The winner's curse under dependence: repairing empirical Bayes using convoluted densities

Stijn Hawinkel, Olivier Thas and Steven Maere

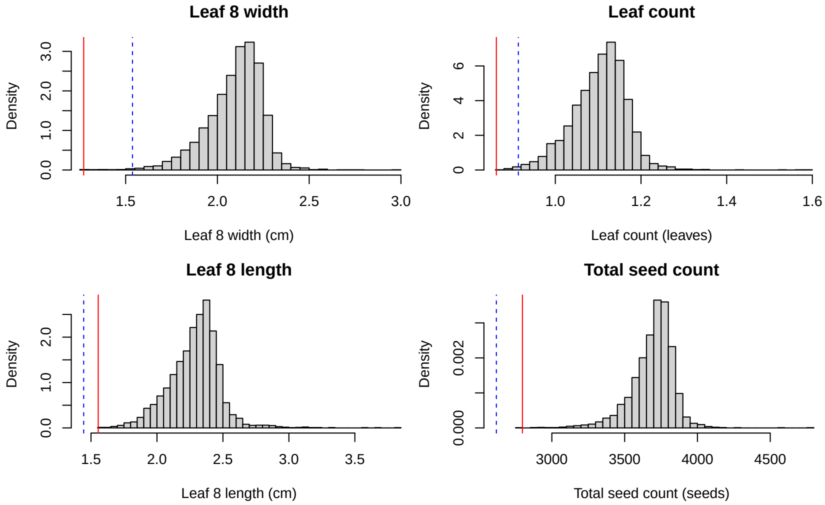
January 24, 2025

VIB-UGENT CENTER FOR PLANT SYSTEMS BIOLOGY

Motivating example: Brassica napus field trial

- Leaf **gene expression** measured in autumn 2016, **phenotypes** in spring 2017 [3]
- Scientific aim: predict phenotypes from gene expression, estimate RMSE (γ) [2]
  - Single gene models (GLS): $\hat{y}_i = \hat{\beta}_0 + \hat{\beta} x_i$
  - Multigene model (elastic net): $\hat{y}_i = \hat{\beta}_0 + \mathbf{x}_i \hat{\boldsymbol{\beta}}$

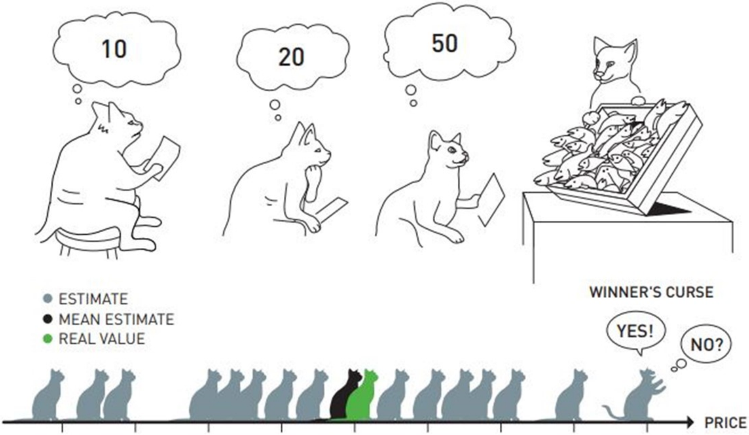
RMSE estimates



Winner's curse

- Only the **most extreme estimates** are of interest
- Estimates $\hat{\gamma}$ are small because
  1) True value γ is small
  2) Estimation error $\hat{\gamma} - \gamma$ is small
- Subset of smallest estimates is **biased**
- $E(\hat{\gamma} - \gamma \mid \hat{\gamma} < c) > 0$ despite $E(\hat{\gamma} - \gamma) = 0$

The auction winner's curse

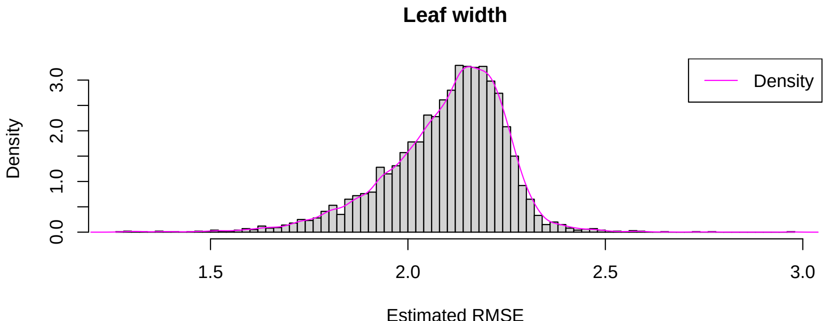


Empirical Bayes: Tweedie's formula [4, 5]

- Bayesian statistics = **immune** to selection bias

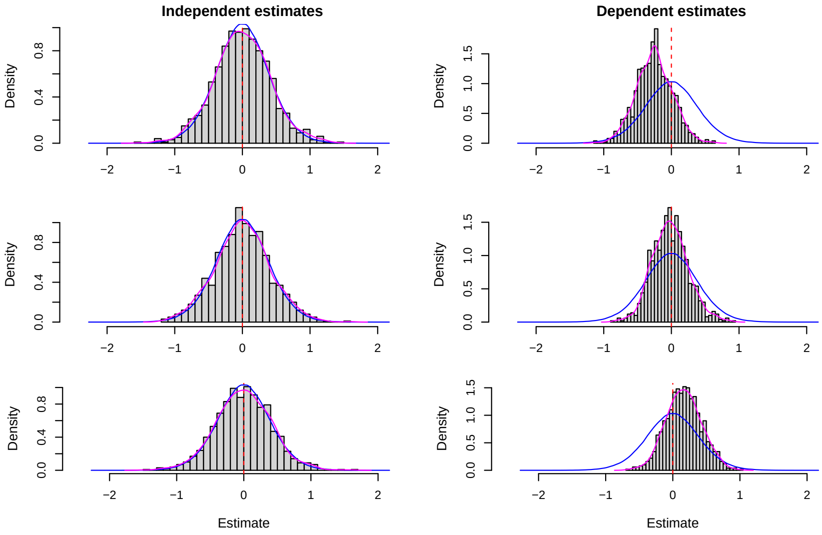
$$E(\gamma_j \mid \hat{\gamma}_j) = \hat{\gamma}_j + \hat{\sigma}^2_{\hat{\gamma}_j} \frac{d\log\left(\hat{f}(\hat{\gamma})\right)}{d\hat{\gamma}}$$

- raw estimate $\hat{\gamma}_j$ and its variance $\hat{\sigma}^2_{\hat{\gamma}_j}$
- $\frac{d\log\left(\hat{f}(\hat{\gamma})\right)}{d\hat{\gamma}}$: derivative of log-density
- No need for **prior density**!



A complication: dependence

- All γ_j's are estimated on the same outcome vector $\mathbf{y}$
  - **Correlated estimates** $=>$ $log(\hat{f}(\gamma))$ is too steep



A theoretical analysis: Hermite polynomials

- Under strong dependence, $\hat{f}(z)$ behaves as a **random function** even as $p \to \infty$ [1, 6]

$$\hat{f}(z) = \phi(z) \sum_{v=0}^{\infty} W_v h_v(z), \tag{1}$$

- $h_v(z)$ the v-th Hermite polynomial, $W_0 = 1$

$$\begin{aligned} E(W_v) &= 0 \text{ if v>1} \\ \text{Var}(W_v) &= \frac{\alpha_v}{v!} = \frac{\int_{-1}^{1} \rho^v dG(\rho)}{v!} \end{aligned} \tag{2}$$

- Dependence introduces **bias** in Tweedie's formula

Our solution: convolution

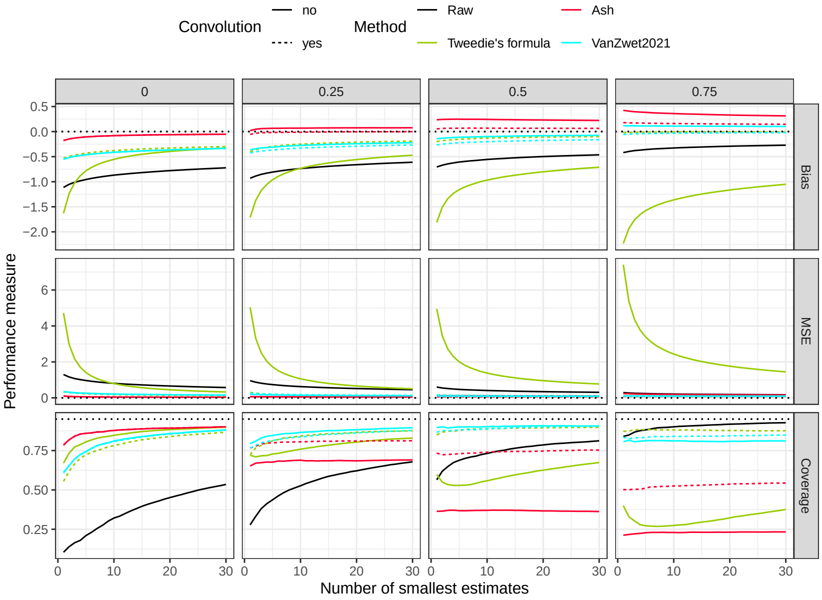
- $\hat{f}(z)$ is too narrow on average

$$E_W\left(\text{Var}_z(z|\mathbf{W})\right) = 1 - \alpha_1 \tag{3}$$

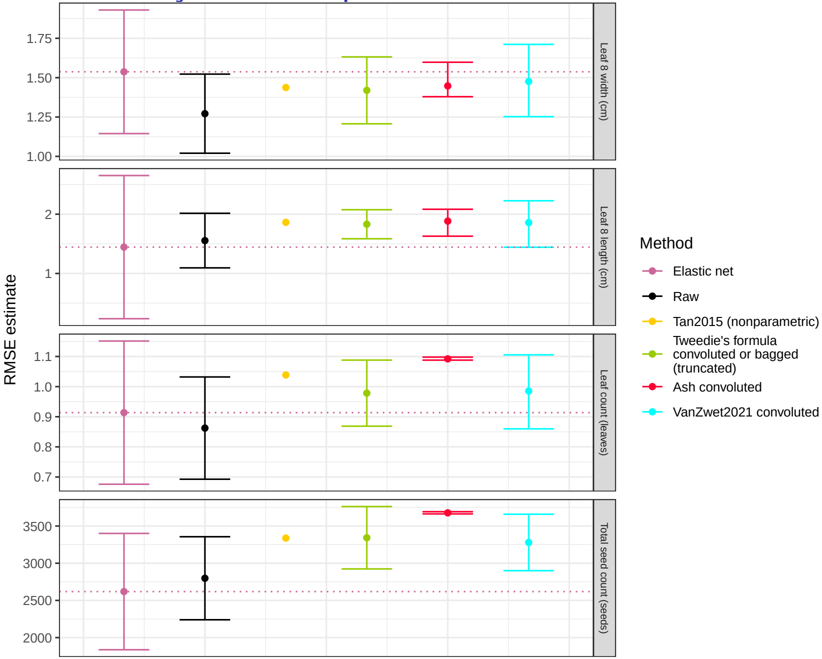
- α_1: average pairwise correlation between the z_j's
- Solution: **convolute** $\hat{f}(z)$ with $N(0, \alpha_1)$

$$\tilde{f}(z) = p^{-1} \sum_{j=1}^{p} r_j(z|z_j, \alpha_1). \tag{4}$$

Simulation study



Real data analysis: B. napus revisited



Conclusions

- Formal proof that **Tweedie's formula** is biased under strong dependence
- Solution: **convolution** with a single parameter normal distribution
- Superiority of single marker gene predictions may be **illusory**

Preprint

HOME | SUBMIT | FAQ | BLOG | ALERTS / RSS | RESOURCES | ABOUT | CHANNELS

Cold Spring Harbor Laboratory

bioRχiv

THE PREPRINT SERVER FOR BIOLOGY

Search

Advanced Search

New Results

Follow this preprint

Previous Next

The winner's curse under dependence: repairing empirical Bayes using convoluted densities

Stijn Hawinkel, Olivier Thas, Steven Maere

doi: https://doi.org/10.1101/2023.09.22.558978

This article is a preprint and has not been certified by peer review [what does this mean?].

0 0 0 0 0 0

Abstract | Full Text | Info/History | Metrics | Preview PDF

Abstract

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 susceptibility to dependence between features since theoretical analyses and comparative studies are few. We prove that estimation through Tweedie's formula is biased in presence of strong dependence, and propose a convolution of its density estimator to restore its competitive performance, which also aids other empirical Bayes methods. Furthermore, we perform a comprehensive simulation study comparing different classes of winner's curse correction methods for point estimates as well as confidence intervals under dependence. We find a bootstrap method by Tan et al. (2015) and empirical Bayes methods with density convolution to perform best at correcting the selection bias, although this correction generally does not improve the feature ranking. Finally, we apply the methods to a comparison of single-feature versus multi-feature prediction models in predicting *Brassica napus* phenotypes from gene expression data, demonstrating that the superiority of the best single-feature model may be illusory.

Posted August 20, 2024.

Download PDF | Email
Print/Save Options | Share
Supplementary Material | Citation Tools
Revision Summary | Get QR code

COVID-19 SARS-CoV-2 preprints from medRxiv and bioRxiv

Subject Area

Bioinformatics

Subject Areas

All Articles

- Animal Behavior and Cognition
- Biochemistry
- Bioengineering
- Bioinformatics
- Biophysics
- Cancer Biology
- Cell Biology
- Clinical Trials*
- Developmental Biology
- Ecology

References

1. Azriel, D. & Schwartzman, A. The Empirical Distribution of a Large Number of Correlated Normal Variables. *J. Am. Stat. Assoc.* **110,** 1217 –1228 (2015).
2. Bates, S., Hastie, T. & Tibshirani, R. Cross-validation: What does it estimate and how well does it do it? *J. Am. Stat. Assoc.* **118,** 1 –22 (2023).
3. De Meyer, S., Cruz, D. F., De Swaef, T., Lootens, P., De Block, J., Bird, K., *et al.* Predicting yield of individual field-grown rapeseed plants from rosette-stage leaf gene expression. *PLoS Comput. Biol.* **19,** 1 –42 (May 2023).
4. Efron, B. Tweedie's Formula and Selection Bias. *J. Am. Stat. Assoc.* **106,** 1602 –1614 (2011).
5. Robbins, H. E. *An Empirical Bayes Approach to Statistics.* in *Breakthroughs in Statistics: Foundations and basic theory* (Springer, 1956), 388–394.
6. Schwartzman, A. Comment on "Correlated z-values and the accuracy of large-scale statistical estimates" by Bradley Efron.