Some extensions of lasso

Outline

- Consistency of lasso for model selection
- Adaptive lasso
- Elastic net
- Group lasso

Consistency of lasso for model selection

- A number of authors have studied the ability of the lasso and related procedures to recover the correct model, as N and p grow. (In contrast to low prediction error, which is an easier goal).
- Many of the results in this area assume a "irrepresentability" condition on the design matrix of the form

$$||(\mathbf{X}_{\mathcal{S}}{}^T\mathbf{X}_{\mathcal{S}})^{-1}\mathbf{X}_{\mathcal{S}}{}^T\mathbf{X}_{\mathcal{S}^c}||_\infty \leq (1-\epsilon) \text{ for some } \epsilon \in (0,1]. \qquad (1)$$

  Here $\mathcal{S}$ indexes the subset of features with non-zero coefficients in the true underlying model, and $\mathbf{X}_{\mathcal{S}}$ are the columns of $\mathbf{X}$ corresponding to those features. Similarly $\mathcal{S}^c$ are the features with true coefficients equal to zero, and $\mathbf{X}_{\mathcal{S}^c}$ the corresponding columns.
- This says that the least squares coefficients for the columns of $\mathbf{X}_{\mathcal{S}^c}$ on $\mathbf{X}_{\mathcal{S}}$ are not too large, that is, the "good" variables $\mathcal{S}$ are not too highly correlated with the nuisance variables $\mathcal{S}^c$.

Adaptive lasso

- The *adaptive lasso* uses a weighted penalty of the form $\sum_{j=1}^{p} w_j|\beta_j|$ where $w_j = 1/|\hat{\beta}_j|^{\nu}$, $\hat{\beta}_j$ is the ordinary least squares estimate and $\nu > 0$.
- The adaptive lasso yields consistent estimates of the parameters while retaining the attractive convexity property of the lasso. Idea is to favor predictors with univariate strength, to avoid spurious selection of noise predictores. Can fit via LARS algorithm.
- When $p > N$, can use univariate regression coefficients in place of full least squares estimates.
- Adaptive lasso recovers the correct model under milder conditions than does the lasso

Elastic net

- In genomic applications, there are often strong correlations among the variables; genes tend to operate in molecular pathways. The lasso penalty is somewhat indifferent to the choice among a set of strong but correlated variables.
- The ridge penalty, on the other hand, tends to shrink the coefficients of correlated variables toward each other
- The *elastic net* penalty is a compromise, and has the form

$$\sum_{j=1}^{p} \left(\alpha|\beta_j| + (1-\alpha)\beta_j^2\right). \tag{2}$$

The second term encourages highly correlated features to be averaged, while the first term encourages a sparse solution in the coefficients of these averaged features. Unlike lasso, more than $\min(N, p)$ coefficients can be nonzero. See Figure 18.5 in text

Group lasso

- In some problems, the predictors belong to pre-defined groups; for example genes that belong to the same biological pathway, or collections of indicator (dummy) variables for representing the levels of a categorical predictor.
- In this situation it may be desirable to shrink and select the members of a group together. The *group lasso* is one way to achieve this.

Group lasso- continued

- Suppose that the p predictors are divided into L groups, with p_ℓ the number in group ℓ. For ease of notation, we use a matrix $\mathbf{X}_\ell$ to represent the predictors corresponding to the ℓth group, with corresponding coefficient vector β_ℓ.

- The grouped-lasso minimizes the convex criterion

$$\min_{\beta \in R^p} \left(||\mathbf{y} - \beta_0 \mathbf{1} - \sum_{\ell=1}^{L} \mathbf{X}_\ell \beta_\ell||_2^2 + \lambda \sum_{\ell=1}^{L} \sqrt{p_\ell} ||\beta_\ell||_2 \right), \tag{3}$$

  where the $\sqrt{p_\ell}$ terms accounts for the varying group sizes, and $||\cdot||_2$ is the Euclidean norm (not squared). Since the Euclidean norm of a vector β_ℓ is zero only if all of its components are zero, this procedure encourages sparsity at the group level.

- Standard group lasso algorithm uses coordinate descent, and assumes that the design matrix in each group is orthonormal (not just orthogonal), and uses simple soft-thresholding.
- This is a restrictive assumption– e.g if we have a categorical predictor coded by indicator variables, we can orthonormalize without changing the problem only if the number of observations in each category are the same.
- for general design matrices, computation is more intricate. see "A note on the group lasso and a sparse group lasso" (Friedman et al) on Tibshirani website