Shrinkage in Sparse Linear Regression

Bryan Liu and Jake A. Soloff

Copas 1983 showed the benefit of Stein-type shrinkage in the context of ordinary linear regression. In ordinary least squares, the magnitude of the predictions were found to be too large on average for test data, and hence shrinkage was found to be beneficial for improving prediction mean squared error.

We are interested in investigating whether these results extend to the case of sparse linear regression. Specifically, suppose we optimize the LASSO objective, instead of the least squares objective. Do we still observe the necessity to shrink our predictions? Or perhaps LASSO over-penalizes and we need to instead *inflate* our predictions? Certainly the answers depend on how we choose the regularization parameter λ: if we consider the extreme cases, $\lambda = 0$ puts us back in the context of OLS (where predictions are too big), whereas when $\lambda \uparrow \infty$ all predictions shrink to zero. We will explore a several typical methods for choosing λ such as cross-validation, AIC, or BIC, and study the relationship between the predicted and actual response on validation data.

Next, we propose a two-step shrinkage procedure. In the first step, we optimize the LASSO objective (and choose λ in some way), and this returns a subset of active covariates. We then run OLS using only this active subset of features, and apply shrinkage post-selection as proposed by Copas.

We will compare the predictive accuracy of OLS, OLS with Copas shrinkage, LASSO, and our proposed two-step shrinkage on simulated data by drawing responses from a sparse linear regression model. We will experiment with the sparsity of the model, as well as the signal-to-noise ratio. In particular, it will be interesting to compare results in the regime in which LASSO can correctly recover the active features to the regime in which in cannot. We then further apply these techniques to the classic SPAM data set where 57 features are used to predict whether an email is SPAM or not SPAM (HAM).

As a natural progression, we also investigate these questions in the context of multiple response linear regression. Instead of treating each response as an independent regression problem, Breiman and Friedman 1996 proposed a method to reduce prediction errors by "pooling" the predictions across multiple responses. This paper again fit in the context of ordinary linear regression, and we propose to investigate the extent to which these ideas can be extended to the the sparse regression setting.

As a concrete application, consider the problem of predicting the fMRI response to natural images for $q = 20$ different voxels in the primary visual cortex (note–this application was the subject of the stat215a final project). A single subject observed a total of $n = 1750$ images, each of shape 128×128. After pre-processing each image using a Gabor wavelet pyramid, $p = 10921$ covariates remain. We have two goals in this problem:

- Find the retinal field of a voxel or a collection of voxels, and
- Predict the response of a voxel to a given image.

In the multi-response linear model, we write $Y = X\beta + \epsilon$, where

- $Y, \epsilon \in \mathbb{R}^{n \times q}$ are the matrix of (fMRI) responses and noise term, respectively,
- $X \in \mathbb{R}^{n \times p}$ is the matrix of predictors (Gabor filter coefficients), and
- $\beta \in \mathbb{R}^{p \times q}$ is the matrix of regression coefficients

Since the support of most of the Gabor filters are fairly isolated *and* a single voxel in the primary visual cortex is understood to only respond to an isolated region of the visual field, it is plausible to suggest that each vector $\beta_{\cdot,k} \in \mathbb{R}^p$, for $k = 1, \ldots, q$ is *sparse*, so one option for estimating β is to solve q LASSO problems independently. If we then see a new image $x \in \mathbb{R}^p$ and predict $\hat{y} = x\hat{\beta}_{\text{LASSO}}$, we wonder if it is possible to obtain a more accurate predictor with $\tilde{y} = B\hat{y}$. As in the single-response case of shrinkage after LASSO, it's no longer clear whether B should shrink predictions. The cross-validation approach for selecting B in Breiman and Friedman 1996 seems to generalize to this context. Our hope is that our earlier exploration of shrinkage in the single-response sparse linear regression can guide our attempts to share information across voxels in this problem.

References:

Copas, J B. 1983. Regression, Prediction and Shrinkage. *Journal of the Royal Statistical Society.* Vol 45, No 3, pp 311-354.

Breiman L, Friedman J H. 1997. Predicting Multivariate Responses in Multiple Linear Regression. *Journal of the Royal Statistical Society.* Vol 59, No 1, pp 311-354.