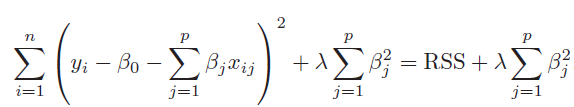
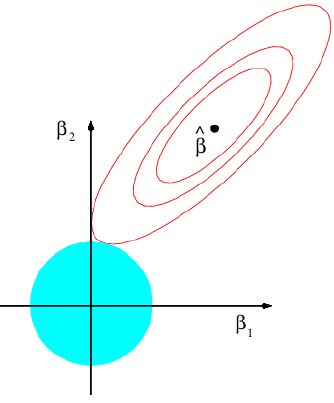
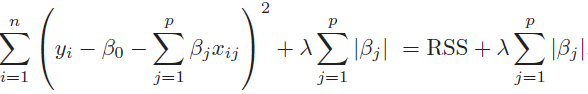
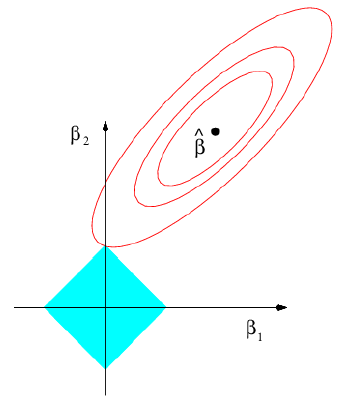
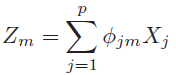
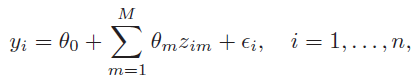
Bias – how well generalizes to other models

Variance – how well it fits current model

* Low bias (linear model) –
  + True relationship between response and predictors linear
* Lowe variance (linear model)
  + (n >> p) – # of observations > # of variables
  + n only slightly greater than p 🡪 overfitting
  + (p > n) 🡪 no unique solution exists, variance infinite
* **Feature selection** – excluding irrelevant variables from regression
  + 3 types:
    - (1) Subset selection
    - (2) Shrinkage – estimated coefficients of predictors shrunk to 0
      * Some become exactly 0 (lasso, not ridge) and can be excluded
    - (3) Dimension reduction –
      * Project p predictors on M-dimensional space (M < p)
      * Compute M different *linear combinations* of variables
      * Use M combinations as predictors
* Subset selection –
  + (1) Best subset selection –
    - Fit separate regression for each possible combination of predictors
      * 🡪 for k predictors
        + i.e. 10 predictors and two variables linear model 🡪
  + (2) Forward stepwise selection –
    - Computationally efficient
    - Start with no predictors
    - Add predictors one-at-a-time until all in model
    - Each step, predictor that gives greatest improvement to model selected
  + (3) Backward stepwise selection –
    - Starts with all predictors, removing one at a time
  + Can NOT use R2 as predictor comparator for various models generated
    - Use either:
      * (1) Cross-validation approach
      * (2) Validation set approach
* Shrinkage –
  + (1) Ridge regression
    - Always apply AFTER scaling the predictors
    - MIN 🡪 
    - λ = *shrinkage penalty* term magnified by coefficients of beta
      * Shrinks estimates of beta close to 0
    - λ = **hyperparameter**
    - L2 norm - 
    - (+) λ 🡪 (-) L2 norm
    - Works best when least squares has high variance:
      * Trade of small increase in bias, for large decrease in variance
    - 
  + (2) Lasso regression
    - Sets some predictors to exactly 0, to be excluded
    - MIN 🡪 
    - L1 norm - 
    - 
      * Intersects on the axis 🡪 coefficients = 0
  + λ = tuning parameter
    - Select using cross-validation
* Dimension reduction (PCA)
  + 
    - Xi – original predictors (# = p)
    - Zi – linear combinations of original predictors (# = M)
    - Φ - constants
    - (M < p)
  + Fit linear regression model:



* + - Θ – regression coefficients
  + PCA:
    - Direction of data along which the observations vary the most
    - Constraint: = 1
    - Each consecutive principal component orthogonal to all previous ones
    - (+) # of predictors (principal components):
      * (-) bias
      * (+) variance
    - Standardize variables unless all on same scale
    - Drawback: *unsupervised* 🡪 response variable not considered when determining principal components
  + Partial Least Squares (PLS):
    - PCS, but uses response variable as well in identifying principal components
* High-dimensional (p > n) data:
  + Problem: perfect fit of data 🡪 overfitting
  + Solution: dimension reduction
* Best training RSS – best subset selection (over backward and forward selection)
* Ridge & Lasso vs. Least Squares:
  + Less flexible
  + Benefit when: (increase in variance) < (increase in bias)