Regularizing Bayesian linear models with an informative prior on R^2

Ben Goodrich* Jonah Gabry* Andrew Gelman* Matthew Stephens†

25 February 2016

Abstract

In this paper we derive an approach for expressing prior beliefs about the location of the R^2, the familiar proportion of variance in the outcome variable that is attributable to the predictors under a linear model. In particular, when there are many predictors relative to the number of observations we would expect the joint prior derived here to work better than placing independent, heavy-tailed priors on the coefficients, which is standard practice in applied Bayesian data analysis but neither reflects the beliefs of the researcher nor conveys enough information to stabilize all the computations.

1 Introduction

Fully making Bayesian estimation of linear models routine for applied researchers requires prior distributions that work well for any data generated according to the assumptions of the likelihood function. Most Bayesian approaches require the researcher to specify a joint prior distribution for the regression coefficients (and the intercept and error variance), but most applied researchers have little inclination to specify all these prior distributions thoughtfully and take a short-cut by specifying one prior distribution that is taken to apply to all the regression coefficients as if they were independent of each other (and the intercept and error variance).

In this paper we derive and demonstrate an approach for expressing prior beliefs about the location of the R^2, the familiar proportion of variance in the outcome variable that is attributable to the predictors under a linear model. Since the R^2 is a well-understood bounded scalar, it is easy to specify prior information about it. In particular, when there are many predictors relative to the number of observations we would expect the joint prior

*Columbia University
†University of Chicago

derived here to work better than placing independent, heavy-tailed priors on the coefficients, which neither reflects the beliefs of the researcher nor conveys enough information to stabilize all the computations.

2 Likelihood

The likelihood for one observation y_i under a linear model can be written as the conditionally normal density

$$f(y_i \,|\, \mu_i, \sigma_\epsilon) = \frac{1}{\sigma_\epsilon \sqrt{2\pi}} \exp\left\{ -\frac{1}{2} \left(\frac{y_i - \mu_i}{\sigma_\epsilon} \right)^2 \right\},$$

where $\mu_i = \alpha + \mathbf{x}_i^\top \boldsymbol{\beta}$ is a linear predictor and σ_ϵ is the standard deviation of the error in predicting the outcome. The likelihood of the entire sample is the product of N individual likelihood contributions and it is well-known that the likelihood is maximized when the sum-of-squared residuals is minimized. This occurs when

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} &= \left(\mathbf{X}^\top \mathbf{X} \right)^{-1} \mathbf{X}^\top \mathbf{y}, \\
\widehat{\alpha} &= \overline{y} - \overline{\mathbf{x}}^\top \widehat{\boldsymbol{\beta}}, \\
\widehat{\sigma}_\epsilon^2 &= \frac{\left(\mathbf{y} - \widehat{\alpha} - \mathbf{X}\widehat{\boldsymbol{\beta}} \right)^\top \left(\mathbf{y} - \widehat{\alpha} - \mathbf{X}\widehat{\boldsymbol{\beta}} \right)}{N},
\end{aligned}$$

where $\overline{\mathbf{x}}$ is a vector that contains the sample means of the K predictors, $\mathbf{X}$ is a $N \times K$ matrix of *centered* predictors, $\mathbf{y}$ is a N-vector of outcomes, and $\overline{y}$ is the sample mean of the outcome.

Taking a QR decomposition of the design matrix, $\mathbf{X} = \mathbf{QR}$, where $\mathbf{Q}^\top \mathbf{Q} = \mathbf{I}$ and $\mathbf{R}$ is upper triangular, we can write the OLS solution for the coefficients as

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{X}^\top \mathbf{X} \right)^{-1} \mathbf{X}^\top \mathbf{y} = \mathbf{R}^{-1} \mathbf{Q}^\top \mathbf{y}.$$

The QR decomposition is often used for numeric stability reasons (see the familiar `lm` function in R), but, as we outline below, it is also useful for thinking about priors in a Bayesian version of the linear model.

3 Priors

The key innovation in this paper is the prior for the parameters in the QR-reparameterized model. To understand this prior, think about the equations that characterize the maximum likelihood solutions before observing the data on $\mathbf{X}$ and especially $\mathbf{y}$.

What would the prior distribution of $\boldsymbol{\theta} = \mathbf{Q}^\top \mathbf{y}$ be? We can write its k-th element as

$$\theta_k = \rho_k \sigma_Y \sqrt{N-1},$$

where ρ_k is the correlation between the kth column of $\mathbf{Q}$ and the outcome, σ_Y is the standard deviation of the outcome, and $1/\sqrt{N-1}$ is the standard deviation of the k column of $\mathbf{Q}$. Then let $\boldsymbol{\rho} = \sqrt{R^2}\mathbf{u}$ where $\mathbf{u}$ is a unit vector that is uniformally distributed on the surface of a hypersphere. Consequently, $R^2 = \boldsymbol{\rho}^\top \boldsymbol{\rho}$ is the familiar coefficient of determination for the linear model.

An uninformative prior on R^2 would be standard uniform, which is a special case of a Beta(a, b) distribution with both shape parameters a and b equal to 1. A non-uniform prior on R^2 is somewhat analogous to ridge regression, which is popular in data mining and produces better out-of-sample predictions than least squares because it penalizes $\boldsymbol{\beta}^\top \boldsymbol{\beta}$, usually after standardizing the predictors. In our case, an informative prior on R^2 effectively penalizes $\boldsymbol{\rho} \top \boldsymbol{\rho}$, which encourages $\boldsymbol{\beta} = \mathbf{R}^{-1}\boldsymbol{\theta}$ to be closer to the origin.

Lewandowski, Kurowicka, and Joe (2009) derives a distribution for a correlation matrix that depends on a single shape parameter $\eta > 0$, which implies the variance of one variable given the remaining K variables is Beta $\left(\eta, \frac{K}{2}\right)$. Thus, the R^2 is distributed Beta $\left(\frac{K}{2}, \eta\right)$ and any prior information about the location of R^2, which we will denote ℓ_{R^2}, can be used to choose a value of the hyperparameter η.

Four ways of choosing η via the specification of ℓ_{R^2} are:

1. ℓ_{R^2} is the prior mode on the $(0, 1)$ interval.

   This is only valid if $K \geq 2$ since the mode of a Beta $\left(\frac{K}{2}, \eta\right)$ distribution is $\left(\frac{K}{2} - 1\right) / \left(\frac{K}{2} + \eta - 2\right)$ and does not exist if $K < 2$.

2. ℓ_{R^2} is the prior mean on the $(0, 1)$ interval, where the mean of a Beta $\left(\frac{K}{2}, \eta\right)$ distribution is $\left(\frac{K}{2}\right) / \left(\frac{K}{2} + \eta\right)$.

3. ℓ_{R^2} is the prior median on the $(0, 1)$ interval.

   The median of a Beta $\left(\frac{K}{2}, \eta\right)$ distribution is not available in closed form, but if $K > 2$ it is approximately equal to $\left(\frac{K}{2} - \frac{1}{3}\right) / \left(\frac{K}{2} + \eta - \frac{2}{3}\right)$. Regardless of whether $K > 2$, we can numerically solve for the value of η that is consistent with a given prior median utilizing.

4. ℓ_{R^2} is some (negative) prior value for $\mathbb{E}\left(\log R^2\right) = \psi\left(\frac{K}{2}\right) - \psi\left(\frac{K}{2} + \eta\right)$, where $\psi(\cdot)$ is the Digamma function. Again, given a prior value for the left-hand side it is easy to numerically solve for the corresponding value of η.

For σ_y we set $\sigma_y = \omega s_y$ where s_y is the sample standard deviation of the outcome and $\omega > 0$ is an unknown scale parameter to be estimated. The only prior for ω that does not

contravene Bayes' theorem in this situation is Jeffreys prior,

$$f(\omega) \propto \frac{1}{\omega},$$

which is proportional to a Jeffreys prior on the unknown σ_y,

$$f(\sigma_y) \propto \frac{1}{\sigma_y} = \frac{1}{\omega \widehat{\sigma}_y} \propto \frac{1}{\omega}.$$

This parameterization and prior makes it easy to work with any continuous outcome variable, no matter what its units of measurement are.

We do not need a prior for σ_ϵ because our prior beliefs about $\sigma_\epsilon = \omega s_y \sqrt{1 - R^2}$ are already implied by our priors beliefs about ω and R^2. That only leaves a prior for $\alpha = \overline{y} - \overline{\mathbf{x}}^\top \mathbf{R}^{-1} \boldsymbol{\theta}$. As a default, an improper uniform prior is possible (the posterior is still proper), but something like a Gaussian prior with mean zero and standard deviation $\sigma_y / \sqrt{N}$ can also work well.

4 Posterior

The previous sections imply a posterior distribution for ω, α, $\mathbf{u}$, and R^2. After fitting the model, the parameters of interest can then be recovered as:

$$\begin{aligned}
\sigma_y &= \omega s_y \\
\sigma_\epsilon &= \sigma_y \sqrt{1 - R^2} \\
\boldsymbol{\beta} &= \mathbf{R}^{-1} \mathbf{u} \sigma_y \sqrt{R^2 (N - 1)}
\end{aligned}$$

When implementing this model, we actually utilize an improper uniform prior on $\log \omega$. Consequently, if $\log \omega = 0$, then the marginal standard deviation of the outcome *implied by the model* is the same as the sample standard deviation of the outcome. If $\log \omega > 0$, then the marginal standard deviation of the outcome implied by the model exceeds the sample standard deviation, so the model overfits the data. If $\log \omega < 0$, then the marginal standard deviation of the outcome implied by the model is less than the sample standard deviation, so the model *underfits* the data or that the data-generating process is nonlinear. Given the regularizing nature of the prior on R^2, a minor underfit would be considered ideal if the goal is to obtain good out-of-sample predictions. If the model badly underfits or overfits the data, then the model should be reconsidered.

5 Conclusion

Priors can be easy or hard for applied researchers to *specify* and easy or hard for applied researchers to *conceptualize*. Traditional shortcut priors on regression coefficients are often

used because they are both easy to specify and to conceptualize. The informative prior on R^2 proposed in this paper is more difficult to conceptualize but with the recent release of the `rstanarm` R package is equally easy to specify.

References

Gabry, J., and Goodrich, B. (2016). rstanarm: Bayesian Applied Regression Modeling via Stan. R package version 2.9.0-3. `http://mc-stan.org/interfaces/rstanarm`

Guan, Y., and Stephens, M. (2011) Bayesian variable selection regression for genome-wide association studies, and other large-scale problems. *Annals of Applied Statistics* **5**(3): 1780–1815.