Bioinformatics, 2022, pp. 1–15

doi: DOI HERE
Advance Access Publication Date: Day Month Year
Paper

PAPER

The winner's curse under dependence: repairing empirical Bayes and a comparative study

Stijn Hawinkel,[1,2] Olivier Thas[3,4,5] and Steven Maere[1,2,*]

[1]Department of Plant Biotechnology and Bioinformatics, Ghent University, Ghent, Belgium, [2]VIB Center for Plant Systems Biology, Ghent, Belgium, [3]Data Science Institute, I-Biostat, Hasselt University, Hasselt, Belgium, [4]National Institute for Applied Statistics Research Australia (NIASRA), University of Wollongong, Wollongong, Australia and [5]Department of Applied Mathematics, Computer Science and Statistics, Ghent University, Ghent, Belgium
[*]Corresponding author: steven.maere@psb.vib-ugent.be
FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

Abstract

Motivation: The winner's curse is a form of selection bias that arises when estimates are obtained for a large number of features, but only a subset of most extreme estimates is reported. It occurs in large scale significance testing as well as in rank-based selection, and imperils reproducibility of findings and follow-up study design. Several methods correcting for this selection bias have been proposed, but questions remain on their relative performance and susceptibility to dependence between features since comparative studies are missing.
Results: We demonstrate how empirical Bayes methods tackling selection bias require a custom density estimator for competitive performance under dependence. Furthermore, we perform a comprehensive simulation study comparing different classes of correction methods for point estimates as well as confidence intervals, while studying the effect of dependence. We find a bootstrap method by Tan et al. [2015] and empirical Bayes methods to perform best overall for correcting the selection bias, although this correction does not always improve the feature ranking. We apply selection bias correction methods to a comparison of the best single-feature prediction model versus the multi-feature model in predicting *Brassica napus* phenotypes from gene expression data, demonstrating that the superiority of the best single-feature model may be illusory.
Availability: All R-code used for the analysis is included in the Supplementary information.
Contact: steven.maere@psb.vib-ugent.be
Supplementary information: Supplementary data are available at *Bioinformatics* online.

Key words: Selection bias, biomarker discovery, simulation, empirical Bayes, dependence

Introduction

Contemporary omics technologies allow to routinely gather datasets containing many features. Commonly all these features are then tested for association with a variable of interest, or some other quantity is estimated for each of these features. Scientific interest then naturally lies with the most extreme estimates. Consequently, often only a small number of most extreme estimates or test statistics (e.g. top 10) are reported and considered for follow-up studies. Yet this subset of extreme estimates is subject to a selection bias called the *winner's curse* [Sun et al., 2011, Zollner and Pritchard, 2007, Palmer and Pe'er, 2017, Efron, 2011]. There are two possible reasons why the estimates are extreme: because their true estimand is extreme or because the random estimation error pushes the estimate towards the extremes. Selecting only the most extreme estimates leads to enrichment of extreme estimation errors, engendering the selection bias. More formally, if all features with estimates for a metric of interest $\hat{\gamma}$ falling below a threshold c are selected (assuming we are interested in small values), the selection bias is defined as $E(\hat{\gamma} - \gamma | \hat{\gamma} < c)$. This bias is usually non-zero and negative, even when the regular estimator is unbiased ($E(\hat{\gamma} - \gamma) = 0$). Selection bias results from using the same dataset for identifying extreme features as well as for providing the corresponding estimates, not unlike data dredging. It has been suggested as a cause for poor reproducibility of published findings and insufficient power in follow-up studies [Sun et al., 2011, Zollner and Pritchard, 2007, Palmer and Pe'er, 2017].
Existing selection bias correction methods are often applied to test statistics in the realm of hypothesis testing [Ghosh et al., 2008, Zhong and Prentice, 2008, Tan et al., 2015, Ferguson et al., 2013, Faye et al., 2011, Bigdeli et al., 2016, Efron, 2011]. Yet also estimation and ordering of raw effect sizes can be of interest as this is applicable beyond hypothesis testing to cases where there is no obvious null value, like predictive modelling. Moreover, even when using test statistics, improving the ordering of top features may improve the success rate of

© The Author 2022. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

subsequent validation efforts. Multiple testing corrections that fix the rate of false findings, e.g. Benjamini-Hochberg adjusted p-values [Benjamini and Hochberg, 1995], are widely adopted, but often not all significant features can be investigated further as resources for follow-up studies are limited. Consequently, it is common practice to only focus on a small set of top test statistics. Conversely, researchers may also consider features falling just short of the significance level when the number of discoveries is small. In such settings it is the number of findings that is fixed, rather than the rate of false findings, and the reliability of the findings might be increased by selection bias correction methods that improve the feature ranking.

Despite the many available methods for selection bias correction, and although many publications do include simulation studies [Tan et al., 2015, Faye et al., 2011, Bigdeli et al., 2016], exhaustive benchmarking studies are lacking. Also the effect of dependence on these procedures has not been thoroughly investigated. The estimates may show dependence for a host of different reasons: co-expression of genes causes correlation between test statistics in differential expression testing, but also reliance on a common random quantity for the calculation of the estimates introduces dependence, e.g. a common outcome variable in prediction problems. [Zuber and Strimmer, 2009] construct correlation adjusted t-statistics ("cat-scores") by decorrelating the test statistics, claiming that this improves feature ranking. Moreover, the empirical Bayes method Tweedie's formula, that is often used to correct for selection bias, has been found to perform poorly in the presence of dependence [Tan et al., 2015, Bigdeli et al., 2016]. Tan et al. [2015] claim that the dependence invalidates Tweedie's formula, but according to Efron [2011] it remains valid and the dependence merely inflates the variability of the resulting estimates. Stephens [2017] notes the independence assumption on the estimates as a weakness of his ash method, but saw no obvious way to relax it. Bigdeli et al. [2016] tackle the problem of correlation for empirical Bayes methods in the context of GWAS studies by splitting the test statistics into non-overlapping sets based on chromosome order, but this is not easily generalizable to other applications.

In this work we investigate the problem with empirical Bayes procedures under dependence, and suggest a solution based on bootstrap aggregating. Next we benchmark these refactored empirical Bayes methods against other selection bias correction methods under dependence. Finally we apply the best performing methods to the *B. napus* dataset, and make recommendations on tackling the winner's curse when estimates may be dependent.

Materials and methods

Correction for selection bias

Given is an $n \times p$ data matrix $\mathbf{X}$ with n samples and p features, possibly with a vector $\mathbf{y}$ of length n that is tested for association with every feature $\mathbf{x}_j$. The quantity of interest for every feature j is indicated by γ_j; we assume that small values are of most interest, e.g. the features with the lowest prediction loss. We distinguish four streams of methods to tackle selection bias: 1) separate estimation, 2) conditional likelihood, 3) bootstrapping and 4) empirical Bayes.

The first method is inspired on the observation that the selection bias springs from using the same data for best feature selection as well as for parameter estimation. An obvious solution is to split the available samples at random in two sets, and use the first set for selecting the features with the smallest estimates, and the second set for providing parameter estimates of the selected features to be reported [Sun and Bull, 2005]. We call this the *separate estimation* procedure.

A second method is conditional likelihood. First γ is estimated for all features using full maximum likelihood (ML) as $\hat{\gamma} = \arg\max_{\gamma} L(\mathbf{x} \mid \gamma)$ with $L(\mathbf{x} \mid \gamma)$ the likelihood of the data $\mathbf{x}$ given parameter value γ (omitting feature subscripts). Next conditional likelihood estimates are found for the subset of features with the smallest estimates, conditional on them falling below some threshold c, which may depend on the other estimates in rank-based selection: $c(\hat{\boldsymbol{\gamma}})$ [Zollner and Pritchard, 2007, Ghosh et al., 2008]. Using Bayes' theorem, the conditional likelihood L_c for features passing the threshold is

$$L_c(\mathbf{x} \mid \gamma) = L(\mathbf{x} \mid \gamma, \hat{\gamma} < c) = \frac{L(\mathbf{x} \mid \gamma) P(\hat{\gamma} < c \mid \gamma, \mathbf{x})}{P(\hat{\gamma} < c \mid \gamma)}. \quad (1)$$

For the features concerned, $P(\hat{\gamma} < c \mid \mathbf{x}, \gamma) = 1$. Assuming normality of the parameter estimate around the true value γ, $P(\hat{\gamma} < c \mid \gamma) = \Phi\left(\frac{c-\gamma}{\sigma_\gamma}\right)$ with Φ the standard normal distribution function and σ_γ the estimate's standard error. This probability shrinks with increasing γ, shifting the peak of the conditional likelihood to larger values and attenuating the selection bias of the ML estimator. The conditional likelihood estimate is then $\hat{\gamma}^{CL} = \arg\max_{\gamma} L_c(\mathbf{x} \mid \gamma)$ [Zollner and Pritchard, 2007]. Conditional ML methods are evidently only applicable to likelihood-based estimation. Moreover, their corrected estimates depend on the cut-off value used or on the number, say h, smallest features considered. This means that, if only the smallest estimate is retained, its corrected estimate is different from the case were the two smallest estimates are retained [Ferguson et al., 2013], unlike for the other three approaches.

A third stream of methods rely on bootstrapping to estimate the selection bias, and then correct the parameter estimates by subtracting the bias. Two distinct bootstrap methods exist: the first one, known as "BR-squared" [Faye et al., 2011], works as follows. For every b = 1, . . ., B:

1. Draw a bootstrap sample by sampling n rows with replacement from $\mathbf{X}$, and if applicable select the corresponding elements from $\mathbf{y}$.
2. Estimate the parameter of interest on both the bootstrap sample (in-sample), yielding $\hat{\gamma}^b_{Dj}$ for feature j, and on the samples not included in the bootstrap sample (out-of-sample), yielding $\hat{\gamma}^b_{Ej}$.
3. Rank the in-sample estimates $\hat{\gamma}^b_{Dj}$ over the bootstrap sample, to yield $\hat{\gamma}^b_{D(k_b)}$, with (k_b) indexing the feature with the k-th smallest value (k-th order statistic) in $\hat{\boldsymbol{\gamma}}^b_D$. Use the same feature ordering to also order $\hat{\gamma}^b_{Ej}$ and $\hat{\gamma}_j$ (the estimates on the real non-bootstrapped data) into $\hat{\gamma}^b_{E(k_b)}$ and $\hat{\gamma}^b_{(k_b)}$, respectively.
4. Set $\hat{q}^b_{(k_b)} = \hat{\gamma}^b_{D(k_b)} - \hat{\gamma}^b_{E(k_b)} + \frac{\hat{\sigma}_{DE(k_b)}}{\hat{\sigma}^2_{D(k_b)}}\left(\hat{\gamma}^b_{D(k_b)} - \hat{\gamma}_{(k_b)}\right)$. $\hat{\sigma}_{DE(k_b)}$ is the empirical covariance between $\hat{\gamma}^b_{D(k_b)}$ and $\hat{\gamma}^b_{E(k_b)}$, and $\hat{\sigma}^2_{D(k_b)}$ is the empirical variance of $\hat{\gamma}^b_{D(k_b)}$, both estimated from all bootstrap samples. The latter term corrects for the fact that $\boldsymbol{\gamma}^b_D$ and $\boldsymbol{\gamma}^b_E$ are estimated from mutually exclusive sets of samples.

Finally, subtract the bias to obtain the corrected estimate $\hat{\gamma}^{\text{BR-squared}}_{(k)} = \hat{\gamma}_{(k)} - 1/B \sum_{b=1}^{B} \hat{q}^b_{(k_b)}$ with $\hat{\gamma}_{(k)}$ the k-th order statistic of $\hat{\boldsymbol{\gamma}}$ (note that this is different from $\hat{\gamma}^b_{(k_b)}$, which is the real data estimate corresponding to the k-th order statistic in

$\hat{\boldsymbol{\gamma}}_D^b$). The second approach, which we will call "Tan2015" [Tan et al., 2015], works as follows:

1. Perform steps 1-3 as for BR-squared, without calculating the out-of-sample estimates.
2. Set $\hat{e}^b_{(k^b)} = \hat{\gamma}^b_{D(k^b)} - \hat{\gamma}_{(k^b)}$ for every b.
3. Subtract the bias: $\hat{\gamma}^{\text{Tan2015}}_{(k)} = \hat{\gamma}_{(k)} - 1/B \sum_{b=1}^{B} \hat{e}^b_{(k_b)}$.

Both bootstrap methods address dependence between estimates by performing the resampling in the first step jointly for all features, i.e sampling entire rows. For the Tan2015 method, also parametric bootstrapping is possible if the dependence structure of the data is known or can be estimated.

A fourth class consists of Bayesian methods, as Bayesian statistics is immune to selection bias [Efron, 2011]. For the two methods considered, the prior distribution is estimated from the data, making them *empirical* Bayes methods. Tweedie's formula is the most popular one for the purpose of selection bias correction [Efron, 2011]. For normally distributed estimators, its estimate of the posterior mean is:

$$\hat{E}(\gamma_j \mid \hat{\gamma}_j) = \hat{\gamma}_j + \hat{\sigma}^2_{\hat{\gamma}_j} \frac{d\hat{l}(\hat{\gamma}_j)}{d\hat{\gamma}_j}, \tag{2}$$

with $\sigma^2_{\hat{\gamma}_j}$ the variance of the estimator $\hat{\gamma}_j$ and $\hat{l} = \widehat{\log f}$ the estimated marginal log density, where

$$f(\hat{\gamma}) = \int_{-\infty}^{\infty} f(\hat{\gamma} \mid \gamma) g(\gamma) d\gamma, \tag{3}$$

and $g(\gamma)$ the prior distribution of the parameters γ over all features. By exploiting the high-dimensionality of the data, the marginal log density l and its derivative can be estimated directly from the observed estimates $\hat{\boldsymbol{\gamma}}$ [Efron, 2011].

Another empirical Bayes procedure which has been applied for selection bias correction [Morrison and Simon, 2018] is adaptive shrinkage (ash) by Stephens [2017]. It assumes a unimodal prior distribution g of which the parameters are estimated from the data, and assumes the estimates $\hat{\gamma}_j$ to be independently and normally distributed with variance $\hat{\sigma}^2_{\hat{\gamma}_j}$. A posterior distribution $p(\gamma_j \mid \hat{\gamma}_j, \hat{\sigma}^2_{\hat{\gamma}_j}) \propto g(\gamma_j) f(\hat{\gamma}_j \mid \gamma_j, \hat{\sigma}^2_{\hat{\gamma}_j})$ is then calculated for each feature, and the posterior mean serves as a point estimate. Note the difference between both empirical Bayes approaches: in the ash method the prior distribution of the parameters $g(\gamma)$ is estimated, whereas Tweedie's formula estimates the marginal distribution of the estimates $f(\hat{\gamma})$, which is a convolution of the prior with the estimator distribution.

Apart from point estimates, confidence intervals for the h smallest estimates are also desirable. The regular confidence intervals do not achieve their nominal coverage for extreme estimates, and custom interval construction is needed [Morrison and Simon, 2018]. For the separate estimation method, the confidence intervals for the selected features are constructed based on the samples from the second group alone. Zollner and Pritchard [2007] and Zhong and Prentice [2008] present a 95% confidence interval for the conditional likelihood estimates consisting of all values γ_j for which twice the difference in log conditional likelihood with $\hat{\gamma}_j^{CL}$ lies below the 95th percentile of a χ^2-distribution with degrees of freedom equal to the number of free parameters of the model. This interval can be asymmetric. Faye et al. [2011] applied a nested bootstrap to construct a symmetric confidence interval for the BR-squared estimates using the bootstrap standard error and relying on normality of the estimator. Morrison and Simon [2018] provide the estimator by Tan et al. [2015] with a confidence interval, also employing a nested bootstrap but using the quantiles of the bootstrap distribution rather than assuming normality of the estimator. This bootstrap interval can be asymmetric to reflect the fact that small estimates are often underestimates. The Bayesian framework of Tweedie's formula also yields an expression for a credible interval on the final estimates, relying on the normality of the posterior distribution. The posterior variance is

$$\text{Var}(\gamma_j|\hat{\gamma}_j) = \sigma^2_{\hat{\gamma}_j} \left(1 + \sigma^2_{\hat{\gamma}_j} \frac{d^2 l(\hat{\gamma}_j)}{d\hat{\gamma}_j^2}\right). \tag{4}$$

and the credible interval for γ_j is found as $\hat{\gamma}_j \pm \Phi(1-\alpha/2)\widehat{\text{Var}}(\gamma_j|\hat{\gamma}_j)^{1/2}$ [Efron, 2011]. Ferguson et al. [2013] also incorporated the variability of the estimator $\frac{d\hat{l}(\hat{\gamma}_j)}{d\hat{\gamma}_j}$ in every point $\hat{\gamma}_j$ through bootstrapping. Since the error on $\hat{\gamma}_j$ and on $\frac{d\hat{l}(\hat{\gamma}_j)}{d\hat{\gamma}_j}$ are approximately independent, this leads to a confidence interval

$$\hat{\gamma}_j \pm \Phi(1-\alpha/2) \left(\widehat{\text{Var}}(\gamma_j|\hat{\gamma}_j) + \widehat{\text{Var}}\left(\frac{d\hat{l}(\hat{\gamma}_j)}{d\hat{\gamma}_j}\right) \sigma^4_{\hat{\gamma}_j}\right)^{1/2}. \tag{5}$$

A credible interval for the ash method can be obtained from the posterior distribution $p(\gamma_j \mid \hat{\gamma}_j, \hat{\sigma}^2_{\hat{\gamma}_j})$ [Stephens, 2017].

Simulation studies

Simulation study 1: mean estimation

A mean vector $\boldsymbol{\mu}$ is generated through p=1,000 independent draws from a normal distribution with mean 0 and variance 0.04. The variance vector $\boldsymbol{\sigma}^2$ is generated by p independent draws from a uniform distribution on [2,10]. Next, multivariate Gaussian data $\mathbf{X}$ of sample size n=50 are drawn with these mean and variance vectors, with correlation between all features equal and set to either 0, 0.25, 0.5 or 0.75. 500 Monte-Carlo instances were generated per correlation strength. The estimand γ_j is the mean μ_j for every feature. Its ML estimate, under the assumption of homoscedastic normality, is $\bar{x}_j = n^{-1}\sum_{i=1}^n x_{ij}$, referred to as the raw estimate. Its standard error $\sigma_{\hat{\gamma}_j}$ is estimated as $\left[(n(n-1))^{-1}\sum_{i=1}^n (x_{ij}-\bar{x}_j)^2\right]^{1/2}$, and a naive confidence interval is constructed as $\hat{\gamma}_j \pm \Phi(1-\alpha/2)\hat{\sigma}_{\hat{\gamma}_j}$.

Separate estimation proceeds by randomly assigning the samples to two equally sized groups, and estimating the γ_j's and σ_j^2's in every group separately. For the conditional likelihood estimation, the cut-off value c is chosen at the h-th smallest raw estimate. For all features falling below the cut-off, the corresponding conditional likelihood (1) is optimised over γ_j to yield the final estimate. A χ^2-distribution with 2 degrees of freedom is used to construct the confidence interval. For the methods relying on bootstrapping, nonparametric bootstrapping proceeds by sampling rows with replacement from $\mathbf{X}$, retaining correlation between parameter estimates. For parametric bootstrapping, a multivariate normal distribution was fit to the data using maximum likelihood, from which n samples were drawn to form the bootstrapped matrix $\mathbf{X}^b$. In both cases B=200 bootstraps are performed. Tweedie's formula with and without density bagging was implemented as described in the previous section, the variance calculation of the derivative of the density in Tweedie's formula in (5) follows Ferguson et al. [2013] with 100 bootstrap samples. The ash method was applied using the *ash* function in the *ashr* R-package [Stephens et al., 2022], both on raw estimates only and using nonparametric and parametric bagging. The *ashci* function was used to construct credible intervals.

For the purpose of feature selection, also test statistics for testing the null hypothesis $H_0 : \gamma_j = 0$ are calculated as $t_j = \frac{\hat{\gamma}_j - 0}{\hat{\sigma}_{\hat{\gamma}_j}} = \frac{\hat{\gamma}_j}{\hat{\sigma}_{\hat{\gamma}_j}}$. The rationale is that the inclusion of estimator variance may improve the feature ranking. The same correction methods as for the raw estimates were applied and the smallest h corrected test statistics were considered as discoveries. For Tweedie's formula a unit variance $\hat{\sigma}^2_{t_j} = 1$ was used. In addition, the correlation adjusted t-statistics (cat scores) implemented in the *shrinkcat.stat* function in the *st* R-package [Opgen-Rhein et al., 2021] were calculated. The *ash* method with equal variances leaves the feature ordering unchanged and was therefore omitted.

As performance measures, we approximate the following quantities by averaging over all Monte-Carlo instances: average estimation error (as approximation of the selection bias), and the average MSE of the $\hat{\gamma}_j$'s for the $h = 1, \ldots, 30$ features with smallest estimates. In addition, the true discovery rate (TDR) is approximated as the average proportion of features that are truly part of the h smallest γ_j values among the h smallest estimates $\hat{\gamma}_j$ or test statistics t_j. The feature selection behaviour is thus always evaluated with respect to the true raw values γ_j, rather than the true test statistics $\frac{\gamma_j}{\sigma_j}$. Further the coverage and width of the confidence and credible intervals for the h smallest point estimates are approximated. Two different notions of coverage exist for a series of selected and ranked estimates. One is the false coverage rate (FCR) by [Benjamini and Yekutieli, 2005], defined as the expected proportion of intervals not covering their true value among an entire set of h selected features. We refer to 1-FCR simply as the coverage, and this is shown unless mentioned otherwise. Yet a method controlling the FCR may still show undercoverage for the intervals around the most extreme estimates, and overcoverage for the less extreme estimates. As an alternative, Morrison and Simon [2018] present the concept of rank conditional coverage (RCC), which is the expected proportion of confidence intervals covering the true value for the feature with rank k point estimate. The empirical Bayes credible intervals are also evaluated according to these frequentist definitions of coverage. Finally, the average MSE of all estimates is calculated per Monte-Carlo instance, and its distribution over the Monte-Carlo instances is shown as a boxplot. Except for the RCC and the MSE of all estimates, all performance measures are thus calculated cumulatively over the h smallest estimates. For the separate estimation method, TDR is shown for the feature selection part of the data, the bias, MSE and confidence intervals are calculated for the final estimates based on the estimation part.

Simulation study 2: cross-validation estimation of predictive performance

An $n \times p$ predictor matrix $\mathbf{X}$ is constructed by independent draws from a standard normal distribution. Elements of an outcome vector $\mathbf{y}$ are drawn from a normal distribution with as mean the corresponding element from the vector $\mathbf{X}\boldsymbol{\beta}$ and variance 1. The entries of the parameter vector $\boldsymbol{\beta}$ are either all 0 (the null scenario), or with entries $\beta_1 = \beta_2 = \ldots = \beta_4 = 0.25$ and all other entries 0 (the alternative scenario). The estimand γ_j in this setting is the out-of-sample root mean squared error (RMSE), which quantifies predictive performance [Hastie et al., 2009] and is estimated through 10-fold cross-validation (CV). A linear model with a single predictor variable estimated by ordinary least squares (OLS) as implemented in the *lm.fit* function in the *stats* R-package is used as a prediction model. The standard error on $\hat{\gamma}_j$ was estimated through either nested CV with 10 inner folds following Bates et al. [2023] (see Supplementary Section 3 for derivation) or through the nonparametric bootstrap [Efron, 1981]. For ash and for building confidence intervals for the separate estimation, only the nested CV SE was used. 500 Monte-Carlo instances were generated per scenario. The true γ values are approximated through Monte-Carlo simulation as detailed in Supplementary Section 1.2.3.

Nonparametric bootstrapping for the bootstrap and empirical Bayes methods proceeds by sampling rows with replacement from $\mathbf{X}$, and taking corresponding elements from $\mathbf{Y}$ (see Supplementary Section 1.2.2 for a minor correction applied to these estimates). Parametric bootstrapping, which retains the dependence structure of the test statistics as required for the Tan2015 method, is not possible in this case without additional assumptions on the joint distribution of the outcome variable and predictors [Morrison and Simon, 2018], and is not included here. However, for bagging the density estimator for the empirical Bayes methods, independence between the estimates is required, so for this aim we can still use the parametric bootstrap. Separate outcome vectors $\mathbf{y}_j$ are then drawn for every feature, based on univariate normal distributions of Y given X_j with parameters estimated from the data using OLS. No nested bootstrap confidence intervals were constructed for the BR-squared and Tan2015 methods as this is computationally too heavy. For the purpose of feature selection, the estimates were decorrelated in the style of the cat scores by premultiplying the estimates by the inverse square root of the estimated correlation matrix of the estimates $\hat{\gamma}_j$. This correlation matrix was estimated on $\boldsymbol{\gamma}^*$ using the *crossprod.powcor.shrink* function in the *st* R-package [Opgen-Rhein et al., 2021].

Simulation study 3: Nonparametric simulation based on B. napus data

For this simulation we use a dataset on 62 field-grown *B. napus* plants of the same genetic background for which the leaf transcriptome was measured in November 2016, and a number of phenotypes were measured at harvest in June 2017 [De Meyer et al., 2023]. Six phenotypes where chosen to be predicted from the transcriptome: leaf 8 width and length, leaf count, total seed count, plant height and total shoot dry weight. The gene expression measurements were *rlog* transformed to stabilize their variance [Love et al., 2014].

In a nonparametric simulation, the samples of the *B. napus* data are split at random into two equally sized groups, and the γ_j of every feature is estimated separately in both groups. In this simulation, only the 1,000 most expressed genes were retained as predictors. Next the selection bias correction methods are applied to the estimates from the first group, and their accuracy is assessed by considering the corresponding estimates from the second group as ground truth [Tan et al., 2015]. We repeat this splitting procedure 20 times, which allows us to assess the performance metrics bias, MSE and TDR of top estimates of the different methods on real data. The quantity of interest γ_j is again the RMSE. The prediction model applied is a linear model estimated by generalised least squares (GLS) with Gaussian autocovariance structure as implemented in the *gls* function in the *nlme* R-package [Pinheiro et al., 2021] to account for spatial autocovariance across the field.

The estimators for γ and its standard error covary (see Supplementary Section 4), which leads to overshrinkage of some

of the large raw estimates $\hat{\gamma}_j$ by Tweedie's formula, even into ranges smaller than the smallest raw estimates (this was never observed in Simulation study 2). Since we are only interested in small γ values, as a solution only the 50% smallest raw $\hat{\gamma}_j$'s are shrunk according to Tweedie's formula, the rest are left unchanged. We will call this the truncated Tweedie estimates, and results are shown only for them.

Case study on **B. napus** field trial

The selection bias correction methods investigated were applied to the single-gene RMSE estimates of the 5,000 most expressed genes in the aforementioned *B. napus* dataset. The prediction models were fitted using generalised least squares (GLS) with Gaussian autocovariance structure [Pinheiro et al., 2021]. In addition, a multigene linear model was fitted using a custom version of elastic net (EN) with mixing parameter $\alpha = 0.5$ which accounts for spatial autocorrelation across the field in the same way as GLS, as implemented in the *pengls* package [Hawinkel et al., 2022b]. As in simulation study 3, the truncated Tweedie estimates are used.

Results

Density bagging repairs empirical Bayes under dependence

The effect of dependence between estimates on empirical Bayes procedures is subject to debate [Tan et al., 2015, Stephens, 2017, Efron, 2011, Morrison and Simon, 2018, Bigdeli et al., 2016]. In our simulations we find that dependence negatively affects Tweedie's formula and ash (Figure S2, see Simulation Study 1 below for details), in line with Tan et al. [2015]. To elucidate why this happens for Tweedie's formula, we simulated 4 Monte-Carlo instances of 1,000 independent and 1,000 dependent estimates $\hat{\gamma}_j$, shown in the first respectively second row of Figure 1 (see Supplementary Section 1.2.1 for details). The distributions in the first row are all very similar and approximate the true marginal distribution of all estimates overlaid in blue. In the second row, the dependence narrows the distribution of estimates and shifts it to higher or lower values, strongly departing from the true marginal distribution; a phenomenon known from the field of multiple testing correction [Efron, 2007, Hawinkel et al., 2022a]. It is easy to envision how this behaviour upsets the performance of Tweedie's formula (2): a naive density estimator will yield wrong derivatives of the log-density, leading to incorrect shrinkage. Hence, a better estimate $\hat{l}$ is needed.

Our solution is to apply a data resampling algorithm which breaks the correlation between the estimates and leads to a better exploration of the marginal log density l. We do this by bootstrapping observations, either nonparametrically or parametrically. Crucially, these bootstrap samples are drawn *independently* for every feature j. As a result, the distribution of the estimates $\hat{\gamma}_j^b$ is much broader than the distribution of the $\hat{\gamma}_j$'s only. Although it may still be shifted, it resembles the true marginal distribution much better (the bottom 2 rows in Figure 1). Bootstrapping has the additional benefit of providing many more observations, which stabilizes the density estimation, although the benefit of attenuating the dependence issue is far greater (see Supplementary Section 1.2.4.1). This paradigm is known under the name of bootstrap aggregating or *bagging* [Breiman, 1996], and has been applied before to density estimation outside the dependence context [Bourel and Cugliari, 2019]. The workflow of the proposed modified Tweedie's formula is then as follows

1. Obtain estimate $\hat{\gamma}_j$ and its standard error $\hat{\sigma}_j$ for every feature j.
2. Resample elements from $\mathbf{x}_j$ with replacement, and if necessary corresponding elements from $\mathbf{y}$ (nonparametric bootstrap) to generate $(\mathbf{x}_j^b, \mathbf{y}^b)$ for every bootstrap instance b. Alternatively, generate $\mathbf{y}^b$ from $\mathbf{x}_j$ using an assumed generative model (parametric bootstrap). Estimate $\hat{\gamma}_j^b$ from $(\mathbf{x}_j^b, \mathbf{y}^b)$. Repeat this process B times, for every feature j independently. Combine all estimates in a $B \times p$ matrix of bootstrap estimates $\boldsymbol{\gamma}^*$.
3. Estimate the marginal log density based on all elements of $\boldsymbol{\gamma}^*$. We use Lindsey's method for density estimation, as this allows a direct estimate of the log density and its derivative [Lindsey, 1974].
4. Plug the estimates from steps 1 and 3 into (2) to obtain empirical Bayes estimates.

Similarly, for the ash method, dependence between estimates narrows the estimate of the prior distribution g, leading to underestimation of the posterior variance. Providing independent bootstrap estimates $\boldsymbol{\gamma}^*$ again widens the estimate $\hat{g}$, approximating g more closely and yielding more realistic posterior uncertainty levels (see Supplementary Figure S1). A low number of bootstrap samples, e.g. B=20, is usually enough to provide a stable, wider density estimate. Figure S2 shows how the inclusion of bagging largely repairs the performance of both empirical Bayes methods under dependence.

Simulation studies

We investigate the performance of the 4 methods for selection bias correction (separate estimation, conditional likelihood, bootstrapping and empirical Bayes) in 3 simulation studies with different estimands γ and data generation procedures. All analyses are run in R 4.2.1, see Supplementary Section 6 for the package versions. The original Tweedie's formula with density estimated on the observed estimates $\hat{\boldsymbol{\gamma}}$ only performs so badly that it is not shown in the main text (see Supplementary Figures S2 and S11).

Simulation study 1: mean estimation

The first simulation study studies selection bias correction for maximum likelihood estimation of the mean of Gaussian data under increasing correlation between features. The results in Figure 2 and Supplementary Figures S3-S4 show how the selection bias of the raw estimates decreases with the number of top features and with increasing correlation between features. This latter effect has been proven by Tan et al. [2015]. The coverage of the naive confidence intervals falls short of the nominal level of 95%. The separate estimation method is immune to selection bias, reduces the MSE of the top estimates with respect to the raw estimates in most cases and its confidence intervals achieve the nominal coverage. Its performance is not affected by dependence, but it has wide confidence intervals and low TDR. Conditional likelihood estimation greatly reduces selection bias of top estimates, but overcorrects when the dependence between the estimates is strong. Its confidence intervals are wider than those of the separate estimation method, and slightly overcover. The BR-squared correction only slightly reduces selection bias and MSE in case of independence between features; its feature selection

performance is also very bad in this case and its confidence intervals are narrow but undercover. As the correlation between features grows stronger, its performance improves over all these performance measures, becoming good under strong correlation. The Tan2015 bootstrap correction works well irrespective of correlation strength, reducing selection bias and MSE of the smallest estimates, and the coverage of its confidence intervals hovers around the nominal level. Yet its feature selection behaviour is often slightly worse than that of the raw estimate. Tweedie's formula with density bagging has low bias and MSE for small estimates for scenarios with strong correlation, but the coverage of its credible intervals falls short of the nominal level. In absence of correlation its selection bias correction performance is mediocre, but it performs well at feature selection. The regular ash method performs well under independence, but overcorrects the selection bias as the correlation grows, resulting in deteriorating feature selection behaviour and narrowing credible intervals with coverage plummeting far below the nominal level. Bagging the prior distribution restores its good performance under dependence, resulting in low bias and MSE and good feature selection, with higher coverage of the credible intervals but still below the nominal level. No substantial difference was found between the coverage and the RCC of any method (Supplementary Figure S5).

Feature selection using test statistics works better than using only raw estimates for all methods, especially for the Tan2015 method and Tweedie's formula under strong dependence. The raw test statistics are among the best performers mostly, only under strong correlation (0.75) do the cat scores provide a slight improvement. Bringing the corrected test statistics back to the scale of the raw estimates by multiplying by the standard errors as suggested by several authors [Ferguson et al., 2013, Ghosh et al., 2008] did not offer a clear improvement for any method (Figure S4).

Simulation study 2: cross-validation estimation of predictive performance

Simulation study 2 applies to cross-validation estimation of predictive performance, where a common outcome variable engenders correlation between the estimates. The results are shown in Figure 3 and Supplementary Figures S7-S11. The separate estimation method has a small positive bias in all cases, because the prediction models are estimated on a dataset with half the sample size, resulting in imprecisely estimated models with higher RMSE. The Tan2015 bootstrap method performs well for bias and MSE of top estimates; the BR-squared method performs worse. Tweedie's formula with bootstrap SEs on γ does well at reducing bias and MSE of extreme estimates, but the credible intervals with nested CV SEs have the best coverage, even when not always reaching the nominal level of 95%. This is because the bootstrap SE estimates are less variable than the nested CV SEs (Supplementary Figure S16), leading to less extreme corrected estimates, but the nested CV estimates of the SE are unbiased whereas the bootstrap SE's are downward biased (Supplementary Figures S16-S17), resulting in undercoverage. Supplementary Figures S9-S10 demonstrate how bagging improves Tweedie's formula in case of independent estimates too, although the benefit from attenuating the correlation issue is far greater. The regular ash method performs poorly in the alternative scenario, likely since its assumption of unimodality of the prior is violated, but in the null scenario it has low bias and MSE of the smallest estimates. Its credible intervals are in all cases very narrow and have a very low coverage. The addition of bagging the prior density made little difference. All correction methods perform worse than the raw estimates in terms of feature selection, except the cat scores which offer a very slight improvement. The correlation between RMSE estimators of different features is close to 1 in this simulation study (Figure S13), which explains why these results are most similar to the ones with the highest correlation of 0.75 in simulation study 1.

Simulation study 3: Nonparametric simulation based on **B. napus** data

In simulation study 3 also predictive performance of a common outcome variable is being estimated for many feature, but now the selection bias correction methods are applied to a real dataset. Figure 4 and Supplementary Figure S12 reveal that in the nonparametric simulation study, the separate estimation method has a slight positive bias and low TDR; its MSE of top estimates varies across the phenotypes and its MSE of all estimates is higher than for the raw estimates. The Tan2015 and ash corrections have the lowest bias and MSE for the top estimates, and ash achieves the lowest overall MSE. Also the truncated Tweedie estimates improve upon the raw estimates for all phenotypes in terms of bias and MSE of top estimates, and for some phenotypes in terms of TDR along with ash. The BR-squared method mostly performs poorly. Bagging was not found to improve the performance of the ash method in this case, although it did yield wider credible intervals. Decorrelating the estimates degrades feature selection accuracy for most phenotypes. This observation is explained by the correlation between estimates which hovers around 0.5 for most phenotypes in this scenario (Figure S13) and is thus lower than in simulation study 2.

Case study on **B. napus** field trial

A relevant application for selection bias correction in presence of dependence is in predictive modelling with omics data. Often, several prediction models are built for the same outcome variable, including single-feature models and multi-feature models. It then often seems as though the best single-feature model outperforms the multi-feature model in terms of predictive power [Zhou et al., 2021, Cruz et al., 2020]. This is also observed in a study by De Meyer et al. [2023], where leaf gene expression as well as several phenotypes were measured for individual field-grown *Brassica napus* (rapeseed) plants. The aim was to build prediction models based on the leaf gene expression in autumn for phenotypes measured the following spring, and estimate their predictive performance. Figure 5 shows the root mean squared error (RMSE) of all single-gene prediction models for predicting six selected outcome phenotypes. Because of common dependence on a single vector of outcome values, these estimates are correlated. In addition, a multi-gene linear model predicting the phenotype concerned from the expression of all genes combined was estimated through elastic net (EN). For four out of six phenotypes considered, namely leaf 8 width, leaf count, plant height and total shoot dry weight, the best single-gene model has a lower estimated RMSE than the multi-gene model. This is a plausible result, as even though the multi-gene model is given at least as much information as the single-gene model, its higher model complexity may degrade prediction accuracy. It is also a desirable result, as a single marker feature is easier

to measure than the ensemble of all features. Yet it may also be a case of winner's curse, so we applied the correction methods investigated to the single-gene RMSEs.

Figure S15 plots the corrected versus the raw estimates, revealing a worrying overcorrection of large raw estimates for Tweedie's formula with nonparametric bagging. For this reason we switched to the truncated Tweedie estimates. The other corrections, on the other hand, appear reliable and no modification was applied, although the ash method exhibits very strong shrinkage for some phenotypes, leading to all features having virtually the same corrected RMSE estimate.

Figure 6 shows the smallest raw and corrected single-gene RMSE estimates as well as the multi-gene estimate, together with their confidence intervals. For leaf count, plant height and total shoot dry weight, the smallest single-gene RMSE estimate is no longer smaller than the EN estimate after correcting for selection bias, but for leaf 8 width it still is. Yet also for phenotypes for which the smallest raw RMSE estimate was not smaller than the EN estimate (leaf 8 length and width and total seed count), the smallest corrected RMSE values are considerably larger than the raw estimate, corroborating our initial suspicion that the latter ones are underestimates. The RMSE estimates of the separate estimation method are mostly the largest with the widest confidence intervals; both observations are due to the smaller sample size on which the RMSEs are estimated. The empirical Bayes credible intervals are narrower than the raw confidence intervals, although often unlikely narrow for the plain ash method and for plant height even for the ash method with density bagging. Often, the feature with the smallest raw RMSE is no longer the smallest one after correction, i.e. the feature ranking has changed (Supplementary Table S1).

Discussion

The collection of most promising estimates in a screening experiment are often biased towards extreme values. Although ubiquitous, this selection bias is often ignored, possibly because of lack of awareness or because existing remedies are not widely known. We have benchmarked four classes of correction methods for selection bias in parametric as well as non-parametric simulation: separate estimation, conditional likelihood, two bootstrap methods (BR-squared and Tan2015) and two empirical Bayes methods (Tweedie's formula and ash), see Table 1 for an overview. Separate estimation is the most drastic solution to selection bias, tackling its root cause by using different parts of the dataset for feature selection and for estimation. It eradicates selection bias, achieves nominal coverage of the confidence intervals and is robust to dependence between estimates. Yet this comes at a high cost in estimator variance and feature selection accuracy, and it may still be biased when the estimate depends on the sample size, as for prediction models. In addition, the result of the analysis is random as a result of the single data split, which is undesirable. Although this approach may be commendable when the sample size is large, splitting the dataset in two is no substitute for confirmation from an independent experiment.

Conditional likelihood is only available for maximum likelihood estimation, but there it greatly reduces selection bias, although its estimates are imprecise and it tends to overcorrect when the estimates are correlated. The two bootstrap methods considered in our comparison, BR-squared and Tan2015, yielded different results. The BR-squared correction performs poorly in absence of correlation between estimates, but recovers in presence of such correlation. In the original publication, it was only applied to SNPs which are naturally correlated, but it appears not to generalise well to settings without correlation. Also its confidence intervals relying on normality of the estimator do not reach the nominal coverage. The bootstrap method by Tan et al. [2015], on the other hand, is a robust method that effectively combats selection bias both in presence and absence of dependence, and its confidence intervals achieve good coverage. Presumably because this work is only available as a preprint, it has received little attention in the literature. Both these methods use the average difference of the ranked estimate of a bootstrap sample with a reference quantity to estimate the selection bias. The BR-squared method uses the estimate from a small, random set of out-of-sample observations as reference, whereas the Tan2015 method uses the estimate from the fixed, larger set of observations from the real dataset. The latter option requires fewer parameters to be estimated, is theoretically better justified and performs best in our simulation study, and hence it should be preferred. Both bootstrap methods do not require an estimate of the standard error and make few distributional assumptions. On the downside, these methods do not work for small sample sizes and they can be computationally demanding, especially for the construction of confidence intervals.

Empirical Bayes methods have previously been found to perform poorly when applied to dependent estimates [Tan et al., 2015, Bigdeli et al., 2016, Ferguson et al., 2013], leading to claims that Tweedie's formula is invalid in this case [Tan et al., 2015] and to concerns over the performance of ash [Stephens, 2017]. Here we show that these performance issues originate from the underestimation of the width of the prior distribution under dependence, and demonstrate a remedy in the form of a simple bootstrap aggregating (bagging) procedure with few repeats. Its key idea is that bootstrap samples are generated independently, breaking the dependence between the corresponding estimates and exploring the entire prior distribution space. After this modification, Tweedie's formula and ash regained a competitive performance in our simulations. Unlike the other methods considered, empirical Bayes methods were not developed specifically for combatting selection bias of extreme estimates, but simply provide a posterior distribution for all estimates, which is easily coupled to downstream analyses. On the downside, empirical Bayes methods rely on accurate estimation of estimator variance and marginal log density (Tweedie's formula) or prior density (ash), and on distributional assumptions, which limit computing time but also render them less robust than bootstrap methods.

When a clear null value and an estimator for the standard error are available, conversion of the raw estimates to test statistics improves ranking of the features and hence also feature selection. In presence of strong correlation, also decorrelating the raw estimates or test statistics can improve feature ranking. On the contrary, methods that work well for reducing selection bias (separate estimation, Tan2015 correction and empirical Bayes methods) can exhibit a poorer feature selection behaviour than the raw estimates. This suggests a trade-off between ranking and estimation.

In a case study on a *B. napus* field trial, we have demonstrated how the apparent superiority of the best single-gene prediction model over the multi-gene prediction model for some phenotypes was most likely a case of selection bias. This result calls for caution in marker selection: a correction for selection bias is needed. Based on our results, we recommend

the Tan2015 bootstrap and empirical Bayes correction methods with density bagging for this purpose.

Competing interests

No competing interest is declared.

Funding

The work of S.H. in the lab of S.M. was supported through a research collaboration with Inari Agriculture NV funded in part by Flanders Innovation & Entrepreneurship (VLAIO, grant HBC.2019.2814).

Author contributions statement

S.H. conceived and conducted the experiments, S.H., O.T. and S.M. wrote and reviewed the manuscript. S.M. supervised this study.

References

S. Bates, T. Hastie, and R. Tibshirani. Cross-validation: What does it estimate and how well does it do it? *J. Am. Stat. Assoc.*, 118(ja):1 – 22, 2023.

Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1 1995.

Y. Benjamini and D. Yekutieli. False Discovery Rate-Adjusted Multiple Confidence Intervals for Selected Parameters. *J. Am. Stat. Assoc.*, 100:71 – 81, 03 2005.

T. B. Bigdeli, D. Lee, B. T. Webb, B. P. Riley, V. I. Vladimirov, A. H. Fanous, K. S. Kendler, and S.-A. Bacanu. A simple yet accurate correction for winner's curse can predict signals discovered in much larger genome scans. *Bioinformatics*, 32 (17):2598 – 2603, 2016.

M. Bourel and J. Cugliari. Bagging of density estimators. *Comput. Stat.*, 34(4):1849 – 1869, 2019.

L. Breiman. Bagging predictors. *Mach. Learn.*, 24(2):123 – 140, 1996.

D. F. Cruz, S. De Meyer, J. Ampe, H. Sprenger, D. Herman, T. Van Hautegem, J. D. Block, D. Inzé, H. Nelissen, and S. Maere. Using single-plant-omics in the field to link maize genes to functions and phenotypes. *Mol. Syst. Biol.*, 16(12): e9667, 2020.

S. De Meyer, D. F. Cruz, T. De Swaef, P. Lootens, J. De Block, K. Bird, H. Sprenger, M. Van de Voorde, S. Hawinkel, T. Van Hautegem, D. Inzé, H. Nelissen, I. Roldán-Ruiz, and S. Maere. Predicting yield of individual field-grown rapeseed plants from rosette-stage leaf gene expression. *PLoS Comput. Biol.*, 19(5):1 – 42, 05 2023.

B. Efron. Nonparametric standard errors and confidence intervals. *Can. J. Stat.*, 9(2):139 – 158, 1981.

B. Efron. Correlation and Large-Scale Simultaneous Significance Testing. *J. Am. Stat. Assoc.*, 102(477):93 – 103, 2007.

B. Efron. Tweedie's Formula and Selection Bias. *J. Am. Stat. Assoc.*, 106(496):1602 – 1614, 2011.

L. L. Faye, L. Sun, A. Dimitromanolakis, and S. B. Bull. A flexible genome-wide bootstrap method that accounts for rankingand threshold-selection bias in GWAS interpretation and replication study design. *Stat. Med.*, 30(15):1898 – 1912, 2011.

J. P. Ferguson, J. H. Cho, C. Yang, and H. Zhao. Empirical Bayes correction for the Winner's Curse in genetic association studies. *Genet. Epidemiol.*, 37(1):60 – 68, 2013.

A. Ghosh, F. Zou, and F. A. Wright. Estimating odds ratios in genome scans: An approximate conditional likelihood approach. *The American Journal of Human Genetics*, 82 (5):1064 – 1074, 2008.

T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman. *The elements of statistical learning: Data mining, inference, and prediction*, volume 2. Springer, 2009.

S. Hawinkel, L. Bijnens, and O. Thas. A resampling-based empirical Bayes method for precise false discovery rate estimation under dependence. *ResearchSquare PrePrint*, 2022a.

S. Hawinkel, S. De Meyer, and S. Maere. Spatial Regression Models for Field Trials: A Comparative Study and New Ideas. *Frontiers in Plant Science*, 13, 2022b.

J. K. Lindsey. Comparison of Probability Distributions. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(1):38 – 47, 1974.

M. I. Love, W. Huber, and S. Anders. Moderated estimation of fold change and dispersion for rna-seq data with deseq2. *Genome Biol*, 15(12):550, 12 2014.

J. Morrison and N. Simon. Rank Conditional Coverage and Confidence Intervals in High-Dimensional Problems. *Journal of computational and graphical statistics*, 27(3):648 – 656, 2018.

R. Opgen-Rhein, V. Zuber, and K. Strimmer. *St: Shrinkage t Statistic and Correlation-Adjusted t-Score*, 2021. R package version 1.2.7.

C. Palmer and I. Pe'er. Statistical correction of the Winner's Curse explains replication variability in quantitative trait genome-wide association studies. *PLoS Genet.*, 13(7): e1006916, 2017.

J. Pinheiro, D. Bates, S. DebRoy, D. Sarkar, and R. C. Team. *Nlme: Linear and Nonlinear Mixed Effects Models*, 2021. R package version 3.1-152.

M. Stephens. False discovery rates: A new deal. *Biostatistics*, 18(2):275 – 294, 2017.

M. Stephens, P. Carbonetto, D. Gerard, M. Lu, L. Sun, J. Willwerscheid, and N. Xiao. *Ashr: Methods for Adaptive Shrinkage, using Empirical Bayes*, 2022. R package version 2.2-57.

L. Sun and S. B. Bull. Reduction of selection bias in genomewide studies by resampling. *Genet. Epidemiol.*, 28 (4):352 – 367, 2005.

L. Sun, A. Dimitromanolakis, L. L. Faye, A. D. Paterson, D. Waggott, and S. B. Bull. BR-squared: A practical solution to the winner's curse in genome-wide scans. *Hum. Genet.*, 129(5):545 – 552, 2011.

K. M. Tan, N. Simon, and D. Witten. Selection bias correction and effect size estimation under dependence. *arXiv preprint arXiv:1405.4251*, 2015.

H. Zhong and R. L. Prentice. Bias-reduced estimators and confidence intervals for odds ratios in genome-wide association studies. *Biostatistics (Oxford, England)*, 9(4): 621 – 634, 2008.

X. Zhou, Y. Kono, A. Win, T. Matsui, and T. S. T. Tanaka. Predicting within-field variability in grain yield and protein content of winter wheat using UAV-based multispectral imagery and machine learning approaches. *Plant Production Science*, 24(2):137 – 151, 2021.

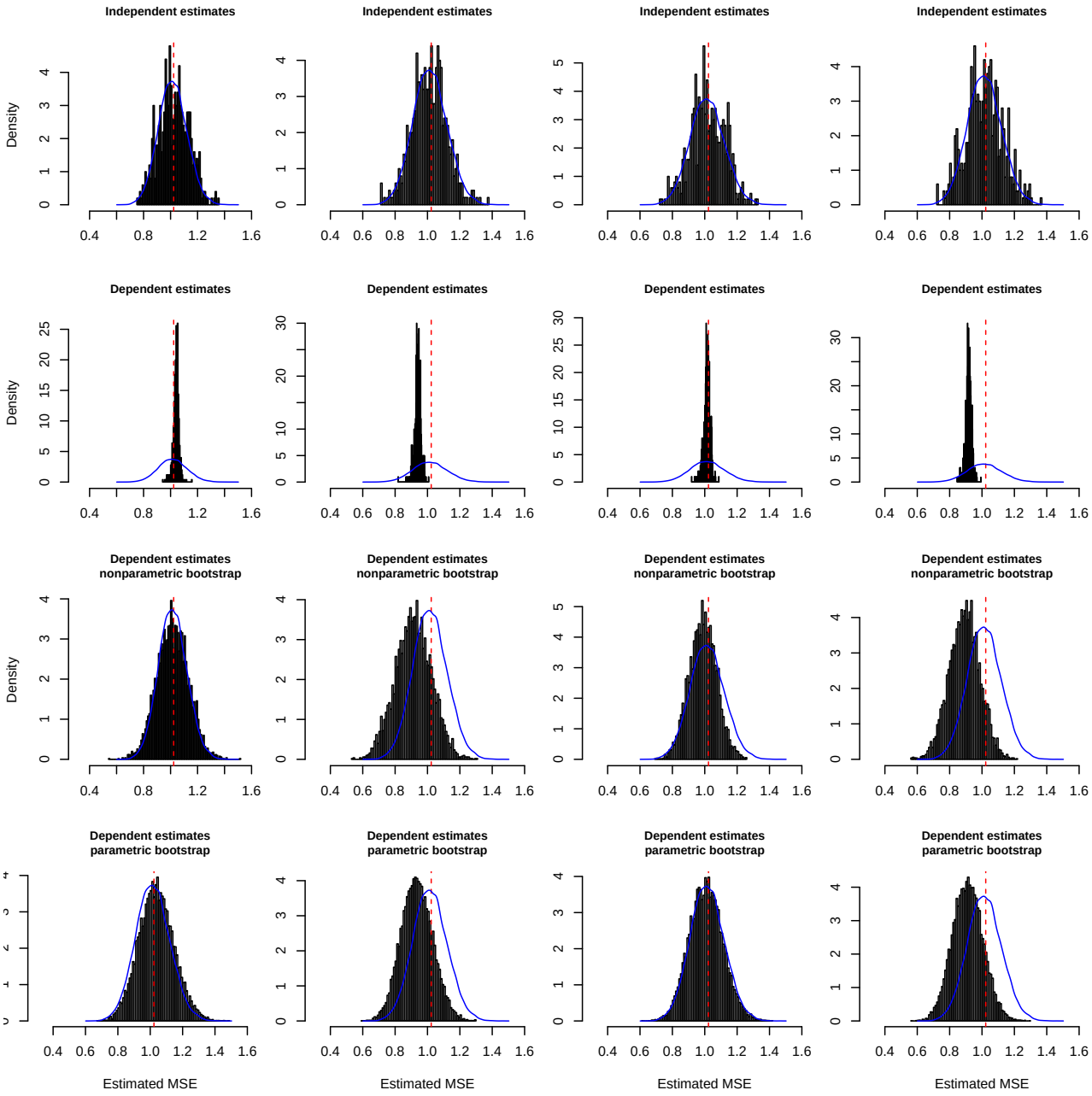

Fig. 1. Histograms of independent estimates (top row) and dependent estimates (second-fourth row), see Supplementary Section 1.2.1 for details. The second row shows the raw estimates, the bottom two rows show the ensemble of bootstrap estimates using nonparametric bootstrapping (third row) and parametric bootstrapping (bottom row). For the columns of the bottom 3 rows, the same Monte-Carlo instances were used, so all values were estimated on the same datasets. The true value is shown as dashed vertical line, the true marginal distribution is overlaid as solid line.

S. Zollner and J. K. Pritchard. Overcoming the winner's curse: Estimating penetrance parameters from case-control data. *Am. J. Hum. Genet.*, 80(4):605 – 615, 2007.

V. Zuber and K. Strimmer. Gene ranking and biomarker discovery under correlation. *Bioinformatics*, 25(20):2700 – 2707, 07 2009.

Figures

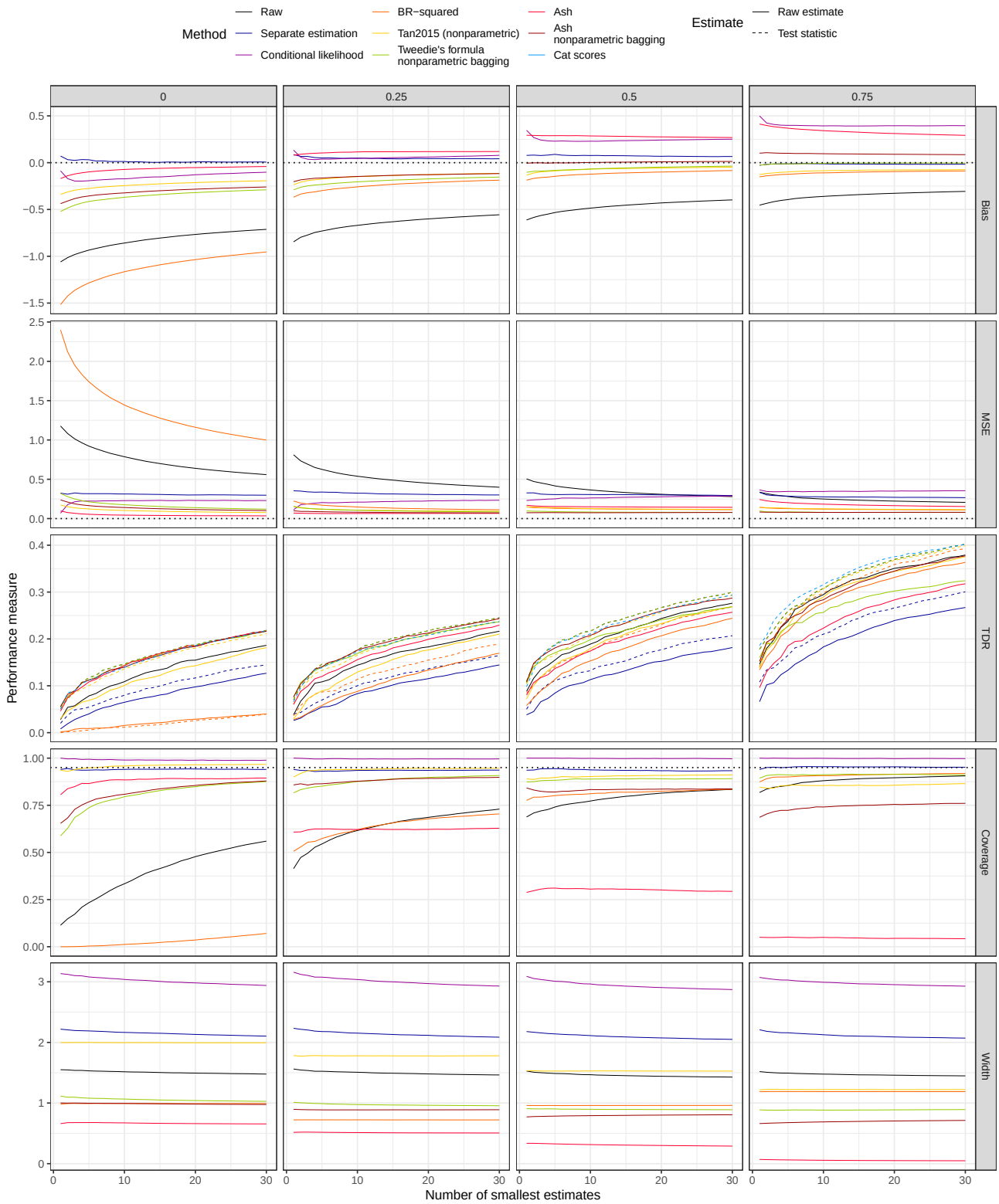

Fig. 2. Simulation results for mean estimation (simulation study 1) for different performance measures (side panels) and different methods (colour) for increasing correlation between features (top panels). The x-axis shows the number of features with smallest estimates considered. For the true discovery rate (TDR) of feature selection, the linetype indicates whether selection is based on raw estimates or on test statistics. The feature selection is the same for conditional likelihood as for the raw estimates, and when using test statistics also identical to ash, so in these cases only the results for the raw estimates are shown. The feature selection behaviour of Tweedie's formula using test statistics is very similar to the one of the raw test statistics, so their lines largely overlap. The dotted horizontal line indicates zero bias or MSE or the nominal coverage of 95%, respectively. All measures shown are averages over all 500 Monte-Carlo instances.

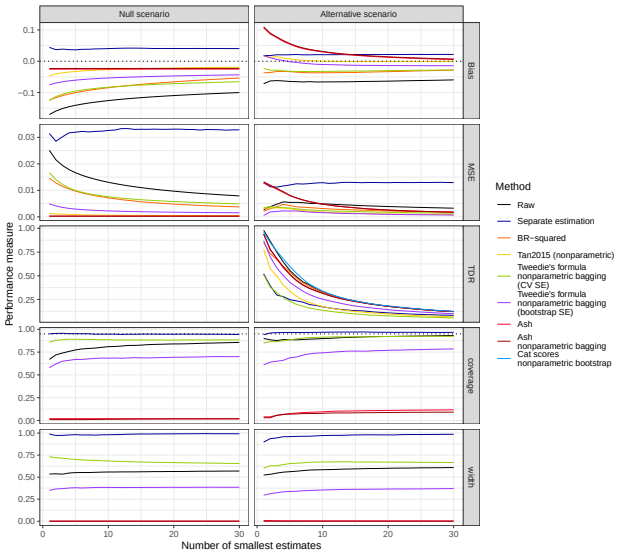

Fig. 3. Parametric simulation results for cross-validation estimation (simulation study 2) of different methods (colour) for the smallest estimates. The x-axis shows the number of top features considered, the y-axis shows the performance measure, the top panels indicate the null or alternative scenario, the side panel indicates which performance measure is shown. Per scenario, the bias (first row) was scaled to the [-1,1] range and the MSE (second row) and confidence interval width (fourth row) to the [0,1] range for comparability. The dotted horizontal line indicates 0 for the bias and the nominal level of 95% for the coverage. Conditional likelihood is not shown as it is not applicable to cross-validation estimation.

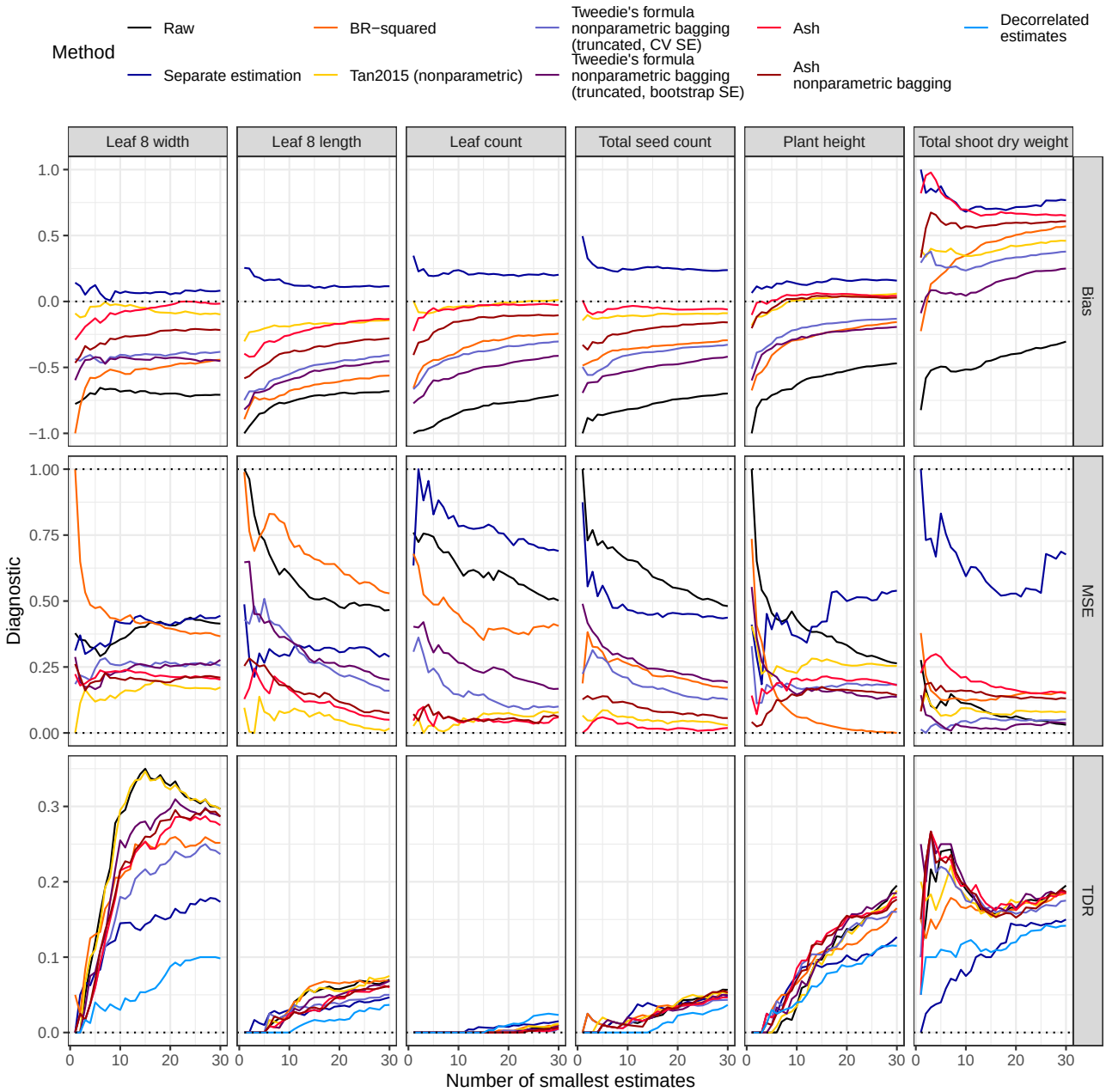

Fig. 4. Nonparametric simulation results based on *B. napus* dataset (simulation study 3) of different methods (colour) for different performance measures (side panels) and phenotypes (top panels). The x-axis shows the number of top features considered, the y-axis shows the performance measure. The bias of the top estimates (top panels) was scaled to the [-1,1] range and the MSE of the top estimates (middle panels) to the [0,1] range per phenotype for legibility. The dotted horizontal lines indicate 0 for all performance measures and also 1 for MSE.

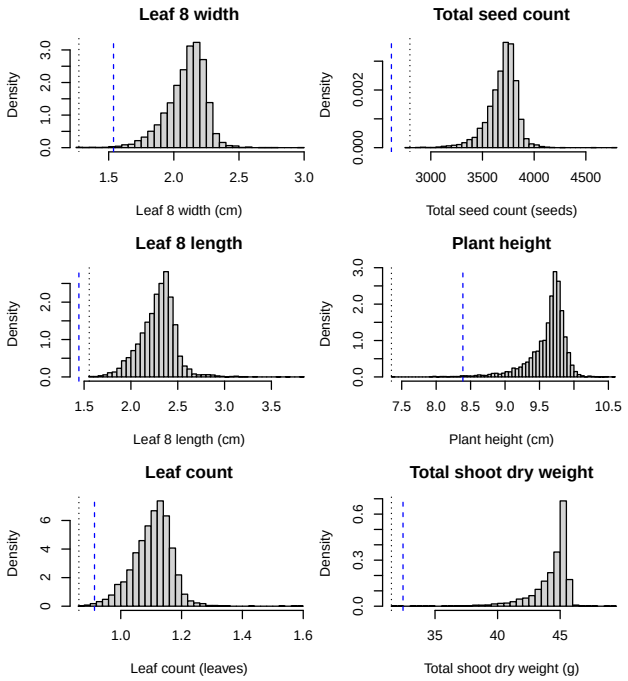

Fig. 5. Histograms of RMSE estimates of single-gene models for predicting different phenotypes (plot titles) of the *Brassica napus* field trial. The RMSE estimate of the multi-gene model is shown as a dashed vertical line, the smallest estimate of the single-gene models is shown as dotted vertical line.

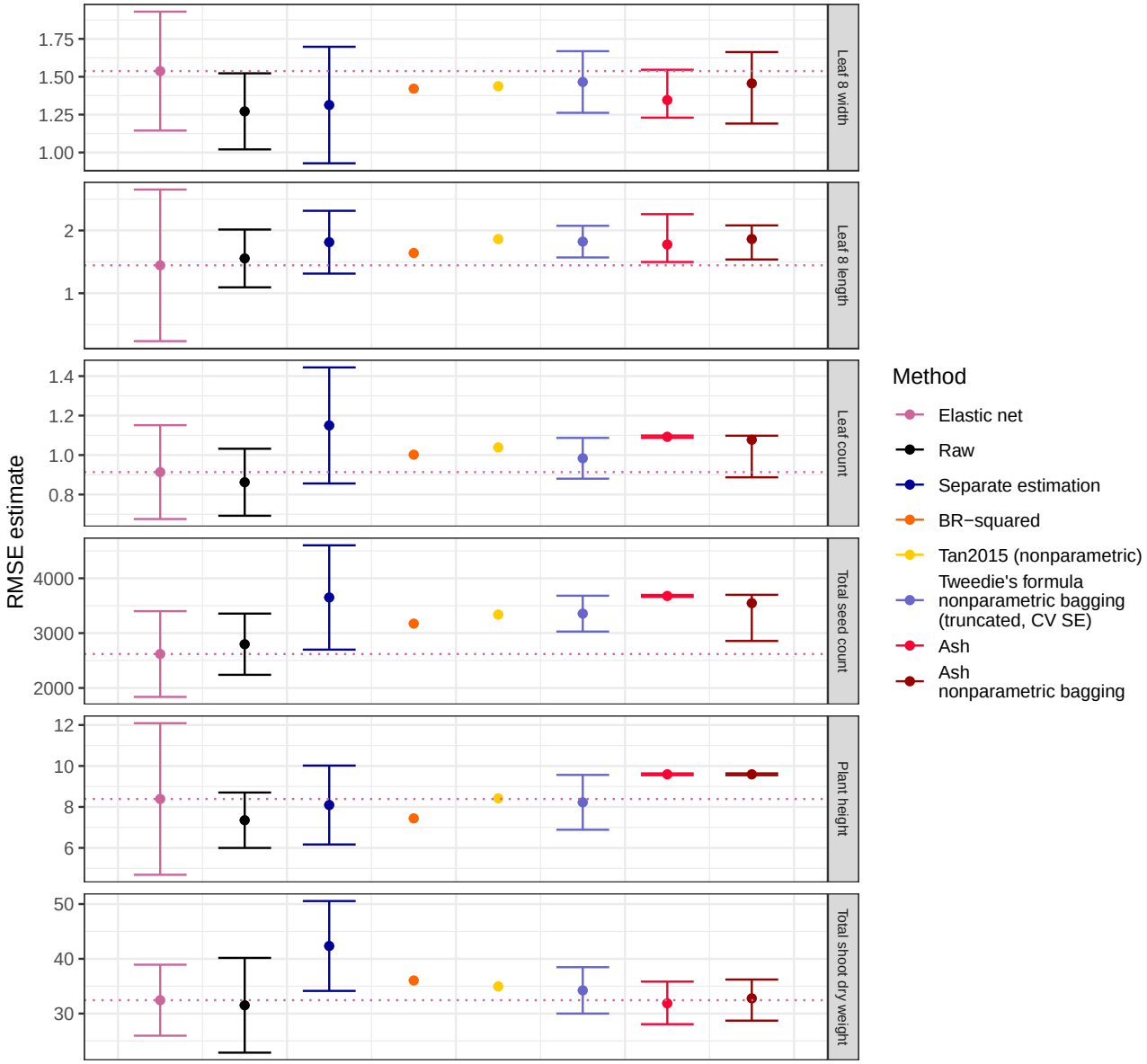

Fig. 6. RMSE point estimates (dots) and 95% confidence intervals (error bars) for the gene with smallest RMSE for different phenotypes of the *B. napus* dataset (side panels) and correction methods (colour). In addition, the elastic net RMSE results for the multi-gene model are shown, the dotted horizontal line corresponds to its point estimate. For the BR-squared and Tan2015 bootstrap methods, no confidence intervals were calculated as this would be too computationally demanding.

Tables

Method	Strengths	Weaknesses
Separate estimation	• Immune to selection bias • Nominal coverage of the confidence interval • Robust to dependence	• Variable estimates • Biased when estimand depends on sample size (e.g. prediction models)
Conditional likelihood	• Reasonable correction of selection bias • Nominal coverage of confidence intervals	• Overcorrects selection bias in presence of dependence between estimates • Variable estimates • Only applicable to likelihood estimation • Computationally intensive • Corrected estimates depend on the number of top features considered
BR-squared	• Reasonable performance under dependence	• Poor performance under independence • Computationally intensive • Ill-suited to small sample size datasets
Tan2015	• Good performance regardless of dependence • No distributional assumptions	• Computationally intensive • Ill-suited to small sample size datasets
Tweedie's formula	• Accurate estimates with reduced selection bias • Good feature selection performance • Robust to dependence thanks to density bagging	• Sensitive to error in estimation of estimator variance and marginal log density
ash	• Accurate point estimates with low selection bias • Good feature selection performance	• Assumption of unimodal prior can be violated • Underestimates posterior estimator variability, even with density bagging
Cat scores	• Improved feature selection under strong dependence using test statistics	• Slight drop in performance in absence of strong dependence • Changeable performance when applied to raw estimates

Table 1. Selection bias correction methods considered in this study with strengths and weaknesses.