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

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

25 February 2016

Abstract

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

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

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. For a sample of size N, the likelihood of the entire sample is the product of the 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

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

where $\overline{\mathbf{x}}$ is a vector of sample means for 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 ordinary least squares (OLS) solution for the regression 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 improved numerical stability (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, which can be thought of as a prior on the correlations between the outcome $\mathbf{y}$ and the columns of the orthogonal matrix $\mathbf{Q}$. To understand this prior, we start with the equations that characterize the maximum likelihood solutions *before* observing the data $(\mathbf{y}, \mathbf{X})$.

Let $\boldsymbol{\theta} = \mathbf{Q}^\top \mathbf{y}$. We can write the k-th element of the vector $\boldsymbol{\theta}$ 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 marginal 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, $\mathbf{u}^\top \mathbf{u} = 1$ implies that the sum of squared correlations is $\boldsymbol{\rho}^\top \boldsymbol{\rho} = r^2$, 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 shape parameters $a = b = 1$. 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 the regression coefficients $\boldsymbol{\beta} = \mathbf{R}^{-1}\boldsymbol{\theta}$ to be closer to the origin.

Consider a correlation matrix among both the outcome and the predictors of our reparameterized model. Lewandowski, Kurowicka, and Joe (2009) derives a distribution for a correlation matrix that depends on a single shape parameter $\eta > 0$ and implies that the conditional variance of one variable given the remaining K variables is distributed Beta $\left(\eta, \frac{K}{2}\right)$. In our case, this means that the conditional variance of $\mathbf{y}$ given the predictors, $1 - r^2$, has a Beta $\left(\eta, \frac{K}{2}\right)$ distribution. From the reflection symmetry of the Beta distribution, r^2 is therefore 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 implying a value for 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 $K > 2$ the median is approximately equal to $\left(\frac{K}{2} - \frac{1}{3}\right) / \left(\frac{K}{2} + \eta - \frac{2}{3}\right)$ (Kerman, 2011). Regardless of whether $K > 2$, we can numerically solve for the value of η that is consistent with a given prior median.

4. ℓ_{r^2} is some (negative) prior value for $E\left(\log r^2\right) = \psi\left(\frac{K}{2}\right) - \psi\left(\frac{K}{2} + \eta\right)$, 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.

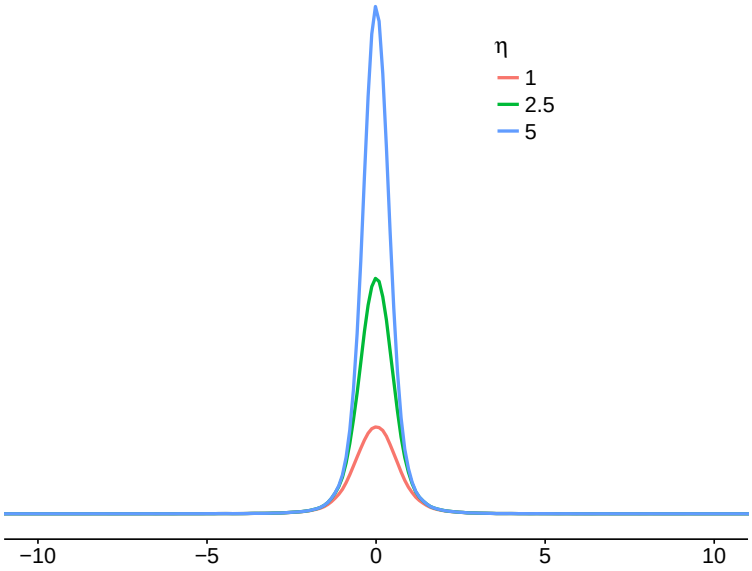

Figure 1: *Implied marginal prior on standardized regression coefficients (computed from 100,000 draws) with* $K = 10$ *and different values for* η*. As* η *increases the prior becomes more concentrated around zero.*

Each of these specifications of ℓ_{r^2} implies a value of the shape parameter η, which is the single hyperparameter of the joint prior on the coefficients. Smaller values for ℓ_{r^2} will correspond to larger values of η, smaller prior correlations among the outcome and predictors, and a prior density for the regression coefficients more concentrated around zero (see Figure 1).

Next we must specify a prior 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 the Jeffreys prior,

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

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

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

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

Finally, we need not directly specify a prior for σ_ϵ because our prior beliefs about σ_ϵ

are already implied by our beliefs about ω and r^2 via the relation

$$\sigma_\epsilon = \omega s_y \sqrt{1 - r^2}.$$

Thus, the only remaining distribution to specify is a prior for $\alpha = \overline{y} - \overline{\mathbf{x}}^\top \mathbf{R}^{-1} \boldsymbol{\theta}$. As a default, an improper uniform prior is possible as the posterior will still be proper.

4 Posterior

The previous sections imply a posterior distribution for ω, α, $\mathbf{u}$, and r^2. After fitting the model, we can recover the parameters of interest from the primitive parameters as:

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

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. Therefore, if $\log \omega > 0$, then the marginal standard deviation of the outcome implied by the model exceeds the sample standard deviation, which is to say that 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 and so the model underfits the data (or 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 Example

We have implemented the proposed prior distribution in the `stan_lm` function in the `rstanarm` R package (Gabry & Goodrich, 2016).[1] Here we provide a brief demonstration.

We will utilize an example from the `HSAUR3` R package, which is used in the third edition of *A Handbook of Statistical Analyses Using R* (Hothorn & Everitt, 2014). The model in section 5.3.1 analyzes an experiment where clouds were seeded with different amounts of silver iodide to see if there was increased rainfall. This effect could vary according to covariates, which (except for time) are interacted with the treatment variable. Most people would probably be skeptical that cloud hacking could explain very much of the variation in rainfall and thus the prior mode of the r^2 would probably be fairly small.

[1] Source code can be found at `https://github.com/stan-dev/rstanarm`.

The frequentist estimator of this model can be replicated in R by executing

```
> data("clouds", package = "HSAUR3")
> mod <- rainfall ~ seeding * (sne+cloudcover+prewetness+echomotion) + time
> ols <- lm(formula = mod, data = clouds)
```

from which we obtain the following estimated coefficients:

```
> round(coef(ols), digits = 3)
          (Intercept)                          seedingyes
               -0.346                              15.683
          sne                                  cloudcover
               0.420                               0.388
          prewetness                           echomotionstationary
               4.108                               3.153
          time                                 seedingyes:sne
               -0.045                              -3.197
          seedingyes:cloudcover                seedingyes:prewetness
               -0.486                              -2.557
          seedingyes:echomotionstationary
               -0.562
```

Note that we have *not* looked at the estimated r^2 or σ for the OLS model. We can estimate a Bayesian version of this model using the `rstanarm` package by prepending `stan_` to the `lm` call and specifying a prior mode for r^2 using the `R2` function:

```
> library("rstanarm")
> post <- stan_lm(formula = mod, data = clouds,
                  prior = R2(location = 0.2, what = "mode"))
> round(coef(post), digits = 3)
          (Intercept)                          seedingyes
               2.303                               6.753
          sne                                  cloudcover
               0.193                               0.162
          prewetness                           echomotionstationary
               1.751                               1.345
          time                                 seedingyes:sne
               -0.019                              -1.407
          seedingyes:cloudcover                seedingyes:prewetness
               -0.204                              -1.068
          seedingyes:echomotionstationary
               -0.264
```

The point estimates from the Bayesian model, which are represented by the posterior medians, appear quite different from the OLS estimates. However, the `log-fit_ratio` (i.e. $\log \omega$) is quite small

```
> summary(post)["log-fit_ratio", "50%"]
          -0.010
```

which indicates that the model only slightly overfits the data when the prior derived above is utilized. Thus, it would be safe to conclude that the OLS estimator considerably overfits the data since there are only 24 observations with which to estimate 12 parameters and no prior information is leveraged. In general, we would expect the prior derived in this paper to be well-suited in situations like the one above, where there are many predictors relative to the number of observations.

6 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 it is equally easy to specify.

References

Gabry, J., & Goodrich, B. (2016). *rstanarm: Bayesian applied regression modeling via Stan.* Retrieved from `http://cran.r-project.org/package=rstanarm` (R package version 2.9.0-3)

Guan, Y., & 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.

Hothorn, T., & Everitt, B. S. (2014). *A handbook of statistical analyses using R* (third ed.). Chapman & Hall/CRC.

Hothorn, T., & Everitt, B. S. (2015). *HSAUR3: A handbook of statistical analyses using R.* Retrieved from `http://cran.r-project.org/package=HSAUR3` (R package version 1.0-5)

Kerman, J. (2011). A closed-form approximation for the median of the beta distribution. *arXiv:1111.0433 [math.ST]*.

Lewandowski, D., Kurowicka, D., & Joe, H. (2009). Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis*, *100*(9), 1989–2001.

R Core Team. (2015). *R: A language and environment for statistical computing. R Foundation for Statistical Computing.* Vienna, Austria. Retrieved from `https://www.R-project.org/`