BAYESIAN APPROACHES

.

1. Never Believe A Model

A deep flaw in all this quantitative behavioral psychology we are performing is that the future will not resemble the past. However, that very element, human psychology, can also contribute to a partial fix. Our knowledge of the world around us invests us with opinions about what is, and is not, plausible based on vague and complex criteria we could never hope to model directly.

Bayesian approaches attempt to make use of our ideas about plausibility in whatever way we can. Though we formally phrase the whole process in terms of probability distributions, it is really just about "encouraging" our fitted models to agree with our prior beliefs about what makes sense.

At times this is implicit, as when we require some historical stability from our parameter sets, while at other times we make the whole process explicit, for example when we have a good idea what model parameters ought to look like.

2. Principles and Terminology

We can address some of the problems arising from overdetermination and ambiguity through the use of *Bayesian inference*, by which we mean updating prior beliefs in light of observed data. We consider *observed quantities* $\boldsymbol{y}$ in light of a model with (generically) some *fixed parameters* which we leave implicit and other parameters $\boldsymbol{\theta}$ whose values we wish to determine.

Take

$$p(\boldsymbol{y}, \boldsymbol{\theta}) = p(\boldsymbol{\theta})p(\boldsymbol{y} \mid \boldsymbol{\theta}) = p(\boldsymbol{y})p(\boldsymbol{\theta} \mid \boldsymbol{y})$$

where we say $p(\boldsymbol{\theta})$ is the *prior density*. $p(\boldsymbol{\theta} \mid \boldsymbol{y})$ is the *posterior density*, $p(\boldsymbol{y} \mid \boldsymbol{\theta})$ is the *likelihood function* and $p(\boldsymbol{y})$ is the *marginal likelihood*. Note that

$$p(\boldsymbol{y}) = \int_{\boldsymbol{\Theta}} p(\boldsymbol{\theta})L(\boldsymbol{\theta})d\boldsymbol{\theta}$$

where we have compressed $p(\boldsymbol{y} \mid \boldsymbol{\theta})$ into $L(\boldsymbol{\theta})$ with implicit dependence on observations $L(\boldsymbol{\theta})$. Bayes' theorem tells us that

$$p(\boldsymbol{\theta} \mid \boldsymbol{y}) = \frac{p(\boldsymbol{\theta})L(\boldsymbol{\theta})}{p(\boldsymbol{y})} \propto p(\boldsymbol{\theta})L(\boldsymbol{\theta})$$

In rare cases, the posterior density is in the same family of distributions as the prior, a situation we describe as *conjugate families*. The examples all come from the *exponential family*, where

$$p(\boldsymbol{y} \mid \boldsymbol{\theta}) = a(\boldsymbol{\theta})b(\boldsymbol{y}) \exp\left[\sum_{j=1}^{J} c_j(\boldsymbol{\theta})d_j(\boldsymbol{y})\right]$$

It is quite easy to see that the multivariate normal is a member of this family.

Given some region $C \subset \boldsymbol{\Theta}$ and a constant $0 < \epsilon < 1$ where

$$\epsilon = \text{Prob}(\boldsymbol{\theta} \in C \mid y) = \int_C p(\boldsymbol{\theta} \mid \boldsymbol{y})$$

we say C is a *Highest posterior density*. Generically for any given ϵ there will be a unique such credible region with the highest possible probability densities in it, denoted the *highest posterior density region*.

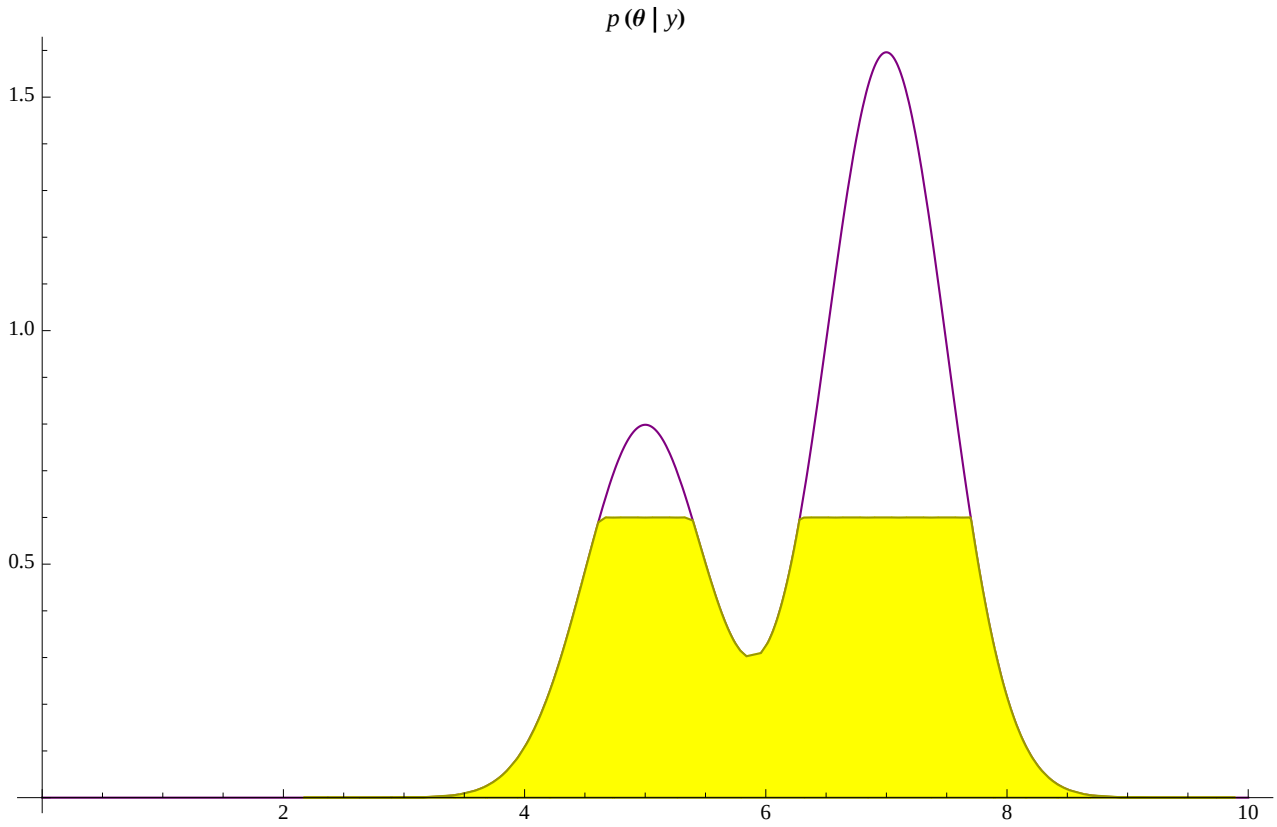

FIGURE 1. Highest posterior density region

. If we wish to make a prediction of some out-of-sample $\hat{\boldsymbol{y}}$, then we can compute

$$\begin{aligned} p(\hat{\boldsymbol{y}} \mid \boldsymbol{y}) &= \frac{p(\hat{\boldsymbol{y}}, \boldsymbol{y})}{p(\boldsymbol{y})} \\ &= \int_{\boldsymbol{\Theta}} \frac{p(\hat{\boldsymbol{y}}, \boldsymbol{y}, \boldsymbol{\theta})}{p(\boldsymbol{y})} d\boldsymbol{\theta} \\ &= \int_{\boldsymbol{\Theta}} p(\hat{\boldsymbol{y}} \mid \boldsymbol{y}, \boldsymbol{\theta}) \frac{p(\boldsymbol{\theta}) p(\boldsymbol{y} \mid \boldsymbol{\theta})}{p(\boldsymbol{y})} d\boldsymbol{\theta} \\ &= \int_{\boldsymbol{\Theta}} p(\hat{\boldsymbol{y}} \mid \boldsymbol{y}, \boldsymbol{\theta}) p(\boldsymbol{\theta} \mid \boldsymbol{y}) \\ &= \mathbb{E}_{\boldsymbol{\theta} \mid \boldsymbol{y}} \left[p(\hat{\boldsymbol{y}} \mid \boldsymbol{y}, \boldsymbol{\theta}) \right] \end{aligned}$$

This expectation is the *Bayesian predictive probability distribution*. Note that if past and future are independent conditional on $\boldsymbol{\theta}$ (as in random sampling) then

$$p(\hat{\boldsymbol{y}} \mid \boldsymbol{y}, \boldsymbol{\theta}) = p(\hat{\boldsymbol{y}} \mid \boldsymbol{\theta})$$

showing that, aside from dependence on the estimated parameters, our probability distribution for $\hat{\boldsymbol{y}}$ does not care what $\boldsymbol{y}$ was.

As the sample count approaches infinity, for "most" cases it turns out that the posterior parameter density converges in distribution to

$$\phi_K\left(\boldsymbol{\theta} \mid \boldsymbol{\theta}_{\mathrm{ML}}, I_{\mathrm{ML}}^{-1}\right)$$

where $\boldsymbol{\theta}_{\mathrm{ML}}$ is the maximum likelihood estimator and I_{ML} is its *information matrix* given by

$$I_{\mathrm{ML}} = \mathbb{E}_{\boldsymbol{y}|\boldsymbol{\theta}_{\mathrm{ML}}}\left[-\frac{\partial^2 L(\boldsymbol{\theta}_{\mathrm{ML}})}{\partial \boldsymbol{\theta}_{\mathrm{ML}} \partial \boldsymbol{\theta}_{\mathrm{ML}}^*}\right]$$

This information matrix is also the basis for some schemes of prior selection, such as Jeffreys' scheme of selecting priors proportional to its square root.

2.1. **Distributional Sampling.** Usually, our problems are too mathematically intractable to have a hope of directly sampling the joint posterior distribution $p(\boldsymbol{\theta} \mid \boldsymbol{y})$, making it hard to find estimates of $\boldsymbol{\theta}$ and its HPD regions. However, we can use *distributional inversion* and *Markov Chain Monte Carlo* (MCMC) techniques to obtain random samples.

2.1.1. *Distributional Inversion.* In the case where a CDF Ψ of the density ψ is available, then we can sample from Ψ simply by forming its inverse and applying it to a sample from $U([0,1])$.

2.1.2. *Gibbs.* In *Gibbs sampling*, we take draws from the *posterior conditional distribution*, which is much more tractable. After taking some "burn-in" draws from the conditional distributions, subsequent draws begin to converge to draws from the overall joint posterior distribution. Let's say we have two variables x and y. We first draw $\tilde{y}$ from $y \mid x$ and then we draw $\tilde{x}$ from $x|\tilde{y}$. Bouncing back and forth numerous times, we converge on a $(\tilde{x}, \tilde{y})$ pair that draws from the joint distribution.

2.1.3. *Metropolis.* If our posterior conditional distributions are themselves somewhat intractable, then we can run an acceptance/rejection algorithm, where (after burn-in) we accept new parameters if the joint probability from a generator distribution and our conditional is sufficiently high.

2.2. **Motivational Background: Online Algorithms.** Academics often have the luxury of defining a fixed data analysis period, studying it, and then leaving it fallow. A common requirement in any ongoing business, however, is to update model fit estimates as new information arrives. While it is certainly true that new calibrations can be calculated *de novo*, we can also take the perspective that previous calibrations provide a bayesian prior, to be updated in light of some new data item.

Let's revisit online, or update, algorithms.

Recall that in ordinary least squares regression, we solve (in principle but not in practice) the normal equations to obtain a parameter estimate

$$\boldsymbol{\beta} = (\boldsymbol{X}^*\boldsymbol{X})^{-1}\boldsymbol{X}^*\boldsymbol{y}$$

The Sherman-Morrison inversion formula

$$(\boldsymbol{A}+\boldsymbol{U}\boldsymbol{V}^*)^{-1} = \boldsymbol{A}^{-1} - \boldsymbol{A}^{-1}\boldsymbol{U}(\boldsymbol{I}+\boldsymbol{V}^*\boldsymbol{A}^{-1}\boldsymbol{U})^{-1}\boldsymbol{V}^*\boldsymbol{A}^{-1}.$$

tells us how, if some new observation $\boldsymbol{x}, y$ arrives, we can update $\boldsymbol{\beta}$. For convenience we define the self-adjoint *prediction error matrix* or *dispersion matrix*

$$\boldsymbol{P} = (\boldsymbol{X}^*\boldsymbol{X})^{-1}$$

so that

$$\boldsymbol{\beta} = \boldsymbol{P}\boldsymbol{X}^*\boldsymbol{y}$$

Define the *prediction error* as

$$h = y - \boldsymbol{x}^*\boldsymbol{\beta}$$

and the *error dispersion* as

$$f = 1 + \boldsymbol{x}^*\boldsymbol{P}\boldsymbol{x}$$

Now we can compute the new dispersion matrix as

$$\begin{aligned}\boldsymbol{P}_{\text{new}} &= \left(\boldsymbol{P}^{-1} + \boldsymbol{x}\boldsymbol{x}^*\right)^{-1} \\ &= \boldsymbol{P} - \boldsymbol{P}\boldsymbol{x}\left(1 + \boldsymbol{x}^*\boldsymbol{P}\boldsymbol{x}\right)^{-1}\boldsymbol{x}^*\boldsymbol{P} \\ &= \boldsymbol{P} - \boldsymbol{P}\boldsymbol{x}f^{-1}\boldsymbol{x}^*\boldsymbol{P}\end{aligned}$$

and our new regression coefficients

$$\begin{aligned}\boldsymbol{\beta}_{\text{new}} &= \boldsymbol{\beta} + \boldsymbol{P}\boldsymbol{x}(1 + \boldsymbol{x}^*\boldsymbol{P}\boldsymbol{x})^{-1}(y - \boldsymbol{x}^*\boldsymbol{\beta}) \\ &= \boldsymbol{\beta} + \boldsymbol{P}\boldsymbol{x}f^{-1}h.\end{aligned}$$

A similar formula applies, of course, when we are *subtracting* rather than adding some observation $\boldsymbol{x}, y$, whence

$$\boldsymbol{\beta}_{\text{reduced}} = \boldsymbol{\beta} - \boldsymbol{P}\boldsymbol{x}(\boldsymbol{x}^*\boldsymbol{P}\boldsymbol{x} - \mathbf{1})^{-1}(y - \boldsymbol{x}^*\boldsymbol{\beta})$$

which allows us to perform efficient *window regression*.

Now let us say that we wish to discount our *old* data relative to new incoming information by some factor $\lambda \in (0, 1]$. We can therefore say that the dispersion matrix of the old data should be multiplied by λ, allowing our update $\boldsymbol{x}, y$ to have full effect. That is to say we obtain

$$\begin{aligned}\boldsymbol{P}_{\text{new}} &= (\lambda\boldsymbol{P} + \boldsymbol{x}\boldsymbol{x}^*)^{-1} \\ &= \frac{1}{\lambda}\left(\boldsymbol{P} - \boldsymbol{P}\boldsymbol{x}\left(\boldsymbol{\lambda} + \boldsymbol{x}^*\boldsymbol{P}\boldsymbol{x}\right)^{-1}\boldsymbol{x}^*\boldsymbol{P}\right)\end{aligned}$$

Our new coefficients are then

$$\begin{aligned}\boldsymbol{\beta}_{\text{new}} &= \boldsymbol{\beta} + \boldsymbol{P}\boldsymbol{x}(\boldsymbol{\lambda} + \boldsymbol{x}^*\boldsymbol{P}\boldsymbol{x})^{-1}(y - \boldsymbol{x}^*\boldsymbol{\beta}) \\ &= \boldsymbol{\beta} + \boldsymbol{P}\boldsymbol{x}{f_\lambda}^{-1}h\end{aligned}$$

This approach is called *discounted least-squares regression* and is commonly seen in control theory[1].

[1] Discounted least-squares regression is equivalent to a subset of *Kalman filter* approaches to estimating state, with a trivial *transition equation* $\beta_{t+1} = \beta_t$ having no torsion and no state disturbance.

3. Automatic Variable Selection/Model Building

Let us postulate that *some* relationship may exist between a collection $\boldsymbol{x} = x_1, \ldots, x_N$ of available explanatory variables and a target dependent variable y, and that, from a Bayesian point of view, we are unwilling to believe that *all* elements of x are important.

If we phrase this in terms of a Bayesian prior, then we will place some probability distribution on variable count or variable inclusion probabilities and integrate to obtain a distribution on the ensemble of models[2]. An explanatory variable that appears in nearly all of the likeliest models is one we will wish to include when making predictions.

Another Bayesian perspective is to place a prior distribution on our regression coefficients. For example, if our prior is that the coefficients themselves have a gaussian likelihood centered at 0, then we obtain a model fit known as the *ridge regression*. This fit, by having a prior with more probability at zero than elsewhere, forces coefficients closer to zero than ordinary least squares would put them. However it does not tend to select variables.

If our prior is that the coefficients themselves have a double exponential likelihood centered at 0, we obtain the *lasso*, which does tend to zero out elements of β.

We can take a simpler perspective, however, and express our prior beliefs algorithmically rather than distributionally. This perspective deprives us of the useful distributional analyses we might make, but provides considerable increases in computational efficiency.

3.1. Forward Stepwise Regression.

The *forward stepwise regression* algorithm begins by defining the model M_0 as

$$y = \beta^0 + \epsilon$$

having residuals

$$\boldsymbol{r}^0 = \boldsymbol{y} - \beta^0$$

Given an existing model with $K \geq 0$ nontrivial predictors $\boldsymbol{x}_{j_1}, \ldots, \boldsymbol{x}_{j_K}$, we compute the correlation of each of the $N - K$ heretofore unused predictors with the residuals $\boldsymbol{r}^K$,

$$\rho_j^K = \mathrm{Corr}(\boldsymbol{x}_j, \boldsymbol{r}^K) \qquad j \in \{1, \ldots, N\} \setminus \{j_1, \ldots, j_K\}$$

Define j_{K+1} to be the index with the maximum available *residual correlation*

$$j_{K+1} = \arg\max_j \rho_j^K \qquad j \in \{1, \ldots, N\} \setminus \{j_1, \ldots, j_K\}$$

and run the regression of y against the set of variables $\boldsymbol{x}_{j_1}, \ldots, \boldsymbol{x}_{j_{K+1}}$, obtaining new residuals $\boldsymbol{r}^{K+1}$ which (by design) are orthogonal to each of the $\boldsymbol{x}_{j_k}$.

Our halting condition for this algorithm has several sensible forms, such that it should stop at a certain K, or at a stage when the maximum available correlation $\rho_j^K < \rho_{\min}$.

Note the similarity to *principal components analysis* (PCA), where we also select variables by maximum covariance explained. The difference, of course, is that in PCA we mix up the variables into factors to achieve maximum possible R^2 per factor.

3.2. The Lasso.

Forward stepwise regression is considered too *greedy* which, in numerical analysis, tends to describe an algorithm designed to choose parameters that are too extreme at any given stage. For example, seizing on a full regression of residuals against, say x_{j_2} can ignore valuable information in some other x_j that happens to have high correlation with x_{j_2}.

[2]Here we are referring to any two cases where the set of explanatory variables differ as being two different models. This is despite the fact that each is a submodel of a more general case where the variable set is the union of the two variable sets found in these submodels.

The *lasso* or *basis pursuit* is an attempt to address this problem by allowing new variables to "participate" as soon as possible. Remarkably, it has both a reasonably simple algorithm and a nice numeric characterization. We begin by normalizing y and all the $\boldsymbol{x_j}$ according to (possibly robust) Z-scores. The lasso is now defined as finding $\boldsymbol{\beta}$ as

$$\boldsymbol{\beta} = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|$$

subject to the constraint

$$T(\boldsymbol{\beta}) := \sum |\beta_j| \leq t$$

This constraint is technically a specific case of *bridge regression*, where we penalize $\sum |\beta_j|^\gamma$ for some γ. The other commonly encountered case is *ridge regression* where $\gamma = 2$. Combining $\sum |\beta_j|$ and $\sum |\beta_j|^2$ penalties results in the *elastic net*.

If $\gamma = 0$ then we are using *best subset selection*. In this case, the solution becomes trickier because our objective function is no longer convex. (Convexity requires $\gamma \geq 1$)

Unlike in ordinary least-squares regression, yet in the same spirit as robust regression, there is no direct solution (Lagrangian techniques are indeed used, though they lack a closed-form solution) but we can still design an iterative algorithm to converge to the answer.

TABLE 1. Lasso and Ridge Summary

	Lasso	Ridge
Coef Prior	double exp	gaussian
Bridge γ	1	2
Effect	Zero least important β_i	Reduce all β_i (proportionally)

3.2.1. *Calculating the Lasso.* The Lasso may be found using *quadratic programming* (say in the form of *conjugate gradient* solvers), at a cost of some considerable complexity. We convert our absolute value constraints to a set of 2^N signed constraints $\sum w_j \beta_j \leq t$ with the w_j exhausting possible combinations of ± 1.

There is a better way to form the lasso. Say we have identified the predictor j_1 with highest explanatory power. Rather than running a full regression to extract every bit of R^2 from its coefficient, we merely increase its coefficient *up to* the point where a second variable j_2 has just as much partial explanatory power on the residuals.

We increase the coefficients in the joint j_1, j_2 direction until some new predictor j_3 comes into play. We add j_3 to our model and continue, halting when our coefficients have reached their t limit. If we make our increases by adding ϵ to the coefficients in little bits, we have reproduced a *stagewise regression*. However, it is possible to be more efficient.

Let's say we presently have a set of predictors x_j "in" the model, $j \in \mathcal{A}$ for our *active set* $\mathcal{A}$, along with the signs s_j we're using on their coefficients. Define

$$\boldsymbol{X}_{\mathcal{A}} = (\dots, s_j \boldsymbol{x}_j, \dots)$$
$$\mathcal{G}_{\mathcal{A}} = \boldsymbol{X}_{\mathcal{A}}^* \boldsymbol{X}_{\mathcal{A}}$$
$$\boldsymbol{A}_{\mathcal{A}} = \left(\boldsymbol{1}^* \mathcal{G}_{\mathcal{A}}^{-1} \boldsymbol{1}\right)^{-\frac{1}{2}}$$
$$\boldsymbol{w}_{\mathcal{A}} = \boldsymbol{A}_{\mathcal{A}} \mathcal{G}_{\mathcal{A}}^{-1} \boldsymbol{1}$$
$$\boldsymbol{u}_{\mathcal{A}} = \boldsymbol{X}_{\mathcal{A}} \boldsymbol{w}_{\mathcal{A}}$$

The $\boldsymbol{u}_{\mathcal{A}}$ makes equal angles with the columns of our selected design matrix $\boldsymbol{X}_{\mathcal{A}}$.

Say our current estimate is $\boldsymbol{\mu}_{\mathcal{A}}$ and we want to update it. The current set of correlations (since we have normalized by z-score) is

$$\boldsymbol{c} = \boldsymbol{X}^*(\boldsymbol{y} - \boldsymbol{\mu}_{\mathcal{A}})$$

and for all $j \in \mathcal{A}$ the value c_j takes on the same globally unique maximal value C.

If we define

$$\boldsymbol{\mu}(\gamma) = \boldsymbol{\mu}_{\mathcal{A}} + \gamma \boldsymbol{u}_{\mathcal{A}}$$

then correlations $c_j(\gamma)$ as a function of γ will be

$$\begin{aligned} c_j(\gamma) &= \boldsymbol{x}_j^* \big(\boldsymbol{y} - \boldsymbol{\mu}(\gamma)\big) \\ &= c_j - \gamma a_j \end{aligned}$$

which for $j \in \mathcal{A}$ satisfies

$$|c_j(\gamma)| = C - \gamma \boldsymbol{A}_{\mathcal{A}}.$$

For any j outside the active set, $j \in \mathcal{A}^c$, we see that $c_j(\gamma)$ takes on its maximal value when

$$c_j - \gamma a_j = C - \gamma \boldsymbol{A}_{\mathcal{A}}$$

or

$$\gamma = \frac{C - c_j}{\boldsymbol{A}_{\mathcal{A}} - a_j}$$

and correlation of the residuals with $-\boldsymbol{x}_j$ takes on a maximal value when

$$\gamma = \frac{C + c_j}{\boldsymbol{A}_{\mathcal{A}} + a_j}$$

Finding the minimum over all j outside the active set, $j \in \mathcal{A}^c$, yields the tiniest possible value of γ such that some new index j joins the active set. Since correlations correspond to angles, this new algorithm is called *least angle regression (LARS)*.

To fully agree with lasso, the LARS algorithm requires a small modification. Our choice of γ may in principle have moved some β_j so far that it changed sign (crossed zero). We define $\boldsymbol{d}$ to be the vector with coordinate values $s_j w_{\mathcal{A},j}$ ($s_j := 0$ outside the active set) so that

$$\begin{aligned} \beta_j(\gamma) &= \beta_j + \gamma d_j \\ \boldsymbol{\mu}(\gamma) &= \boldsymbol{X}\boldsymbol{\beta}(\gamma) \end{aligned}$$

and solve for zero crossings (if any) by computing

$$\bar{\gamma}_j = -\frac{\beta_j}{d_j}$$

where defined. The minimum such $\bar{\gamma}_j$ defines our next solution point, at which we will remove the corresponding predictor index from the active set and add something else.

3.2.2. *Computational Cost.* At step k of the LARS algorithm, we have to obtain $m - k$ inner products, and then effectively invert $\mathcal{G}_k$. This is equivalent to updating a Cholesky decomposition of $\mathcal{G}_k$ to one of $\mathcal{G}_{k+1}$. The overall cost therefore agrees with that of a Cholesky decomposition of the design matrix. That is to say, for m variables our costs are $O(m^3 + nm^2)$.

The LARS algorithm is efficient when the number of independent variables is not too large. As the variable count grows, other optimization schemes, such as stochastic gradient descent, tend to dominate its efficiency.

3.3. **The Elastic Net.** Another perspective on the elastic net is that forms an arbitrary combination of Lasso and Ridge, by computing

$$\boldsymbol{\beta} = \arg\min_{\boldsymbol{\beta}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\| + \alpha_1 \sum |\beta_j|^1 + \alpha_2 \sum |\beta_j|^2.$$

There is a computable correspondence between LASSO and elastic net solutions, so we can use LARS to find elastic net coefficients as well.

3.4. **The Garrote.** Commonly in finance, we also require each $\beta_j \geq 0$. In this case, the technique is called a *positive LASSO* or a *garrote*.

3.5. **Automated Metaparameter Selection.** Following Fu (1998) and Tibshirani (1996) we can use generalized cross validation (GCV) to choose one or both of the metaparameters γ and λ. To do so, we start with the GCV loss function for our fits

$$\mathrm{GCV}(\gamma, \lambda) = \frac{\mathrm{RSS}}{n(1 - p(\lambda)N^{-1})^2}$$

where $p(\lambda)$ is the effective parameter count.

To find $p(\lambda)$ we begin with our bridge fit variables $\hat{\boldsymbol{\beta}}$ and a lasso fit $\hat{\boldsymbol{\beta}}_L$. We form the matrix

$$D = \mathrm{diag}(2\gamma^{-1}|\hat{\beta}_j|^{\gamma-2})$$

We define W as the generalized inverse of D, and n_0 as the number of zeros in $\hat{\boldsymbol{\beta}}_L$. Now

$$p(\lambda) = \mathrm{trace}\left(X\left(X^*X + \lambda W\right)^{-1} X^*\right) - n_0.$$

3.6. **Bayesian Perspective.** From a Bayesian point of view, if we are penalizing $\sum |\beta_j|^\gamma$ then we can view our fitting procedure as one of minimizing

$$(\boldsymbol{\beta}|\,\boldsymbol{Y}) \sim \mathrm{Const} \times \exp\left(-\frac{1}{2}\left(\mathrm{RSS} + \lambda \sum |\beta_j|^\gamma\right)\right)$$

which (by symmetry) corresponds to a prior distribution of each β_j of

$$\pi(\beta_j; \lambda, \gamma) = \frac{\gamma 2^{-(1+\gamma^{-1})}\lambda^{\gamma^{-1}}}{\Gamma(\gamma^{-1})} \exp\left(-\frac{1}{2}\left|\frac{\beta_j}{\lambda^{\gamma^{-1}}}\right|^\gamma\right)$$

which conveniently corresponds to a gaussian if $\gamma = 2$, and to a LaPlace double exponential distribution if $\gamma = 1$.

We can think of a ridge regression as forcing two highly correlated independent variables to have similar coefficients. In contrast LASSO will "pick" just one of them.

	SUN	NAP	SKY	SPY	PALL	XLE	USO	PBPB
OLS	-0.038192	-0.034847	-0.002651	0.083976	0.023637	0.811738	-0.078405	0.000478
AIC	0.000000	0.000000	0.000000	0.079643	0.000000	0.713096	0.000000	0.000000
BIC	0.000000	0.000000	0.000000	0.079643	0.000000	0.713096	0.000000	0.000000
LARS	-0.025133	-0.027691	0.000000	0.085127	0.013258	0.781491	-0.051430	0.000000
LASSO	-0.025233	-0.027747	-0.000000	0.085182	0.013341	0.781647	-0.051606	0.000000
Elastic	-0.025233	-0.027747	-0.000000	0.085182	0.013341	0.781647	-0.051606	0.000000
Pos BIC	0.000000	0.000000	0.000000	0.093837	0.000000	0.727290	0.000000	0.000000
Pos LARS	0.000000	0.000000	0.000000	0.096711	0.004405	0.730234	0.000000	0.000000
Pos LASSO	0.000000	0.000000	0.000000	0.096766	0.004455	0.730249	0.000000	0.000000
Pos Elast	0.000000	0.000000	0.000000	0.096766	0.004455	0.730249	0.000000	0.000000

3.7. **Fully Bayesian Linear Regression.** Let us now revisit our favorite linear models

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

Assume errors are i.i.d. normal with variance σ^2 and set the *error precision*

$$h = \frac{1}{\sigma^2}$$

We can express the likelihood function in terms of our classical estimate

$$\boldsymbol{\beta}_{\text{OLS}} = (\boldsymbol{X}^*\boldsymbol{X})^{-1}\boldsymbol{X}^*\boldsymbol{y}$$

and its SSE as

$$L(\boldsymbol{\beta}, h) = \frac{h^{N/2}}{(2\pi)^{N/2}} \exp\left[-\frac{h}{2}\left(\text{SSE} + (\boldsymbol{\beta} - \boldsymbol{\beta}_{\text{OLS}})^*\boldsymbol{X}^*\boldsymbol{X}(\boldsymbol{\beta} - \boldsymbol{\beta}_{\text{OLS}})\right)\right]$$

The *conjugate prior* of a normal distribution for $\boldsymbol{\beta}$ and h is a *normal-gamma distribution* with PDF

$$p_{\text{NG}}(\boldsymbol{\beta}, h \mid \mathring{\boldsymbol{\beta}}, \mathring{\boldsymbol{Q}}, \mathring{s}^{-2}, \mathring{\nu}) = \phi(\boldsymbol{\beta}; \mathring{\boldsymbol{\beta}}, \mathring{\boldsymbol{Q}})p^{\Gamma}(h; \mathring{s}^{-2}, \mathring{\nu})$$

which, while mathematically convenient, is unnecessarily restrictive for practical work. However, it does allow for a simple *Bayes factor* statistic for comparing alternative regression models A and B from normal-gamma priors

$$BF^{\text{B}}_{\text{A}} = \frac{\Gamma\left(\frac{\nu_B}{2}\right)\left(\mathring{\nu}_B\mathring{s}^2_B\right)^{\frac{\mathring{\nu}_B}{2}}\left(\frac{\|\Sigma_B\|}{\|\mathring{Q}_B\|}\right)^{\frac{1}{2}}\left(\nu_B s^2_B\right)^{\frac{\nu_B}{2}}}{\Gamma\left(\frac{\nu_A}{2}\right)\left(\mathring{\nu}_A\mathring{s}^2_A\right)^{\frac{\mathring{\nu}_A}{2}}\left(\frac{\|\Sigma_A\|}{\|\mathring{Q}_A\|}\right)^{\frac{1}{2}}\left(\nu_A s^2_A\right)^{\frac{\nu_A}{2}}}$$

which approaches the ratio of SSE in special cases.

3.7.1. *Gibbs for linear regressions.* For a linear regression, we can observe Gibbs sampling working as follows. Let the priors be that regression coefficients have a multivariate normal distribution while errors have variance σ^2 drawn from the *inverse gamma distribution* (i.e. the distribution such that the inverse of its argument has the gamma distribution)

$$\boldsymbol{\beta} \sim \Phi(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{V}_{\boldsymbol{\beta}})\sigma^2 \sim \text{IG}(a, b)$$

Our density looks hard to sample from, since the likelihood function

$$L(\boldsymbol{\beta}, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^*(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right)$$

is already complex and the density

$$p(\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{y}) \propto$$
$$L(\boldsymbol{\beta}, \sigma^2) \exp\left(-\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\mu}_{\boldsymbol{\beta}})^* \boldsymbol{V}_{\boldsymbol{\beta}}{}^{-1}(\boldsymbol{\beta} - \boldsymbol{\mu}_{\boldsymbol{\beta}})\right) \sigma^{-2(a+1)} \exp\left(-\frac{1}{b\sigma^2}\right)$$

is worse. For Gibbs sampling, however, we find that defining

$$\boldsymbol{D}_{\boldsymbol{\beta}} = \left(\frac{1}{\sigma^2}\boldsymbol{X}^*\boldsymbol{X} + \boldsymbol{V}_{\boldsymbol{\beta}}{}^{-1}\right)^{-1}$$

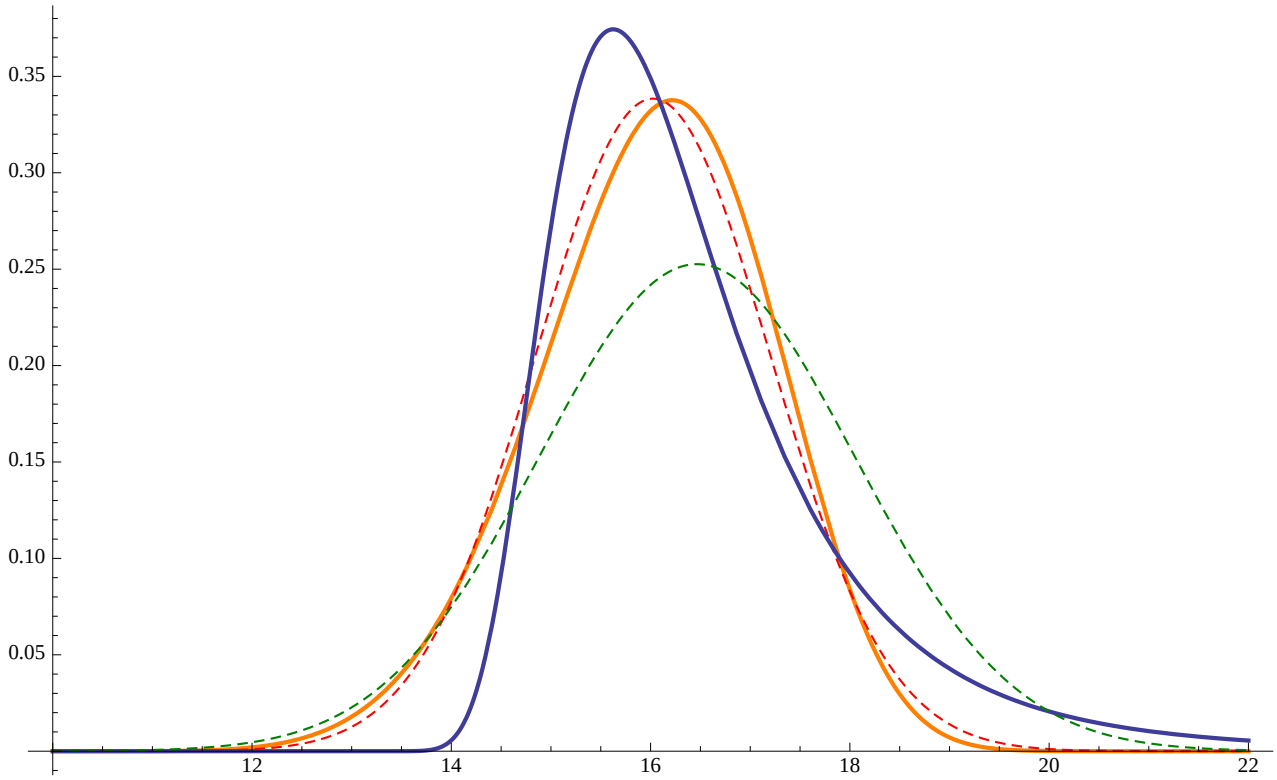

FIGURE 2. Gamma distribution (blue) and Inverse Gamma distribution (orange) (with gaussian matching means and standard deviations, dashed).

and

$$d_{\boldsymbol{\beta}} = \frac{1}{\sigma^2}\boldsymbol{X}^*\boldsymbol{y} + \boldsymbol{V_\beta}^{-1}\boldsymbol{\mu_\beta}$$

we find the conditional density for $\boldsymbol{\beta}$ is

$$\boldsymbol{\beta} \mid \sigma^2, \boldsymbol{y} \sim \Phi(\boldsymbol{D_\beta} d_{\boldsymbol{\beta}}, \boldsymbol{D_\beta})$$

For σ^2 the density is

$$\sigma^2 \mid \boldsymbol{\beta}, \boldsymbol{y} \sim \text{IG}\left(\frac{n}{2} + a, \left(b^{-1} + \frac{1}{2}(\boldsymbol{y} - \boldsymbol{X\beta})^*(\boldsymbol{y} - \boldsymbol{X\beta})\right)^{-1}\right)$$

Alternating draws from these distributions will yield joint draws for $\boldsymbol{\beta}, \sigma^2 \mid \boldsymbol{y}$.

3.7.2. *Linear regression regimes.* Let's now consider a univariate *segmented regression* model where at some *changepoint* time λ, the regression switched from one economic regime to another. That is to say, we have

$$y \sim \Phi\left(\theta_1 + \theta_2 x_t, \sigma^2\right)$$

when $t \leq \lambda$ and

$$y \sim \Phi\left(\beta_1 + \beta_2 x_t, \tau^2\right)$$

otherwise. We'll take priors

$$\begin{aligned}\boldsymbol{\theta} = \begin{pmatrix}\theta_0\\ \theta_1\end{pmatrix} &\sim \Phi(\boldsymbol{\mu}_{\boldsymbol{\theta}}, \boldsymbol{V}_{\boldsymbol{\theta}})\\ \boldsymbol{\beta} = \begin{pmatrix}\beta_0\\ \beta_1\end{pmatrix} &\sim \Phi(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{V}_{\boldsymbol{\beta}})\\ \sigma^2 &\sim \mathrm{IG}\,(a_1, a_2)\\ \tau^2 &\sim \mathrm{IG}\,(b_1, b_2)\\ \lambda &\sim \mathrm{U}\,(0, T)\end{aligned}$$

Here the likelihood function is separable into products of gaussian densities with parameters dependent on sample data lying within the range $t \leq \lambda$ or not. We obtain $\boldsymbol{D}_{\boldsymbol{\beta}}, d_{\boldsymbol{\beta}}$ and $\boldsymbol{D}_{\boldsymbol{\theta}}, d_{\boldsymbol{\theta}}$ as above for the constant-parameter case, plus an expression for the density of λ

$$p(\lambda \mid \boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2, \tau^2, y) \propto \prod_{t \leq \lambda} \phi(y_t, \theta_0 + \theta_1 x_t, \sigma^2) \prod_{t > \lambda} \phi(y_t, \beta_0 + \beta_1 x_t, \tau^2)$$

which requires renormalization to be usable.

We can also run Gibbs sampling for *restricted regressions*, where parameters are constrained to take values within intervals, simply by truncating and renormalizing the marginals.

3.7.3. *Bayesian Approaches to Dependent Errors.* Extending the Bayesian linear regression model to the case where errors may have nontrivial interdependence of errors, we can consider

$$\epsilon \sim \Phi(\mathbf{0}, h^{-1}\boldsymbol{\Omega})$$

where $\boldsymbol{\Omega}$ is a covariance matrix. Here we collect terms involving $\boldsymbol{\beta}$ and h to form conditional posteriors as above (depending on our choice of priors for these two parameters). The general form of the conditional posterior for $\boldsymbol{\Omega}$ will be

$$\frac{p(\boldsymbol{\Omega})}{\sqrt{\|\boldsymbol{\Omega}\|}} \exp\left(-\frac{1}{2}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right)^* h\boldsymbol{\Omega}\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\right)\right)$$

where the form of the unconditional density of $\boldsymbol{\Omega}$ is our choice, written $p(\boldsymbol{\Omega})$. For $\boldsymbol{\beta}$ we first obtain a symmetric diagonal decomposition of its pseudo inverse, i.e. we find $\boldsymbol{P}$ such that $\boldsymbol{P}\boldsymbol{\Omega}\boldsymbol{P}^* = \boldsymbol{I}$, and then orthogonalize our data by taking

$$\begin{aligned}\tilde{\boldsymbol{y}} &= P\boldsymbol{y}\\ \tilde{\boldsymbol{X}} &= P\boldsymbol{X}\\ \tilde{\boldsymbol{\epsilon}} &= P\boldsymbol{\epsilon}\end{aligned}$$

so that we can consider, for any given $\boldsymbol{\Omega}$, that we have the transformed model

$$\tilde{\boldsymbol{y}} = \tilde{\boldsymbol{X}}\boldsymbol{\beta} + \tilde{\boldsymbol{\epsilon}}$$

Taking

$$\begin{aligned}\hat{\boldsymbol{\beta}}(\boldsymbol{\Omega}) &= \left(\tilde{\boldsymbol{X}}^*\tilde{\boldsymbol{X}}\right)^{-1}\tilde{\boldsymbol{X}}^*\boldsymbol{y}\\ \bar{\boldsymbol{V}} &= \left(\mathring{\boldsymbol{V}}^{-1} + h\boldsymbol{X}^*\boldsymbol{\Omega}^{-1}\boldsymbol{X}\right)^{-1}\\ \bar{\boldsymbol{\beta}} &= \bar{\boldsymbol{V}}\left(\mathring{\boldsymbol{V}}^{-1}\mathring{\boldsymbol{\beta}} + h\boldsymbol{X}^*\boldsymbol{\Omega}^{-1}\boldsymbol{X}\hat{\boldsymbol{\beta}}(\boldsymbol{\Omega})\right)\end{aligned}$$

we again have a multivariate normal conditional for $\boldsymbol{\beta}$,

$$\boldsymbol{\beta} \mid \boldsymbol{y}, h, \boldsymbol{\Omega} \sim \Phi(\bar{\boldsymbol{\beta}}, \bar{\boldsymbol{V}})$$

3.8. **Multivariate and Selection Models.** Let's now consider a multivariate case where we specify

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

as

$$y_i = \beta_0 + \sum_{j=1}^{p} \theta_j x_{i,j} + \epsilon_i$$

and want to figure out which of the β_j are zero. For this, we want a prior with

- a large point mass at zero, denoting an unimportant independent variable
- a very flat distribution ("slab" after Alan Miller) outside zero, indicating we are willing to believe a wide range of coefficient values if the variable is significant

Here an auxiliary indicator variable $\boldsymbol{I}$ is useful, so that

$$\boldsymbol{\theta} = \boldsymbol{I} \circ \boldsymbol{\beta}$$

for "effect size" coefficients β_j. Running a Gibbs sampler gives us estimates of these variables, with the mean value of I indicating inclusion probabilities.

A real problem for estimating these models is that it is hard to distinguish $\beta_j \approx 0$ from $I_j \approx 1$. For some priors (e.g. ones where each variable's inclusion probability is independent) this is less important than for others (where the overall independent variable count is constrained).

One good choice, therefore, is to bypass inclusion with *adaptive shrinkage* in the *Bayesian Lasso*, and use a "slab and hump" prior with a narrow hump around zero. In this case we set a threshold c so that we say a variable is "in" the model if $|\beta_j| > c$. We take an exponential distribution for the variance τ_j^2 of our beta prior, i.e. $\beta_j \mid \tau_j^2 \sim \Phi(0, \tau_j^2)$ and $\theta_j = \beta_j$. By choosing the parameter μ of the exponential distribution large enough, we obtain a strong spike. Alternatively we can just take $P(\tau_j^2) \propto 1/\tau_j^2$ (with a bound).

```
model {
  for (i in 1:n) {
    mean[i]<-alpha+inprod(X[i,],beta)
    y[i]~dnorm(mean[i],tau)
  }
  for (j in 1:p) {
    ind[j]~dbern(pind)
    betaT[j]~dnorm(0,taub)
    beta[j]<-ind[j]*betaT[j]
  }
  alpha~dnorm(0,0.0001)
  tau~dgamma(1,0.001)
  taub~dgamma(1,0.001)
  pind~dbeta(2,8) }
```