Consensus Selection for $\ell_1$-regularized models

Logan Grosenick

EE364b: Convex Optimization II Class Project

Problem

Methods that use an $\ell_1$-norm to encourage model sparsity—such as the lasso, total variation denoising, soft thresholding, and compressed sensing—are now widely applied in a variety of fields. However, it remains unclear how to sensibly combine several such models resulting from multiple fits to resampled data. Because resampling approaches have been shown to be of great utility in reducing model variance and improving model selection, a method able to generate a single solution from multiple fits to resampled data is desirable.

Solution: Consensus Selection with ADMM

Consensus models have been proposed in cases where a very large number of training examples require fitting a single problem in a distributed way [BPC+11]. However, one can use the same procedure to arrive at a single consensus model for many $\ell_1$-regularized fits to non-independent samples of the same data (generated using subsampling, bootstrapping, or cross-validation, for example). A convenient and efficient means of fitting such models in the Alternating Direction Method of Multipliers (ADMM) [BPC+11].

Case study: Consensus Selection Lasso (CSL)

Consider the simple data generating function

$$y = X^T \beta + \epsilon, \epsilon \sim \mathcal{N}(0, I), i = 1, \ldots, n,$$

with data $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times p}$, and model coefficients $\beta \in \mathbb{R}^p$.

The lasso [Tib96] solves the estimation problem

$$\hat{\beta} = \arg \min \{ \frac{1}{2} \| y - X \beta \|^2_2 + \lambda \| \beta \|_1, \quad \beta \in \mathbb{R}^p \},$$

a bi-criterion problem that trades off a least-squares fit to the data with an $\ell_1$-norm penalty on the model coefficients. The latter is a heuristic for encouraging model sparsity, as it is a convex relaxation of the $\ell_0$-norm. The consensus is the average of the $\ell_1$-regularized fits to non-independent samples of the same data (generated using subsampling, bootstrapping, or cross-validation, for example).

The ADMM algorithm for the consensus selection lasso is

$$\beta_{k+1}^j = \left( X_j^T X_j + \rho I \right)^{-1} \left( X_j^T y_j + \rho \left( \beta_k^j - \bar{\beta}_j \right) \right),$$

$$\bar{\beta}_j = \frac{1}{m} \sum_{i=1}^m \beta_{k+1}^j,$$

where $\bar{\beta}_j$ is the average of the $\ell_1$-regularized fits to non-independent samples of the same data (generated using subsampling, bootstrapping, or cross-validation, for example).

CSL performance

Consensus selection for the lasso was compared to two recent methods for fitting and model selection in $\ell_1$-regularized problems: the median probability model [BB04] and Stability Selection [MB10]. Performance was measured by (i) correct variable selection and (ii) generalization performance on new (out-of-sample) data.

Problem instances each with $n = 200$ observations on $p = 1000$ correlated variables were generated according to the model (1), with $X_i \sim \mathcal{N}(0, \Sigma)$ and $\Sigma$ generated from an autoregressive process such that $\Sigma_{i,j} = 0.8^{\mid i-j \mid}$. All variables were drawn with replacement and fit over a grid of 100 $\lambda$ values. The median probability model and stability selection methods used the glmnet package in R to fit the lasso. The models were then given by:

- Median probability model: take the median of the coefficients over the fits
- Stability Selection: choose only those coefficients that have a maximum probability of appearing in the model of more than threshold $\tau$ over the 50 fits. Refit the resulting model to only the chosen variables. The results shown here used $\tau = 0.6$.

Regularization in ADMM

Interpretations of the CSL ADMM algorithm

1. The first two ADMM steps (3) and (4) are a Tikhonov (or "ridge") regression [Tik43] and a soft-thresholding step, respectively. Together these are equivalent to elastic net regression [Zhi05].
2. $\beta = \frac{1}{m} \sum_{i=1}^m \beta_i$ is an average of bootstrapped estimates, also known as a "bagged" estimate of $\beta$ [Bre96], and thus is a (reduced variance) approximation to the mean of the posterior distribution.
3. Consensus selection with ADMM (i) aggregates MAP fits in order to approximate the posterior mean with prior $P(\beta) \sim \mathcal{N}(0, \rho I)$, then sequentially estimates $\beta_k$ with Gaussian prior $P(\beta_k) \sim \mathcal{N}(\mu_k, \varphi_k I)$, so that each estimate is biased towards the difference between the most recent sparse fit and the integrated residual $\rho_k$.
4. The combination of the prior in update (3) and the addition of the bagged dual variable in update (4) mean that the elements of $\beta$ are differentially adjusted before thresholding. This is reminiscent of reweighting or "adaptive" $\ell_1$ methods.

Conclusions

In the case of the lasso, consensus selection is competitive with state of the art methods for model selection, and yields fits that generalize well to new data. It admits several interpretations that relate the stages of the algorithm to existing estimation methods, and that help explain its positive performance on sparse, correlated data. These results are easily extensible to more general $\ell_1$-penalized models, and in cases where it is possible to split the problem across variables, the ADMM formulation allows scaling to very large problems [BPC+11].

References