Penalized Utility Bayes Estimators

P. Richard Hahn and Rob McCulloch (Chicago Booth)
Carlos M. Carvalho and David Puelz (UT Austin)

FSU, September 2015
Decoupling Shrinkage and Selection

This is a talk about variable selection.

- “Decoupling shrinkage and selection in Bayesian linear models: A posterior summary perspective.” JASA 2015

- “Optimal ETF selection for passive investing.”

- “Variable selection in nonlinear regression with penalized predictive utility.”
Three reasons for variable selection

- Because we’re scientists and we test hypotheses!
- Because fewer variables are faster to compute with!
- Because thinking hard about fewer things is easier than thinking hard about more things.

There are probably others. The important point is that they are distinct reasons.
Consider the (Bayesian) linear model with Gaussian errors:

\[ Y = X\beta + \epsilon, \]
\[ \epsilon \sim N(0, \sigma^2 I). \]  

(1)

for a given prior \( \pi(\beta, \sigma^2) \).

How do the three motivations for sparsity manifest in this canonical setting?
This is my motivation for sparsity

Coefficients:    Estimate  Std. Error  t value  Pr(>|t|)
(Intercept)     -8.211e+06  8.519e+04  -96.383  < 2e-16 ***
X.trim350       2.657e+04   1.160e+04    2.290   0.022024 *
X.trim400      -1.275e+04  1.649e+04   -0.773   0.439459
X.trim420       4.954e+04  1.178e+04    4.205   2.62e-05 ***
X.trim430       2.662e+04  1.177e+04    2.263   0.023662 *
X.trim500       2.935e+04  1.177e+04    2.494   0.012623 *
X.trim550      -4.942e+03  1.078e+04   -0.458   0.646705
X.trim55 AMG    2.823e+04  1.178e+04    2.397   0.016542 *
X.trim600      -4.777e+03  1.079e+04   -0.415   0.678100
X.trim63 AMG    4.445e+04  1.080e+04    4.117   3.84e-05 ***
X.trim65 AMG    6.142e+03  1.080e+04    5.677   0.505024
X.trimunsp      2.666e+04  1.081e+04    2.466   0.013657 *
X.conditionNew  3.513e+04   2.284e+02  153.819  < 2e-16 ***
X.conditionUsed 3.513e+04   2.284e+02  153.819  < 2e-16 ***
X.isOneOwner    -5.043e+02   1.725e+02   -2.924   0.003459 **
X.mileage      -1.324e-01   2.522e-03   -52.488  < 2e-16 ***
X.year          4.103e+03   4.224e+01    97.134  < 2e-16 ***
X.colorBlack   -4.381e+02   6.660e+02   -0.658   0.510685
X.colorBlue     6.830e+02   7.000e+02    0.976   0.329230
X.colorBronze   3.997e+03   3.460e+03    1.155   0.247937

Residual standard error: 10740 on 39391 degrees of freedom
Multiple R-squared:  0.9429, Adjusted R-squared:  0.9428
F-statistic:  9706 on 67 and 39391 DF,  p-value:  < 2.2e-16
Our third view of variable selection can be motivated by a utility function that balances predictive ability and parsimony:

\[ \mathcal{L} \left( \tilde{Y}, \gamma \right) = n^{-1} || \tilde{Y} - X\gamma ||_2^2 + \lambda ||\gamma||_0. \]

Think about a situation where in order to predict, you have to pay for the covariate values.
The role of the posterior distribution

After we have seen data $Y$, our expected utility can be expressed in terms of the posterior distribution over parameters $\theta = (\beta, \sigma^2)$.

$$E_\theta \mathcal{L}(\theta, \gamma) = E_\theta E_{\tilde{Y} | \theta} \lambda \| \gamma \|_0 + \| \tilde{Y} - X \gamma \|_2^2,$$

$$= \lambda \| \gamma \|_0 + E_\theta E_{\tilde{Y} | \theta} \| X \beta + \tilde{\epsilon} - X \gamma \|_2^2,$$

$$= \lambda \| \gamma \|_0 + E_\theta E_{\tilde{\epsilon} | \sigma^2} \tilde{\epsilon}^t \tilde{\epsilon} + 2 \tilde{\epsilon}^t (X \beta - X \gamma) + \| X \beta - X \gamma \|_2^2,$$

$$= \lambda \| \gamma \|_0 + n \bar{\sigma}_2 + \| X \bar{\beta} - X \gamma \|_2^2 + \text{constant}.$$

The past data enters into our utility consideration by defining the posterior predictive distribution.
The DSS expected loss can be written as:

$$E_{\theta} \mathcal{L}(\theta, \gamma) = \lambda ||\gamma||_0 + ||X\bar{\beta} - X\gamma||^2_2.$$ 

The counting penalty $||\gamma||_0$ is hard to handle computationally...as an initial approximation, we consider the 1-norm instead.

$$E_{\theta} \mathcal{L}(\theta, \gamma) = \lambda ||\gamma||_1 + ||X\bar{\beta} - X\gamma||^2_2.$$
1. Obtain \( \pi(\beta, \sigma^2 \mid Y) \) and calculate \( \hat{Y}_{Bayes} = X\bar{\beta} \).
   (Use any prior you want.)

2. Feed \( \hat{Y}_{Bayes} \) to lars to extract our optimal action, i.e. the sparse summary:

\[
\beta_\lambda \equiv \arg \min_\gamma \lambda \|\gamma\|_1 + n^{-1} \|X\bar{\beta} - X\gamma\|_2^2.
\]

However, the devil is in the details...


Posterior summary plots

How much predictive deterioration is a result of sparsification?

Practically, how different is $\beta_\lambda$ from the optimal $\bar{\beta}$?

Variation Explained

$$\rho^2_\lambda = \frac{n^{-1}||X\beta||^2}{n^{-1}||X\beta||^2 + \sigma^2 + n^{-1}||X\beta - X\beta_\lambda||^2}$$

Excess Error

$$\psi_\lambda = \sqrt{n^{-1}||X\beta_\lambda - X\beta||^2 + \sigma^2 - \sigma}.$$ 

These quantities are random variables!
“Find the smallest model such that with probability (blank) I give up less than (blank) in predictive ability.”
DSS vs. MPM

<table>
<thead>
<tr>
<th></th>
<th>DSS(5)</th>
<th>DSS(4)</th>
<th>DSS(3)</th>
<th>DSS(2)</th>
<th>DSS(1)</th>
<th>MPM (Robust)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>0.08</td>
</tr>
<tr>
<td>Sex</td>
<td>●</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>0.98</td>
</tr>
<tr>
<td>BMI</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>0.99</td>
</tr>
<tr>
<td>MAP</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>–</td>
<td>–</td>
<td>0.99</td>
</tr>
<tr>
<td>TC</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>0.66</td>
</tr>
<tr>
<td>LDL</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>0.46</td>
</tr>
<tr>
<td>HDL</td>
<td>●</td>
<td>●</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>0.51</td>
</tr>
<tr>
<td>TCH</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>0.26</td>
</tr>
<tr>
<td>LTG</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>●</td>
<td>–</td>
<td>0.99</td>
</tr>
<tr>
<td>GLU</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>–</td>
<td>0.13</td>
</tr>
</tbody>
</table>

Similar, but not the same.
Q: Why not just run lasso and be done with it?

A:

1. For what $\lambda$?

2. By “preconditioning”, we tend to get sparser solutions.

3. We don’t have to use the same design matrix as the observed data.
Q: Won’t you get unwanted “double shrinkage”?  

A: By using a version of the adaptive lasso (using posterior draws to determine the adaptive penalty) we appear not to.

\[
\beta_\lambda \equiv \arg \min_\gamma \sum_j \frac{\lambda}{|\beta_j|} |\gamma_j| + n^{-1} \|X\bar{\beta} - X\gamma\|_2^2. \tag{2}
\]
Minimal double shrinkage

We actually observe “re-inflation” ($\ell_0$ approximation in action).
Q: Can you prove anything about it?

A:

1. If $\lambda_n \to 0$ and the posterior converges, then sure.

2. For $\lambda > 0$ fixed, we are proudly inconsistent for $\beta$!

3. We conjecture that penalized Bayes estimators have super-convergence properties.
Q: Can you do anything other than linear models?

A: Glad you asked!
Logistic regression

Starting with

\[ L(\tilde{Y}, \gamma) = \lambda \| \gamma \|_0 - n^{-1} \log \left[ f(\tilde{Y}, X, \gamma) \right], \]

\[ = \lambda \| \gamma \|_0 + n^{-1} \sum_{i=1}^{n} \left( \tilde{Y}_i X_i \gamma - \log (1 + \exp (X_i \gamma)) \right), \]

taking expectations gives

\[ L(\gamma) = \lambda \| \gamma \|_0 + n^{-1} \sum_{i=1}^{n} (\bar{\pi}_i X_i \gamma - \log (1 + \exp (X_i \gamma))), \]

where \( \bar{\pi}_i \) is the posterior mean probability that \( \tilde{Y}_i = 1. \)
Logistic regression excess error

We can also generalize our definition of excess error

\[ \psi_\lambda = \sqrt{n^{-1} \mathbb{E} \left( \| \tilde{Y} - \hat{Y}_\lambda \|^2 \right)} - \sqrt{n^{-1} \mathbb{E} \left( \| \tilde{Y} - \mathbb{E}(\tilde{Y}) \|^2 \right)} \]

\[ = \sqrt{n^{-1} \sum_i \pi_i - 2\pi_\lambda,i \pi_i + \pi^2_{\lambda,i} - \sqrt{n^{-1} \sum_i \pi_i (1 - \pi_i)}} \]

in a way that applies to GLMs.
Using the multivariate normal log-likelihood gives an optimization problem

\[
\mathcal{L}(\tilde{X}, \Gamma) = \lambda \|\Gamma\|_0 - \log \det(\Gamma) - \text{tr}(n^{-1}\tilde{X}\tilde{X}'\Gamma)
\]

\[
E \left( \mathcal{L}(\tilde{X}, \Gamma) \right) = \mathcal{L}(\Gamma) = \lambda \|\Gamma\|_0 - \log \det(\Gamma) - \text{tr}(\tilde{\Sigma}\Gamma)
\]

that can be approached with the graphical lasso technology.
ETFs for passive investing

Given past returns on a set of target asset, \( \{R_j\}_{j=1}^q \), and a collection of exchange traded funds (ETFs), \( \{X_i\}_{i=1}^p \), write:

\[
R_j = \beta_{j1}X_1 + \cdots + \beta_{jp}X_p + \epsilon_j, \quad \epsilon_j \sim N(0, \sigma^2).
\]

In real life \( p \approx 200 \). Do we need this many to recapitulate the co-movement of the target assets?
Posterior parsimony plot

Model size refers here to edges in the precision matrix.
Nonlinear DSS

In the linear case we had:

\[
E_{\tilde{Y}} \mathcal{L}(\tilde{Y}, \gamma) = \lambda \|\gamma\|_0 + \|X\bar{\beta} - X\gamma\|_2^2
\]

For nonlinear regression with additive error we similarly have

\[
E_{\tilde{Y}} \mathcal{L}(\tilde{Y}, \gamma) = \lambda \text{dim}(\gamma) + \|\hat{f}(X) - \gamma(X)\|_2^2
\]

So we need to 1) fit a nonlinear Bayesian regression and 2) define an “action space” for \(\gamma\).
1. Inference: Choose a model for $f(X)$. Do this once with all the $X$’s. (fit)

2. Decision/summary: Use utility (you want a simple model) to find simple models that approximate $f(X)$ well. (fit the fit)

We need fast, flexible ways to fit the fit.

Inference is all in step 1 — no worries about over-fitting!
Bayesian Additive Regression Trees represent an unknown function as a sum of regression trees: $f(X) = \sum_{l=1}^{L} g_l(X)$. 

![Diagram of BART detour](image-url)
Each individual tree defines a step function.

In the case of binary responses, a link function can be used and the levels in each partition correspond to probabilities.
Sums of many trees result in step functions with very fine partitions.

The BART model is computationally attractive compared to comparable (Bayesian) non-linear regression methods, such as Gaussian processes.

Ensembles of trees are slowly making their way into more traditional (not purely predictive) statistical settings, such as causal inference, e.g. “Bayesian nonparametric modeling for causal inference” (Hill, 2012).
We use huge regression trees

Are you *sure* we don’t have to worry about over-fitting?
Predicting used car prices

We have a database of $\approx 20K$ used car sales transactions. The sales price $Y$ is a nonlinear function of vehicle attributes $X$, such as engine size, MSRP (new), mileage, color, trim level, date of sale, price of gas, etc.

In all, we have upwards of 100 predictor variables.

Can we do vehicle pricing with a much smaller number of attributes?
Suppose that you have 25 variables and I provide you with a specific 3-variable model that gets you within 5% the predictive performance of the true (possibly non-sparse) model, with 90% posterior probability.

Do you care that it isn’t the very best among all 3-variable models?

On the other hand, if I search (heuristically) and find no such 3-variable model, it doesn’t mean there isn’t one, just that my heuristic didn’t find it.
Perhaps statistical uncertainty helps our cause...it makes finding a good-enough model easier.

Provided we took great care in our initial modeling, and take seriously our posterior uncertainty, we can use that uncertainty to ease the burden of our optimization step.

The utility function then effectively serves to “break ties”. People have been using priors (improperly) in this role for years.
Take-aways

- Utility functions can enforce inferential preferences that are not prior beliefs. (statistical versus practical significance)

- Optimization theory can be used in the service of posterior summary.

- Posterior draws can be used to benchmark candidate summaries. (satisficing)

- There are interesting open problems about other statistical properties of PUBEs.