will be more biased than the OLS ones but have lower variance. (Bias-Variance tradeoff)
- Ridge regression will work best in situations where the OLS estimates have high variance. Such situations include $p \approx n$, $p > n$, multicollinearity, etc. $p>n$, $p \approx n$
- Ridge regression also has substantial computational advantages over best subset selection: for any given $\lambda$, we only need to fit one model and the computations turn out to be very simple

$$\hat{\beta}_R = (X'X + \lambda I)^{-1} X'y$$

# The Lasso

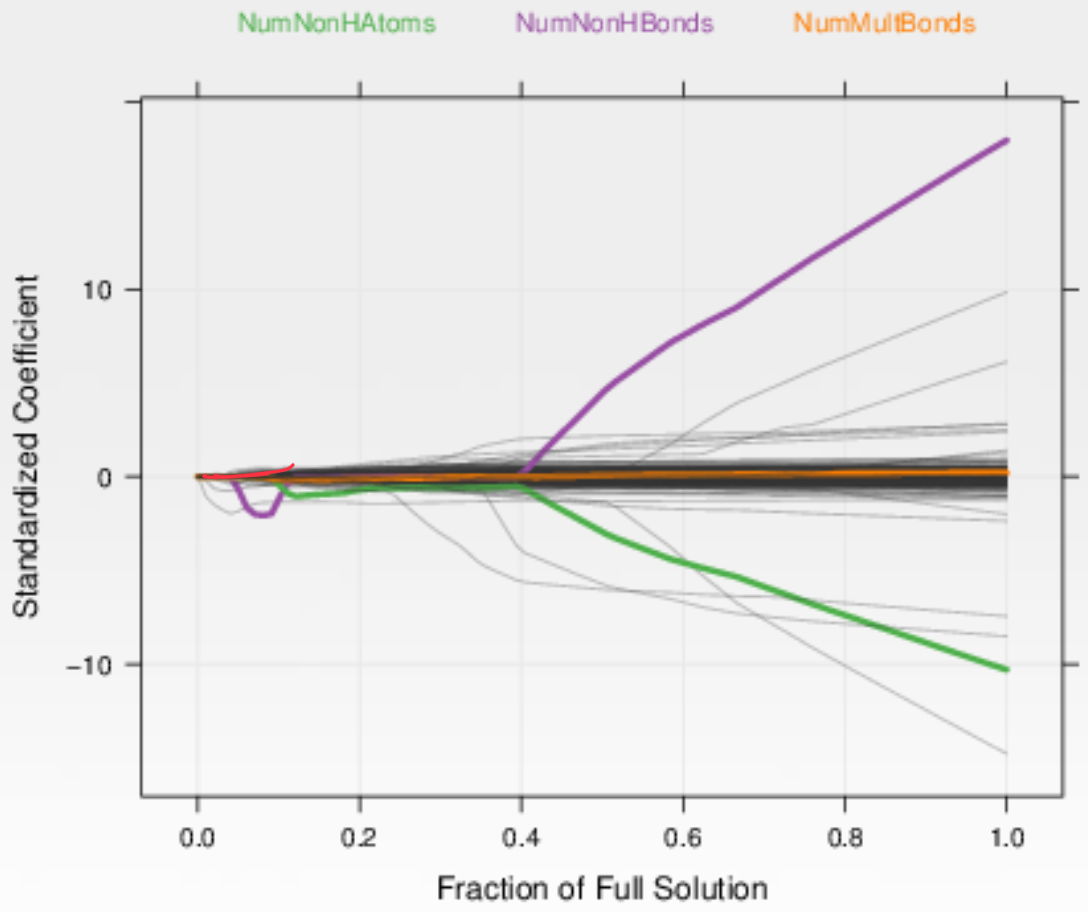
- Ridge regression isn't perfect. The penalty term will never force any of the coefficients to be exactly zero. Thus, the final model will include all variables, which makes it harder to interpret.
- The *lasso* overcomes this disadvantage by minimizing

$$\sum_{i=1}^{n}\left(Y_i - \beta_0 - \sum_{j=1}^{p}\beta_j X_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = \text{RSS} + \lambda\sum_{j=1}^{p}|\beta_j|,$$

where $\lambda$ is a tuning parameter and selecting a good value of $\lambda$ for the lasso is critical.

- Lasso can force some of the coefficient estimates to be exactly equal to zero, yielding *sparse* models—that is, models that involves only a subset of the variables.

# Lasso regression for solubility data



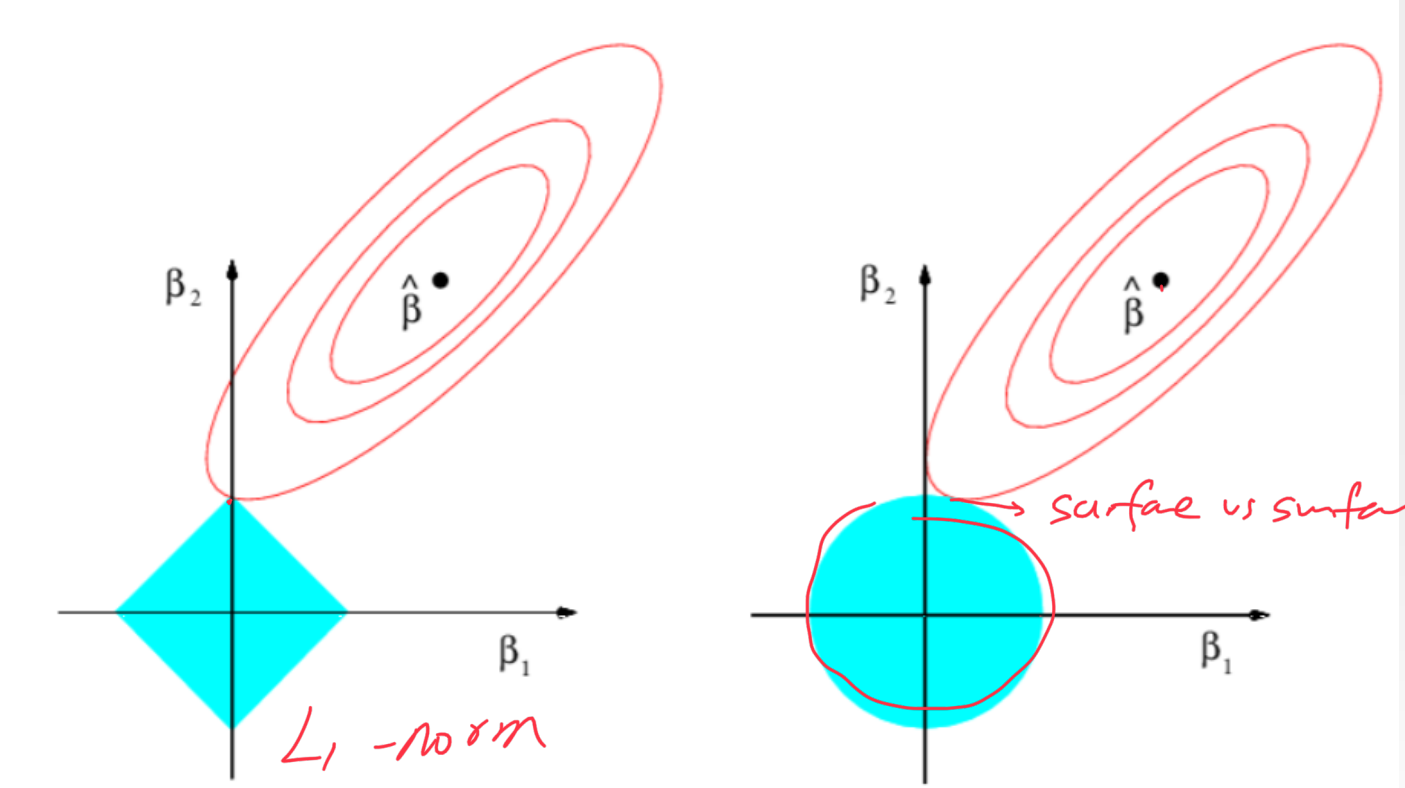
# Ridge regression for solubility data

- The standardized ridge regression coefficients are displayed.
- **Caution**: It is necessary to standardize variables before using Lasso and Ridge Regression
- Depending on the value of $\lambda$, the Lasso can produce a model involving only a subset of the variables.

Variable Selection

- The notation $||\beta||_1 = \sqrt{\sum_{j=1}^{p} |\beta_j|}$ is the $l_1$ $norm$ of $\beta$.

# Why can lasso do variable selection?




- Contours of the error and constraint functions for the lasso(left) and ridge regression (right)

# Lasso vs. ridge regression

- Neither ridge regression nor the lasso will universally dominate the other.
- The lasso tends to perform better in a setting where a relatively small number of predictors have substantial coefficients. (sparse model)
- Ridge regression will perform better when the response is a function of many predictors, all with coefficients of roughly equal size.
- However, the number of predictors that is related to the response is never known a priori for real data sets. A technique such as cross-validation can be used to determine which approach is better on a particular data set.
- Unlike ridge regression, the lasso performs variable selection, resulting in models that are easier to interpret

feature selection

prediction

# Elastic net

- The elastic net is a generalization of the lasso model. This model combines the two types of penalties

$$\sum_{i=1}^{n}\left(Y_i - \beta_0 - \sum_{j=1}^{p}\beta_j X_{ij}\right)^2 + \lambda_1 \sum_{j=1}^{p}|\beta_j| + \lambda_2 \sum_{j=1}^{p}\beta_j{}^2,$$

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