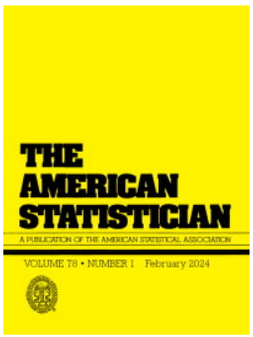
Taylor & Francis
Taylor & Francis Group



The American Statistician

ISSN: 0003-1305 (Print) 1537-2731 (Online) Journal homepage: www.tandfonline.com/journals/utas20

Out-of-Sample R^2: Estimation and Inference

Stijn Hawinkel, Willem Waegeman & Steven Maere

To cite this article: Stijn Hawinkel, Willem Waegeman & Steven Maere (2024) Out-of-Sample R^2: Estimation and Inference, The American Statistician, 78:1, 15-25, DOI: 10.1080/00031305.2023.2216252

To link to this article: https://doi.org/10.1080/00031305.2023.2216252

THE AMERICAN STATISTICIAN
2024, VOL. 78, NO. 1, 15–25: General
https://doi.org/10.1080/00031305.2023.2216252

Out-of-Sample R^2: Estimation and Inference

Stijn Hawinkel [a,b], Willem Waegeman[c], and Steven Maere [a,b]

[a]Department of Plant Biotechnology and Bioinformatics, Ghent University, Ghent, Belgium; [b]VIB Center for Plant Systems Biology, Ghent, Belgium; [c]Department of Data Analysis and Mathematical Modelling, Ghent University, Ghent, Belgium

ABSTRACT

Out-of-sample prediction is the acid test of predictive models, yet an independent test dataset is often not available for assessment of the prediction error. For this reason, out-of-sample performance is commonly estimated using data splitting algorithms such as cross-validation or the bootstrap. For quantitative outcomes, the ratio of variance explained to total variance can be summarized by the coefficient of determination or in-sample R^2, which is easy to interpret and to compare across different outcome variables. As opposed to in-sample R^2, out-of-sample R^2 has not been well defined and the variability on out-of-sample $\hat{R}^2$ has been largely ignored. Usually only its point estimate is reported, hampering formal comparison of predictability of different outcome variables. Here we explicitly define out-of-sample R^2 as a comparison of two predictive models, provide an unbiased estimator and exploit recent theoretical advances on uncertainty of data splitting estimates to provide a standard error for $\hat{R}^2$. The performance of the estimators for R^2 and its standard error are investigated in a simulation study. We demonstrate our new method by constructing confidence intervals and comparing models for prediction of quantitative *Brassica napus* and *Zea mays* phenotypes based on gene expression data. Our method is available in the R-package *oosse*.

ARTICLE HISTORY
Received February 2023
Accepted May 2023

KEYWORDS
Bootstrap; Coefficient of determination; Cross-validation; Prediction; Standard error

1. Introduction

Predictive model performance is evaluated through loss functions, which quantify the discrepancy between predicted and observed values. For quantitative outcomes, the most popular loss function is the squared error loss, defined as $l\{y_i, m(\mathbf{x}_i)\} = \{y_i - m(\mathbf{x}_i)\}^2$, with y_i the observed outcome in sample i=1,…,n, with n the sample size, and $m(\mathbf{x}_i)$ the outcome predicted by a model m using a set of predictors $\mathbf{x}_i$. Most often, summary statistics of the loss distribution are reported, such as the mean squared error (MSE), which is estimated as $n^{-1}\sum_{i=1}^{n} l\{y_i, m(\mathbf{x}_i)\}$ (Hastie et al. 2009). Yet unlike misclassification loss for categorical outcomes, the MSE is not easy to interpret for researchers unfamiliar with the specific prediction problem, as it depends on the degree of variability of the dataset under study as well as on the measurement unit used. For this reason, and to allow for comparison between different outcome variables, the MSE is often compared to the total variance in the sample: $n^{-1}\sum_{i=1}^{n} l(y_i, \bar{y})$ with $\bar{y}$ the average outcome, here called mean of squares total (MST) in analogy to the MSE. This gives rise to the coefficient of determination or R^2, which is defined as (Kvålseth 1985)

$$R^2 = 1 - \frac{\text{MSE}}{\text{MST}}$$

The R^2 measure is unitless and hence comparable across outcome variables with different units (Valbuena et al. 2019). It is often employed as a goodness-of-fit statistic of a model to a given sample. In this case, values $m(\mathbf{x}_i)$ and $\bar{y}$ are predicted using a model trained on the entire sample, so including observation i, and R^2 ranges from 0 to 1 and can be interpreted as the proportion of variance explained by the model (Anderson-Sprecher 1994). Modifications have been proposed to penalize for model complexity of general linear models, for example, adjusted R^2 (Wherry 1931), to better guide model building. The R^2 measure has also been extended to generalized linear models (Zhang 2017; Cameron and Windmeijer 1997; Nagelkerke 1991) survival analysis (Verweij and Houwelingen 1993). For linear models, a standard error and confidence intervals have been derived for in-sample R^2 (Cohen, West, and Aiken 2014).

Yet these R^2 values are intended as goodness-of-fit diagnostics and model building aides *within* a single dataset. Modern prediction models, for example random forests (Breiman 2001), are more flexible and make fewer assumptions about the type of association between outcome and predictors than linear models. Moreover, linear models may be applied to high-dimensional omics datasets with many more predictors than observations. Since both these scenarios may cause overfitting, in this case the in-sample loss is a poor measure for predictive performance,

CONTACT Steven Maere *stmae@psb.ugent.be* Department of Plant Biotechnology and Bioinformatics, Ghent University, Ghent, Belgium; VIB Center for Plant Systems Biology, Ghent, Belgium.

Supplementary materials for this article are available online. Please go to *www.tandfonline.com/r/TAS*.

which is instead evaluated on data not used to train the model. Ideally this is an independent test dataset, but more often the out-of-sample loss is estimated on the same dataset using data splitting algorithms such as cross-validation (CV) (Bates, Hastie, and Tibshirani 2023) or the bootstrap (Efron and Tibshirani 1997). Thereby the loss is estimated on a different part of the dataset than was used for building the model. As a consequence, the aforementioned results on in-sample R^2 and its standard error and confidence interval no longer hold. Out-of-sample R^2 values for instance lie in the interval $]-\infty, 1]$, instead of $[0, 1]$ for in-sample R^2. Still, for statistical inference, a formal definition of out-of-sample R^2 is needed, as well as some measure of uncertainty on the point estimate $\hat{R}^2$. Here we define out-of-sample R^2 as a model comparison of the prediction model at hand with a baseline prediction model that ignores covariate information, provide an unbiased estimator for it and exploit recent advances in the field of data splitting algorithms (Bates, Hastie, and Tibshirani 2023) to present a Standard Error (SE) for out-of-sample $\hat{R}^2$. We validate the estimators for out-of-sample R^2 and its standard error in a simulation study. Subsequently, we illustrate how these estimators can be used for comparing predictability of outcome variables and for building confidence intervals on real omics datasets of *Brassica napus* and *Zea mays* field trials. The proposed methodology is available in the R-package *oosse* at *https://cran.r-project.org/web/packages/oosse*.

2. Out-of-Sample R^2: Definition and Inference

2.1. Out-of-Sample R^2 as a Model Comparison

We propose to regard out-of-sample R^2 as a comparison of two out-of-sample prediction models, just like the regular in-sample R^2 is a comparison of two in-sample models (Anderson-Sprecher 1994; Campbell and Thompson 2008). Call m_d a prediction model that is trained on a dataset $\mathbf{d} = (\mathbf{y}, \mathbf{x})$ of size n, which can be used to make predictions for y given $\mathbf{x}$: $m_d(\mathbf{x})$. To score this prediction model m_d, we are interested in the expected squared error loss for a hypothetical out-of-sample observation Y_{oos} with respect to its predicted value $m_D(\mathbf{X}_{\text{oos}})$, averaged over all possible training datasets $\mathbf{D} = (\mathbf{Y}, \mathbf{X})$ of fixed size n drawn from the same population as $\mathbf{d}$. More formally, we are looking for

$$E_{\mathbf{D}}\left(E_{(Y,\mathbf{X})_{\text{oos}}}\left[\{Y_{\text{oos}} - m_D(\mathbf{X}_{\text{oos}})\}^2 \mid \mathbf{D}\right]\right), \tag{1}$$

with the inner expectation running over all possible out-of-sample observations and the outer expectation over all possible training datasets. We work under the common scenario where no independent test set of out-of-sample observations is provided, and only a single observed dataset $\mathbf{d}$ of size n is available, on which all estimation needs to be based. For this purpose, we assume that out-of-sample observations are drawn from the same population as $\mathbf{D}$. Here and in what follows, all variables without subscript oos belong to $\mathbf{d}$.

The first model in the comparison, referred to as the null model (Anderson-Sprecher 1994), simply uses the average outcome of the observed data $\bar{Y}$ as prediction for all out-of-sample observations, ignoring available predictors. The expected loss of this model is the out-of-sample MST that can be estimated analytically from the vector of outcome values $\mathbf{y}$ as derived below, relying on the equality $E(Y) = E(\bar{Y}) = E(Y_{\text{oos}})$:

$$\begin{aligned}
MST &= E_{\mathbf{D}}\left\{E_{Y_{\text{oos}}}\left(\left[Y_{\text{oos}} - \bar{Y}\right]^2 \mid \mathbf{D}\right)\right\} \\
&= E_{\mathbf{D}}\left\{E_{Y_{\text{oos}}}\left(\left[\{Y_{\text{oos}} - E(Y)\} + \{E(Y) - \bar{Y}\}\right]^2 \mid \mathbf{D}\right)\right\} \\
&= E_{\mathbf{D}}\left\{E_{Y_{\text{oos}}}\left([Y_{\text{oos}} - E(Y_{\text{oos}})]^2\right.\right. \\
&\qquad \left.\left. + \left[E(Y) - \bar{Y}\right]^2 + 2[Y_{\text{oos}} - E(Y)][E(Y) - \bar{Y}] \mid \mathbf{D}\right)\right\} \\
&= E_{\mathbf{D}}\left\{\text{var}(Y) + \left[E(\bar{Y}) - \bar{Y}\right]^2\right. \\
&\qquad \left. + 2E_{Y_{\text{oos}}}\left[\{Y_{\text{oos}} - E(Y)\}\left\{E(Y) - \bar{Y}\right\} \mid \mathbf{D}\right]\right\} \\
&= E_{\mathbf{D}}\left\{\text{var}(Y) + \left[E(\bar{Y}) - \bar{Y}\right]^2\right. \\
&\qquad \left. + 2\left[E(Y) - \bar{Y}\right]E_{Y_{\text{oos}}}\left[Y_{\text{oos}} - E(Y_{\text{oos}}) \mid \mathbf{D}\right]\right\} \\
&= \text{var}(Y) + \text{var}(\bar{Y}) + 0 = \frac{n+1}{n}\text{var}(Y).
\end{aligned} \tag{2}$$

This result nicely illustrates how the expected loss is a sum of variability of the estimator around the expected value and the variability of the observations around the expected value. An unbiased estimator for the population variance $\text{var}(Y)$ is provided by $(n-1)^{-1}\sum_{i=1}^{n}(y_i - \bar{y})^2$. The estimator for the out-of-sample MST then inflates this estimator by a factor (n+1)/n to account for the variability in the estimation of $E(Y)$ through $\bar{Y}$.

The second model in the comparison is the prediction model that makes use of the covariate information. Since it is a more complicated model, no analytical expression for its expected out-of-sample loss (the MSE) is available, so that, for want of independent test data, it needs to be estimated through data splitting algorithms. Here we discuss cross-validation and the 0.632 bootstrap (Efron and Tibshirani 1997), but other options are possible, for example, a single split of the available data in a training and a test dataset.

In cross-validation (CV), the samples of $\mathbf{d} = (\mathbf{y}, \mathbf{x})$ are divided into K equally sized folds of n/K observations, assuming for simplicity that K divides n. The set of samples in fold $k = 1, \ldots, K$ is indicated by $\mathbf{d}_k$, the other observations as $\mathbf{d}_{\setminus k}$. For all k, the model is trained on $\mathbf{d}_{\setminus k}$ yielding model m_k. The squared error loss of this fold is estimated as $\frac{K}{n}\sum_{i\in\mathbf{d}_k}\left\{y_i - m_k(\mathbf{x}_i)\right\}^2$, and the overall estimate of the MSE becomes $n^{-1}\sum_{k=1}^{K}\sum_{i\in\mathbf{d}_k}\left\{y_i - m_k(\mathbf{x}_i)\right\}^2$. Repeating the splitting into folds reduces the variability of the estimate $\widehat{\text{MSE}}$; the final estimate is then the average $\widehat{\text{MSE}}$ over the repeated splits. The procedure outlined above is what we refer to as simple CV. For nested CV, we refer to Bates, Hastie, and Tibshirani (2023), who provide an estimate of $\text{var}(\widehat{\text{MSE}})$, as well as a correction for the fact that m_k is trained on a dataset of size $\frac{n(K-1)}{K}$ rather than n.

The 0.632 bootstrap is an alternative way of estimating the MSE (Efron and Tibshirani 1997). In this case, the samples are split by resampling n samples with replacement. A sample has a probability of $1 - \exp(-1) \approx 0.632$ of being contained in this bootstrap sample, hence the name. This resampling step is repeated B times, with $\mathbf{d}_b$ indicating the set of included samples (possibly containing the same sample several times), m_b the model trained on this set and $\mathbf{d}_{\setminus b}$ the set containing the unique

excluded samples, with $b = 1, \ldots, B$. Call N_i^b the number of times sample i is included in $\mathbf{d}_b$, and $J_i^b = I(N_i^b = 0)$ with I() the indicator function. The 0.632 bootstrap estimate of the MSE is then given by

$$n^{-1}\exp(-1)\sum_{i=1}^{n}\left[\{y_i - m_d(\mathbf{x}_i)\}^2\right] + n^{-1}\left\{1-\exp(-1)\right\}\sum_{i=1}^{n}\left(\frac{\sum_{b=1}^{B}\left[J_i^b\left\{y_i - m_b(\mathbf{x}_i)\right\}^2\right]}{\sum_{b=1}^{B}J_i^b}\right),$$

with m_d trained on the set of all samples. The bootstrap 0.632 estimate is thus a weighted average of in-sample and out-of-sample error, but can be written as a weighted sum over all samples. A standard error on the 0.632 estimate, as well as variations on this estimator, are provided by Efron and Tibshirani (1997). Bates, Hastie, and Tibshirani (2023) show that CV and the 0.632 bootstrap indeed estimate the quantity in (1); CV is known to be unbiased whereas the bootstrap is slightly biased for MSE estimation (Braga-Neto and Dougherty 2004; Jiang and Simon 2007; Kohavi 1995; Molinaro, Simon, and Pfeiffer 2005).

Out-of-sample R^2 is then a population parameter defined as

$$R^2 = 1 - \frac{E_{\mathbf{D}}\left(E_{(Y,\mathbf{X})_{\text{oos}}}\left[\{Y_{\text{oos}} - m_D(\mathbf{X}_{\text{oos}})\}^2 \mid \mathbf{D}\right]\right)}{E_{\mathbf{D}}\left[E_{Y_{\text{oos}}}\left\{\left(Y_{\text{oos}} - \bar{Y}\right)^2 \mid \mathbf{D}\right\}\right]}, \quad (3)$$

and depends on the sample size n of the data $\mathbf{D}$ on which the prediction model is trained, on the prediction model and on the joint distribution of the outcome and predictors. Note that out-of-sample R^2 does not belong to the observed dataset $\mathbf{d}$, but rather represents an expectation over all datasets $\mathbf{D}$ that could be drawn from the population. The value of out-of-sample R^2 reflects a comparison of the null model with the more elaborate prediction model: when it is smaller than 0, the null model achieves the best out-of-sample predictions; when it is larger than 0, the elaborate prediction model achieves the lowest out-of-sample squared error loss. In the latter case, R^2 can be interpreted as the proportion of the null model's squared error loss explained by the elaborate model. For estimation of out-of-sample R^2, we plug in the aforementioned estimators of the out-of-sample squared errors MST and MSE based on the observed data:

$$\hat{R}^2 = 1 - \frac{\widehat{\text{MSE}}}{\widehat{\text{MST}}} = 1 - \frac{\widehat{\text{MSE}}}{(n+1)/\{n(n-1)\}\sum_{i=1}^{n}(y_i - \bar{y})^2}. \quad (4)$$

This expression can be seen as the prediction equivalent of the forecasting out-of-sample R^2 by Campbell and Thompson (2008).

2.2. The Pooling and Averaging Estimators for R^2

In (4) above, the squared error losses are calculated and summed over all observations before combining them into the final R^2 estimate. This is called the *pooling* strategy in the field of multivariate loss function estimation (e.g., area-under-the-curve (AUC) estimation), as opposed to the *averaging* approach whereby the performance measure is calculated for every left-out fold separately, and then averaged over the folds to obtain the final estimate (Bradley 1997). The latter approach is not applicable for 0.632 bootstrap estimation of the MSE, but can be employed when using cross-validation. In this case, the MSE and MST are estimated for every left-out fold separately as follows. The MSE is estimated per fold as in Section 2.1. For the MST, we consider the following two estimation strategies: either as in (2) based on the training folds, adapting n to be the sample size of the training folds, or as the empirical variance of the left-out fold

$$\frac{\sum_{i\in\mathbf{d}_k}(y_i - \bar{y}_{\mathbf{d}_k})^2}{\left\{\sum_{i=1}^{n} I(i \in \mathbf{d}_k)\right\} - 1},$$

with $\bar{y}_{\mathbf{d}_k}$ the average of the left-out fold $\mathbf{d}_k$ (Valbuena et al. 2019). In both cases, these estimates are combined into an R^2 estimate for every fold separately and then averaged. We call the two presented estimators "averaging R^2 with training MST" and "averaging R^2 with test MST," respectively.

2.3. The Standard Error of $\hat{R}^2$

As a measure of uncertainty on the estimate $\hat{R}^2$, we derive an expression for its standard error (SE) $\left(\text{SE}(\hat{R}^2) = \text{var}(\hat{R}^2)^{1/2}\right)$. For this, we rely on asymptotic normality of the estimator $\hat{R}^2$. According to the first order delta method (Gauss 1823),

$$\text{var}(\hat{R}^2) \approx \left(\nabla R^2\right)^T \begin{bmatrix} \text{var}(\widehat{\text{MSE}}) & \text{cov}(\widehat{\text{MSE}}, \widehat{\text{MST}}) \\ \text{cov}(\widehat{\text{MSE}}, \widehat{\text{MST}}) & \text{var}(\widehat{\text{MST}}) \end{bmatrix} \nabla R^2. \quad (5)$$

The gradient is found as $\left(\nabla R^2\right)^T = \left(\frac{\partial R^2}{\partial \text{MSE}}, \frac{\partial R^2}{\partial \text{MST}}\right) = \left(-\frac{1}{\text{MST}}, \frac{\text{MSE}}{\text{MST}^2}\right)$, and is evaluated at the estimates $\widehat{\text{MSE}}$ and $\widehat{\text{MST}}$. An estimate of $\text{var}(\widehat{\text{MSE}})$ for cross-validation estimation of the MSE was provided by Bates, Hastie, and Tibshirani (2023), and for 0.632 bootstrap estimation of the MSE by Efron and Tibshirani (1997). The estimate of $\text{var}(\widehat{\text{MST}})$ is given by $\frac{2}{(n-1)}\widehat{\text{MST}}^2$ (Harding, Tremblay, and Cousineau 2014).

The covariance $\text{cov}(\widehat{\text{MSE}}, \widehat{\text{MST}})$ is usually considerable and positive, since the MSE and MST are estimated on the same outcome vector $\mathbf{y}$. It can be decomposed as $\text{cov}(\widehat{\text{MSE}}, \widehat{\text{MST}}) = \text{cor}(\widehat{\text{MSE}}, \widehat{\text{MST}})\left\{\text{var}(\widehat{\text{MSE}})\text{var}(\widehat{\text{MST}})\right\}^{1/2}$. $\text{cor}(\widehat{\text{MSE}}, \widehat{\text{MST}}) = \rho$ cannot be derived analytically, such that it is estimated either via nonparametric or parametric bootstrap, or via jackknifing as follows. Bootstrap samples of size n are either drawn nonparametrically, by sampling entries with replacement from the data $\mathbf{d}$, or parametrically by assuming a parametric model, estimating the corresponding parameters from the data and drawing $\mathbf{y}$ from the model parameterized by these estimates while keeping $\mathbf{x}$ fixed. The MSE and MST are then estimated for every bootstrap sample in the same way as for the original sample. In jackknifing, each observation is dropped in turn, and the MSE and MST are estimated on the remaining samples of size $n-1$, leading to a number of jackknife estimates equal to the sample size n. For the bootstrap and jackknife samples, simple rather than nested CV is used for MSE estimation to reduce computation time, but the splitting into CV folds is repeated as for the MSE estimation on the observed dataset. The empirical correlation between the bootstrap or jackknife estimates of the MSE and MST is used as estimate of the correlation between the estimators for MSE and MST. When using the 0.632 bootstrap for MSE estimation

in combination with bootstrapping for the estimation of the correlation between MSE and MST, we refer to the former as inner bootstrap samples and to the latter as outer bootstrap samples.

2.4. Inference on R^2

Given the standard error estimate, one-sided approximate z-tests can be used to test the null hypothesis that $R^2 \leq 0$, that is, whether the prediction model is significantly better than the null model. Another popular application of standard errors is the construction of confidence intervals, again relying on asymptotic normality of the estimator $\hat{R}^2$. The confidence interval is then constructed as

$$\hat{R}^2 \pm \widehat{\text{var}}(\hat{R}^2)^{1/2} \Phi^{-1}\left(\frac{\alpha}{2}\right) \tag{6}$$

with α the significance level and Φ^{-1} the inverse cumulative distribution function of the standard normal distribution. The upper bound of the confidence interval is truncated at 1.

As an alternative to the delta method SE from (5), the standard deviation of R^2 estimates across nonparametric or parametric bootstrap samples (obtained in the same way as for estimating ρ) can be used as an estimate of the SE (further referred to as the bootstrap SE). As alternatives for the confidence intervals based on the delta method SE, we consider percentile and bias-corrected and accelerated (BCa) bootstrap confidence intervals constructed based on the distribution of the bootstrap R^2 estimates (DiCiccio and Efron 1996), as well as confidence intervals constructed from the bootstrap SE in the same way as for the delta method SEs in (6). This latter method was chosen rather than the bootstrap-t method (DiCiccio and Efron 1996) since this would require calculating standard errors for every bootstrap instance and hence be too computationally intensive.

3. Simulation Study

We conduct a simulation study in which we apply the proposed methodology for R^2 estimation on a one-dimensional prediction model (ordinary least squares, OLS) and a high-dimensional prediction model (elastic net, EN). We study the performance of the estimators for the MST and MSE, and compare the pooling and averaging estimators for R^2. Next we compare our delta method estimator for the SE on $\hat{R}^2$ and corresponding confidence intervals with competitor methods based on the bootstrap SE and percentile and BCa confidence intervals.

3.1. Setup

In the simulation study, observations y_i were drawn from the following model:

$$Y_i \sim N(\mathbf{X}_i\boldsymbol{\beta}, \sigma^2). \tag{7}$$

$\mathbf{X}$ is the $n \times p$ design matrix and $\boldsymbol{\beta}$ is the vector of coefficients of length p; $\sigma^2 = 1$ is the residual variance. We investigate a one-dimensional scenario (p=1) and a high-dimensional scenario (p=1000). All elements of $\mathbf{X}$ were drawn independently from the standard normal distribution. For the one-dimensional scenario, we consider the following sample sizes n: 20, 30, 50, and 100, and β is set to either 0, 0.5, 1, or 1.5. For the high-dimensional scenario, we consider sample sizes 30, 50, 75, and 100, and the first 10 entries of $\boldsymbol{\beta}$ are all set to either 0, 0.5, or 1; all other entries are 0. In the one-dimensional scenario, 1000 Monte Carlo instances are generated, and in the high-dimensional scenario 100.

In the one-dimensional scenario, the outcome was predicted using OLS using the *lm.fit* function in the *stats* R-package. In the high-dimensional scenario, the outcome was predicted using EN (Zou and Hastie 2005) with fixed mixing parameter 0.5 using the *glmnet* function from the *glmnet* R-package (Friedman, Hastie, and Tibshirani 2010). The penalty parameter was tuned through an inner loop of 10-fold CV as implemented in the *cv.glmnet* function.

The out-of-sample MSE was estimated via either 10-fold CV as in Bates, Hastie, and Tibshirani (2023), or via the 0.632 bootstrap as in Efron and Tibshirani (1997). Corresponding standard errors were calculated via nested CV with nine inner folds (Bates, Hastie, and Tibshirani 2023) or via empirical influence functions (Efron and Tibshirani 1997), respectively. The number of splits into cross-validation folds was repeated as recommended by Bates, Hastie, and Tibshirani (2023), with 1, 25, 100, or 200 splits in the one-dimensional scenario, or 5, 25, or 100 splits in the high-dimensional scenario. The number of bootstrap samples for 0.632 bootstrap estimation of the MSE was varied between 25, 100, and 200. For estimating the correlation between MSE and MST estimators through bootstrapping, either 10, 50, 100, or 500 bootstrap instances were used.

The true out-of-sample MSE is not known from the generative model (7) alone, as it depends on the accuracy of the estimation of m. Instead, it is approximated through Monte Carlo simulation, by generating 5000 datasets according to (7) with the same array of sample sizes n and coefficients β as specified above, fitting the prediction model to these datasets and evaluating their predictive performance on an independent test dataset drawn from (7) with sample size 10,000. This yields 5000 precise estimates of the MSE; their average provides an approximation of the true out-of-sample MSE. The approximated true out-of-sample MSE in combination with the true MST given by $\frac{\sigma^2(n+1)}{n}$ is then used to approximate true out-of-sample R^2 of this parameter setting and prediction method using (3) (see Tables S1 and S2). One-sided approximate z-tests were used to test whether $R^2 \leq 0$ in all scenarios with approximated true R^2 values below 0, and the proportion of times the null hypothesis was rejected was taken as an estimate of the Type I error. A significance level of 5% was used. Coverage of the 95% confidence intervals is approximated as the proportion of Monte Carlo instances for which the confidence interval includes the approximated true R^2. In addition, the average width of the confidence interval is calculated.

The true SE of $\hat{R}^2$ is approximated differently, as it also depends on the data splitting algorithm and its settings (number of CV splits or number of bootstraps). It is approximated by the standard deviation of the $\hat{R}^2$ values of the algorithm and parameter settings concerned over all Monte Carlo instances from the simulation, see Figures S1–S3, S19–S20. The true corre-

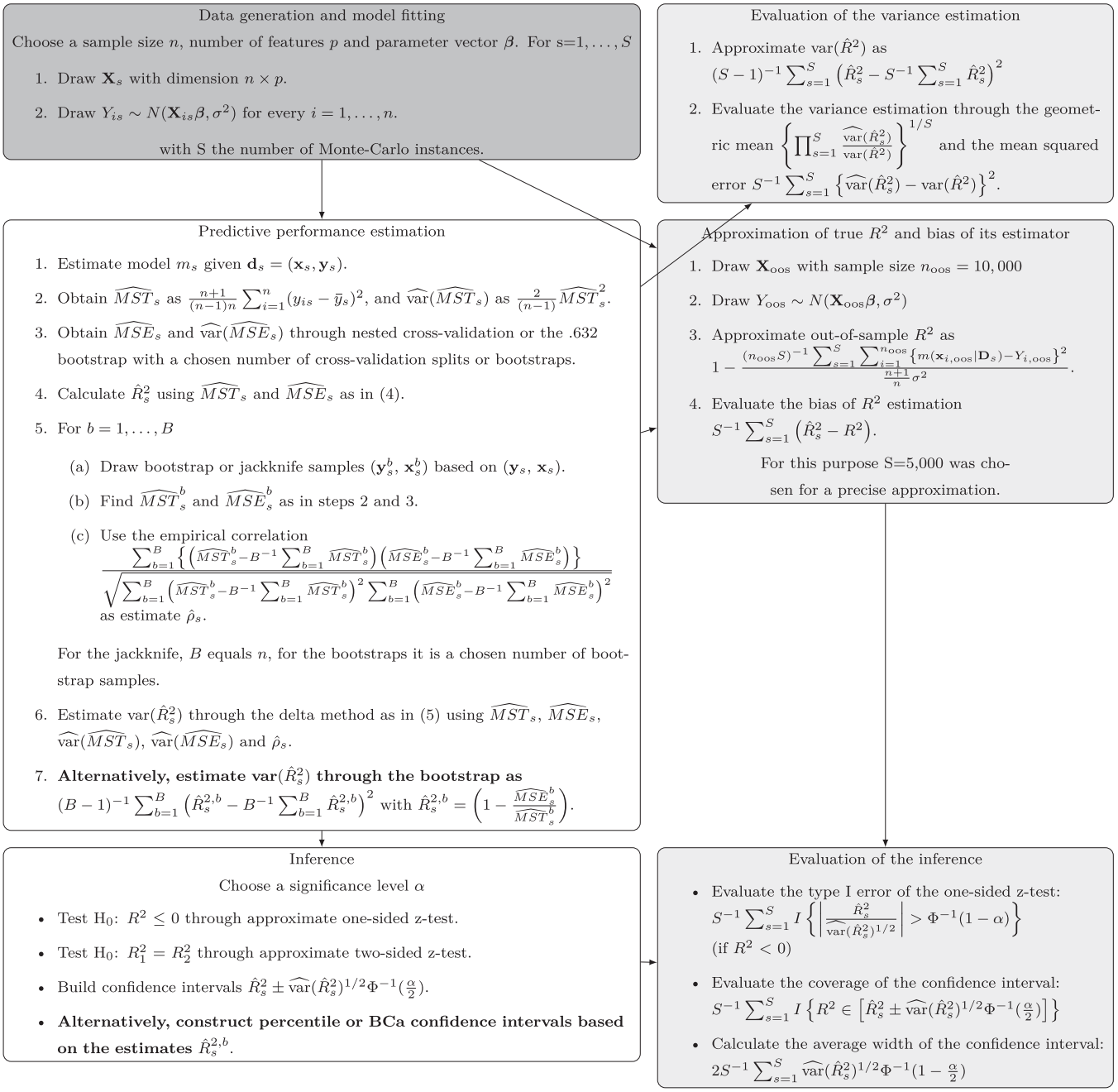


Figure 1. Schematic overview of the simulation study. The top left panel (in dark gray) describes the repeated data generation. The center and bottom left panels (in white) describe the estimation of R^2 and $\text{var}(\hat{R}^2)$ and statistical inference on either simulated or real data. The right panels (in light gray) describe the approximation of true R^2 and $\text{var}(\hat{R}^2)$, as well as performance evaluation, which is only possible for simulated data where the ground truth is known. Arrows indicate input of one block into another. Bold text indicates competitor methods that depart from the methodology proposed in the article.

lation ρ between $\widehat{\text{MSE}}$ and $\widehat{\text{MST}}$ was approximated analogously as the empirical correlation of these quantities over the same Monte Carlo instances. The accuracy of the different SE estimators was assessed by calculating the ratio of the estimated to the approximated true SE, and taking the geometric mean over all Monte Carlo instances. Also the MSE of the estimated SE with respect to the approximated true SE was calculated, which reflects both the bias and variance of the estimators. The accuracy of the estimation of ρ was assessed graphically using boxplots of the estimates $\hat{\rho}$. The whole pipeline of data generation, R^2 estimation and evaluation is shown in Figure 1.

3.2. Results

The unbiasedness of the MST estimator (2) is demonstrated numerically in Figure S6. In the one-dimensional scenario, MSE estimation through 0.632 bootstrap is downward biased, whereas the cross-validation estimation with bias correction proposed by Bates, Hastie, and Tibshirani (2023) is unbiased (Figure S4). The estimation of $\text{var}(\widehat{\text{MSE}})$ through CV or 0.632 bootstrap suffers from a small upward bias at small sample sizes, but this bias decreases as the sample size grows (Figures S7–S8).

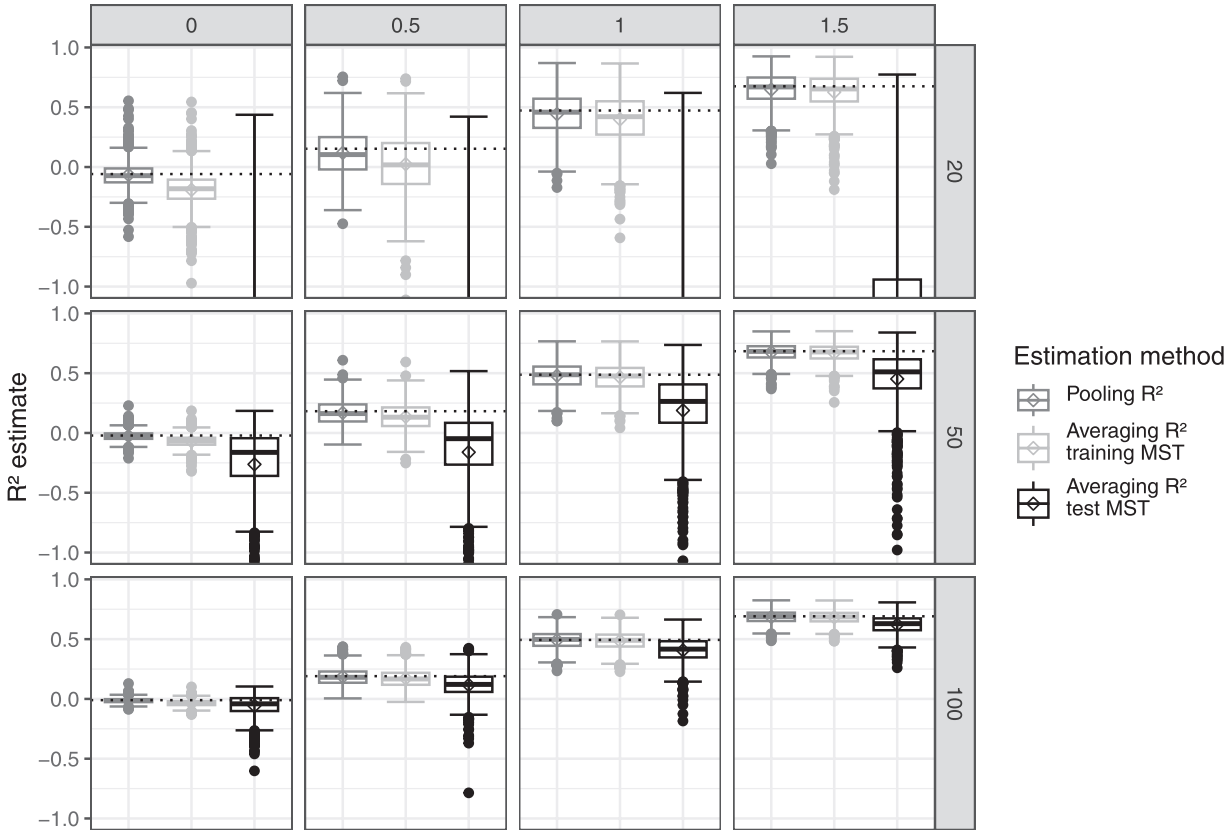


Figure 2. Boxplots of R^2 values estimated in three different ways in the one-dimensional scenario using cross-validation (grayscale, see Section 2.2) for different sample sizes (side panels) and effect sizes (top panels). Simple, nonnested CV is used to estimate the MSE, the rest of the setup is the same as for the one-dimensional scenario simulations with exclusion of the sample size 30. Diamonds indicate means; the plot was truncated at −1 for legibility. The dotted horizontal line indicates true R^2 approximated through Monte Carlo simulation.

For the high-dimensional scenario, Figures S22 reveals an upward bias in MSE estimation for large sample sizes and large effect sizes. For the CV, this bias results from the fact that in k-fold CV the model is trained on a dataset of size $n\frac{(k-2)}{k}$ in the nested CV scheme, rather than the models trained on sample size n for which CV is attempting to estimate the performance. The bias correction proposed in Appendix C of Bates, Hastie, and Tibshirani (2023) worked fine in the one-dimensional case, as evident from Figure S4, but fails for the high-dimensional case, presumably because we are in the proportional, sparse regime with both n and p going to infinity, as mentioned by Bates, Hastie, and Tibshirani (2023). This bias could be reduced by increasing the number of folds (Bates, Hastie, and Tibshirani 2023).

The performance of the different alternative estimators of R^2 is shown in Figure 2 for the one-dimensional scenario, and in Figure S21 for the high-dimensional scenario. In the one-dimensional scenario, the pooling R^2 is an unbiased estimator of true R^2 as defined in (3) with low variance. The averaging R^2 with training MST also has a low variance, but is downward biased for weak signal strengths. The averaging R^2 with test MST estimator is very variable and dramatically downward biased for smaller sample sizes, and even at a sample size of 100 some of the bias persists. For this reason, we choose to work with the pooling estimator (4). In the high-dimensional scenario (Figure S21), similar behavior is seen, except that also the pooling estimator for R^2 is slightly biased for the stronger effect sizes and larger sample sizes in this case because the estimator for the MSE is biased. Yet the pooling estimator remains the best estimator with least bias. The normality assumption for this pooling estimator $\hat{R}^2$, required for the delta method approximation of the SE and for construction of the confidence intervals, is assessed in Figures S36–S39. Some departures from normality can be seen, especially at small sample sizes.

The accuracy of the SE($\hat{R}^2$) estimation is shown in Figure 3 and Figure S9 for cross-validation in the one-dimensional scenario. The bias for delta method standard errors is mostly positive (i.e., conservative) and decreases with effect size. The nonparametric bootstrap method performs best for the estimation of the correlation ρ between $\widehat{\text{MSE}}$ and $\widehat{\text{MST}}$ (Figure S10). Bootstrap standard errors are more accurate than delta method standard errors in the null scenario, but tend to be downward biased (i.e., liberal) as the signal strength increases, especially the parametric bootstrap. The coverage of the confidence intervals (Figure 3) is close to the nominal level for the delta method SE with nonparametric bootstrap or jackknife estimation of ρ, but the intervals based on the bootstrap (bootstrap SE, percentile and BCa) show undercoverage in some scenarios. When the predictors are not predictive of the outcome at all, all confidence intervals except the BCa intervals have a coverage above the nominal level of 95%. The delta method SE confidence intervals are generally wider than the bootstrap confidence intervals (Figure S11). All methods control the Type I error of the approximate one-sided z-test below the significance level of 5% (Figure S12).

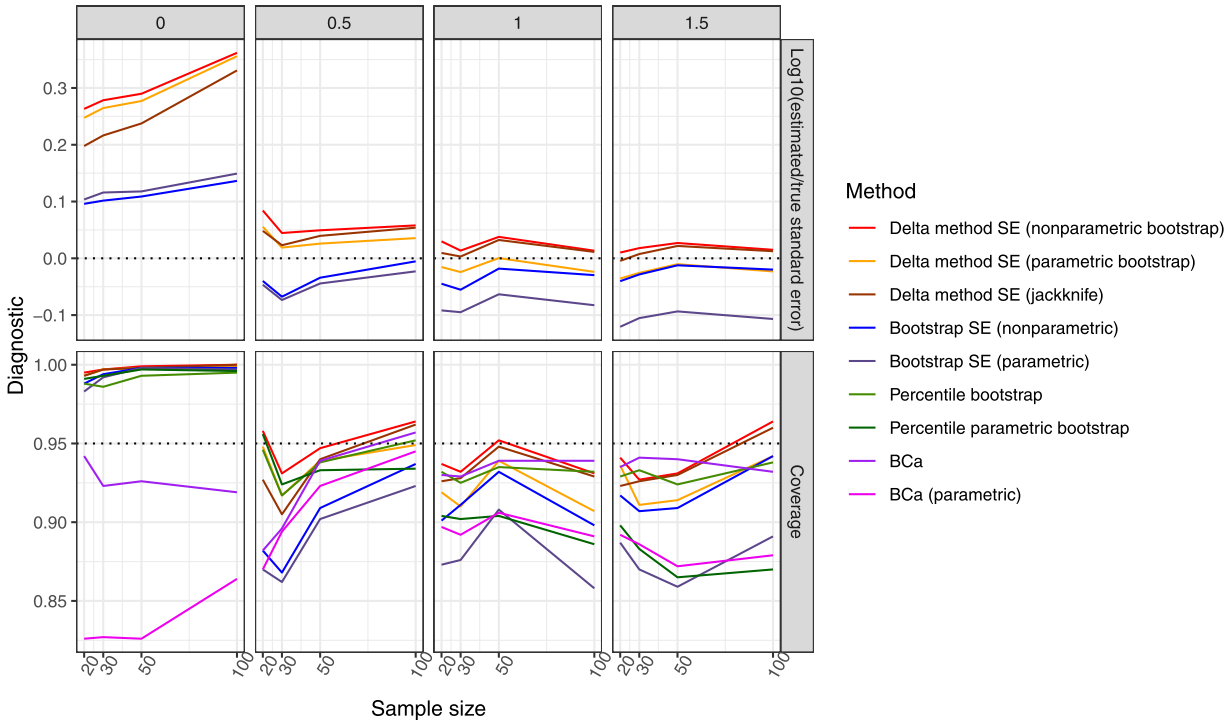


Figure 3. Diagnostics for the one-dimensional simulation scenario using cross-validation: log10 of the geometric mean of the ratio of estimated to approximated true standard error (SE) of $\hat{R}^2$ (top panels) and coverage of the confidence intervals (bottom panels) as a function of estimation method (color), sample size (x-axis) and effect size (columns) for 500 bootstraps and 200 repeats of the cross-validation splits. The dotted lines indicate unbiased R^2 estimation and the nominal coverage of 95%, respectively.

Similar conclusions are reached for the 0.632 bootstrap (Figures S13–S18), even though it is slightly upward biased for R^2 estimation (Figure S5).

In the high-dimensional scenario when using CV, the parametric bootstrap yields the best estimate of the SE of $\hat{R}^2$ in the null scenario of no predictive value, but underestimates the SE for stronger signal strengths, especially at large sample sizes (Figure 4 and Figure S24). As the sample size and effect size increase, the delta method SE with nonparametric bootstrap or jackknife estimation of ρ and nonparametric bootstrap SE perform best, converging on the true value (Figure S24). These findings are also reflected in the coverages of the confidence intervals, which lie above the nominal level of 95% in the null scenario for all methods except the percentile bootstrap and BCa intervals (Figure 4 and Figure S27). As the signal strength increases, only the delta method SE confidence intervals with ρ estimation using nonparametric bootstrap or jackknife and nonparametric bootstrap SE confidence intervals maintain a coverage close to the nominal level (Figure 4 and Figure S27), although the low coverage of the other methods is partly caused by the bias in the MSE estimation at high sample sizes mentioned above (Figures S22–S23). For every sample size, the confidence intervals are much wider than for the one-dimensional scenario (Figure S28). The Type I error is not controlled at the significance level for the bootstrap SE methods (Figure S29). Similar results are found for 0.632 bootstrap estimation of R^2 (Figures S30–S35).

In both one- and high-dimensional scenarios, the correlation between $\widehat{\text{MSE}}$ and $\widehat{\text{MST}}$ is close to 1 when the predictors have no predictive value, but decreases and in some cases becomes negative as the effect size increases (Figures S10, S15, S26, and S32). The negative correlations are likely an effect of the randomness in the design matrix, where designs with more variable predictors lead to more variance in the outcome (high $\widehat{\text{MST}}$), but also to better parameter estimates and hence better predictions (low $\widehat{\text{MSE}}$).

4. Case Study: Predictability of *Brassica napus* and *Zea mays* Phenotypes

Gene expression and phenotypes were measured for 62 *Brassica napus* ssp *napus* (rapeseed) plants (De Meyer et al. 2023), and 60 *Zea Mays* (maize) plants (Cruz et al. 2020). For each crop, five phenotypes were considered for prediction: leaf 8 width, number of branches, number of leaves, root width and number of seeds for *B. napus* and leaf 16 blade length, leaf 16 blade width, husk leaf length, ear length and plant height at the time of leaf sampling for *Z. mays*. The gene expression counts were *rlog* transformed using a negative binomial regression model prior to analysis to stabilize their variance (Love, Huber, and Anders 2014). Only the 5000 genes with the highest expression were retained for model fitting. The outcome phenotypes were predicted from these 5000 genes through EN with the same settings as in the simulation study, and 10-fold CV with 100 repeats of the split into folds was used to estimate the corresponding MSE. The SE on the resulting $\hat{R}^2$ was calculated using the delta method as in Section 2.3 with jackknife estimation of ρ. In addition, the SE was estimated using the nonparametric bootstrap with 50 bootstrap samples. A one-sided approximate z-test was performed to test whether $R^2 \leq 0$, and confidence intervals were constructed.

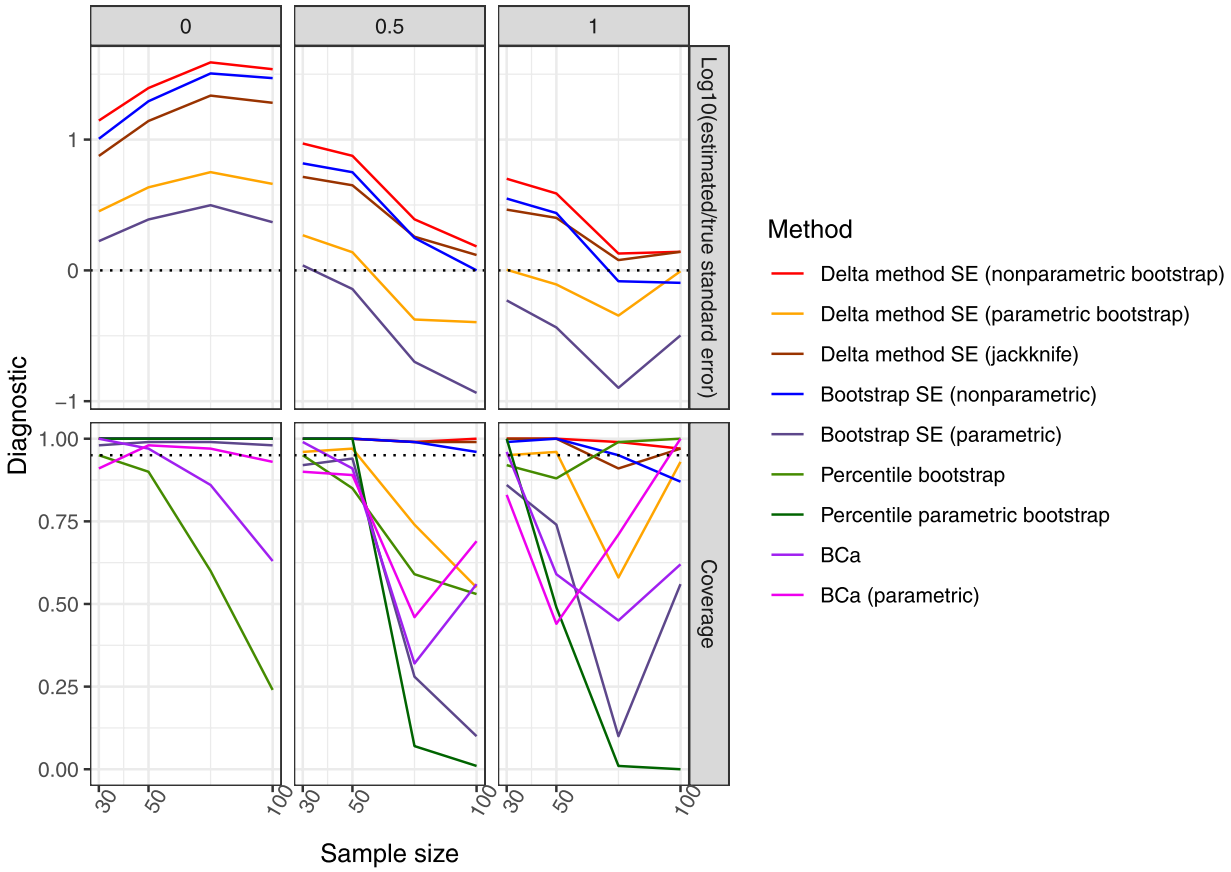


Figure 4. Diagnostics for the high-dimensional simulation scenario using cross-validation: log10 of the geometric mean of the ratio of estimated to approximated true standard error (SE) of $\hat{R}^2$ (top panels) and coverage of the confidence intervals (bottom panels) as a function of estimation method (color), sample size (x-axis) and effect size (columns) for 100 bootstraps and 100 repeats of the cross-validation splits. The dotted lines indicate unbiased R^2 estimation and the nominal coverage of 95%, respectively.

Table 1. Estimated R^2, corresponding delta method and nonparametric bootstrap standard error (SE), one-sided *p*-value and lower and upper bound of the 95% confidence interval for 5 *Brassica napus* phenotypes.

	$\hat{R}^2$	Delta method				Bootstrap			
		SE	*p*-value	2,5%	97,5%	SE	*p*-value	2,5%	97,5%
Leaf 8 width	0.72	0.07	<0.0001	0.58	0.86	0.09	<0.0001	0.54	0.91
Total branch count	0.11	0.17	0.26	−0.22	0.44	0.16	0.26	−0.22	0.43
Number of leaves	0.19	0.25	0.23	−0.30	0.67	0.15	0.12	−0.12	0.49
Root system width	−0.02	0.17	0.55	−0.35	0.31	0.22	0.53	−0.45	0.41
Number of seeds	0.42	0.20	0.02	0.03	0.82	0.11	<0.0001	0.21	0.63

Table 2. Estimated R^2, corresponding delta method and nonparametric bootstrap standard error (SE), one-sided *p*-value and lower and upper bound of the 95% confidence interval for 5 *Zea mays* phenotypes.

	$\hat{R}^2$	Delta method				Bootstrap			
		SE	*p*-value	2,5%	97,5%	SE	*p*-value	2,5%	97,5%
Leaf 16 blade length	0.30	0.23	0.10	−0.15	0.74	0.16	0.03	−0.02	0.61
Leaf 16 blade width	0.49	0.21	0.01	0.07	0.90	0.09	<0.0001	0.30	0.67
Husk leaf length	0.30	0.29	0.15	−0.27	0.87	0.19	0.06	−0.08	0.68
Ear length	−0.02	0.18	0.54	−0.38	0.34	0.24	0.53	−0.48	0.44
Plant height	−0.01	0.15	0.54	−0.30	0.27	0.19	0.53	−0.38	0.35

NOTE: For leaf 16 blade length with bootstrap SE, the *p*-value is a significant whereas the confidence interval includes 0; this is because the *p*-value is one-sided but the confidence interval is two sided.

Estimated R^2 values, standard errors, *p*-values and confidence intervals of the *B. napus* and *Z. mays* data are shown in Tables 1 and 2, respectively. For *B. napus*, leaf 8 width and number of seeds have an R^2 significantly different from 0 according to the one-sided approximate z-test based on the delta method SE; for *Z. mays* leaf 16 blade width is significant according to this test. The bootstrap SE's are mostly smaller, as in the simulations, and yield narrower confidence intervals and smaller *p*-values, but with similar conclusions of the corresponding significance tests.

Table 3. Approximate two-sided z-statistic for difference in R^2 of *Brassica napus* phenotypes in columns and rows, with corresponding *p*-value calculated using the delta method SE between brackets.

	Leaf 8 width	Total branch count	Number of leaves	Root system width
Total branch count	−2.98 (0.0029)			
Number of leaves	−2.39 (0.017)	0.27 (0.79)		
Root system width	−2.45 (0.014)	−0.39 (0.70)	−0.61 (0.54)	
Number of seeds	−1.59 (0.11)	1.69 (0.091)	0.88 (0.38)	1.41 (0.16)

NOTE: For instance, the top left entry is the total branch count $\hat{R}^2$ minus leaf 8 width $\hat{R}^2$, divided by a standard error estimate for this difference.

A relevant scientific question is the comparison of R^2 estimates for different phenotypes within the same dataset. If the design matrix were fixed, the different estimates would be independent and the variance of the difference would simply equal the sum of the variances of both R^2 estimates. However, if the design matrix is random, as is the case here with the gene expression measurements, the MSE estimates of two phenotypes correlate. The variance on the estimator for the difference between R^2 of two different phenotypes a and b then equals

$$\begin{aligned}&\mathrm{var}(\hat{R}_a^2 - \hat{R}_b^2)\\ &\quad = \mathrm{var}(\hat{R}_a^2) + \mathrm{var}(\hat{R}_b^2) - 2\mathrm{cor}(\hat{R}_a^2,\ \hat{R}_b^2)\sqrt{\mathrm{var}(\hat{R}_a^2)\mathrm{var}(\hat{R}_b^2)}.\end{aligned}$$

The correlation $\mathrm{cor}(\hat{R}_a^2,\ \hat{R}_b^2)$ was estimated by the bootstrap: samples from the gene expression matrix were sampled with replacement 50 times, and corresponding entries of the phenotypes a and b are used to estimate R^2. The empirical correlation of these 50 bootstrap estimates is then used as an estimate of $\mathrm{cor}(\hat{R}_a^2,\ \hat{R}_b^2)$. The test statistics and p-values using the delta method SE for the approximate two-sided z-test of the null hypothesis of zero difference between R^2 values for *B. napus* are shown in Table 3. The leaf 8 width is the only phenotype with an $\hat{R}^2$ significantly different from some other phenotypes: total branch count, number of leaves and root system width.

A further research question could be whether the predictability of certain phenotypes differs between species. Since they are calculated on independent datasets, the variance of the difference between two R^2 estimators is simply the sum of the variances. Hence, the approximate two-sided z-statistic for the difference between R^2 values of, for example, *B. napus* leaf 8 width and *Z. mays* leaf 16 blade width is, using the delta method SE's, $\frac{0.72-0.49}{\sqrt{0.07^2+0.21^2}} = 1.06$, so not significant ($p$-value = 0.28). Yet the standardized difference in predictability between *B. napus* leaf 8 width and *Z. mays* plant height $\frac{0.72-(-0.01)}{\sqrt{0.07^2+0.15^2}} = 4.52$ is indeed significant (p-value = 6.2×10^{-6}).

5. Discussion

Research into R^2-like measures has generally focused on in-sample performance, yet R^2 is also frequently employed to score out-of-sample prediction. Here we have formally defined out-of-sample R^2 as one minus the ratio of prediction loss of a particular prediction model to the prediction loss of the null model ignoring covariate information. The resulting R^2 estimate then has a clear interpretation: when it is larger than 0, the prediction model is useful for out-of-sample prediction. When it is smaller than 0, the average of the observed data is a better predictive model, either because the predictors do not contain enough information on the outcome vector, because the sample size is too small to allow for accurate model fitting, or because the model is ill-suited to the prediction task. Out-of-sample R^2 is useful for reporting predictive modeling results to an audience that is not familiar with the measurement units of the outcome variable, or with the MSE to be expected from a good prediction model. In addition, standardizing the MSE to R^2 is necessary when comparing predictability of different outcome variables. We demonstrated how out-of-sample R^2 can be estimated using data splitting algorithms. We found that the pooling R^2 estimator, which separately estimates the squared error losses of the null and prediction models and only then combines them into a final estimate $\hat{R}^2$, is unbiased. Hence this pooling estimator should be preferred to averaging estimators that calculate $\hat{R}^2$ values in every cross-validation fold separately and then average over the folds, which suffer from bias. Unlike the 0.632 bootstrap, cross-validation in combination with the pooling estimator provides almost unbiased estimates of R^2, in agreement with previous findings (Braga-Neto and Dougherty 2004; Jiang and Simon 2007; Kohavi 1995; Molinaro, Simon, and Pfeiffer (2005)), and should be the preferred way to estimate predictive performance. Some downward estimation bias appears in high-dimensional settings with strong signal, but this can be countered by a sufficiently high number of folds if computationally feasible. Out-of-sample R^2 can easily be extended to any prediction setting where a loss function can be evaluated for every observation separately, for example, deviance of generalized linear models. The only possible complication is that no more analytical expression for (the equivalent of) the out-of-sample MST may be available in more complicated settings, in which case it needs to be estimated through cross-validation as well (see Figure S6). Finally, it is good to remember that R^2 estimated through resampling algorithms does not apply specifically to the model trained on the dataset at hand, but rather to all models trained on datasets randomly drawn from the same population of interest.

Like any parameter estimated from a finite sample, estimates obtained through data splitting algorithms are uncertain. Yet by lack of estimators for R^2-like measures' standard errors, often only their point estimates are reported. These alone do not allow simple statistical questions to be answered, such as whether the predictive model significantly reduces the loss with respect to the null model. One way to answer this question is by repeatedly permuting the outcome variable, and each time refitting the predictive model and calculating predictive performance, thus, building a null distribution of the performance estimate (Cruz et al. 2020; De Meyer et al. 2023). Such per-

mutation methods have the downside of being computationally demanding, and of not stating a clear null value for R^2. The null hypothesis tested by these permutation methods is that the predictors are not predictive of the outcome, which implies some unknown R^2 value below 0 (see Supplementary Section 3). Also, permutation methods only yield p-values, but no standard errors.

As an alternative, we have provided a standard error for the out-of-sample R^2 estimated through cross-validation or the 0.632 bootstrap by building on recent advances in standard error estimation for loss functions. Our method is also easily extendable to R^2 values estimated on independent test data. The standard errors on $\hat{R}^2$ allow for testing the null hypothesis H_0: $R^2 \leq 0$, which mirrors the interpretable alternative hypothesis H_a: $R^2 > 0$ that indicates that the predictive model significantly improves upon the null model without predictors. The standard error on $\hat{R}^2$ can also be used for the construction of confidence intervals and for comparing predictability of different outcome variables with possibly different units. Comparison of different prediction models for the same outcome variable, on the other hand, is better done directly on the MSE values, as this does not require the approximation of the variance of a ratio through the delta method.

The delta method standard error estimators we provided are upward biased in some settings. Possible explanations are a poor approximation by the delta method at low sample sizes when the departure from normality of the estimator $\hat{R}^2$ is strongest (see Supplementary Section 2), and the difficulty in estimating the correlation ρ between $\widehat{\text{MSE}}$ and $\widehat{\text{MST}}$. Yet the delta method standard errors, using nonparametric or jackknife estimates of ρ, are still preferable to bootstrap standard errors, which can be downward biased and lead to loss of Type I error control and lower than nominal coverage of the confidence intervals. The variance of the estimator of out-of-sample R^2 is small when there is hardly any predictive value in the predictors, but all methods considered overestimate this variance. Fortunately, this bias decreases as the predictive value increases, promising good power to detect truly predictable outcomes. Nevertheless, the estimator variance of $\hat{R}^2$ was found to be considerable in the high-dimensional scenario, supposedly because of the variability in the model fitting of high-dimensional, penalized models. This cautions against overinterpretation of (subtle differences between) $\hat{R}^2$ values estimated for such models. Hence, we encourage reporting standard errors and confidence intervals for diagnostics of predictive models to provide insight into the reproducibility of the result, guide follow-up study design, and allow for model comparison.

Supplementary Materials

R-package The R-code for calculation of out-of-sample $\hat{R}^2$ and its standard error is available in the R-package *oosse* from CRAN (*https://cran.r-project.org/web/packages/oosse*). (url)

R-code R-code for running all simulations and analyses is available at *https://github.com/maerelab/Rsquared*. (url)

Supplementary material Exhaustive simulation results, proofs, and software versions. (pdf)

Funding

The work of SH in the lab of SM was supported through a research collaboration with Inari Agriculture NV funded in part by Flanders Innovation & Entrepreneurship (VLAIO, grant HBC.2019.2814). W.W. received funding from the Flemish Government under the "Onderzoeksprogramma Artificiële Intelligentie (AI) Vlaanderen" programme. The funders had no role in study design, data analysis, decision to publish, or preparation of the manuscript.

ORCID

Stijn Hawinkel https://orcid.org/0000-0002-4501-5180
Steven Maere https://orcid.org/0000-0002-5341-136X

References

Anderson-Sprecher, R. (1994), "Model Comparisons and R2," *The American Statistician*, 48, 113–117. [15,16]

Bates, S., Hastie, T., and Tibshirani, R. (2023), "Cross-Validation: What Does It Estimate and How Well Does It Do It?" *Journal of the American Statistical Association*, 118, 1–22. [16,17,18,19,20]

Bradley, A. P. (1997), "The Use of the Area under the ROC Curve in the Evaluation of Machine Learning Algorithms," *Pattern Recognition*, 30, 1145–1159. [17]

Braga-Neto, U. M., and Dougherty, E. R. (2004), "Is Cross-validation Valid for Small-Sample Microarray Classification?" *Bioinformatics (Oxford, England)*, 20, 374–380. [17,23]

Breiman, L. (2001), "Random Forests," *Machine Learning*, 45, 5–32. [15]

Cameron, A. C., and Windmeijer, F. A. G. (1997), "An R-squared Measure of Goodness of Fit for Some common Nonlinear Regression Models," *Journal of Economics*, 77, 329–342. [15]

Campbell, J. Y., and Thompson, S. B. (2008), "Predicting Excess Stock Returns Out of Sample: Can Anything Beat the Historical Average?" *The Review of Financial Studies*, 21, 1509–1531. [16,17]

Cohen, P., West, S. G., and Aiken, L. S. (2014), *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, New York: Psychology Press. [15]

Cruz, D. F., De Meyer, S., Ampe, J., Sprenger, H., Herman, D., Van Hautegem, T., et al. (2020), "Using Single-plant-omics in the Field to Link Maize Genes to Functions and Phenotypes," *Molecular Systems Biology*, 16, e9667. [21,23]

De Meyer, S., Cruz, D. F., De Swaef, T., Lootens, P., Block, J. D., Bird, K., et al. (2023), "Predicting Yield of Individual Field-Grown Rapeseed Plants from Rosette-Stage Leaf Gene Expression," *PLoS Computational Biology*, 19, e1011161. [21,23]

DiCiccio, T. J., and Efron, B. (1996), "Bootstrap Confidence Intervals," *Statistical Science*, 11, 189–228. [18]

Efron, B., and Tibshirani, R. (1997), "Improvements on Cross-validation: The .632+ Bootstrap Method," *Journal of the American Statistical Association*, 92, 548–560. [16,17,18]

Friedman, J., Hastie, T., and Tibshirani, R. (2010), "Regularization Paths for Generalized Linear Models via Coordinate Descent," *Journal of Statistical Software*, 33, 1–22. [18]

Gauss, C. F. (1823), "*Theoria Combinationis Observationum Erroribus Minimis Obnoxia*," Göttingen: Dieterich. [17]

Harding, B., Tremblay, C., and Cousineau, D. (2014), "Standard Errors: A Review and Evaluation of Standard Error Estimators Using Monte Carlo Simulations," *The Quantitative Methods for Psychology*, 10, 107–123. [17]

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Vol. 2), New York: Springer. [15]

Jiang, W., and Simon, R. (2007), "A Comparison of Bootstrap Methods and An Adjusted Bootstrap Approach for Estimating the Prediction Error in Microarray Classification," *Statistics in Medicine*, 26, 5320–5334. [17,23]

Kohavi, R. (1995), "A Study of Cross-Validation and Bootstrap for Accuracy Estimation and Model Selection," in *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, Vol. 2 (IJCAI'95), Montreal, Quebec, Canada. San Francisco, CA: Morgan Kaufmann Publishers Inc., pp. 1137–1143. [17,23]

Kvålseth, T. O. (1985), "Cautionary Note about R2," *The American Statistician*, 39, 279–285. [15]

Love, M. I., Huber, W., and Anders, S. (2014), "Moderated Estimation of Fold Change and Dispersion for RNA-seq Data with DESeq2," *Genome Biology*, 15, 550. [21]

Molinaro, A. M., Simon, R., and Pfeiffer, R. M. (2005), "Prediction Error Estimation: A Comparison of Resampling Methods," *Bioinformatics*, 21, 3301–3307. [17,23]

Nagelkerke, N. J. D. (1991), "A Note on a General Definition of the Coefficient of Determination," *Biometrika*, 78, 691–692. [15]

Valbuena, R., Hernando, A., Manzanera, J. A., Görgens, E. B., Almeida, D. R. A., Silva, C. A., et al. (2019), "Evaluating Observed versus Predicted Forest Biomass: R-squared, Index of Agreement or Maximal Information Coefficient?" *European Journal of Remote Sensing*, 52, 345–358. [15,17]

Verweij, P. J. M., and Houwelingen, H. C. V. (1993), "Cross-Validation in Survival Analysis," *Statistics in Medicine*, 12, 2305–2314. [15]

Wherry, R. J. (1931), "A New Formula for Predicting the Shrinkage of the Coefficient of Multiple Correlation," *The Annals of Mathematical Statistics*, 2, 440–457. [15]

Zhang, D. (2017), "A Coefficient of Determination for Generalized Linear Models," *The American Statistician*, 71, 310–316. [15]

Zou, H., and Hastie, T. (2005), "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society*, Series B, 67, 301–320. [18]