BAYESIAN STATISTICS 9,
J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid,
D. Heckerman, A. F. M. Smith and M. West (Eds.)
© Oxford University Press, 2010

Subordinators, Adaptive Shrinkage and a Prequential Comparison of Three Sparsity Methods

B. Clarke & C. Severinski
University of Miami, USA *Unaffiliated, Canada*
bclarke2@med.miami.edu cody.severinski@telus.net

Summary

Theorem 3 of Polson and Scott appears to generalize to include adaptive shrinkage methods which often have the oracle property. However, as effective as sparsity methods can be for certain ranges of sample size and number of terms (in an additive model), initial examples suggest shrinkage methods do not perform well prequentially when $n \geq p$.

Keywords and Phrases: Prediction, Sparsity

1. THE PS CLASS AND ADAPTIVITY

Penalized methods have been around for years. Roughly, the selection of the error term corresponds to choosing the likelihood while the selection of a penalty corresponds to choosing the prior. or instance, LASSO corresponds to assuming each y_i is an independent outcome of $Y_i \sim N(x_i\beta, \sigma^2)$ where $\beta = (\beta_1, \dots, \beta_p)$, $\sigma > 0$, and the the design points x_i are p-dimensional. The penalty term therefore corresponds to the product of p independent double exponential densities with shape parameter λ. Other shrinkage penalties have analogous interpretations and different penalties will favor different regression functions.

Polson & Scott (2010, Section 5.1) asks: "What are we assuming about β when we use a penalty function?" They answer this question by representing a class of penalties in terms of stochastic processes called subordinators. Thus nonzero β's represent jumps. This seems to be the first time that a whole class of penalties has been characterized and Polson & Scott (2010) show a correspondence among estimators, priors, penalties, subordinators, and mixtures of normals for this class.

Now recall that the oracle property, see Donoho & Johnstone (1994). This requires consistent variable selection and asymptotically optimal estimation of the

B. Clarke is Professor of Statistics in Dept. of Medicine, Center for Computational Sciences, and the Dept. of Epi. and Public Health at the University of Miami. C. Severinski just finished his Master's degree. No grant was awarded to do this job.

parameters in the correct model. Some penalized methods have the oracle property, some don't. The literature reveals that often, but not always, the difference between a version of a penalized method that has the oracle property and a version that does not is the property of 'adaptivity'. Roughly, 'adaptive' refers the inclusion of one decay parameter for each parameter of interest. For instance, the LASSO penalty is $\lambda \sum_{j=1}^{p} |\beta_j|$ and the Adaptive LASSO penalty is $\sum_{j=1}^{p} \lambda_j |\beta_j|$. Analogous changes for Elastic Net and COSSO also give the oracle property. However, SCAD is oracle because of the form of the penalty function and only requires a single decay parameter. Nevertheless, it seems less common to get the oracle property without adaptivity than with it.

It is seen in Polson & Scott (2010, Example 2, Section 4.1) that LASSO is in the Polson-Scott class but that Theorem 3 as stated does not include adaptive penalty methods. However, the idea of the proof of Theorem 3 (and discussions with the authors) suggest some form of the following conjecture may be provable.

Theorem 1 (Conjectured extension of Theorem 3 to adaptivity). *Let T_s be a subordinator (cadlag, stationary, independent increment process), $s \in [0, \nu]$, Laplace exponent $\psi(t)$, and marginal g at time ν and consider p more independent subordinators $T_{j,s}$ with $s \in [0, \nu]$, Laplace exponents ψ_j and marginals g_j. Suppose $T_s = T_{1,s_1} + \ldots + T_{p,s_p}$, where each $s_j = s_j(s)$ is an increasing function. Then, the cgf of T_s leads to the penalty*

$$w(\beta, \nu) = \nu \sum_{j=1}^{p} s_j \psi_j(\beta_j^2).$$

Moreover, if g is the marginal for T_ν and the $T_{j,s_j(\nu)}$'s are integrable where the g_j's are the marginals for the $T_{j,s_j(\nu)}$'s, then the penalized LSE is the posterior mode under the prior:

$$p(\beta_j) \propto e^{-\psi_j(\beta_j^2)} = \int_0^\infty N(\beta_j | 0, T_{j,s_j(\nu)})^{-1}) \left[T_{j,s_j(\nu)}^{-1} g(T_\nu) \right] dT_{j,s_j(\nu)}.$$

The idea of the proof is to mimic the earlier proof in Polson and Scott but to evaluate the the subordinators for each parameter at different times. It is possible that a linear combination of the subordinators for each parameter would also give a form of the result. If some version of this conjecture is true, then we suggest a similar modification of Theorem 4 in Polson & Scott (2010) can be found for adaptive penalties.

2. PREDICTIVE COMPARISON

Despite the theory and the new representation of penalty terms in terms of stochastic processes, it is urgent to ask what the 'sparse' models found by penalized methods are good for. The answer seems to be: They are sometimes good for model identification but rarely for prediction at least when $p \gg n$.

Penalized methods seem to scale up better than branch-and-bound when a true model really is sparse i.e., has few non-zero terms relative to both p and n. Indeed, these models essentially never include the case $p \gg n$ since the oracle property always requires an assumption like $p = \mathcal{O}(n^{1-\alpha})$ for some $\alpha \geq 0$. Even when $p = \mathcal{O}(n^{1-\alpha})$, it is unclear how to obtain SE's for parameters set equal to zero let alone other assessments of model uncertainty. This is important because penalized methods combine model selection and parameter estimation in one procedure.

An important point seen in the graphs below is that even when a sparse model is 'pretty good', predictive performance need not be. This is unsurprising because model identification and prediction are usually conceptually disjoint goals. Even worse, the contexts where sparse methods are used usually do not satisfy the hypothesis that the true model really is sparse. So, we can be quite sure that the 'sparse' model will neglect terms that contribute predictively but cannot be identified with the existing n for the chosen p or may indeed be 'crowded out' by the variables already included. This principle is dramatized in van der Linde (2010, Examples 4.6.1 and 4.6.2). She notes that even when a model that fits a test set of data well is found it can be predictively poor: Other models may fit equally well and be as physically plausible.

To illustrate this, consider the simple signal plus noise model

$$Y_i = f(X_i) + \epsilon_i \quad \epsilon_i \sim N(0, \sigma^2) \quad X_i \sim \mathrm{Unif}[a, b] \quad i = 1, \ldots, n, \tag{1}$$

where the draws of X_i and ϵ_i are independent. Let us examine the predictive performance of penalized methods when f is chosen to be one of three functions $-\log(1-x)$, the tooth given by $x + 9/(4\sqrt{2\pi})e^{-16(2x-1)^2}$, and a normalized version of the Mexican hat proportional to $(1 - x^2/\sigma^2)e^{-x^2/2\sigma^2}$ with $\sigma = .2$. Note that these three functions are in order of increasing difficulty. Now penalized methods we will compare are Ridge Regression, LASSO, the SCAD penalty and their stacking average, Wolpert (1992). Note that Ridge is the least sophisticated shrinkage method while LASSO (more sophisticated) is included in Theorems 3 and 4 in Polson & Scott (2010) while SCAD is the most sophisticated. Here, it is taken as a proxy for ALASSO (for which we conjecture an extension of Theorems 3 and 4 hold) because the software for ALASSO was difficult to use.

Our goal is to give a prequential comparisons of the predictors these four penalized methods generate, see Dawid (1984). To this end, the class of models we will use to approximate (1) consists of elements of the Legendre basis or the Fourier basis. These test function–basis pairs are an M-open setting.

The simulation results shown in Figure 1 show a single run of $n = 30$ data points. Using a burn in of 10 points, X_{11} was generated. Then, $f(x_{11})$ was added to the outcome generated by ϵ_{11} for form y_{11}. Then, using only $x_1, \ldots, x_{11}, y_1, \ldots, y_{11}$ $\hat{f}_{11}$ was found using one of the three penalized methods with the first p elements of one of the bases (plus the constant term) using R packages (lars for lasso, SIS for SCAD; ridge was coded from scratch); the decay parameters were determined within the package or by add-on functions as needed. Then x_{12} was generated and $\hat{y}_{12} = \hat{f}(x_{12})$ was found where $\hat{f}$ was the estimated regression function. Then, $Y_{12} = y_{12}$ was generated and the process continued up to $n = 30$.

Figure 1 shows six plots of single runs of this procedure for the three functions and two choices of basis elements for $n = 30$ and $p + 1 = 51$ (The $+1$ corresponds to the constant term; for the Fourier basis we used the first 25 sine / cosine pairs). For $-\log(1-x)$, $(a, b) = (-1, 1)$ and $\sigma^2 = 1/4$. For tooth, $(a, b) = (0, 1)$ and $\sigma^2 = 1/25$, and for Mexhat $(a, b) = (-3, 3)$ and $\sigma^2 = 25$. In each case, the value of σ^2 was chosen based on the local and global behavior of the function. Simulations with comparable values of σ^2 yielded comparable results. It is seen that the Legendre basis gives a clearly better fit than the Fourier basis for $-\log(1-x)$ because all methods miss the rise in $-\log$ for Fourier. The Legendre and Fourier basis give

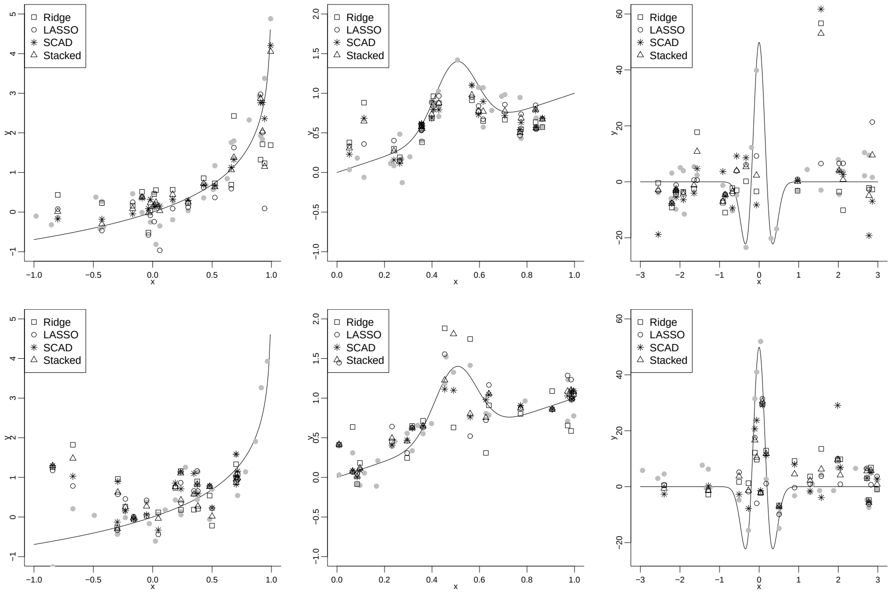

Figure 1: In each panel, the grey dots are a scatterplot of the data for the function indicated. The first column is $-\log(1-x)$, the second column is tooth and the third column is MexHat. The panels in the top row use the Legendre basis and the panels in the bottom row use the Fourier basis.

roughly comparable (and poor) fits for tooth because none of the methods really captures the shape of the tooth: Legendre misses the peak and Fourier just gives increased scatter. For Mexhat, Fourier seems to give a better fit because all four methods detect the central mode, though none of them detect the secondary modes.

In Figure 2 we see the aggregate behavior of the four methods for 20 runs. The first, third and fourth rows show the cumulative average MSE for the $-\log(1-x)$, tooth, and MexHat functions. It is seen that as n increases the MSE curves level off. The second row shows a bias-variance decomposition for the MSE averaged over sequences of selections of both X_k and ϵ_k for $k \leq i$ for the $-\log(1-x)$ function under the Legendre basis from the upper left panel. Specifically, the second row shows plots of the average predictuals for each time step on the left i.e., $(1/20)\sum_{j=1}^{20}(y_{i,j} - \hat{y}_{i,j})$ for each $i = 1, \ldots n$ and the average SD of the predictions on the right, i.e., $(1/19)\sum_{j=1}^{20}(y_{i,j} - \bar{\hat{y}}_i)^2$ where $\bar{\hat{y}}_i = (1/20)\sum_{j=1}^{20}\hat{y}_{i,j}$ for each $i = 1, \ldots, n$ is the average over the predictions made at the i-th time step. It is seen that the SCAD and Stacking SD curves are routinely the lowest indicating the least variability and that the LASSO and Ridge average predictual curves usually exhibit the highest and lowest values. Thus, in this case, Stacking and SCAD appear to do best in terms of smallest variance and in terms of smallest bias. Similar bias-variance decompositions can be done for the other five cases. We also verified that

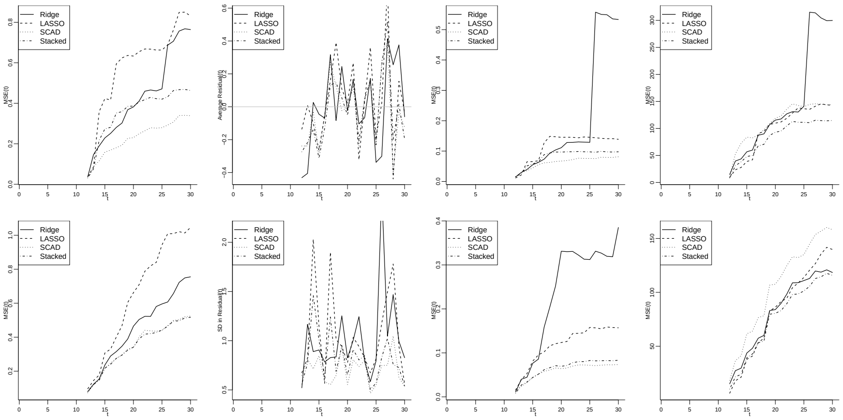

Figure 2: The first column show the MSE for the four methods for 20 runs for the $\log(1-x)$ function using the Legendre basis (top) and the Fourier basis (bottom). The second column of panels shows the plot of the predictuals and the SD's averaged over the 20 runs. The third and fourth columns are the same as the first but for the tooth and Mexhat functions respectively.

the average coefficients of the three terms (Ridge, LASSO, and SCAD) in Stacking were generally all non-zero, although there were also a few cases where the weight Stacking put on SCAD increased to one and the other weights decreased to zero with n. The results are summarized in Table 1 where an asterisk indicates that two methods are indistinguishable.

Table 1: *Ranking of four methods for three functions.*

Funct.	basis	order	Funct.	basis	order	Funct.	basis	order
-log	Leg.	SCAD Stack RR LASSO	tooth	Leg.	SCAD* Stack* LASSO RR	Mex	Leg.	Stack LASSO* SCAD* RR
-log	Four.	Stack* SCAD* RR LASSO	tooth	Four.	SCAD* Stack* LASSO RR	Mex	Four.	Stack* RR* LASSO SCAD

3. CONCLUSIONS

It is seen that SCAD was the best single method. This is not a surprise since SCAD was the only method with the oracle property method. Overall, Stacking performed as well as SCAD.

As function complexity went up, fit and prediction deteriorated. Moreover, a look at the scatterplots of the methods suggests none of the methods perform well prequentially. This is corroborated by the bias-variance analysis which show that the biases are an appreciable proportion of the range of the functions and the SD's are often quite large (relative to the range of the function). Indeed, in some cases where SCAD performed well, it did so by ignoring the peaks and troughs of the function and only capturing the flat portions of the function well. We admit that using a spline basis or a wavelet basis might be better able to model local modes that Legendre or Fourier.

Essentially, our results suggest that sparse methods may only be prequentially good in the rare case that a sparse model really is true. Otherwise put, we should not expect the models obtained from sparse methods – even good methods with the oracle property – to perform well predictively without further validation.

REFERENCES

Dawid, AP. 1984. Present position and potential developments: Some personal views: Statistical theory: The prequential approach. *Journal of the Royal Statistical Society. Series A (General)*, **147**(2), 278–292.

Donoho, D.L., & Johnstone, J.M. 1994. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, **81**(3), 425.

Polson, Nicholas G., & Scott, James G. 2010. Shrink Globally, Act Locally: Sparse Bayesian Regularization and Prediction. *In:* Bernardo, J.M., Bayarri, M.J., Berger, J.O., Dawid, A.P., Heckerman, D., Smith, A.F.M., & West, M. (eds), *Bayesian Statistics 9*. Oxford University Press.

van der Linde, A. 2010. Reduced rank regression models with latent variables in Bayesian functional data analysis. Submitted to *Bayesian Analysis*.

Wolpert, D.H. 1992. On the connection between in-sample testing and generalization error. *Complex Systems*, **6**(1), 47.