Using causal inference and machine learning tools for biomarker discovery in high-dimensional biology settings

Nima Hejazi

31 May 2024

Department of Biostatistics,
T.H. Chan School of Public Health,
Harvard University

nshejazi
nhejazi
nimahejazi.org

Precision Medicine Symposium
Steno Diabetes Center Aarhus

Joint work with A. Hubbard, M. van der Laan, P. Boileau

Preview

1. Computational biology research produces complex data, but statistical methods are tied to modeling assumptions challenging to verify from substantive knowledge.
2. *Model misspecification* can seriously undermine the scientific value of common, parametric statistical modeling approaches.
3. Semi-parametric inference facilitates construction of robust estimators that are compatible with machine learning.
4. Variance moderation strengthens hypothesis testing strategies, reducing false positives and preserving power under instability.

A common problem

- *Question*: What factors causally impact a health outcome of interest (e.g., cancer, death).
- *Experiment*: Assign[?] patients to novel therapy vs. standard of care (or exposure) and then evaluate outcome's occurrence.
- *Goal*: Deepen mechanistic insights — how does the therapy or exposure biologically operate? Identify intervention points.
- Combine tools from $\star$-omics and molecular biology, clinical trials, causal inference, (bio)statistics, epidemiology.

Meet the data: Smoking and DNA methylation

- *Question*: Which CpG sites, or larger functional units (e.g., "CpG islands") are affected by long-term smoking?
- *Why?* Attempt to understand how smoking induces regulatory and functional changes that relate to disease (e.g., cancer).
- *Study*: Observational exposure study of 253 individuals (172 smokers, 81 non-smokers) and 450,000 CpG sites assayed.
- *Goal*: Characterize biological mechanisms, signatures, or biomarkers derived from or attributable to smoking.

Data structure and notation

- Structural causal model (SCM) (Pearl 2000) describing data-generating process for a single unit O:

$$L = f_L(U_L); A = f_A(L, U_A); Y = f_Y(A, L, U_Y).$$

- f_L, f_A, f_Y are unknown but deterministic functions; U_L, U_A, U_Y are exogenous (unobserved) random errors.
- $Y = (Y_b : b = 1, \ldots B)$ is a vector of biomarker outcomes (e.g., $B = 450,000$ CpG sites).
- Temporal ordering between variables: L (sex-at-birth, age), A (smoking, benzene), Y_b (biomarker measurement for site b).
- Data on a single study unit $O = (L, A, Y)$, with $O \sim P_0 \in \mathscr{M}$, of which we observe n i.i.d. copies, $O_1, \ldots, O_n$.

Hypothetical interventions and causal inference

- *Static* interventions consider replacing f_A with an assigned value $a \in \mathscr{A}$ deterministically: "What if everyone smoked?"
- Generates counterfactual RV $Y(a) = (Y_b(a), b : 1, \ldots B)$: the expression of the B biomarkers if A had been set to a.
- Viewed as *potential outcomes* (POs) (Rubin 2005), $Y_b(1)$ when setting $A = 1$ and $Y_b(0)$ when setting $A = 0$.
- Note $Y_b = AY_b(1) + (1 - A)Y_b(0)$ — partial visibility is the *fundamental problem of causal inference* (Holland 1986).
- Causal inference yields interpretable, scientifically well-aligned estimands, e.g., the average treatment effect (ATE).

A familiar workhorse: the linear model

- The linear model is *flexible* — linearity in parameters!
- Flexible: transformations (X_j^2), interactions ($X_j X_k$).
- For biomarker Y_b, fit *working* linear model, $\mathbb{E}_0[Y_b \mid \mathbf{X}] = \mathbf{X}\beta$; if $X_1 \equiv A$ is the exposure, then β_1 measures its impact on Y.
- Under this working model, β_1 is a *conditional* effect measure, whose interpretation depends on $\mathbf{X} \setminus X_1$, and which coincides with the ATE only under randomization.
- Test the contrast of interest with a standard t-test:

$$t_b = \frac{\hat{\beta}_b - \beta_{b,H_0}}{\hat{\sigma}_b}$$

Variance moderation to the rescue?!

- When sample size is small, σ_b^2 may be so small (by chance) that even small effect sizes $(\hat{\beta}_b - \beta_{b,H_0})$ yield large t_b.
- False positives! Many biomarkers flagged relevant despite small effect size, only since variance is even smaller still.
- Can we do better? A **moderated** t-test (Smyth 2004):

$$\tilde{t}_b = \frac{\hat{\beta}_b - \beta_{b,H_0}}{\tilde{\sigma}_b} \quad \text{where} \quad \tilde{\sigma}_b^2 = \frac{\sigma_b^2 d_b + \sigma_0^2 d_0}{d_b + d_0}$$

- Helps reduce erroneously large t_b by "averaging out" low variance across each of the many biomarkers.

Variable importance measures as target parameters!

- The statistical functional identifying the ATE ($\psi_{b,0}$) may be used as an estimand to assess variable importance:

$$\theta^{\text{ATE}} = \mathbb{E}[Y_b(1) - Y_b(0)]$$
$$\psi_{b,0} := \Psi_b(P_0) = \mathbb{E}_{L,0}[\mathbb{E}_0[Y_b \mid A = 1, L] - \mathbb{E}_0[Y_b \mid A = 0, L]]$$

- $\psi_{b,0}$ is a mapping, $\Psi_b(P_0)$, that depends on the underlying true (but unknown) distribution $P_0 \in \mathscr{M}$ — model-agnostic!
- The causal parameter θ^{ATE} is identified by the statistical functional $\psi_{b,0}$ under some assumptions (no unmeasured confounding, positivity), i.e., $\theta^{\text{ATE}} \equiv \psi_{b,0}$.

Locally efficient estimation

- An estimator $\hat{\psi}_b$ is asymptotically linear if it admits the form

$$\hat{\psi}_b - \psi_{b,0} = \frac{1}{n}\sum_{i=1}^{n} D_b(O_i; P_0) + o_P\left(\frac{1}{\sqrt{n}}\right),$$

  where $D_b(O; P_0)$ is the efficient influence function (wrt $\mathscr{M}$), whose asymptotic variance at P_0 is the *efficiency bound*.

- $D_b(O; P_0)$ helpful to construct efficient estimators. For ATE,

$$D_b(O_i; P_0) = \left[\frac{2A_i - 1}{g_0(L_i)}\right](Y_{b,i} - \overline{Q}_{0,b}(A_i, L_i)) + [\overline{Q}_{0,b}(1, L_i) - \overline{Q}_{0,b}(0, L_i)] - \psi_{b,0},$$

  where $g_0(L) = \mathbb{P}_0(A = 1 \mid L)$ is the propensity score and $\overline{Q}_{0,b}(A, L) = \mathbb{E}_0[Y_b \mid A, L]$ the conditional outcome mean.

Constructing locally efficient estimators

- Examining $D_b(O; P_0)$, we know we must estimate $g_0(L)$ and $\overline{Q}_{0,b}(A, L)$, but how to do this?
- No need to try to exactly specify functional forms or assume we know the underlying true data-generating distribution P_0.
- Machine learning of $g_0(L)$ and $\overline{Q}_{0,b}(A, L)$, e.g., by ensembling, stacking (van der Laan et al. 2007, Breiman 1996).
- One-step estimator (Bickel et al. 1993) uses "debiasing" based on an additive correction: $\hat{\psi}_b^+ = \hat{\psi}_b + n^{-1} \sum_{i=1}^n \hat{D}_b(O_i)$.
- A valid variance estimator: $\hat{\mathbb{V}}(\hat{\psi}_b^+) = n^{-1} \sum_{i=1}^n \hat{D}_b^2(O_i)$, but its small-sample behavior may be erratic (asymptotically valid).

Moderated test statistics with efficient influence functions

- Moderated t-statistic of Smyth (2004) can be paired with locally efficient estimators:

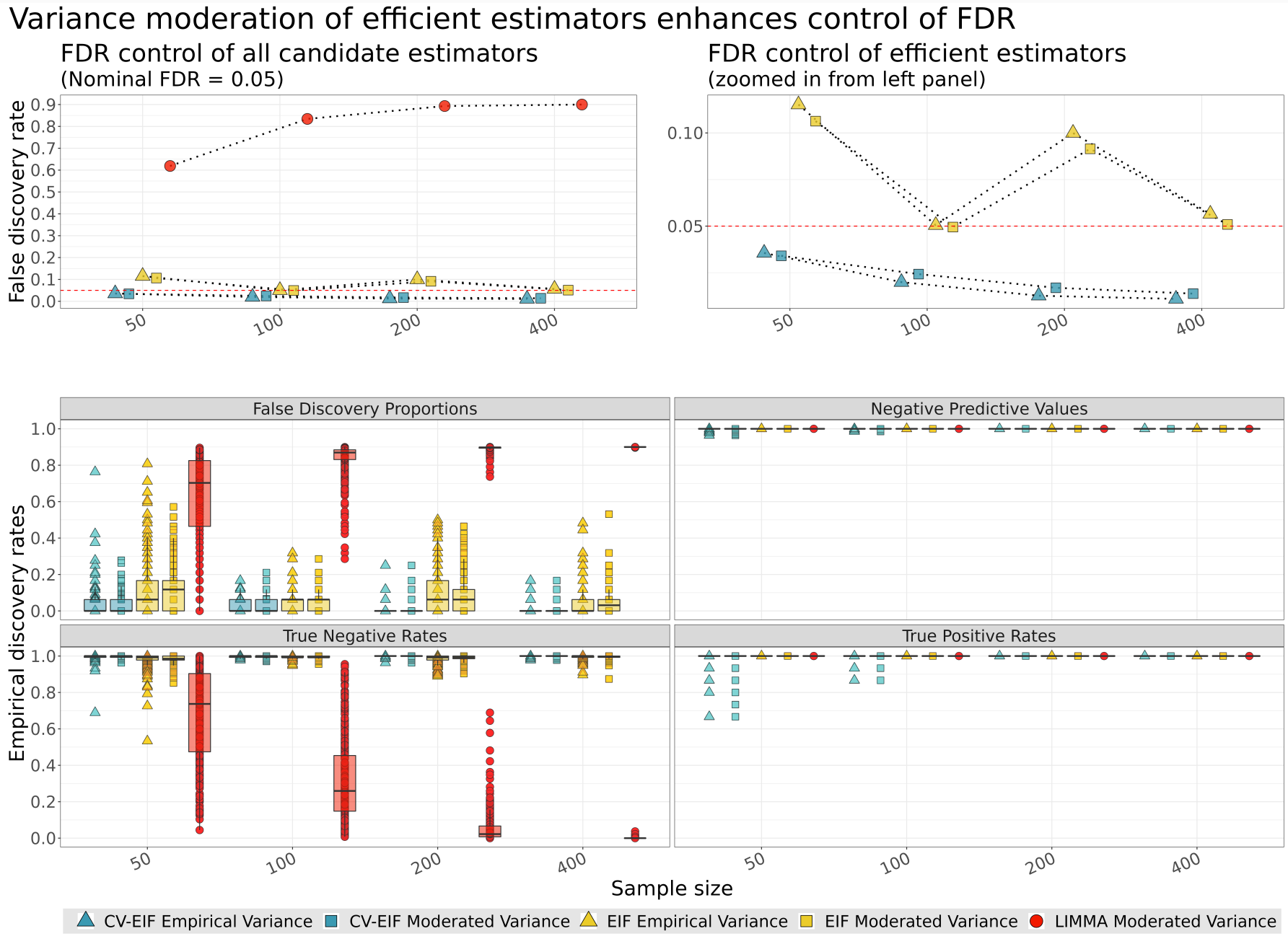
$$\tilde{t}_b = \frac{\hat{\psi}_b^+ - \cancel{\psi_{b,0}}^{H_0:\ \psi_{b,0}=0}}{\tilde{\sigma}_b},$$

  where the *moderated* (influence function) variance is

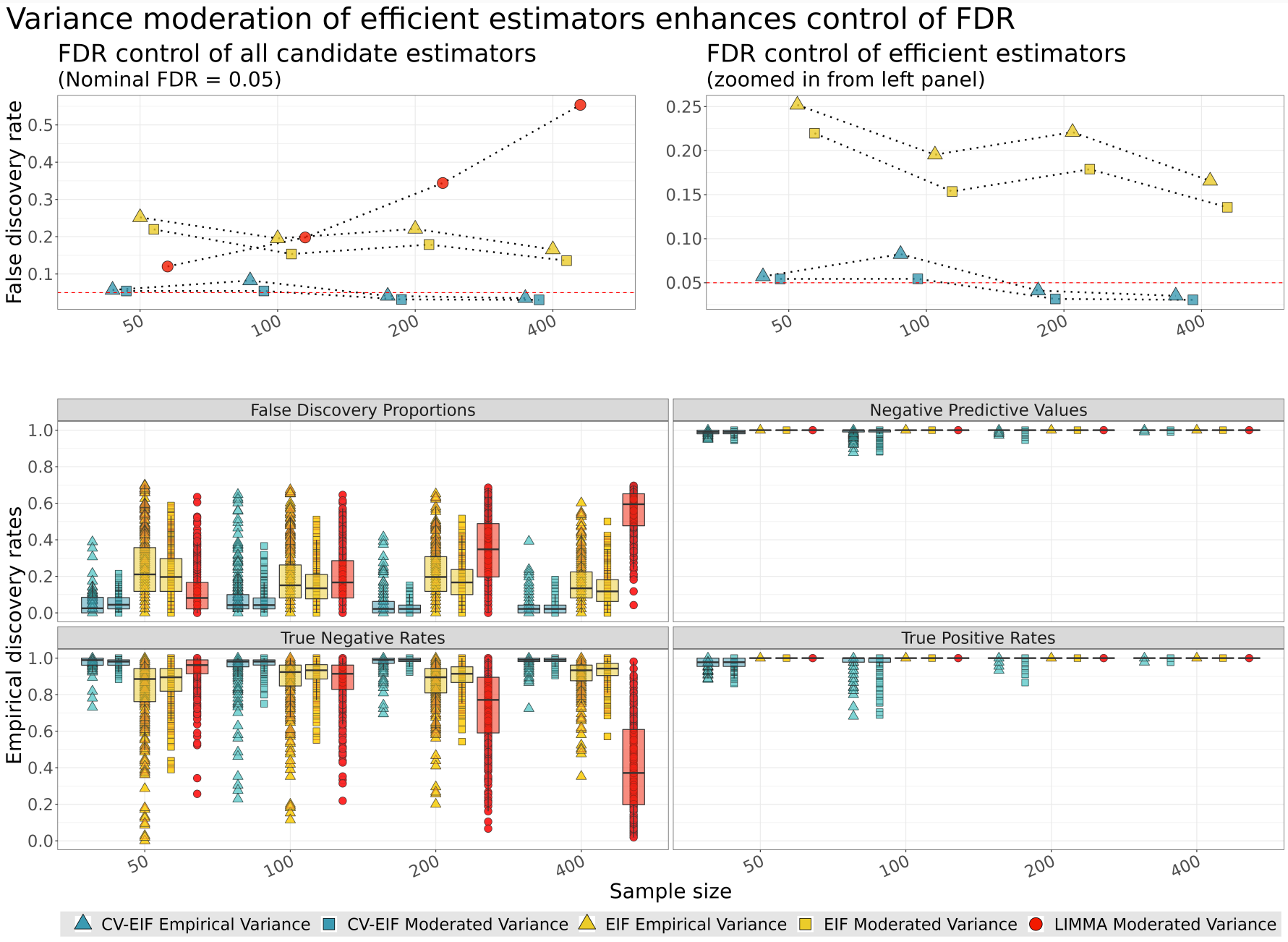
$$\tilde{\sigma}_b^2 = \frac{\hat{\sigma}_b^2 d_b + \hat{\sigma}_0^2 d_0}{d_b + d_0}$$

- Preserves robust variance estimator while adding stability by "averaging out" potentially erratic variance across biomarkers.
- Avoid model misspecification while stabilizing inference.

Numerical study: Rare effect (10%) without positivity issues



Numerical study: Common effect (30%) with positivity issues



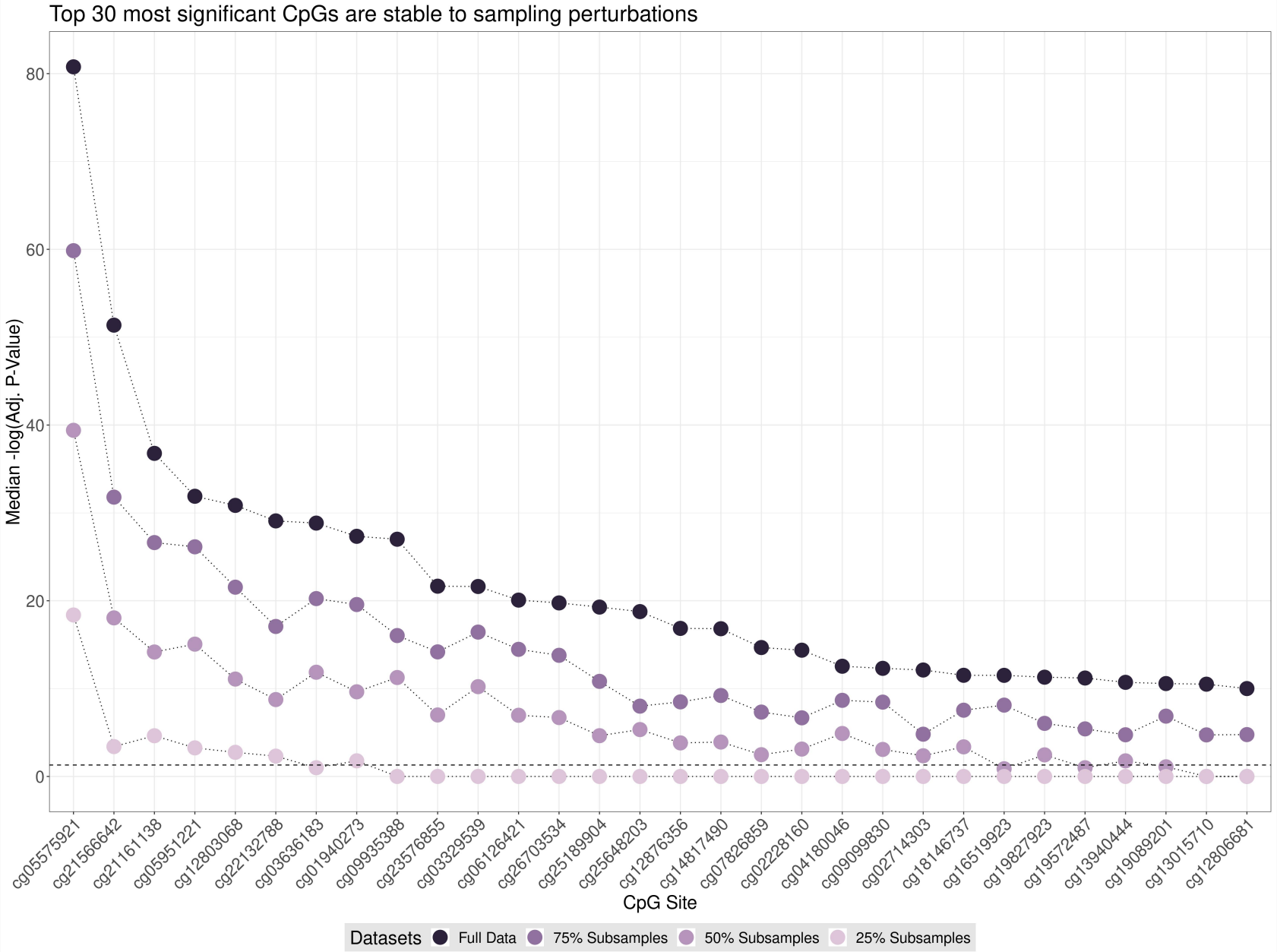
Differential expression analysis procedure

1. Apply a filtering procedure to reduce the set of candidate biomarkers (Tuglus and van der Laan 2009) *optionally*.
2. For each biomarker, generate an efficient estimate $\hat{\psi}_b$ of $\psi_{0,b}$ with EIF $\hat{D}_b(O_i)$ after estimating nuisances $(g_0, \overline{Q}_{0,b})$.
3. Apply variance moderation across the EIF estimates, yielding *moderated* $\tilde{\sigma}_b^2$, to be used for stabilized hypothesis testing.
4. Hypothesis testing from moderated test statistics can be made "optimistic" (MVN) or conservative (logistic).
5. Apply a multiple testing correction for accurate simultaneous inference across all B biomarkers, e.g., by controlling the False Discovery Rate (Benjamini and Hochberg 1995).

Applying the differential expression procedure

1. Filtered set of 450,000 CpG sites measured down to 2537 CpG sites by covariate adjustment in a working linear model.
2. For each candidate CpG (among 2537), estimate $(g_0, \overline{Q}_{0,b})$ by super (ensemble) learning, then construct efficient estimate $\hat{\psi}_b$ of ATE from the EIF $\hat{D}_b(g_n, \overline{Q}_{n,b})$.
3. Variance moderation based on estimated EIF, yielding *moderated* $\tilde{\sigma}_b^2$, for stabilized hypothesis testing across CpGs.
4. Applied Holm's procedure to control the family-wise error rate (FWER), tagging 1173 CpG sites as differentially methylated.
5. Top CpG sites (cg05575921) located in AHRR gene, previously identified in at least 30 epigenome-wide association studies on impact of smoking exposure on blood and lung tissues.

Ranking differentially methylated CpGs



Open-source software: `R/biotmle`

- `R` package for differential expression or methylation analysis based on *model-agnostic*, efficient estimators of the ATE.
- Incorporates machine learning and allows cross-validation.
- Statistical inference based on semi-parametric efficiency theory *and* variance moderation.
- Where can you find it?
  - `https://github.com/nhejazi/biotmle`
  - `https://bioconductor.org/packages/biotmle`

Review

1. Computational biology research produces complex data — why should statistical inference be tied to challenging-to-verify modeling assumptions?
2. *Model misspecification* seriously undermines scientific value of common, parametric statistical modeling approaches.
3. Semi-parametric inference allows the construction of robust estimators that readily bring machine learning into the process.
4. Variance moderation strengthens hypothesis testing strategies, reducing false positives and preserving power under instability.

References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300.

Bickel, P. J., Klaassen, C. A., Ritov, Y., and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press.

Breiman, L. (1996). Stacked regressions. *Machine learning*, 24(1):49–64.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960.

Pearl, J. (2000). *Causality*. Cambridge university press.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331.

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1):1–25.

Tuglus, C. and van der Laan, M. J. (2009). Modified FDR controlling procedure for multi-stage analyses. *Statistical Applications in Genetics and Molecular Biology*, 8(1).

van der Laan, M. J., Polley, E. C., and Hubbard, A. E. (2007). Super Learner. *Statistical applications in genetics and molecular biology*, 6(1).

Thank you

https://nimahejazi.org

https://github.com/nhejazi

https://twitter.com/nshejazi

https://doi.org/10.1177/0962280221146313

https://arxiv.org/abs/1710.05451