Cross-validation strategies

Evangelina López de Maturana & Oscar González-Recio

Topics

Why we need to evaluate the performance of statistical models

Designs

Description

5-fold CV

Why to evaluate the model performance? Model assessment

Predictive models are important tools to provide estimates of:

- patient outcome (Harrell et al, 1996)
- Yet-to-be observed outcomes

The **apparent performance** of the model on this **training** set will be better than the performance in another data set, even if the latter test set consists of patients from the same population (**OPTIMISM**)

The **generalization performance** of a learning method relates to its **prediction capability on independent** test data (Hastie et al, 2008)

Many prediction models **perform poorly** when assessed in **external** validation studies

Model assessment in the genomic prediction setting

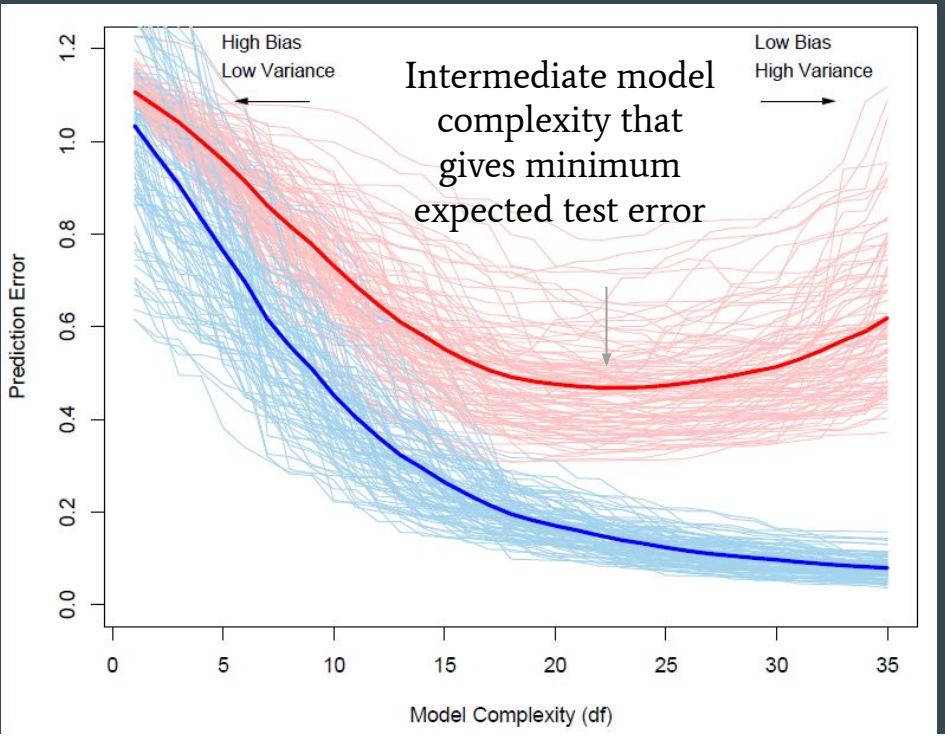
A critical question in the genomic prediction is how **accurately** a model predicts

It has to be **informative** to the potential end user of the model (e.g., breeder, clinician)

Ideally, if we had enough data, we would set aside a validation set and use it to assess the performance of our prediction model

However, since data are often scarce, this is usually not possible

Interplay between bias, variance and model complexity



$$L(Y, \hat{f}(X)) = \begin{cases} (Y - \hat{f}(X))^2 & \text{squared error} \\ |Y - \hat{f}(X)| & \text{absolute error.} \end{cases}$$

Training error is the average loss over the training sample

It decreases when complexity increases

A model with 0 training error is overfit→ it will poorly generalize

Training error is not a good estimate of the test error

(Hastie et al, 2008)

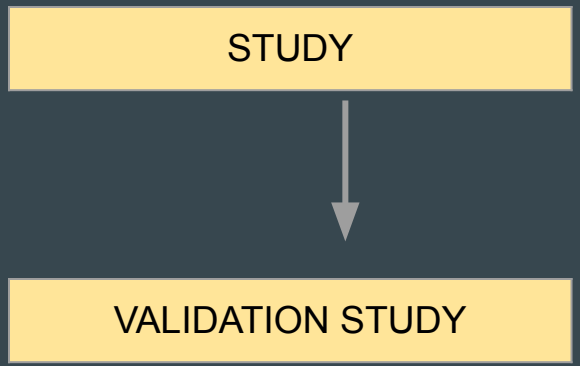
How to evaluate the model performance? Model selection

We may want to estimate the performance of different models in order to choose the best one

- External validation
- Internal validation

How to evaluate the model performance?

External validation (Two stages study)



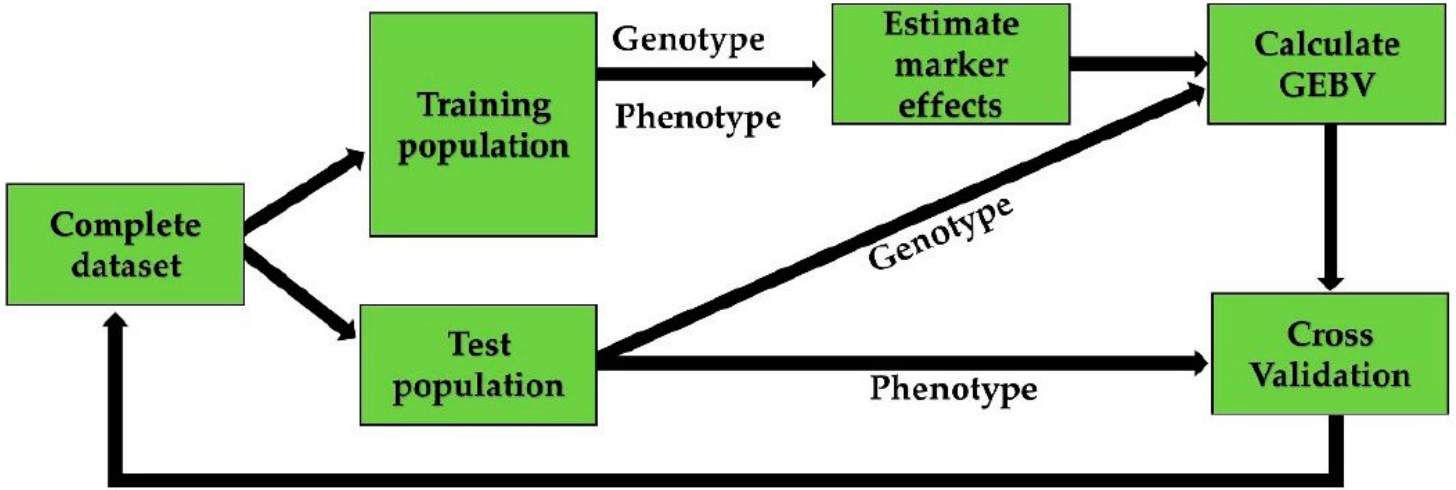
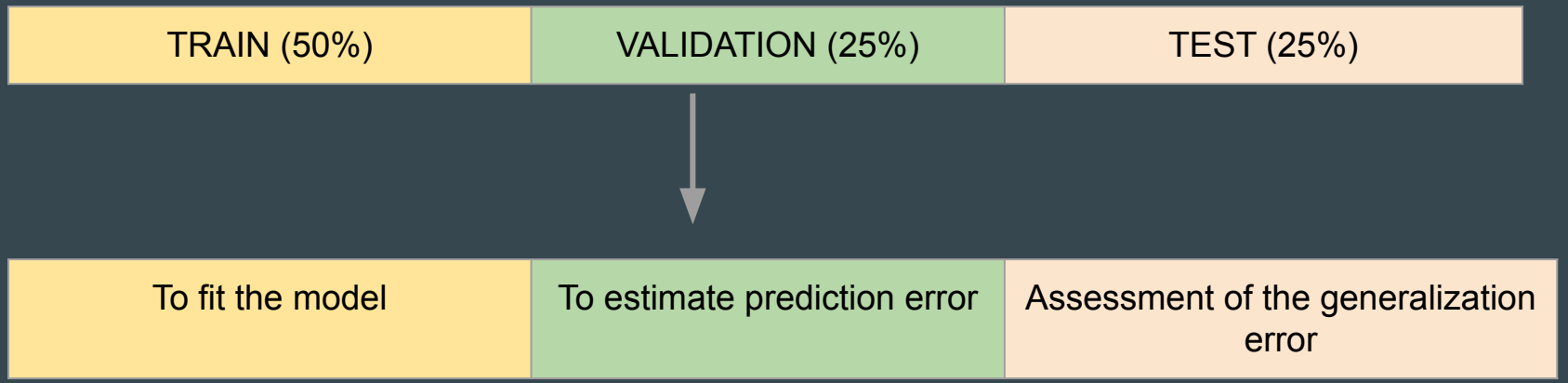


Figure 1. Overview of genomic selection with cross validation using a training population to estimate marker effects in order to get a genomic estimated breeding value (GEBV) of lines in the test-population.

Agronomy 2019, 9, 95; doi:10.3390/agronomy9020095

How to evaluate the model performance?

In a data-rich situation, the best approach is to randomly split the data as follows



Types of validation if insufficient data

It is difficult to give a general rule regarding the size of the training set:

- It depends on the signal-to-noise ratio of the underlying function
- The complexity of the models being fit to the data

Types of validation if insufficient data

These general tools can be used with any loss function

Internal validation:

- Split sample methods
- Cross-validation
- Bootstrap resampling:
  - Regular bootstrap
  - 0.632 Bootstrap
  - 0.632+ Bootstrap

Randomly split the dataset (split half-cross validation)

A straightforward approach is to randomly split the training data in two parts:

- Training: to develop the model
- Testing: to measure its performance

With this split-sample approach, model performance is determined on similar, but independent, data (Picard and Berk, 1990)

However, in the absence of sufficient sample size, independent validation is misleading and should be dropped as a model evaluation step (Steyerberg et al, 2001)

Cross-validation

Probably the most widely used method for estimating prediction error

K-fold cross-validation uses part of the available data to fit the model, and the remaining data to test it

Data is randomly splitted into K roughly equal-sized parts

For the k^{th} part, we fit the model to the other K –1 parts of the data, and calculate the prediction error of the fitted model when predicting the kth part of the data

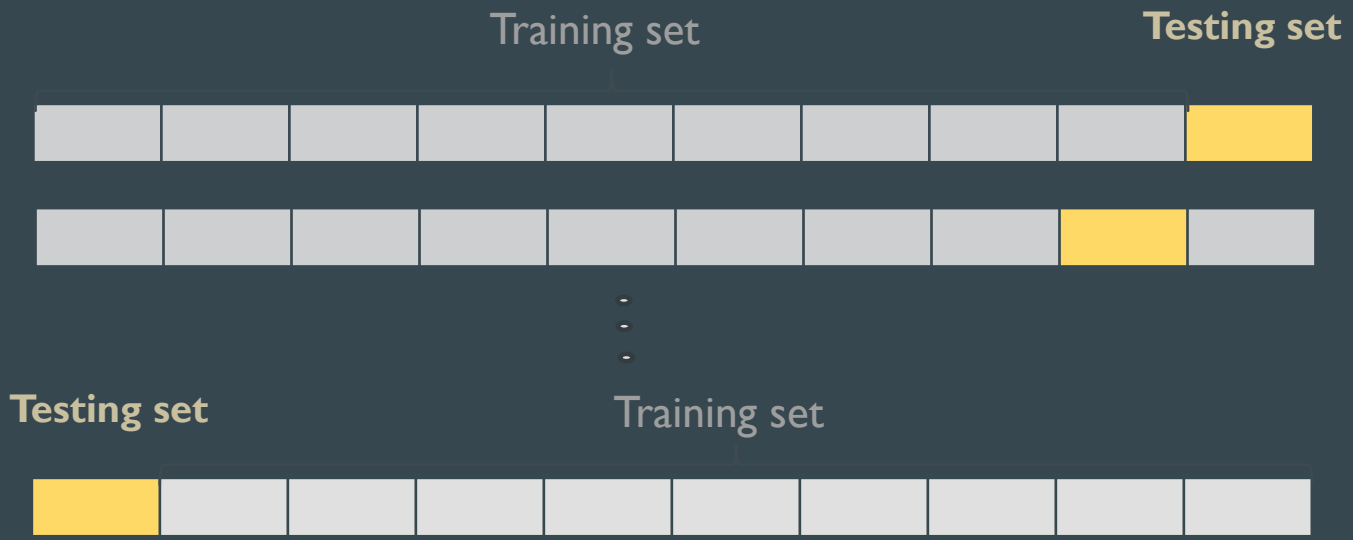
We do this for k = 1, 2, . . . ,K and combine the K estimates of prediction error

Cross-validation

It provides a nearly unbiased estimate of the future error rate (Efron and Tibshirani, 1997)

However, the low bias of cross-validation is often paid for by high variability

K-fold cross-validation



N-times K-fold cross-validation

To improve the stability of the cross-validation, the whole procedure can be repeated several times, taking new random subsamples

Example: n times k-fold crossvalidation

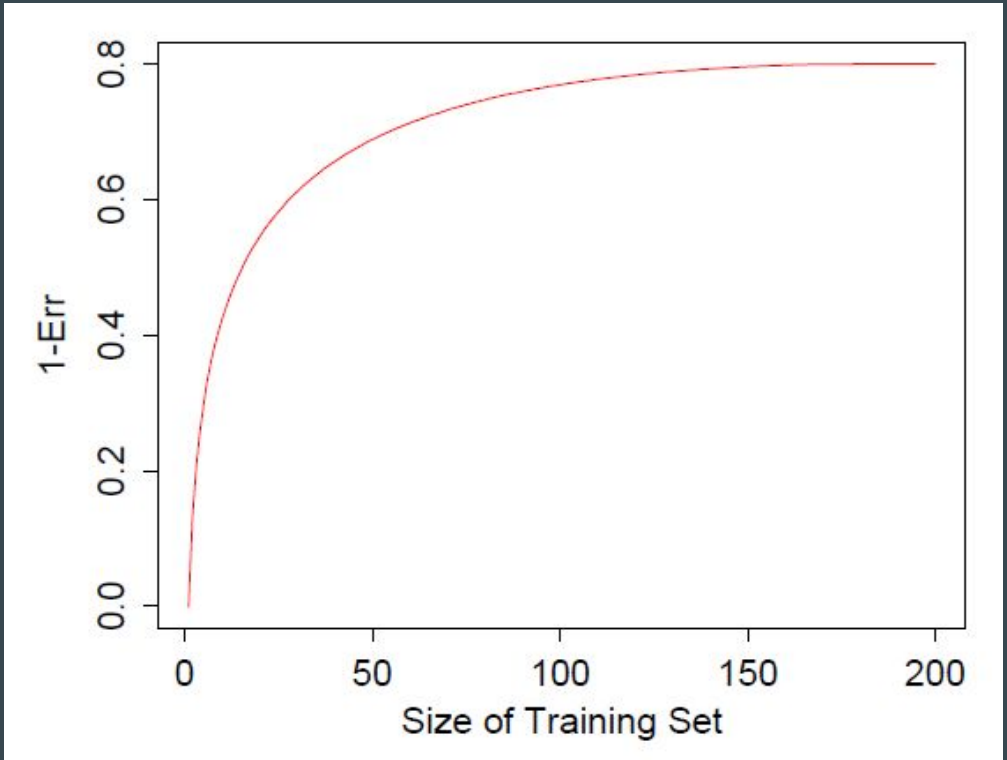
Cross-validation

Usually, K is equal to 5 (5 fold-crossvalidation) or 10 (10 fold-crossvalidation)

If K=N, the process is called leave-one-out cross-validation (equivalent to the jack-knife technique, Efron and Tibshirani, (1983))

For the ith observation the fit is computed using all the data except the i-th

K-fold crossvalidation



Hastie et al (2008)

What value for K?

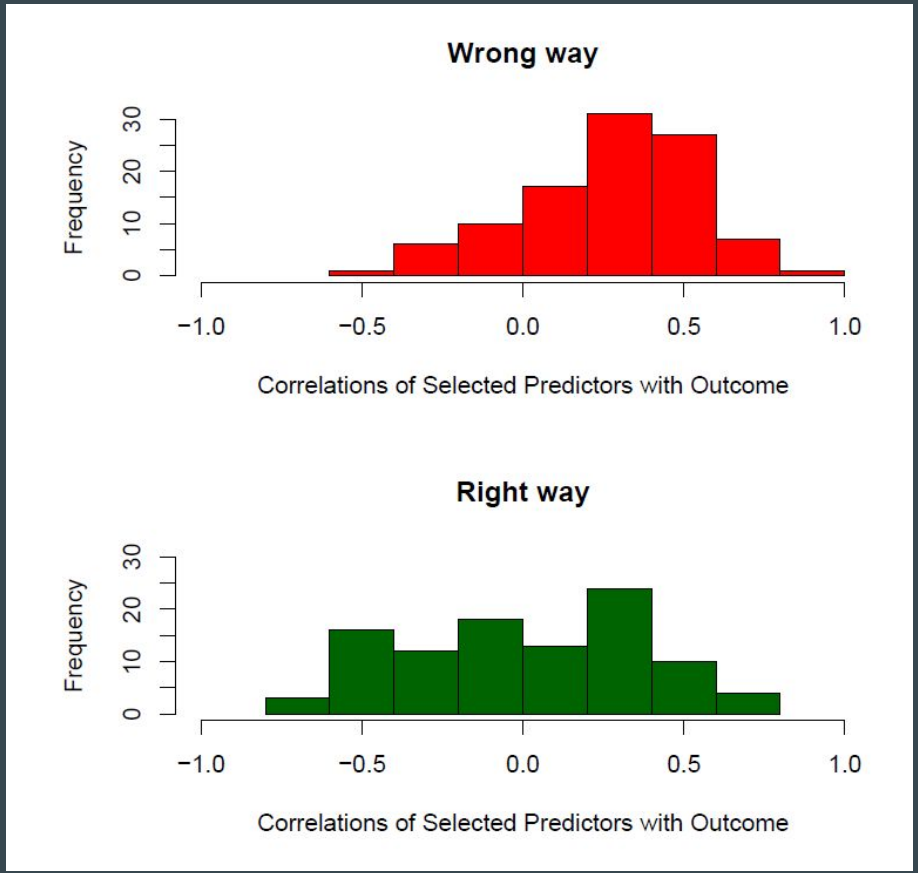
- If the learning curve has a considerable slope at the given training set size, five- or tenfold cross-validation will overestimate the true prediction error
- If K=5, CV has lower variance, although bias can be a problem (it depends on the size of the training set)
- Leave-one-out cross-validation has low bias but can have high variance
- Considerable computational burden of leave-one-out cross-validation
- Overall, five- or tenfold cross-validation are recommended as a good compromise: see Breiman and Spector (1992)and Kohavi (1995)

Wrong way of performing cross-validation

1. In a classification problem, we find a subset of “good” predictors that show association with the outcome
2. Using just this subset of predictors, we build a multivariate classifier
3. Then, we perform a cross-validation to estimate:
    a. the unknown parameters
    b. the prediction error of the final model

Why do you think this is not correct?

Unfair advantage of the predictors



Hastie et al (2008)

Right way of performing CV

1. Randomly divide the sample in K folds
2. For each fold
    a. Find a subset of features associated with the outcome using all the samples but those in fold k (univariate analysis)
    b. Build a multivariate classifier with selected features using all the samples but those in fold k
    c. Use the classifier to predict the outcome for the samples in fold k

Samples must be "left out" before any selection or filtering steps are applied, unless the filtering does not involve the class labels

Bootstrapping

It consists of drawing samples with replacement from the original data set, of the same size as the original data set (Efron and Tibshirani, 1983), to create many simulated samples

This process allows for the calculation of standard errors, confidence intervals, and hypothesis testing

Bootstrap procedures can substantially reduce the variability of error rate predictions

Bootstrapping

Models as estimated in the bootstrap sampling may be fitted in bootstrap samples and original samples, and then compute the difference → estimate of the optimism in the apparent performance

estimated performance = apparent performance - average(bootstrap performance - test performance)

Bootstrapping

Subjects not included in the bootstrap sample:

- Leave-one out bootstrap:
  - it solves the overfitting problem of the previous case
  - It suffers from the training-set-size bias
- Testing in the out of bag (PESSIMISTIC)
- 0.632 Bootstrap
- 0.632+ Bootstrap

0.632 Bootstrap

The average number of distinct observations in each bootstrap sample contains ~ 63% of the original data (some records appear more than once and other not at all) and around 27% of records are kept out of bag (OOB samples)

- Bias will behave ~ to twofold cross-validation

  estimated performance is a weighted combination of apparent and test estimated performance

estimated performance= 0.368 x apparent performance + 0.632 x test performance

bootstrap sample

Ind not sampled in bootstrap

0.632+ Bootstrap

It is an extension of the .632 method

The weights for the estimated performance are dependent on the amount of overfitting

estimated performance= (1-w) x apparent performance + w x test performance

The weight w is determined by the relative overfitting R: $w = .632 / (1-0.368 \times R)$

R is the ratio of the difference in test and apparent performance to the difference between 'no information' and apparent performance

$$R = (\text{test performance} - \text{apparent performance}) / (\text{"no information" performance} - \text{apparent performance})$$

'No information' performance could be approx as the average of the performance in the original sample where the outcome was randomly permuted, repeated as often as the number of bootstraps (Steyerberg et al., 2001)

Take-home messages

Leave-one-out cross-validation is reasonably unbiased but can suffer from high variability in some problems

5-fold or 10-fold cross-validation exhibits lower variance but higher bias when the error rate curve is still sloping at the given training set size

The leave-one-out bootstrap has low variance but sometimes has noticeable bias

.632+ estimator is the best overall performer, combining low variance with only moderate bias

Method		Training sample	Test sample	Estimated performance	Repetitions
Apparent		Original	Original	Original sample	1
Split-sample	50%	50% of original	Independent: 50% of original	Test	1
	33%	66.67% of original	Independent: 33.33% of original	Test	1
Cross-validation	50%	50% of original	Independent: 50% of original	Average(test)	2
	10%	90% of original	Independent: 10% of original	Average(test)	10
	10×10%	90% of original	Independent: 10% of original	Average(test)	100
Bootstrapping	Regular	Bootstrap	Original	Apparent - average(bootstrap-test)	100[a]
	.632	Bootstrap	Independent: subjects not sampled in bootstrap	0.368 × Apparent + 0.632 × average(test)	100[a]
	.632+	Bootstrap	Independent: subjects not sampled in bootstrap	(1-w) × Apparent + w × average(test)[b]	100[a]

[a]100 bootstrap samples were drawn for EPV 5, 10 or 20, while 50 samples were used for EPV 40 or 80.

[b]The weight w was calculated as: $w = .632 / (1 - .368 \times R)$, with R =(test performance − apparent performance) /("no information" performance − apparent performance)[8](see text).

E.W. Steyerberg et al., 2001

Bibliography

E. W. Steyerberg et al. / Journal of Clinical Epidemiology 54 (2001) 774–781

Hastie, Tibshirani and Friedman. The elements of statistical learning (2008)

Topics

Why we need to evaluate the performance of statistical models

Designs

Description

5-fold CV

Hands-on

CV.R