Efficient estimation and correction of selection-induced bias with order statistics

Yann McLatchie and Aki Vehtari
Aalto University, September 18, 2023

https://arxiv.org/abs/2309.03742

Selection-induced bias

Suppose that we choose the model with the highest LOO-CV elpd point estimate from a collection of K candidate models,

$$j^* = \underset{k=1,\dots,K}{\arg\max}\ \widehat{\mathrm{elpd}}_{\mathrm{LOO}}\big(\mathrm{M}_k \mid y\big), \tag{1}$$

and define selection-induced bias concretely as

$$\mathrm{bias}\big(\mathrm{M}_1, \dots, \mathrm{M}_K \mid y\big) = \widehat{\mathrm{elpd}}_{\mathrm{LOO}}\big(\mathrm{M}_{j^*} \mid y\big) - \mathrm{elpd}\big(\mathrm{M}_{j^*} \mid y\big). \tag{2}$$

Bias grows with K

Simulate $n = 100$ data points sampled from

$$y = X\beta + \epsilon \tag{3}$$

$$\epsilon \sim \text{normal}(0, \sigma^2 I), \sigma^2 = 1 - \beta_\Delta^2 \tag{4}$$

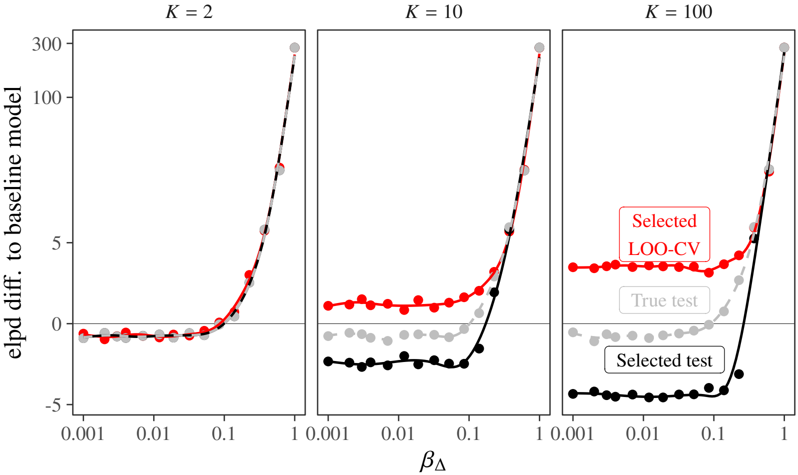
$$\beta = (1, \beta_\Delta, 0, \ldots, 0), \tag{5}$$

Compare $K - 1$ one-predictor models of the form

$$\text{M}_k : y_i \mid \beta_1, \beta_k, \tau \sim \text{normal}(\beta_1 + X_{i,k}\beta_k, \tau^2), \tag{6}$$

to baseline model: $\text{M}_{\text{base}} : y_i \mid \beta_1, \tau \sim \text{normal}(\beta_1, \tau^2)$.

Bias grows with K



Our proposal: modelling the elpd difference

Suppose we have K models, which we compare by elpd:

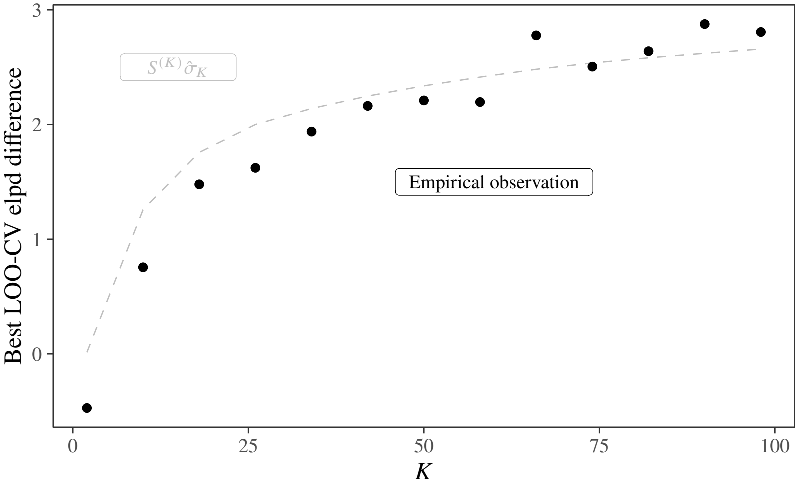
1. fit a half-normal distribution to the upper-tail of K elpd difference point estimates;
2. estimate its standard deviation by MLE, $\hat{\sigma}_K$;
3. compute the expected maximum from K equally-un-predictive models using the maximum order statistic,

$$S^{(K)}\hat{\sigma}_K, \tag{3}$$

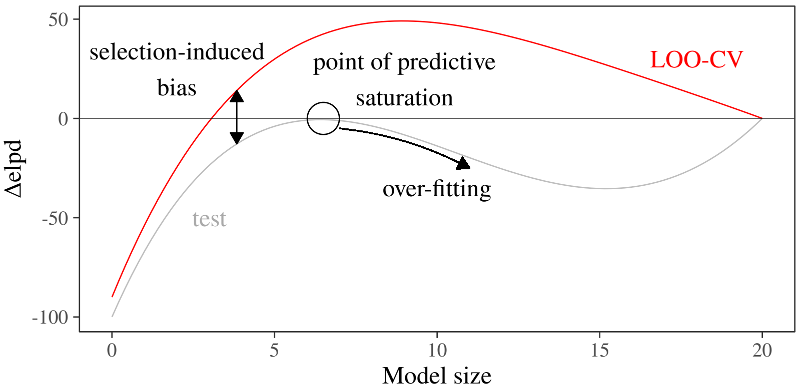
where, for $X_i \sim \text{normal}(0, 1)$

$$S^{(K)} := \mathbb{E}\left[\max_{1 \leq i \leq K} X_i\right]. \tag{4}$$

Our proposal: modelling the elpd difference



Bias compounds in forward search



Correcting bias in forward search

We correct for bias along the search path according to:

$$\Delta\widehat{\text{elpd}}^{(k)}_{\text{corrected}} = \begin{cases} \Delta\widehat{\text{elpd}}^{(k)}_{\text{LOO}} - \widehat{\text{bias}}^{(k)}, \text{ if } |\Delta\widehat{\text{elpd}}^{(k)}_{\text{LOO}}| < S^{(k)}\hat{\sigma}_k \\ \Delta\widehat{\text{elpd}}^{(k)}_{\text{LOO}}, \text{ otherwise.} \end{cases} \tag{3}$$

We produce an estimate of selection induced bias, denoted $\widehat{\text{bias}}^{(k)}$, building on our order statistics-based heuristic:

$$\widehat{\text{bias}}^{(k)} = 1.5 \times S^{(k)}\hat{\sigma}_k. \tag{4}$$

Simulated experiment

Simulate for $p = 100$ predictors:

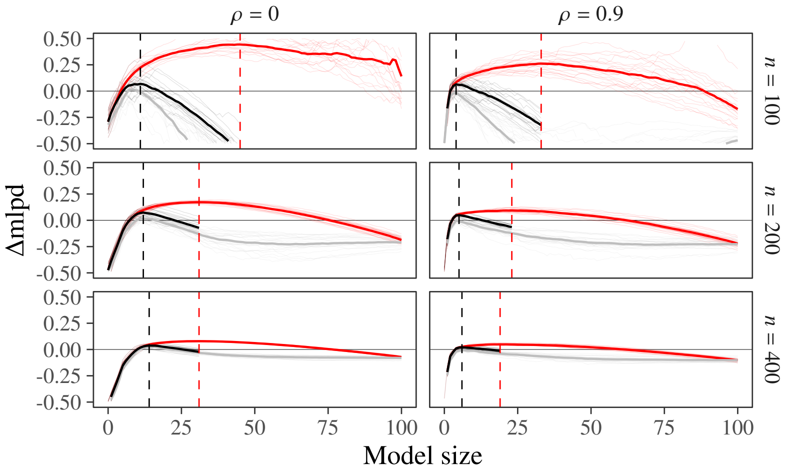
$$x \sim \text{normal}(0, R) \tag{5}$$

$$y \sim \text{normal}(w^T x, \sigma^2), \tag{6}$$

where the matrix $R \in \mathbb{R}^{p \times p}$ is 5×5 block diagonal, having within-block correlation $\rho = \{0, 0.9\}$. Only the first 15 predictors influence the target y: $(w_{1:5}, w_{6:10}, w_{11:15}) = (\xi, 0.5\xi, 0.25\xi)$, and zero otherwise. Set $\xi = 0.59$ and $\sigma^2 = 1$ to fix $R^2 = 0.7$. We simulate $n = \{100, 200, 400\}$ data points according to this data-generating process (DGP).

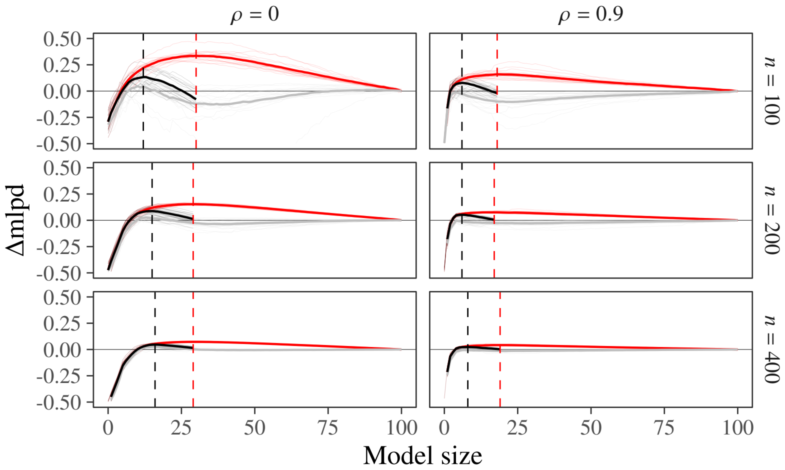
Simulated experiment

Gaussian priors:

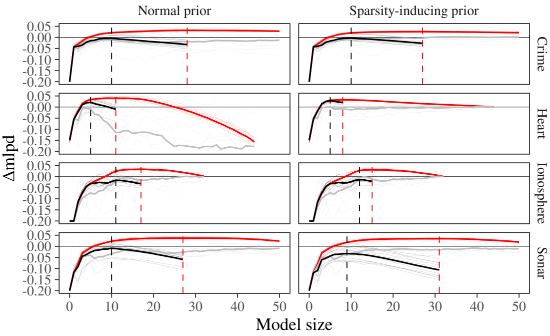


Simulated experiment

R2D2 priors:

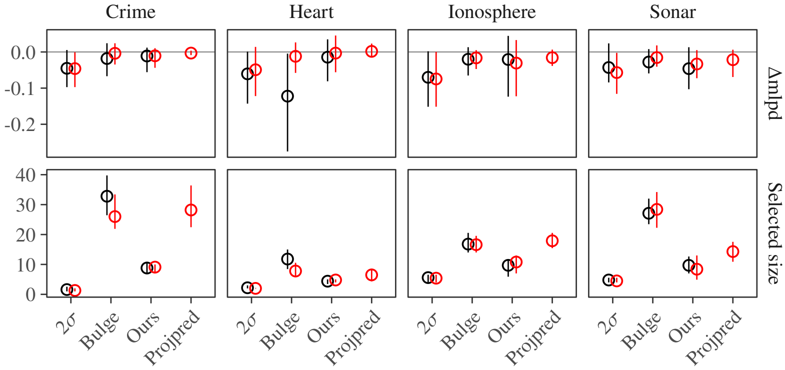


Real-world experiments



Real-world experiments

R2D2 priors in red; Gaussian priors in black.



Recommendations

1. In the two-model case: if the models are not nested, combine them by model averaging or stacking; ensure the models' respective priors are reasonable (goes for all scenarios) and select the more complex of the two; or, keep them both as a set of best models.
2. In the many-model case: all of the recmmendations above, *and* test for clearly predictive models using order statistics $S^{(K)}\hat{\sigma}_K$.
3. In forward search: first try projpred if the model space is large and the observation family allows efficient projection, otherwise LOO-CV forward search can be useful, and we can correct for selection-induced bias in an online fashion.