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

Ben Goodrich Jonah Gabry Andrew Gelman

Columbia University

29 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 of these prior distributions thoughtfully and take a shortcut by specifying a single prior distribution that is taken to apply to all 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 directly 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. Our work shares some common themes with other research. For example, Guan and Stephens (2011) proposes a sparcity-inducing prior on the coefficients of a linear model that has implications for the R^2. In particular, their prior relaxes the assumption of prior independence between the parameter representing the typical magnitude of the non-zero coefficients and the parameter governing

sparcity, which allows for the a priori expectation that the proportion of variance explained does not necessarily increase markedly with model complexity.

Our approach does not induce sparcity in that all coefficients will be nonzero, but does lead to substantial shrinkage towards zero. The degree of shrinkage depends on the number of estimated effects and, in particular, on prior information provided by the researcher about R^2 . Since the R^2 is a well-understood bounded scalar it is easy to specify prior information about it and, for most applied problems, researchers will have enough familiarity with their subject matter to have some a priori knowledge about the extent to which their predictors will account for variation in the outcome. 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 regression coefficients in the linear model, which neither reflects the beliefs of the researcher nor conveys enough information to stabilize all the computations.

2 QR-reparameterized likelihood

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

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

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 it 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 maximum likelihood estimate for the regression coefficients — the ordinary least squares (OLS) solution — 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 Specification of the joint prior distribution

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.

3.1 Prior for R^2

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 only on a single shape parameter $\eta > 0$ and implies that the conditional variance of one variable given the remaining K variables has a Beta distribution with parameters $a = \eta$ and $b = \frac{K}{2}$. In our case, this means that the conditional variance of $\mathbf{y}$ (given the predictors) is

$$(1 - R^2) \sim \text{Beta}\left(\eta, \frac{K}{2}\right).$$

From the reflection symmetry of the Beta distribution it follows immediately that our prior on the R^2 itself must be

$$R^2 \sim \text{Beta}\left(\frac{K}{2}, \eta\right).$$

Any available 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.

The following are four ways of implying the value of η by taking ℓ_{R^2} to be (a) the prior mode of R^2, (b) the prior median of R^2, (c) the prior mean of R^2, and (d) the prior mean of $\log R^2$:

(a) $\ell_{R^2} \in (0, 1)$ is the prior mode, where the mode of a Beta $\left(\frac{K}{2}, \eta\right)$ distribution is $\left(\frac{K}{2} - 1\right) / \left(\frac{K}{2} + \eta - 2\right)$. This is only valid for $K \geq 2$ as otherwise the mode does not exist. If the mode does exist then we obtain

$$\eta = \frac{\frac{K}{2}\left(1 - \ell_{R^2}\right) + 2\ell_{R^2} - 1}{\ell_{R^2}}.$$

(b) $\ell_{R^2} \in (0, 1)$ is the prior mean, where the mean of a Beta $\left(\frac{K}{2}, \eta\right)$ distribution is $\left(\frac{K}{2}\right) / \left(\frac{K}{2} + \eta\right)$. In this case we have

$$\eta = \frac{\frac{K}{2}\left(1 - \ell_{R^2}\right)}{\ell_{R^2}}.$$

(c) $\ell_{R^2} \in (0, 1)$ is the prior median. The median is not available in closed form, but if $K > 2$ the median of a Beta $\left(\frac{K}{2}, \eta\right)$ is approximately equal to $\left(\frac{K}{2} - \frac{1}{3}\right) / \left(\frac{K}{2} + \eta - \frac{2}{3}\right)$ (Kerman, 2011). However, even if $K \leq 2$, we can numerically solve for the value of η that is consistent with a given prior median.

(d) $\ell_{R^2} \in (-\infty, 0)$ is the prior expectation of $\log R^2$, which can be expressed in terms of the Digamma function $\psi(\cdot)$ as $E\left(\log R^2\right) = \psi\left(\frac{K}{2}\right) - \psi\left(\frac{K}{2} + \eta\right)$. Again, given a prior value for the left-hand side we can numerically solve for the corresponding value of the shape hyperparameter η.

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).

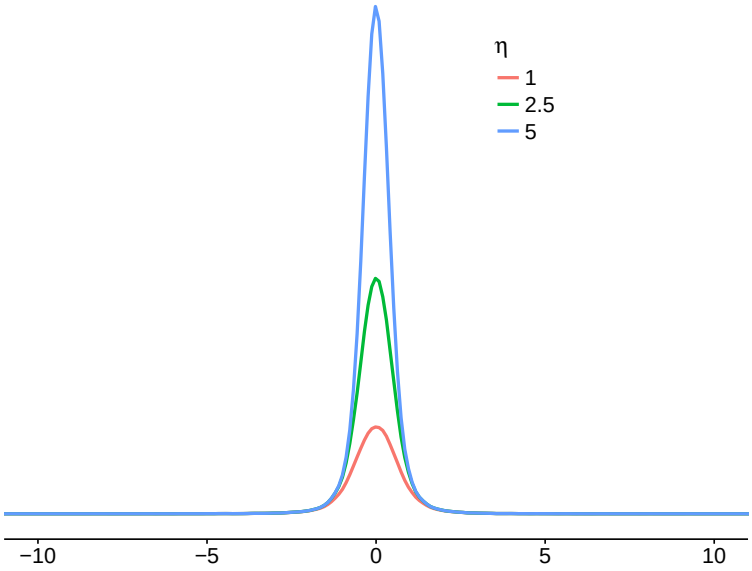

Figure 1: *Implied marginal prior for one of the K standardized regression coefficients (computed from 100,000 draws). Here the value of K is fixed at 10 and the plotted densities correspond to different values of the hyperparameter η. For larger values of η the prior is more concentrated around zero.*

3.2 Log fit-ratio and prior for σ_y

Let $\sigma_y = \omega s_y$, where s_y is the sample standard deviation of the outcome and $\omega > 0$ is an unknown scale parameter to be estimated. We use the scale-invariant Jeffreys prior $f_\omega(\omega) \propto 1/\omega$, 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 is the only prior that does not contravene Bayes' theorem in this situation, as any other prior would result in the marginal standard deviation of the outcome being a function of the estimated standard deviation of the outcome. This parameterization and prior also makes it easy to work with any continuous outcome variable, regardless of the unit of measurement.

When implementing the model we prefer to work with ω on the log scale so we use the flat prior $f_\phi(\phi) \propto 1$, where $\phi = \log \omega$, which is equivalent to the Jeffreys prior on ω itself. We refer to ϕ as the *log fit-ratio* since it is the logarithm of the ratio of the marginal standard deviation of the outcome implied by the model to the observed sample standard deviation,

$$\phi = \log \omega = \log \frac{\sigma_y}{s_y}.$$

We can interpret the log fit-ratio as a measure of underfitting or overfitting. If $\phi = 0$, then the marginal standard deviation of the outcome is the same as the sample standard deviation of the outcome. If $\phi > 0$, 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. Otherwise $\phi < 0$, in which case 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.

3.3 Prior for α

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 $f_\alpha(\alpha) \propto 1$ is possible as the posterior will still be proper.

4 Posterior

The previous sections imply a joint posterior distribution for the primitive parameters $\left(\phi, \alpha, \mathbf{u}, R^2\right)$. As demonstrated in Section 5, draws from the posterior can be obtained via Markov chain Monte Carlo (MCMC). We can then easily recover the parameters of primary interest from the primitive parameters as

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

where these computations are performed for each posterior draw of the parameters.

5 Example

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

[1] Source code for the `rstanarm` package can be found at `https://github.com/stan-dev/rstanarm`. The Stan code used by the `stan_lm` function is available in the `lm.stan` file in the `exec` directory.

demonstration.

We will utilize an example from the `HSAUR3` R package, which is the companion R package to 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 should be fairly small.

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:

```
> s <- summary(ols)$coef
> round(s[, 1:2], 2)
                                 Estimate Std. Error
(Intercept)                         -0.35       2.79
seedingyes                          15.68       4.45
sne                                  0.42       0.84
cloudcover                           0.39       0.22
prewetness                           4.11       3.60
echomotionstationary                 3.15       1.93
time                                -0.04       0.03
seedingyes:sne                      -3.20       1.27
seedingyes:cloudcover               -0.49       0.24
seedingyes:prewetness               -2.56       4.48
seedingyes:echomotionstationary     -0.56       2.64
```

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, which will use Stan (Stan Development Team, 2016) to draw from the posterior distribution via MCMC. To fit the model we simply prepend `stan_` to the `lm` call and specify a prior mode for R^2 using the `R2` function.[2]

[2]The `what` argument to the `R2` function can take the values `mode`, `mean`, `median`, and `log`, corresponding to the four methods for choosing η via the specification of ℓ_{R^2} detailed in Section 3.1.

```
> library("rstanarm")
> post <- stan_lm(formula = mod, data = clouds,
                  prior = R2(location = 0.2, what = "mode"))
> print(post)
                                   Median MAD_SD
(Intercept)                         2.4    2.2
seedingyes                          6.5    3.5
sne                                 0.2    0.7
cloudcover                          0.2    0.2
prewetness                          1.8    2.8
echomotionstationary                1.3    1.5
time                                0.0    0.0
seedingyes:sne                     -1.3    1.0
seedingyes:cloudcover              -0.2    0.2
seedingyes:prewetness              -1.1    3.4
seedingyes:echomotionstationary    -0.2    2.0
sigma                               2.6    0.4
log-fit_ratio                       0.0    0.1
R2                                  0.3    0.1
```

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 ($\log \omega$) is estimated to be about 0, which indicates a good fit of the model to the data. 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. Although the proposed informative prior on R^2 is certainly more difficult to conceptualize, with the implementation in the `rstanarm` package it is now equally easy to specify.

In this paper we have only focused on the most basic linear regression models. However, the proposed framework is flexible enough to accomodate more complicated scenarios. For instance, a natural extension would be to allow for the stratification of observations and the shrinking of R^2 values within each stratum toward a common global value.[3] We plan

[3]See Gelman and Pardoe (2006) for an example of how R^2 can be defined at each level of a hierarchical model.

to implement this feature in future versions of `rstanarm`.

References

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

Gelman, A., & Pardoe, I. (2006). Bayesian measures of explained variance and pooling in multilevel (hierarchical) models. *Technometrics*, *48*(2), 241–251.

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.* R package version 1.0-5. `http://cran.r-project.org/package=HSAUR3`.

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.* `https://www.R-project.org/`. Vienna, Austria.

Stan Development Team. (2016). *Stan, Version 2.9.0.* `http://mc-stan.org`.