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

Ben Goodrich* Jonah Gabry* Andrew Gelman*

28 March 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

*Columbia University, New York.

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

Our approach results in a regularizing prior on the coefficients but does not induce sparcity in the sense of setting coefficients equal to exactly 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.

The paper is organized as follows. We begin in Section 2 with a brief review of the linear model with a QR decomposition applied to the design matrix. Section 3 covers the specification of our proposed joint prior distribution for the parameters in the QR-reparameterized model. In Section 4 we show how to recover the parameters of interest from the posterior distribution of the primitive parameters of the reparameterized model. Finally, in Section 5 we demonstrate our implementation of the proposed model in the `rstanarm` R package and then conclude with a brief discussion of possible extensions to the work presented here.

2 QR-reparameterized likelihood

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

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

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 centered 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 least squares solution — as

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

It is helpful for intuition to notice that when substituting $\mathbf{QR}$ for $\mathbf{X}$ in the linear model we obtain another linear model

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \implies \mathbf{y} = \mathbf{Q}\boldsymbol{\theta} + \boldsymbol{\epsilon},$$

where the relationship between $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ is given by

$$\boldsymbol{\beta} = \mathbf{R}^{-1}\boldsymbol{\theta}.$$

That is, we have moved from a regression of $\mathbf{y}$ on $\mathbf{X}$ to a regression of $\mathbf{y}$ on the orthogonal matrix $\mathbf{Q}$.

The QR decomposition is often used in this way for improved numerical stability (as in 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. To understand this prior, we start with the equations that characterize the maximum likelihood solutions *before* observing the data $(\mathbf{y}, \mathbf{X})$.

Let $\boldsymbol{\theta} = \mathbf{R}\boldsymbol{\beta} = \mathbf{Q}^\top\mathbf{y}$ and $\mathbf{Q}_k$ denote the kth column of $\mathbf{Q}$. We can write the kth element of the vector $\boldsymbol{\theta}$ as

$$\theta_k = \text{Corr}(\mathbf{y}, \mathbf{Q}_k)\,\frac{\text{sd}(\mathbf{y})}{\text{sd}(\mathbf{Q}_k)} = \rho_k\,\sigma_y\,\sqrt{N-1},$$

where ρ_k is the correlation between $\mathbf{Q}_k$ and the outcome, σ_y is the marginal standard deviation of the outcome, and $1/\sqrt{N-1}$ is the standard deviation of $\mathbf{Q}_k$.

We will return to σ_y in Section 3.2 and for now focus on specifying a prior distribution for the vector of correlations $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_K)$, that is, a prior on the correlations between $\mathbf{y}$ and the columns of $\mathbf{Q}$. Here we assume that $\boldsymbol{\rho}$ has a spherical distribution represented by $\boldsymbol{\rho} = \sqrt{R^2}\,\mathbf{u}$, where $\mathbf{u}$ is a unit vector uniformally distributed on the surface of a hypersphere. Consequently, $\mathbf{u}^\top\mathbf{u} = 1$ implies that the sum of squared correlations is $\boldsymbol{\rho}^\top\boldsymbol{\rho} = R^2$. This is the familiar coefficient of determination for the linear model, which can be interpreted as the proportion of variance in $\mathbf{y}$ attributable to $\mathbf{X}$.[1]

[1] The R^2 is the same regardless of whether we use $\mathbf{Q}\theta$ or $\mathbf{X}\beta$.

3.1 Prior for R^2

An uninformative prior on R^2 would be standard uniform, which is a special case of a $\mathsf{Beta}\,(a, b)$ 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 will effectively penalize $\boldsymbol{\rho}^\top \boldsymbol{\rho}$, 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 $b = \frac{K}{2}$. This means that the conditional variance of $\mathbf{y}$ (given the predictors) is

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

and from the reflection symmetry of the Beta distribution it follows that our prior on the R^2 itself is

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

Any available prior information about the location of R^2, which we will denote ℓ_{R^2}, can be used to choose a value for 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.

The mode of a $\mathsf{Beta}\left(\frac{K}{2}, \eta\right)$ distribution is $\left(\frac{K}{2} - 1\right) / \left(\frac{K}{2} + \eta - 2\right)$, which exists when the model has at least three predictors.[2] If the mode does exist then

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

The relationship between η and the prior mode ℓ_{R^2} for several different values of K is shown in Figure 1.

(b) $\ell_{R^2} \in (0, 1)$ is the prior mean.

[2] For the mode to exist both shape parameters must be greater than 1. If there are fewer than three predictors then $K \leq 2 \implies K/2 \leq 1$ and this condition is not satisfied.

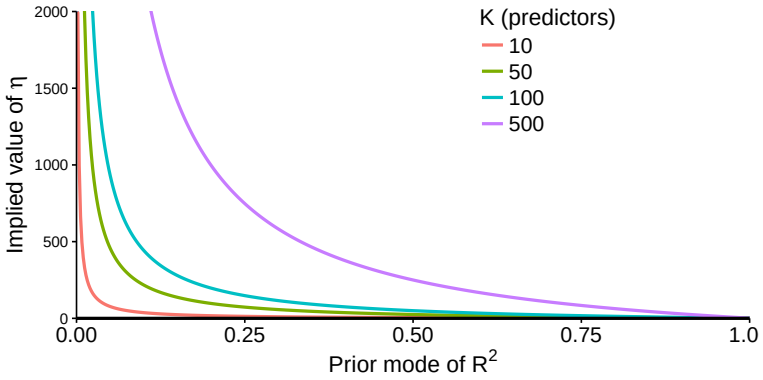

Figure 1: *The value of the shape parameter η as a function of the prior mode of R^2 for different values of K, the number of predictors. In order to make the plotted curves easily visible we zoom vertically to the region between 0 and 2000 on the y-axis.*

The mean of a Beta $\left(\frac{K}{2}, \eta\right)$ distribution is $\left(\frac{K}{2}\right) / \left(\frac{K}{2} + \eta\right)$. Solving for η we obtain

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

(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 value for the prior median.

(d) $\ell_{R^2} \in (-\infty, 0)$ is the prior expectation of $\log R^2$.

The expectation $\mathbb{E}\left(\log R^2\right)$ can be expressed in terms of the Digamma function $\psi(\cdot)$ as $\mathbb{E}\left(\log R^2\right) = \psi\left(\frac{K}{2}\right) - \psi\left(\frac{K}{2} + \eta\right)$. 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 (Figure 1), smaller prior correlations among the outcome and predictors, and a prior density for the regression coefficients more concentrated around zero (Figure 2).

3.2 Prior for the marginal standard deviation

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

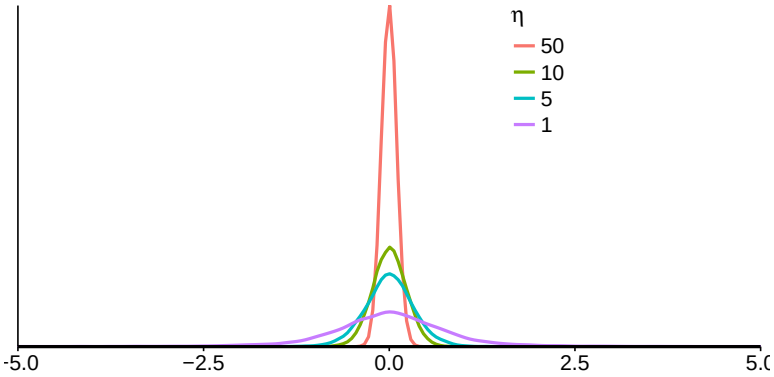

Figure 2: *Implied marginal prior for one of the K standardized regression coefficients (computed from 10,000 draws). Here the value of K is fixed at 100 and the plotted densities correspond to different values of the hyperparameter η. As the value of η increases the prior becomes increasingly concentrated around zero.*

$f_\omega(\omega) \propto 1/\omega$, which is proportional to a Jeffreys prior on the unknown σ_y,

$$f_{\sigma_y}(\sigma_y) \propto \frac{1}{\sigma_y} = \frac{1}{\omega s_y} \propto \frac{1}{\omega}.$$

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 combination of 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 ϕ as a measure of underfitting or overfitting. There are three general possibilities:

1. If $\phi = 0$, then the marginal standard deviation of the outcome implied by the model is the same as the sample standard deviation of the outcome ($\sigma_y = s_y$).

2. If $\phi > 0$, then the marginal standard deviation of the outcome implied by the model exceeds the sample standard deviation ($\sigma_y > s_y$). That is, the model overfits the data.

3. If $\phi < 0$, then the marginal standard deviation of the outcome implied by the model is less than the sample standard deviation ($\sigma_y < s_y$). Either 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 the model intercept

We need not directly specify a prior for σ_ϵ because our prior beliefs about σ_ϵ are fully implied by our beliefs about ω and R^2 via the relations

$$\sigma_\epsilon = \sigma_y\sqrt{1 - R^2} = \omega s_y\sqrt{1 - R^2},$$

which follow from the definition of R^2 and the definition of σ_y in terms of ω and s_y from Section 3.2. Thus, the only remaining distribution to specify is the prior for the model intercept

$$\alpha = \overline{y} - \overline{\mathbf{x}}^\top \mathbf{R}^{-1}\boldsymbol{\theta}.$$

As a default, a flat (improper uniform) prior $f_\alpha(\alpha) \propto 1$ is possible as the resulting posterior distribution will be proper. Other choices of f_α are also viable, for instance a zero-mean normal with scale $\sigma_y/\sqrt{N}$.

4 Posterior

The previous sections imply a joint posterior distribution for the primitive parameters

$$f_{post}\left(\phi, \alpha, \mathbf{u}, R^2 \,|\, \mathbf{y}, \mathbf{X} = \mathbf{QR}, \eta\right).$$

In Section 5 we discuss sampling from this posterior using Markov chain Monte Carlo (MCMC). Here, assume we have already obtained a sample of size S from f_{post} consisting of MCMC draws

$$\left(\phi, \alpha, \mathbf{u}, R^2\right)^{(s)} = \left(\phi^{(s)}, \alpha^{(s)}, \mathbf{u}^{(s)}, R^{2(s)}\right), \quad s = 1, \ldots, S.$$

Although some of these parameters are themselves useful (e.g, the log fit-ratio, ϕ), we would like to make inferences with the posterior distribution of the parameters of interest $\boldsymbol{\beta}$, σ_y, and σ_ϵ. Following our definitions, we can obtain a sample from the posterior of $(\boldsymbol{\beta}, \sigma_y, \sigma_\epsilon)$ from the MCMC draws of the primitive parameters by computing

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

for each draw $s = 1, \ldots, S$.

5 Example

We have implemented the proposed model and prior distribution in the `stan_lm` function in the `rstanarm` R package (Gabry & Goodrich, 2016). In this section we provide a brief 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 least squares estimator of this model can be replicated in R by executing

```
> library(arm)  # we prefer arm's display() for printing lm() results
> data("clouds", package = "HSAUR3")
> mod <- rainfall ~ seeding * (sne+cloudcover+prewetness+echomotion) + time
> mle <- lm(formula = mod, data = clouds)
```

from which we obtain the following estimated coefficients:

```
> display(mle)
lm(formula = mod, data = clouds)
                              coef.est coef.se
(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
---
n = 24, k = 11
residual sd = 2.20, R-Squared = 0.72
```

Note that we have *not* looked at the estimated R^2 or σ for the least squares model. We can estimate a Bayesian version of this model using the `rstanarm` package, which calls Stan (Stan Development Team, 2016) to draw from the posterior distribution of the primitive parameters via MCMC and then uses the equations in Section 4 to obtain the estimates to be returned to the user.[3] To fit the model we simply prepend `stan_` to the `lm` call and specify a prior mode for R^2 using the `R2` function. 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")
> R2prior <- R2(location = 0.2, what = "mode")
> post <- stan_lm(formula = mod, data = clouds, prior = R2prior)
> print(post)
stan_lm(formula = mod, data = clouds, prior = R2prior)

Estimates:
                                  Median MAD_SD
(Intercept)                        2.5    2.2
seedingyes                         6.6    3.7
sne                                0.2    0.6
cloudcover                         0.2    0.2
prewetness                         1.6    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             -0.9    3.5
seedingyes:echomotionstationary   -0.2    2.0
sigma                              2.6    0.4
log-fit_ratio                      0.0    0.1
R2                                 0.3    0.1

Sample avg. posterior predictive distribution of y (X = xbar):
         Median MAD_SD
mean_PPD 4.5    2.7
```

The point estimates from the Bayesian model, which are represented by the posterior medians, appear quite different from the least squares estimates. However, the log fit-ratio is estimated to be about 0, which indicates a good fit of the model to the data.[4] It would be

[3]The `rstanarm` package is available from CRAN and source code can be found at `https://github.com/stan-dev/rstanarm`. The Stan code relevant to this paper is available in the `lm.stan` file in the `exec` directory.

[4]The uncertainty estimates labeled `MAD_SD` are proportional to the median absolute deviation (MAD) from the posterior median. The `rstanarm` package reports `MAD_SD` rather than the raw posterior standard deviations because the former will be more robust for long-tailed distributions.

safe to conclude that the least squares 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 for regression coefficients are often used because they are both easy to specify and to conceptualize. In comparison, the informative prior on R^2 proposed in this paper is difficult to conceptualize, but with its 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.[5] We plan 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.

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

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`.