Group Lasso Standardization

Stephen Berg

June 21, 2017

The unstandardized group lasso problem solves

$$arg\ min_\beta\ -l(X\beta) + \lambda \sum_{i=1}^{G} \sqrt{p_i} \cdot \|\beta_i||_2$$

where p_i is the number of parameters in group i and β_i refers to the subset in group i of parameters in β.

Simon and Tibshirani recommend solving

$$arg\ min_\beta\ -l(X\beta) + \lambda \sum_{i=1}^{G} \sqrt{p_i} \cdot \|X_i\beta_i||_2$$

where the l2 norm of the predictions is penalized rather than the l2 norm of the coefficients themselves. They show that this is equivalent to solving

$$arg\ min_\theta\ -l(U\theta) + \lambda \sum_{i=1}^{G} \sqrt{p_i} \cdot \|\theta_i||_2$$

where the relationship between U and X is the following: $U_i R_i = X_i$ for all i, and R_i^{-1} orthonormalizes the columns of X_i. Thus $U_i\theta_i = X_i\beta_i = U_i R_i \beta_i$, so $\beta_i = R_i^{-1}\theta_i$.

They do not mention centering or subtracting out the mean of the columns of X before doing the standardization, but this seems appropriate, suggesting the modification

$$arg\ min_\beta\ -l(X\beta) + \lambda \sum_{i=1}^{G} \sqrt{r_i} \cdot \|P_{0^\perp} X_i\beta_i||_2$$

where $P_{0^\perp}$ is the projection of $X_i\beta_i$ onto the orthogonal complement of the unpenalized groups, and r_i denotes the rank of $P_{0^\perp} X_i$. The modification of p_i to r_i is useful in the case of k-level factors. Suppose X_{ij} is a k-level factor, with $j = 1, ..., k$, and the only unpenalized covariate is the intercept. Then

$$X_i = \begin{bmatrix} X_{i1} & X_{i2} & ... & X_{i(k-1)} & \left(1 - X_{i1} - ... - X_{i(k-1)}\right) \end{bmatrix}$$

and

$$P_{0^\perp} X_i = \begin{bmatrix} P_{0^\perp} X_{i1} & P_{0^\perp} X_{i2} & ... & P_{0^\perp} X_{i(k-1)} & \left(-P_{0^\perp} X_{i1} - ... - P_{0^\perp} X_{i(k-1)}\right) \end{bmatrix}$$

which has rank $k-1$. Under the modified penalty with an unpenalized intercept, including all k levels of a factor or including any $k-1$ levels will produce the same model.

To standardize the model with penalty $||P_{0^\perp} X_i \beta_i||$, we need to orthonormalize the columns of $P_{0^\perp} X_i$ for each penalized group i. This can be done in the following way. Let

$$Q_i D_i Q_i^T = X_i^T P_{0^\perp}^T P_{0^\perp} X_i = X_i^T P_{0^\perp} X_i.$$

Then the columns of

$$P_{0^\perp} X_i Q_i D_i^{-1/2}$$

are orthonormal.

It remains to efficiently compute $X_i^T P_{0^\perp} X_i$, particularly in the case where X is sparse and X_0 is low rank. Suppose $X_0 = U_0 R_0$, where the columns of U_0 are orthonormal and R_0 is a $p_0 \times p_0$. Then

$$P_{0^\perp} = diag(n) - U_0 U_0^T = diag(n) - X_0 R_0^{-1} R_0^{-1^T} X_0^T$$

and

$$X_i^T P_{0^\perp} X_i = X_k^T X_k - (X_k^T X_0)(R_0^{-1} R_0^{-1^T})(X_0^T X_k)$$

where the last term is parenthesized in a way that takes advantage of both the possible sparsity of X (to compute $X_k^T X_0$ rather than $P_{0^\perp} X_0$) and the low rank of X_0 ($R_0^- 1 R_0^- 1^T$ is $p_0 \times p_0$).