Multivariate Empirical Bayes Launch Rule: A Practical Case Study

Bartek Kozakowski · Kenneth Hung · Jingang Miao · Houssam Nassif

Meta

The 2026 Conference on Digital Experimentation at MIT

Lessons from fitting, validating and choosing a multi-metric launch rule

Use precise auxiliary metrics to estimate a noisy primary effect. Validate both predictions and decision stability.

01 Model and decision

Five metrics: one noisy primary and four precise auxiliaries.
Fit a three-component Gaussian prior over true effects θ_i using historical experiments and expectation maximization [12].

$$X_i \mid \theta_i \sim N(\theta_i, \Sigma_i),\ \Sigma_i = D_i R D_i$$

D_i contains standard errors; R contains error correlations.
The measurement model accounts for noise in observed effects.

Launch if $P = Pr(\theta_p > 0 \mid X) \geq t$, or $E[\theta_p \mid X] \geq m$. [10]

X_i: observed effects; θ_i: true effects; θ_p: primary effect. t: probability threshold; m: minimum effect.

02 Validation design

Use separate experiments for fitting and validation. Split each experiment in two A/B validation cohorts into disjoint halves [5, 8].

$$X_a = \theta + \varepsilon_a;\ X_\beta = \theta + \varepsilon_\beta$$

Same true effect; independent errors. Split differences check noise.
Cross-products check the prior second moment.

Predict half B from half A. Compare held-out signs and sizes.

03 Lessons learnt

Check noise before fitting a prior

Use split-half differences to check standard errors without relying on the prior.

Store in-experiment error correlations

A/A error correlations may not represent the A/B experiments used for launch decisions.

Check decision stability

Continue fitting and refit plausible priors to check whether the same experiments still launch.

Constrain the prior

Sharing covariance structure between mixture components reduces parameters and makes the fitted prior more stable.

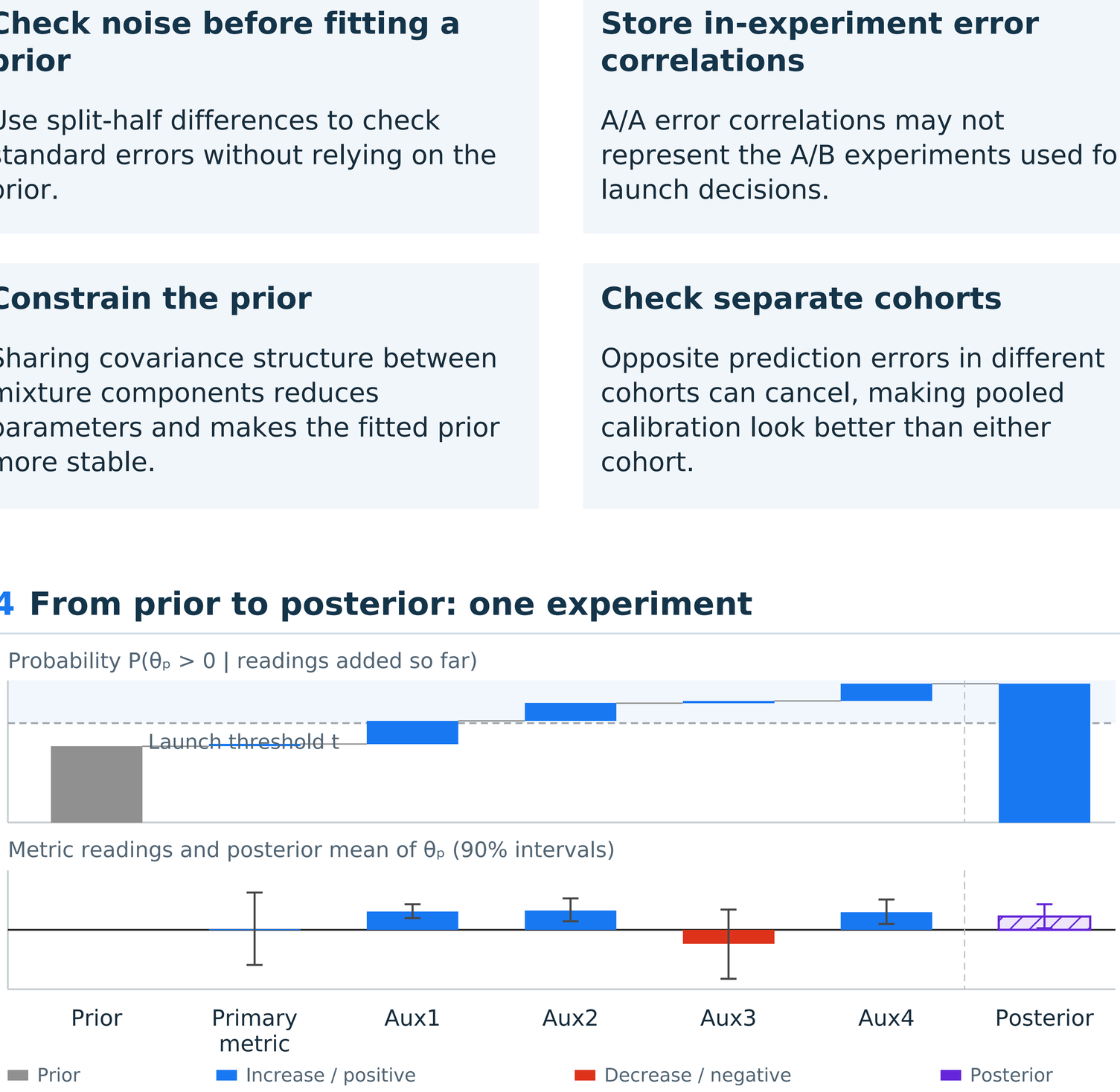
Check separate cohorts

Opposite prediction errors in different cohorts can cancel, making pooled calibration look better than either cohort.

Compare with the primary-only model

Check whether auxiliary metrics improve held-out predictions beyond those using the primary metric alone.

04 From prior to posterior: one experiment



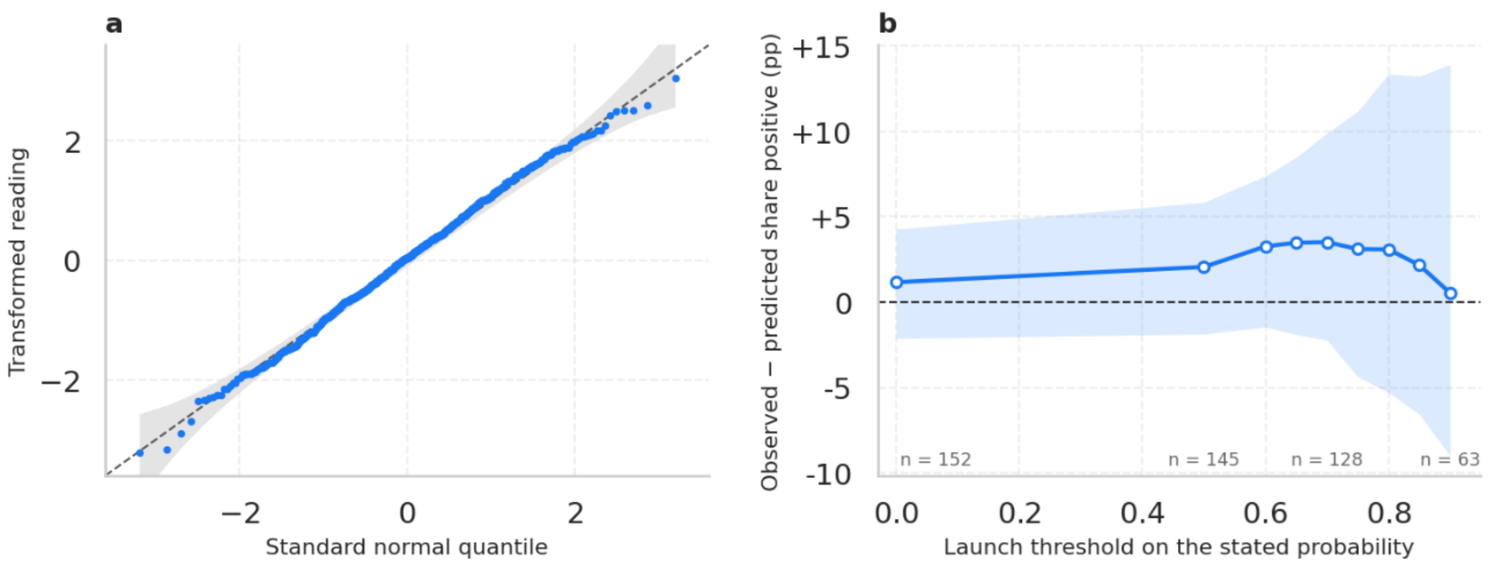
Add information, reduce uncertainty

Start with the fitted prior and add readings in order. The posterior gives both P and $E[\theta_p \mid X]$, with reduced uncertainty.

A negative reading can increase P

Aux3 is negative, yet P rises. With negative cross-metric correlations, it can indicate that another positive reading understates its true effect. Together, the readings strengthen evidence for a positive primary effect.

05 Training fit and pooled validation



a: Training-set fit

Transformed primary readings are compared with their expected quantiles. Points near the diagonal support the fitted distribution.

b: Held-out sign calibration

Compare observed and predicted positive held-out shares. Zero means agreement. Pooled agreement extends through t = 0.90; bands are 90% clustered intervals.

Selected references (paper numbering)

[5] Coey & Cunningham (2019). WWW, 285–295.
[8] Kessler et al. (2025). arXiv:2512.03366.
[10] Stephens (2017). Biostatistics 18(2):275–294.
[12] Urbut et al. (2019). Nature Genetics 51:187–195.