Regularization and Shrinkage Methods

April 14, 2013

MATH 287D Spring 2013: Statistical Learning
University of California, San Diego
Instructor: Jelena Bradic

http://math.ucsd.edu/~jbradic
Introduction
  Overview of Linear Regression
  Shrinkage Examples
Linear Regression Model

Assumes that $f(x) = E[Y|X = x]$ is linear in $x$, namely

$$E[Y|X = x = (x_1, \ldots, x_p)] = \beta_0 + \sum_{j=1}^{p} \beta_j x_j$$

The response $Y$ is numerical.
The predictors $X_j$ can be:

- Quantitative inputs or transformations of quantitative inputs.
- Basis expansion; e.g. $X_2 = X_1^2$, $X_3 = X_1^3$.
- Indicator variables coding the levels of a qualitative variable; e.g. $X_j = 1_G = j$.
- Interaction variables; e.g. $X_3 = X_1 \cdot X_2$. 
Least Squares Method

▶ We have a sample \((x_i, y_i), i = 1, \ldots, N\).
▶ Least squares assumes that \(Y|X = x \sim N(\beta_0 + \beta^T x, \sigma^2)\), and computes the maximum likelihood

\[
\hat{\beta} = \arg \min_{\beta} \text{RSS}(\beta) = \arg \min_{\beta} \sum_{i=1}^{N} (y_i - \beta_0 - \beta^T x_i)^2.
\]

Normal equations

\[
X^T (y - X\beta) = 0
\]

where \(y = (y_1, \ldots, y_N)^T\) is the response vector and \(X\) is the design matrix, \(N\)-by-(\(p + 1\)) with rows \(1^T, x_1^T, \ldots, x_N^T\).

If \(X\) is full-rank, i.e. nonsingular, \(\hat{\beta}\) is well-defined / unique

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

▶ The fitted values are

\[
\hat{y} = X = X(X^T X)^{-1} X^T = Hy
\]

where \(H = X(X^T X)^{-1} X^T\) is called the hat matrix.
Least Squares: Distribution of Estimates

- If indeed $Y | X = x \sim \mathcal{N}(\beta_0 + \beta^T x, \sigma^2)$, then
  
  $\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2 (X^T X)^{-1})$  
  $\hat{\sigma}^2 \sim \frac{\sigma^2}{N - p - 1} \chi^2_{N-p-1}$

  Gauss-Markov Theorem LS estimator has the smallest variance among all unbiased, linear estimators.

- Associated $t$-distributions (testing significance of single features):
  
  $\frac{x^T - x^T \beta}{\sqrt{x^T (X^T X)^{-1} x}} \sim t_{N-p-1}$

- Associated $F$-distributions (testing significance of group of features):
  
  $p$ features::
  
  $\frac{\|X \hat{\beta} - X \beta\|^2 / (p + 1)}{\hat{\sigma}^2} \sim F_{p+1,N-p-1}$

  $p_1$ features::
  
  $\frac{(RSS_0 - RSS_1) / (p_1 - p_0)}{RSS_1 / (N - p_1 - 1)} \sim F_{p_1-p_0,N-p_1-1}$

  $RSS_1$ residual sum of squares for the bigger model with $p_1 + 1$ parameters  
  $RSS_0$ residual sum of squares for the nested smaller model with $p_0 + 1$ parameters  
  $p_1 - p_0$ features are set to zero
Multiple Regression via Univariate Regression

Suppose $p = 1$ ..... then ...

$$\hat{\beta} = \frac{\sum_{i=1}^{N} x_i y_i}{\sum_{i=1}^{N} x_i^2} = \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{\langle \mathbf{x}, \mathbf{x} \rangle}$$

Suppose $p \geq 1$ ..... then ...

- If columns of $\mathbf{x}$ are orthogonal, i.e. $\langle \mathbf{x}_j, \mathbf{x}_k \rangle = 0$ then

$$\hat{\beta}_j = \frac{\langle \mathbf{x}_j, \mathbf{y} \rangle}{\langle \mathbf{x}_j, \mathbf{x}_j \rangle} \text{ — univariate estimators}$$

- Observational data, differently from balanced designed experiments, do not obey orthogonality principle. Maybe we can orthogonalize ....

Suppose $p \geq 2$ ... then

$$\hat{\beta}_1 = \frac{\langle \mathbf{x} - \bar{\mathbf{x}} \mathbf{1}, \mathbf{y} \rangle}{\langle \mathbf{x} - \bar{\mathbf{x}} \mathbf{1}, \mathbf{x} - \bar{\mathbf{x}} \mathbf{1} \rangle}$$

- Step 1: regress $\mathbf{x}$ on $\mathbf{1}$ to produce the residual $\mathbf{z} = \mathbf{x} - \bar{\mathbf{x}} \mathbf{1}$
- Step 2: regress $\mathbf{y}$ on residual $\mathbf{z}$

Multiple regression coefficient $\hat{\beta}_j$ represents the additional contribution of $\mathbf{x}_j$ on $\mathbf{y}$, after $\mathbf{x}_j$ has been adjusted for $\mathbf{x}_0, \mathbf{x}_1, \cdots, \mathbf{x}_{j-1}, \mathbf{x}_{j+1}, \cdots, \mathbf{x}_p$

What if some $\mathbf{x}_k$ is highly correlated to some other $\mathbf{x}_l$ ? (unstable estimators, residual close to zero, Z-scores very small)
Stepwise Selection based on $F$-Tests

- **Forward stepwise selection** starts with the intercept and sequentially adds the most significant variable; stops when no variable not yet included is significant ‘enough’.
  - After an effect has been added, all effects in the current model are checked to see if any of them should be removed.
  - Then the process continues until a stopping criterion is met.
  - The traditional criterion for effect entry and removal is based on their $F$-statistics and corresponding p-values, which are compared with some specified entry and removal significance levels; however, these statistics may not actually follow an $F$-distribution so the results might be questionable.

- **Backward stepwise selection** starts with the full model and sequentially removes the least significant variable; stops when all variables already included are significant ‘enough’.

- **Hybrid stepwise selection** is a mixture of the previous two.
Sparse Linear Regression

Setup of the problem:

\[ y = x\theta^* + w \quad \text{with } \theta^* \text{ sparse} \]

Lasso estimator

\[ \hat{\theta} = \operatorname{arg\ min}_\theta \frac{1}{n} \sum_{i=1}^n (y_i - x_i^T \theta)^2 + \lambda_n \sum_{j=1}^p |\theta_j| \]

Structured (inverse) Covariance Matrices

Setup of the problem:

\[ \mathbf{x} = (x_1, \cdots, x_n) : x_i \in \mathbb{R}^p, \] with sparse \( \text{cov}(\mathbf{x}) = \mathbf{\Sigma} \in \mathbb{R}^{p \times p} \) or \( \text{cov}^{-1}(\mathbf{x}) = \mathbf{\Theta}^* \in \mathbb{R}^{p \times p} \)

Estimator

\[ \hat{\Theta} = \arg \min_{\Theta} \left( \frac{1}{n} \mathbf{x}_i \mathbf{x}_i^T \right) + \lambda_n \sum_{b \in B} \| \Theta_b \|_F \]

Low Rank Matrix Approximation

Setup of the problem:

$$\Theta^* \sim UDV \quad \text{with} \quad U \in \mathbb{R}^{p_1 \times r}, V \in \mathbb{R}^{p_2 \times r} \quad \text{and} \quad r \ll \min p_1, p_2$$

Estimator

$$\hat{\Theta} = \arg \min_{\Theta} \frac{1}{n} \sum_{i=1}^{n} (y_i - \langle x_i, \Theta \rangle)^2 + \lambda_n \min \{p_1, p_2\} \sum_{j=1}^{\min \{p_1, p_2\}} \sigma_j(\Theta)$$

Setup of the problem: Samples from discrete markov random field (Ising or Potts Models)

\[
P_\theta(x_1, \cdots, x_p) = \frac{1}{Z(\theta)} \exp \left\{ \sum_{j \in V} \theta_j(x_j) + \sum_{(j,k) \in E} \theta_{jk}(x_j, x_k) \right\}
\]

Estimator is defined given empirical marginal distributions \(\hat{\mu}_j, \hat{\mu}_{jk}\)

\[
\hat{\Theta} = \arg \min_{\Theta} \sum_{j \in V} E_{\hat{\mu}_j}[\theta_j(x_j)] + \sum_{(j,k) \in E_{\hat{\mu}_{jk}}} [\theta_{jk}(x_j, x_k)] - \log Z(\theta) + \lambda_n \sum_{(j,k)} \|\theta_{jk}\|_F
\]

Sparse Principal Components Analysis

Setup of the problem:

\[ \Sigma = ZZ^T + D \] with leading eigenspace with sparse columns

Estimator

\[ \hat{\Theta} = \arg \min_{\Theta} -\langle \langle \Theta, \hat{\Sigma} \rangle \rangle + \lambda_n \sum_{(j,k)} |\theta_{jk}| \]