Discussion of ***Automatic Relevance Determination with Statistical Guarantees*** by Zihe Liu and Feng Liang (***LL***)

David Draper and Erdong Guo

Department of Statistics
University of California, Santa Cruz

draper@ucsc.edu
Google Scholar: search on Draper, David
eguo1@ucsc.edu

OBayes 2022

7 Sep 2022

An Example of the Problem Addressed by LL

We're in the familiar ***multiple linear regression*** setting:

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{e}\,, \quad \text{in which} \tag{1}$$

- $\boldsymbol{y}$ is an $(n \times 1)$ vector of real-valued ***outcomes***,
- $\boldsymbol{X}$ is an $(n \times p)$ matrix of real-valued ***predictors*** (with all $(n \times p)$ entries in $\boldsymbol{X}$ regarded as fixed known constants),
- β is a $(p \times 1)$ vector of real-valued ***regression coefficients***, and
- $(\boldsymbol{e} \,|\, \sigma^2) \sim N_n(\boldsymbol{0}, \sigma^2{}_n I_n)$ is an $(n \times 1)$ vector of real-valued ***latent variables***, describing (not explaining) the lack of perfect fit between $\boldsymbol{y}$ and $\boldsymbol{X}\hat{\beta}$ arising from the vector $\hat{\beta}$ of coefficient estimates produced by your prediction algorithm.

Here $0 < \sigma < \infty$ and n and p are ***finite positive integers***.

The case $(n > p)$ was first solved in an ***optimal mean-squared-error*** manner (minimizing $||\boldsymbol{y} - \boldsymbol{X}\hat{\beta}||_2^2$) by Gauss (1785) (in which $||\cdot||_2$ is the L_2 norm), and this solution was good enough to solve many practical problems for about the next 200 years.

But What About When $(p > n)$ or $(p \gg n)$?

Example: We (Draper, Guo, et al., 2022) have recently been working with a data set compiled on $\boldsymbol{n = 950}$ representative women, each with one of 5 types of breast cancer; for each woman we have $\boldsymbol{p = 39{,}868}$ binary predictors: indicators of presence or absence of particular mutations on particular genes.

The goal is to ***optimally predict breast cancer type*** from the binary $\boldsymbol{X}$ variables.

From problem context, most of the entries in the optimal $\hat{\beta}$ will be 0; this defines a ***variable selection*** problem with ***sparsity***.

The ***LL*** paper is about these $(p \gg n)$ settings, in which Gauss's results form part, but not all, of the solution.

LL build upon the empirical Bayes variable selection method called ***Automatic Relevance Detection*** (ARD: Mackay, 1995; Neal, 1995).

Automatic Relevance Detection

The ARD idea begins with the desire to use the independence prior

$$p(\boldsymbol{\beta}) = \prod_{j=1}^{p} p_j(\beta_j) \quad \text{with} \quad (\beta_j | r_j) \overset{I}{\sim} N(0, r_j^2)\,; \tag{2}$$

here all of the r_j are finite nonnegative constants but some (perhaps many) are 0; this will induce ***sparsity*** because $(r_j = 0)$ iff $(\beta_j = 0)$.

However, this defines an ***impossible specification problem***: how can you specify the hyperparameter vector $\boldsymbol{r} = (r_1, \ldots, r_p)$ intelligently in a fully *a priori* manner?

This motivates an empirical Bayes approach in which we learn the r_j from the data by optimizing the ***evidence*** (marginal likelihood) with respect to the r_j, which ***LL*** call the ***relevance parameters***:
$(r_j = 0)$ iff ($\boldsymbol{x_j}$ is irrelevant to the prediction process).

LL go one step further and estimate $\boldsymbol{r}$ with a ***variational*** approach, yielding what they call ***Variational ARD*** (VARD).

Variational ARD

In VARD the prior in equation (2) is ***approximated*** by the variational distribution

$$q(\beta) = \prod_{j=1}^{p} q_j(\beta_j) \quad \text{with} \quad (\beta_j | \mu_j, \phi_j) \overset{I}{\sim} N(\mu_j, \phi_j^2)\,; \tag{3}$$

here all of the μ_j are finite real constants and the ϕ_j are finite nonnegative constants, but (as before) some (perhaps many) of the ϕ_j are 0, again inducing ***sparsity***.

The next move in variational Bayes is typically to define an ***evidence lower bound (ELBO) function*** to be minimized: for ***LL*** this is

$$\mathcal{L}(\boldsymbol{r}, \boldsymbol{\mu}, \boldsymbol{\phi}^2, \sigma^2) = -E_q[\log p(\boldsymbol{y} \mid \beta)] + \alpha\, KL(q||p)\,, \tag{4}$$

in which $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)$, $\boldsymbol{\phi}^2 = \left(\phi_1^2, \ldots, \phi_p^2\right)$, $E_q \log p(\boldsymbol{y} \mid \beta)]$ is the expectation with respect to $q(\beta)$ of the log likelihood function induced by the sampling model $\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{e}$ in equation (1), $KL(q||p)$ is the KL divergence between q and p, and $\alpha \geq 0$ is a hyperparameter introduced by ***LL*** to ***increase the flexibility*** of the variational approximation.

Variational ARD (continued)

After simplifications based on marginal optimizations, ***LL*** are led to the following ***penalized regression problem***: for fixed $\alpha \geq 0$, and pretending that σ is known (this can easily be remedied later), minimize

$$L(\boldsymbol{\mu}) = ||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\mu}||_2^2 + c_1 \sum_{j=1}^{p} g(\mu_j)\,, \tag{5}$$

in which ($c_1 = 2\,n > 0$) and

$$g(\mu) = \frac{\alpha\,\sigma^2}{2\,n}\left\{\frac{n\,\tau(\mu)}{\alpha\,\sigma^2} - \log\left[1 - \frac{n\,\tau(\mu)}{\alpha\,\sigma^2}\right]\right\}\,; \tag{6}$$

here

$$\tau(\mu) = \sqrt{\frac{\mu^4}{2} + \frac{\alpha\,\sigma^2}{n}\,\mu^2} - \frac{\mu^2}{2}\,. \tag{7}$$

We have considered this problem from a ***Bayesian*** point of view, to examine the prior on $\boldsymbol{\mu}$ implied by equation (5).

A Bayesian Shrinkage Prior Hidden Inside VARD

Consider the ***multiple linear regression problem*** specified by

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\mu} + \boldsymbol{e}^* \quad \text{with} \quad (\boldsymbol{e}^* \,|\, \sigma_*^2) \sim N_n(\boldsymbol{0}, \sigma_*^2 \, {}_n I_n)\,, \tag{8}$$

and for simplicity assume temporarily that $(\sigma_* \in (0, \infty))$ is known (as with ***LL***'s work, this can easily be remedied later); here the likelihood function $\ell(\boldsymbol{\mu} \,|\, \boldsymbol{y})$ is induced from the sampling model $(\boldsymbol{y} \,|\, \boldsymbol{\mu}) \sim N_n(\boldsymbol{\mu}, \sigma_*^2 \, {}_n I_n)$.

Give $\boldsymbol{\mu}$ a prior distribution $p(\boldsymbol{\mu})$; then (after simplification) we have

$$\begin{aligned} p(\boldsymbol{\mu} \,|\, \boldsymbol{y}) &= c\, \ell(\boldsymbol{\mu} \,|\, \boldsymbol{y})\, p(\boldsymbol{\mu})\,, \quad \text{from which} \\ -2 \log p(\boldsymbol{\mu} \,|\, \boldsymbol{y}) &= c - 2 \log \ell(\boldsymbol{\mu} \,|\, \boldsymbol{y}) - 2 \log p(\boldsymbol{\mu}) \\ &= c + \frac{||\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\mu}||_2^2}{\sigma_*^2} - 2 \log p(\boldsymbol{\mu})\,. \end{aligned} \tag{9}$$

As usual, finding the ***Maximum a Posteriori*** (MAP) estimate of $\boldsymbol{\mu}$ is equivalent to minimizing $-2 \log p(\boldsymbol{\mu} \,|\, \boldsymbol{y})$.

Bayesian Shrinkage Prior (continued)

So the ***Bayesian MAP estimate*** of $\boldsymbol{\mu}$ in this problem satisfies

$$\begin{aligned}\hat{\boldsymbol{\mu}}_{MAP} &= \underset{\boldsymbol{\mu}\in\mathbb{R}^p}{\arg\min}\left[c + \frac{||\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\mu}||_2^2}{\sigma_*^2} - 2\,\log p(\boldsymbol{\mu})\right] \\ &= \underset{\boldsymbol{\mu}\in\mathbb{R}^p}{\arg\min}\left[||\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\mu}||_2^2 - c_2\,\log p(\boldsymbol{\mu})\right]\,, \qquad (10)\end{aligned}$$

in which $(c_2 = 2\,\sigma_*^2 > 0)$. Comparing this with ***LL***'s ***penalized regression estimate***

$$\hat{\boldsymbol{\mu}}_{LL} = \underset{\boldsymbol{\mu}\in\mathbb{R}^p}{\arg\min}\left[||\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\mu}||_2^2 + c_1\sum_{j=1}^{p} g(\mu_j)\right]\,, \qquad (11)$$

we find that ***LL***'s ***implied prior*** satisfies

$$\begin{aligned}-\,c_2\,\log p(\boldsymbol{\mu}) &= c_1\sum_{j=1}^{p} g(\mu_j)\,, \quad \text{from which} \\ p(\boldsymbol{\mu}) &= \prod_{j=1}^{p}\exp\left[-c_3\,g(\mu_j)\right] \quad \text{with} \quad c_3 > 0\,. \qquad (12)\end{aligned}$$

Bayesian Shrinkage Prior (continued)

$$p(\boldsymbol{\mu}) = \prod_{j=1}^{p} \exp\left[-c_3\, g(\mu_j)\right] \quad \text{with} \quad c_3 > 0\,. \tag{13}$$

This is a prior in which the μ_j are ***IID***, each with marginal PDF of the form $\exp\left[-c\, g(\mu_j)\right]$ for some $(c > 0)$;

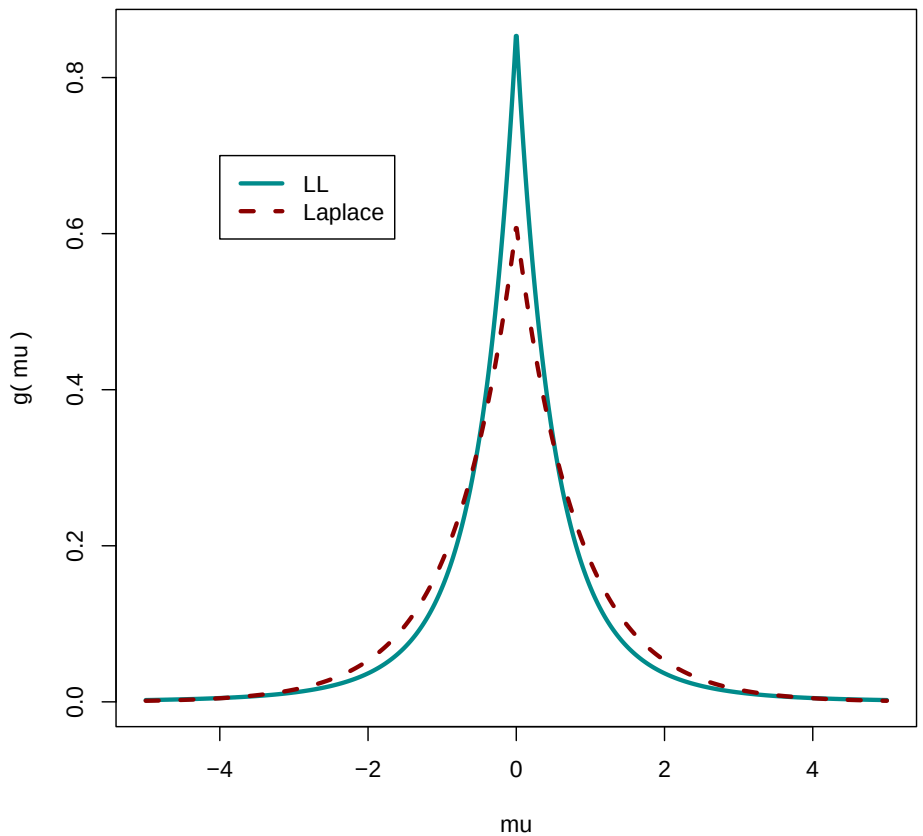
guess what these marginal densities look like:

Bayesian Shrinkage Prior (continued)

$$p(\boldsymbol{\mu}) = \prod_{j=1}^{p} \exp\left[-c_3\, g(\mu_j)\right] \quad \text{with} \quad c_3 > 0\,. \tag{13}$$

This is a prior in which the μ_j are ***IID***, each with marginal PDF of the form $\exp\left[-c\, g(\mu_j)\right]$ for some $(c > 0)$;

guess what these marginal densities look like:



Simulation Results

No wonder the ***LL*** method is an ***attempted improvement on the lasso***, which has the ***Laplace distribution*** as its implied prior.

Simulation Results: We have replicated some of ***LL***'s simulation findings, and we've also done some new simulations of our own in Cases 1 and 3 of their Table 1; in what follows s is the number of nonzero β_j in the data-generating model for the simulations, β_S is a vector identifying the s non-zero β_j, and M is the number of our simulation replications.

Case 1: $n = 200,\ p = 800,\ s = 10,$
$\beta_S = (1, -2, 3, -4, 5, -6, 7, -8, 9),\ \sigma = 1,\ M = 1{,}000$

Our result 1: We agree that the lasso, as initially proposed by Tibshirani (1996), has a best-possible true positive rate (TPR) of 100% = 1 but a high false discovery rate (FDR): we got an FDR of about 0.569:

		truth (case 1)		
		include	don't	
lasso	include	10	13.18	23.18
	don't	0	776.82	776.82
	total	10	790	800

Simulation Results (continued)

Our result 2: But why are we comparing the 1996 lasso with the 2022 ***LL*** proposal? Lots of improvements have been made to the 1996 version; a fairly recent example is the ***relaxed lasso*** (Hastie et al., 2017): TPR = 1 (***LL***'s method: 1), FDR = 0.005 (***LL***'s method: 0.036)

		truth (case 1)		
		include	don't	
relaxed lasso	include	10	0.05	10.05
	don't	0	789.95	789.95
	total	10	790	800

In Case 1 the relaxed lasso is about as fast as the ***LL*** method, with the same TPR and with an FPR that's ***smaller by a multiplicative factor of about 7***.

Simulation Results (continued)

Case 3: $n = 100$, $p = 400$, $s = 20$,
$\beta_S = \texttt{rep}(\ \log 100 \doteq 4.6,\ 20\ )$, $\sigma = 5$, $M = 1{,}000$

truth (case 3)

		include	don't	
relaxed lasso	include	19.79	23.37	43.16
	don't	0.21	356.63	356.84
	total	20	380	400

Our result 3: Here the relaxed lasso performs a bit better than the ***LL*** proposal on TPR (relaxed lasso 0.990, ***LL*** proposal 0.948) but much worse on FDR (relaxed lasso 0.545, ***LL*** proposal 0.324).

But why are we using variational approximations when we can go full Bayes with a superb shrinkage option: the ***horseshoe prior*** (Carvalho et al., 2010).

Simulation Results (continued)

Our result 4: With a burn-in of 1,000 iterations and a monitoring run of 10,000 in each simulation replication,
the horseshoe performance in Case 3 was outstanding:

		truth (case 3) include	truth (case 3) don't	
horseshoe	include	19.68	0.12	19.8
	don't	0.32	379.88	380.2
	total	20	380	400

Horseshoe TPR $= 0.984$ (***LL*** proposal 0.948),
horseshoe FDR $= 0.006$ (***LL*** proposal 0.324).

The main drawback with the horseshoe is that ***the MCMC takes a long time***: 4 seconds per 1,000 iterations (clocktime to produce the table above: 47.3 seconds) versus 0.1 seconds for the entire ***LL*** proposal.

n Moderate, p MUCH Bigger

Example: Draper, Guo, et al. (2022): data set compiled on $\boldsymbol{n = 950}$ representative women, each with one of 5 types of breast cancer; for each woman we have $\boldsymbol{p = 39{,}868}$ binary predictors.

The ***relaxed lasso (fitting a multinomial logit model) does surprisingly well here***, and in only 9 clock seconds: here's the confusion matrix on the validation sample with a $\left(\frac{2}{3}, \frac{1}{3}\right)$ modeling/validation split:

	True					
Predicted	1	2	3	4	5	Total
1	156	13	0	2	6	177
2	11	41	0	3	1	56
3	0	0	51	0	2	53
4	3	1	0	19	0	23
5	2	1	1	1	3	8
Total	172	56	52	25	12	317

```
 Percent Correct:  0.8517
```

Extreme gradient (tree) boosting (Chen and Guestrin, 2016) performs even better (88% percent correct), but that's for another talk.

Concluding Comments and Questions For the Authors

- We would be interested to see additional simulations with much larger $\frac{p}{n}$ ratios (e.g., in the genetic example above this ratio was 42, versus 5 in the most extreme of ***LL***'s cases).
- Optimizing over one coordinate at a time is only guaranteed to find the global minimum if the target function is convex; is this true of the original ELBO function $\mathcal{L}(\boldsymbol{r}, \boldsymbol{\mu}, \boldsymbol{\phi}^2, \sigma^2)$? (We couldn't tell from the paper whether this is true.)

  Or perhaps what's going on is that ***LL*** substitute a convex target function as an approximation to the ELBO and hope that the substitution will give good results.
- To produce truly useful asymptotics in problems of this type, p and n must both be allowed to go to infinity; do we understand correctly that ***LL*** are assuming that

$$\frac{\log p}{n} = O(1)? \tag{14}$$