Week 11: Nov. 7 - Nov. 9

Bayesian Modeling and Ridge Regression

Recall(?) ridge regression is a form of penalized regression such that:

$$\hat{\tilde{\beta}}_R = \text{arg min}_{\hat{\tilde{\beta}}} \ ||\tilde{y} - X\hat{\tilde{\beta}}||_2^2 + \lambda||\hat{\tilde{\beta}}_R||_2^2 \rightarrow \hat{\tilde{\beta}}_R = \left(X^X + \lambda I\right)^{-1} X^T \tilde{y}. \tag{1}$$

A more familiar (and perhaps simpler) form would be:

$$\hat{\tilde{\beta}}_R = \text{arg min}_{\hat{\tilde{\beta}}} \ \sum_{i=1}^{n} (y_i - \beta_1 x_1 + \ldots \beta_p x_p)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \tag{2}$$

As λ gets large all of the values are shrunk toward zero. As λ goes to 0, the ridge regression estimator results in OLS estimator. It can be shown that ridge regression results better predictive ability than OLS by reducing variance of the predicted values at the expense of bias. Note that typically the X values are assumed to be standardized, so that the intercept is not necessary.

An alternative form of penalized regression is known as LASSO. The LASSO uses an L1 penalty such that:

$$\hat{\tilde{\beta}}_L = \text{arg min}_{\hat{\tilde{\beta}}} \ ||\tilde{y} - X\hat{\tilde{\beta}}||_2^2 + \lambda||\hat{\tilde{\beta}}_L||_1, \tag{3}$$

the L1 penalty results in $||\tilde{x}||_1 = |x_1| + \cdots + |x_m|$, which minimizes the absolute differences. LASSO can also be written as:

$$\hat{\tilde{\beta}}_R = \text{arg min}_{\hat{\tilde{\beta}}} \ \sum_{i=1}^{n} (y_i - \beta_1 x_1 + \ldots \beta_p x_p)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \tag{4}$$

The nice feature of LASSO, relative to ridge regression, is that coefficients are shrunk to 0 providing a way to do *variable selection.* One challenge with LASSO is coming up with proper distributional assumptions for inference about variables.

Bayesian Regression vs. Ridge Regression

Consider the following prior $p(\tilde{\beta}) = N(0, I_p\tau^2)$. How does this relate to ridge regression? First compute the posterior distribution for $\tilde{\beta}$.

$$p(\tilde{\beta}|-) \propto \exp\left[-\frac{1}{2}\left(\frac{1}{\sigma^2}\tilde{\beta}^T X^T X\tilde{\beta} - \frac{1}{\sigma^2}\tilde{\beta}^T X^T\tilde{y} + \tilde{\beta}^T\frac{I_p}{\tau^2}\tilde{\beta}\right)\right] \quad (5)$$

$$\propto \quad (6)$$

Thus $Var(\tilde{\beta}|-) =$
and $E(\tilde{\beta}|-) =$

Q: does this look familiar?

Define:

Note that in a similar fashion we can use a specific prior on $\tilde{\beta}$ to achieve LASSO properties. It is also important to clarify the differences between classical ridge regression (and Lasso) with the Bayesian analogs. In the Bayesian case we can still easily compute credible intervals to account for the uncertainty in our estimation. Interval calculations for inference are difficult in the these settings, particularly for Lasso.

Bayesian Model Selection

Recall, we discussed common model selection techniques:

-
-
-

-
-

In modeling, there is an additional layer of uncertainty the comes from the 'assumed model'. This is part of the model assumptions (which includes the prior in Bayes). Classical settings do not provide a coherent way to address this uncertainty, However, in a Bayesian framework we have a coherent way to talk about model selection. Specifically, given a prior on the model space we can compute posterior probability for a given model.

In model selection for linear regression the goal is to decide which covariates to include in the model. To do this, we introduce a parameter $\tilde{z}$, where
Then define $\beta_i =$. Note the $b_i's$ are the real-values regression coefficients. For now we will ignore the intercept (standardizing the covariates). The regression equation now becomes:

$$y_i = \tag{7}$$

Again thinking about the regression model as a conditional expectation, then for $p = 3$:

$$\begin{aligned} E[y|\tilde{x}, \tilde{b}, \tilde{z} = (1, 0, 1)] &= \\ E[y|\tilde{x}, \tilde{b}, \tilde{z} = (0, 1, 0)] &= \end{aligned}$$

Note that the vector $\tilde{z}_a$ defines a model and is interchangable with the notation M_a. Now the goal is a probabilistic statement about $\tilde{z}_a$, specifically:

$$Pr(\tilde{z}_a|\tilde{y}, X) = \tag{8}$$

where X is a matrix of observed covariates. Of course, this requires a prior on z_a, which we will see momentarily.

For model comparison between model a and model b, an alternative way to express this is through the following (familiar) ratio:

(9)

Now the question is, how do we think about the priors for $\tilde{z}$ or equivalently for M_a?

Bayesian Model Comparison

The posterior probability for model a is a function of

In a regression setting the marginal probability is computed by integrating out the parameters as:

$$p(\tilde{y}|X, \tilde{z}_a) = \tag{10}$$
$$= \tag{11}$$

where $\tilde{\beta}_a$ is a $p_{z_a} \times 1$ vector containing the p_{z_a} elements in M_a. In general this integration is very difficult, particularly when p, the dimension of $\tilde{\beta}$, is large. However, recall Zellner's g-prior had a form that facilitates efficient integration. Under this prior $p(\tilde{\beta}_a|\sigma^2, \tilde{z}_a) =$

It can be shown that integrating out $\tilde{\beta}_a$,
is fairly straightforward. This leaves:

$$p(\tilde{y}|X, \tilde{z}_a) = \tag{12}$$

Due to the form of the priors, this can also be easily integrated such that the marginal probability is:

$$p(\tilde{y}|X, \tilde{z}_a) = \pi^{-n/2}\frac{\Gamma([\nu_0 + n]/2}{\Gamma(\nu_0/2}(1+g)^{-p_{z_a}/2}\frac{(\nu_0\sigma_0^2)^{\nu_0/2}}{(\nu_0\sigma_0^2 + SSR_g^z)^{(\nu_0+n)/2}}, \tag{13}$$

where $SSR_g^z =$.

Given the marginal likelihoods, we can compute the posterior probability of a given model, M_a as:

$$Pr(M_a = \tilde{y}, X) = Pr(\tilde{z}_a|\tilde{y}, X) = \tag{14}$$

Using this formulation we can choose the most probable model or a set of the most probable models.

Bayesian Model Averaging

A powerful tool in Bayesian modeling, particularly for predictive settings, is Bayesian model averaging. Rather than choosing a single model, or set of models, we will average across models according to their posterior probability.

Assume we are interested in some quantity of interest Δ, which can be computed as a function of the posterior distribution. Then:

$$p(\Delta|X, \tilde{y}) = \tag{15}$$

.

For example let Δ represent the posterior predictive distribution for $\tilde{y}^*$, given X^*. Then the model averaged posterior predictive distribution can be written as,

$$p(\tilde{y}^*|, X^*X, \tilde{y}) = \tag{16}$$

This model averaged prediction is a special type of an ensemble method, which have nice predictive properties *and* in this case account for uncertainty in the model selection process.

General Model Selection and Averaging

In many cases, the g-prior framework is too restrictive or the models will not allow closed form solutions for the marginal likelihood. For instance consider the following model:

$$\tilde{y}|- \sim \tag{17}$$

where $H(\phi)$ is a correlation matrix. Finding the marginal (integrated) likelihood analytically would be very difficult in this case.

In situations like this, where model selection is often conducted using MCMC. Micaela will give us an overview of Gibbs sampling for linear regression, but the basic idea is that each iteration you:

-
-
-

where for each z_i the $Pr(z_i = 1|\tilde{y}, X, \tilde{z}_{-i})$ can be computed and $\tilde{z}_{-i}$ is all of the elements excluding i.