Analytical approximation for LOO-R2 standard error

Leevi Lindgren *

June 9, 2022

1 Introduction

R-squared, or R2, is a typical measure used to assess how well a model fits data. Gelman et al. (2019) propose a Bayesian version of the R2 as the classical R2 might get values larger than 1 for Bayesian regression models. See details from the paper. A nice property of Bayesian R2 is that we get a distribution of R2 values "for free", as it is computed using posterior predictive mean values of the model. This means that we can quantify the uncertainty of the estimator easily.

If (Bayesian) R2 is computed from the same data that was used to fit the model, it will give an overestimate of the predictive performance on a new, unobserved data. As often we don't have independent test data, we can use cross-validation to estimate the out-of-sample behavior of the R2. Vehtari et al. (2016) propose an efficient and stable method for leave-one-out cross-validation using Pareto smoothed importance sampling. In the paper, log predictive density is used as the utility describing the predictive performance, but the method can easily be extended for other utilities as well, such as R2.

After computing LOO-R2 estimate, we obviously want to quantify the uncertainty of the estimator. The uncertainty in LOO-R2 comes from not knowing the future data distribution (Vehtari and Ojanen, 2012). One way to do it is to use Bayesian bootstrap (Rubin, 1981). However, this report describes how to quantify the uncertainty using an analytical Taylor approximation approach.

The idea is to note that LOO-R2 is defined as a function of a ratio of two random variables. We can then use Taylor approximation of the function and the mean, variance, and covariance of the two random variables to approximate the standard error of the LOO-R2 estimator.

This approach was motivated by the work by Hastings (1970) who proposes an approximation for the variance of the ratio of two random variables. However, the formula provided by Hastings (1970) contains two typos: on page 8, in the formula for the variance of the ratio, $\bar{Z}$ is missing the "bar" symbol and computed variance $s_{\bar{Z}}$ is missing the second power. The missing second power

*I thank Aki Vehtari and Nikolas Siccha for their comments.

caused the author of this report to mistake it for the standard deviation instead of variance, which was followed by weird simulation results.

2 Taylor approximation for a function of two variables

Let $f(x, y) : \mathbb{R}^2 \to \mathbb{R}$ be a function of two (random) variables. The first order Taylor approximation of a function of two variables around point x_0, y_0 is given by

$$f(x, y) \approx f(x_0, y_0) + \frac{\partial}{\partial x} f(x_0, y_0)(x - x_0) + \frac{\partial}{\partial y} f(x_0, y_0)(y - y_0) \tag{1}$$

Now, let's take two random variables X and Y and do the approximation around the expected values $(\mu_X, \mu_Y) = (E[X], E[Y])$. Using (1) we get

$$f(X, Y) \approx f(\mu_X, \mu_Y) + \frac{\partial}{\partial x} f(\mu_X, \mu_Y)(X - \mu_X) + \frac{\partial}{\partial y} f(\mu_X, \mu_Y)(Y - \mu_Y) \tag{2}$$

2.1 Taylor approximation for the expected value

Given that terms containing partial derivatives in (2) go to zero under expectation, we get the following, simple, first order approximation for the expected value of $f(X, Y)$:

$$E[f(X, Y)] \approx f(\mu_X, \mu_Y) \tag{3}$$

2.2 Taylor approximation for variance

For variance

$$\begin{aligned} \operatorname{var}(f(X, Y)) &= E\left[\left(f(X, Y) - E[f(X, Y)]\right)^2\right] \\ &\approx E\left[\left(f(X, Y) - f(\mu_X, \mu_Y)\right)^2\right] \end{aligned}$$

Next, we plug in (2) for $f(X, Y)$ which yields (we write $\frac{\partial}{\partial x} f(\mu_X, \mu_Y) = \frac{\partial f}{\partial x}$ for notational simplicity):

$$\begin{aligned} \operatorname{var}(f(X, Y)) &\approx E\left[\left(\frac{\partial f}{\partial x}(X - \mu_X) + \frac{\partial f}{\partial y}(Y - \mu_Y)\right)^2\right] \\ &= E\left[\left(\frac{\partial f}{\partial x}\right)^2 (X - \mu_X)^2 + 2\frac{\partial f}{\partial x}\frac{\partial f}{\partial y}(X - \mu_X)(Y - \mu_Y) + \left(\frac{\partial f}{\partial y}\right)^2 (Y - \mu_Y)^2\right] \\ &= \left(\frac{\partial f}{\partial x}\right)^2 \operatorname{var}(X) + 2\frac{\partial f}{\partial x}\frac{\partial f}{\partial y}\operatorname{cov}(X, Y) + \left(\frac{\partial f}{\partial y}\right)^2 \operatorname{var}(Y) \end{aligned} \tag{4}$$

3 Approximation for the standard error of LOO-R2

LOO-R2 is defined as

$$\mathrm{R2}_{loo} = 1 - \frac{\mathrm{var}(\hat{e}_{loo})}{\mathrm{var}(y)} \tag{5}$$

where $\hat{e}_{loo} = y - \hat{y}_{loo}$. Note that the nominator and denominator of the second term in (5) can be interpreted as the mean squared error of the LOO predictions and mean squared error of predicting the data with its mean, respectively. So we write (5) as

$$\mathrm{R2}_{loo} = 1 - \frac{\mathrm{MSE}_{\hat{e}}}{\mathrm{MSE}_{y}} \tag{6}$$

We can then compute the estimator for the variance of both using the same approach as in Sivula et al. (2022) and Vehtari et al. (2016):

$$\mathrm{var}(\mathrm{MSE}_{\hat{e}}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\hat{e}_{loo,i}^2 - \mathrm{MSE}_{\hat{e}}\right)^2 \tag{7}$$

and

$$\mathrm{var}(\mathrm{MSE}_{y}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left((y_i - \hat{y})^2 - \mathrm{MSE}_{y}\right)^2, \tag{8}$$

where $\bar{y}$ is the sample mean of observations y.

To utilize the Taylor approximation, we need to compute partial derivatives of function $f(x, y) = 1 - \frac{x}{y}$. With a simple calculus, we get

$$\frac{\partial}{\partial x} f(x, y) = -\frac{1}{y} \tag{9}$$

$$\frac{\partial}{\partial y} f(x, y) = \frac{x}{y^2}. \tag{10}$$

Substituting these into (4) yields

$$\mathrm{var}(f(X, Y) \approx \frac{1}{\mu_Y^2} \left(\mathrm{var}(X) - 2\frac{\mu_X}{\mu_Y}\mathrm{cov}(X, Y) + \left(\frac{\mu_X}{\mu_Y}\right)^2 \mathrm{var}(Y)\right). \tag{11}$$

To use this expression for the mean squared errors, we also need the covariance between $\mathrm{MSE}_{\hat{e}}$ and MSE_{y}. This can be estimated in a similar way as the variances:

$$\mathrm{cov}(\mathrm{MSE}_{\hat{e}}, \mathrm{MSE}_{y}) = \frac{1}{n(n-1)} \sum_{i=1}^{n} \left(\hat{e}_{loo,i}^2 - \mathrm{MSE}_{\hat{e}}\right) \left((y_i - \hat{y})^2 - \mathrm{MSE}_{y}\right). \tag{12}$$

Putting all pieces together, variance, and consequently the standard error, of LOO-R2 estimator can then be approximated by letting $\mu_X = \text{MSE}_{\hat{e}}$ and $\mu_Y = \text{MSE}_y$ and then substituting with (7), (8) and (12) into (11):

$$\text{var}(\text{R2}_{loo}) = \frac{1}{\text{MSE}_y^2}\left(\text{var}(\text{MSE}_{\hat{e}}) - 2\frac{\text{MSE}_{\hat{e}}}{\text{MSE}_y}\text{cov}(\text{MSE}_{\hat{e}}, \text{MSE}_y) + \left(\frac{\text{MSE}_{\hat{e}}}{\text{MSE}_y}\right)^2 \text{var}(\text{MSE}_y)\right). \tag{13}$$

4 Simulation experiments

To illustrate behaviour of the proposed analytical approximation, we run a set of simulations with a few different configurations. We proceed as follows: choose number of observations n, number of predictors p and observational noise σ and then, draw predictor matrix $X \in \mathbb{R}^{n \times p}$ and $\beta \in \mathbb{R}^p$ from $N(0, 1)$ and $\epsilon \in \mathbb{R}^n$ from $N(0, \sigma^2)$. Then we compute $y = X\beta + \epsilon$ for which we with linear model using R-package rstanarm (Goodrich et al., 2020). We repeat this for 100 for each set of parameters. See underlying code from `https://github.com/LeeviLindgren/loo-r2-se`.

Figure 1 plots standard error obtained from the Taylor approximation and Bayesian bootstrap for different values of n, σ, and p. The diagonal line represents points where x and y-axis values are equal and the actual simulation results are represented with the dots. Dots' color indicates the underlying LOO-R2 estimate obtained by taking the mean of BB samples.

Figure 2 runs similar experiments, but we focus for lower values of n, 5 or 10. Not surprisingly, we see more dispersion in the standard error estimates. One pattern seems to be that when the R2 estimate is small (darker dots) compared to the standard error, we start to bias in the Taylor approximation.

Finally, we run three additional simulations, where we set the prior of β as $N(0, 10)$ and let the number of predictors be either 20, 50, or 100 and the number of observations be $n = 50$. With the wide prior and high number of predictors compared to n, models are likely to overfit the data badly, and we should be observing R2 values close to 1. From figure 3 we observe that when the number of predictors is 50 or 100, Taylor approximation seems to produce downward biased estimates.

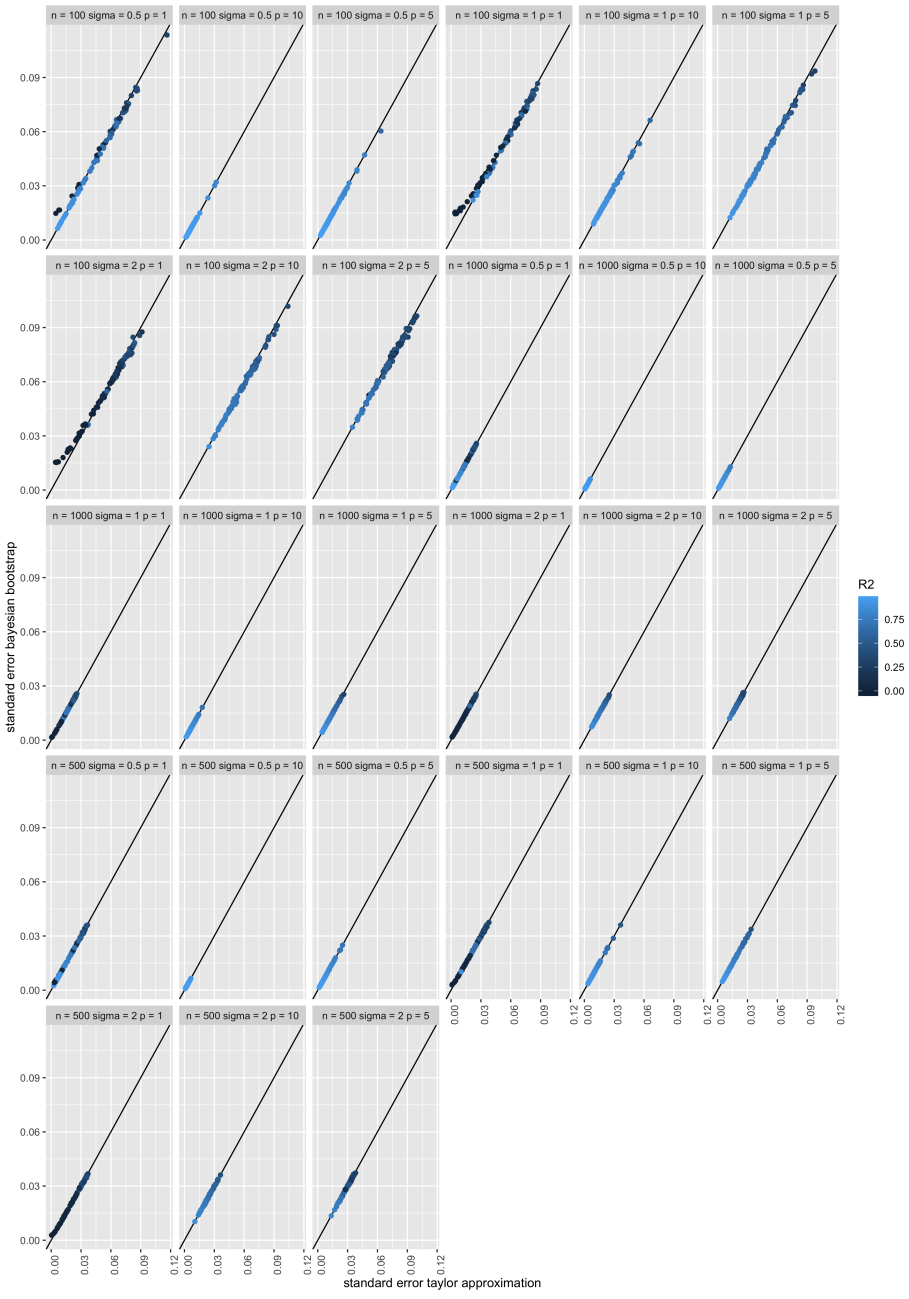

Figure 1: Simulation results for larger values of n.

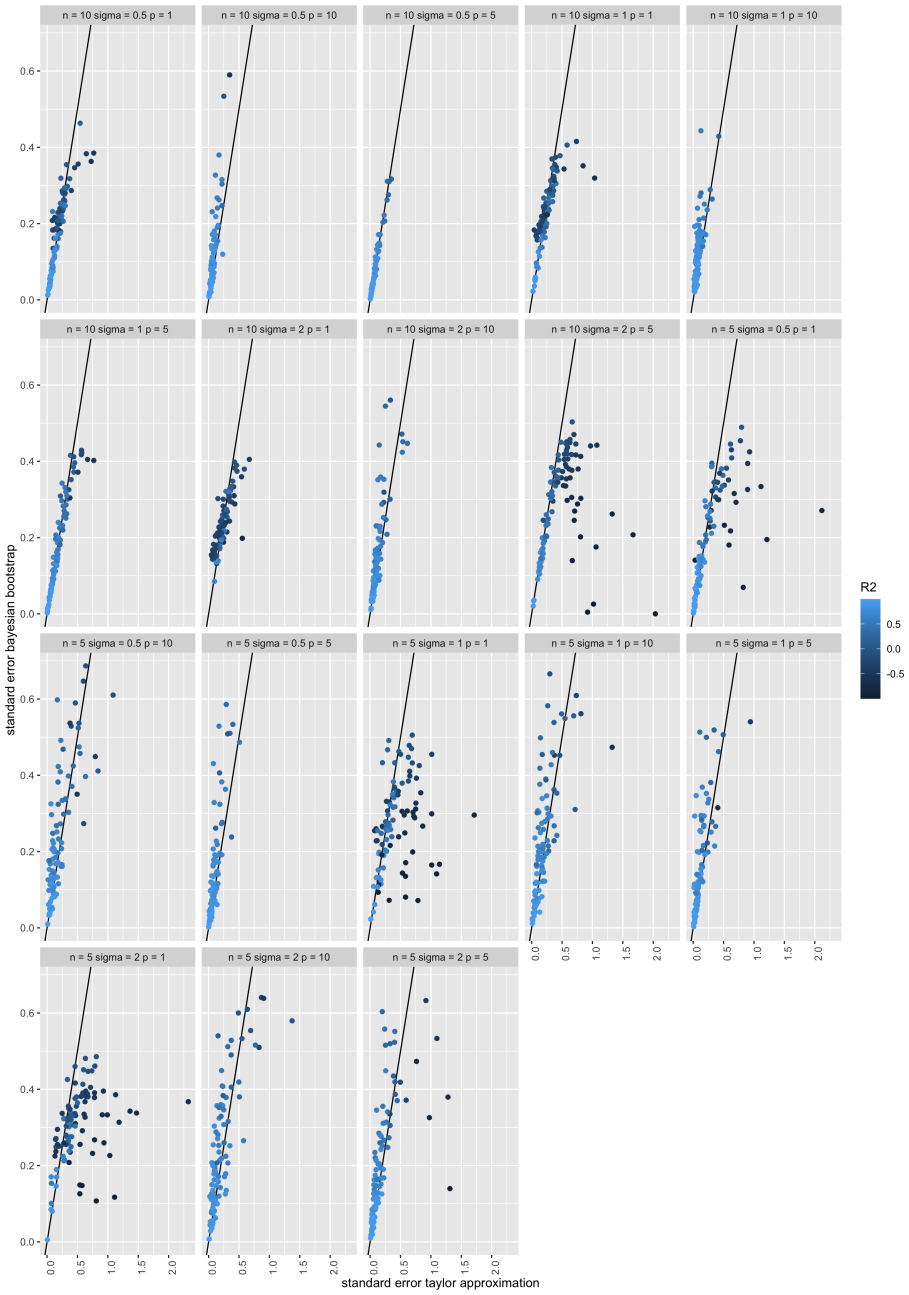

Figure 2: Simulation results for lower values of n.

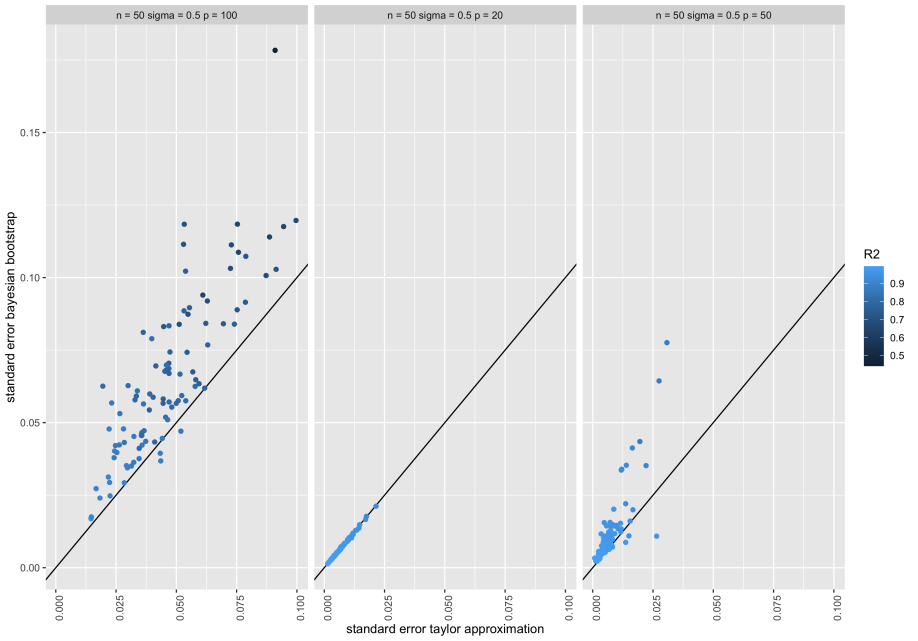

Figure 3: Simulation results for high number of predictors.

References

Andrew Gelman, Ben Goodrich, Jonah Gabry, and Aki Vehtari. R-squared for Bayesian Regression Models. *The American Statistician*, 73(3):307–309, July 2019. ISSN 0003-1305. doi: 10.1080/00031305.2018.1549100. URL `https://doi.org/10.1080/00031305.2018.1549100`. Publisher: Taylor & Francis _eprint: https://doi.org/10.1080/00031305.2018.1549100.

Ben Goodrich, Jonah Gabry, Imad Ali, and Sam Brilleman. rstanarm: Bayesian applied regression modeling via Stan., 2020. URL `https://mc-stan.org/rstanarm`. R package version 2.21.1.

W Keith Hastings. Monte carlo sampling methods using markov chains and their applications. 1970.

Donald B. Rubin. The Bayesian Bootstrap. *The Annals of Statistics*, 9(1):130–134, January 1981. ISSN 0090-5364, 2168-8966. doi: 10.1214/aos/1176345338. URL `https://projecteuclid.org/journals/annals-of-statistics/volume-9/issue-1/The-Bayesian-Bootstrap/10.1214/aos/1176345338.full`. Publisher: Institute of Mathematical Statistics.

Tuomas Sivula, Måns Magnusson, Asael Alonzo Matamoros, and Aki Vehtari. Uncertainty in Bayesian Leave-One-Out Cross-Validation Based Model Com-

parison. *arXiv:2008.10296 [stat]*, March 2022. URL http://arxiv.org/abs/2008.10296. arXiv: 2008.10296.

Aki Vehtari and Janne Ojanen. A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, 6(none):142–228, January 2012. ISSN 1935-7516. doi: 10.1214/12-SS102. URL https://projecteuclid.org/journals/statistics-surveys/volume-6/issue-none/A-survey-of-Bayesian-predictive-methods-for-model-assessment-selection/10.1214/12-SS102.full. Publisher: Amer. Statist. Assoc., the Bernoulli Soc., the Inst. Math. Statist., and the Statist. Soc. Canada.

Aki Vehtari, Andrew Gelman, and Jonah Gabry. Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *arXiv:1507.04544 [stat]*, September 2016. doi: 10.1007/s11222-016-9696-4. URL http://arxiv.org/abs/1507.04544. arXiv: 1507.04544.