Semiparametric Estimation with Robust Empirical Bayes Inference and Supervised Clustering in High-Dimensional Biological Exposure Studies

NIMA S. HEJAZI, MARK J. VAN DER LAAN, MARTYN T. SMITH & ALAN E. HUBBARD

School of Public Health
UNIVERSITY OF CALIFORNIA, BERKELEY

OVERVIEW & MOTIVATIONS

- A general approach for deriving stable variance estimates for data-adaptive semiparametric estimators is introduced.
- Variance moderation uniformly improves variance estimates, with a negligible effect asymptotically.
  - curbing the error rate of tests relative to classical approaches
  - and facilitating *supervised clustering* from derived association profiles.
- Illustrate how the proposal applies for any asymptotically linear estimator through the lens of targeted maximum likelihood.
- The `biotmle` R package [1] implements the variance moderation procedure, leveraging state-of-the-art machine learning.

DATA: BENZENE BIOMARKERS

- There is a pressing need for model-free, technology-agnostic statistical methods for analyzing multiple kinds of exposome data.
- We consider data generated by a study of occupational exposure, using the *Illumina Human Ref-8 BeadChips* platform.
- Data on baseline confounders and exposure collected for $n = 125$ subjects/participants, with $22,000$+ gene expression measures.
- Covariates (W): age, sex, smoking status.
- Exposure (A): degree of Benzene exposure (none, $<$ 1ppm, $>$ 5ppm).
- Outcome ($Y = (Y_b : b = 1, \ldots, B)$): gene expression measures vector.

METHODOLOGY I: SEMIPARAMETRIC VARIANCE MODERATION

- Let observed data $O = (W, A, Y) \sim P_0 \in \mathcal{M}$, where W represents potential baseline confounders, A the exposure of interest, and $Y = (Y_b, b = 1, \ldots, B)$ a vector of potential biomarkers.
- We consider, as an example, the *average treatment effect* (ATE), as the causal parameter of interest, which is identified by the observed data parameter:

$$\Psi_b(P_0) = \mathbb{E}_W[Q_0^b(A = 1, W) - Q_0^b(A = 0, W)], \tag{1}$$

  where $Q_0^b(A, W) \equiv \mathbb{E}_{P_0}(Y_b \mid A, W)$ and may be estimated via *ensemble machine learning* [4, 5, 6].
- Like the estimator $\hat{\beta}$ in a linear model $m(A, W \mid \beta)$, $\Psi_b(P_n)$ is *asymptotically linear* (for Ψ_b) [7]:

$$\sqrt{n}(\Psi_b(P_n) - \Psi_b(P_0)) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} D_b(O_i) + o_p(1). \tag{2}$$

- Ψ_b has efficient influence function (EIF), relative to the nonparametric model $\mathcal{M}$:

$$D_b(P_0)(o) = \left(\frac{I(a = 1)}{g(1 \mid w)} - \frac{I(a = 0)}{g(0 \mid w)} \right) \cdot \left[y_b - Q_0^b(a, w) \right] + \left(Q_0^b(1, w) - Q_0^b(0, w) - \Psi_b(P_0)(o) \right). \tag{3}$$

- A moderated test statistic [8, 2] may be constructed for use with asymptotically linear estimators:

$$\widetilde{t}_b = \frac{\sqrt{n}(\Psi_b(P_n) - \psi_{\text{null}})}{\widetilde{S}_{b,n}^2} \quad \text{where} \quad \widetilde{S}_{b,n}^2 = \frac{d_0 S_0^2 + d_b S_b^2(D_{b,n})}{d_0 + d_b}, \tag{4}$$

  $\{S_b^2, d_b\}$: var. EIF and df for b^{th} biomarker; $\{S_0^2, d_0\}$: var. EIF and df for other $(B - 1)$ biomarkers.

METHODOLOGY II: SUPERVISED DISTANCE MATRICES

- Let $\phi : (W, A, Y) \mapsto D_b(P_0)(O)$ be the EIF transformation, where $D_b(P_0)(O_i)$ is the contribution of subject i to the estimate of the biomarker-specific target parameter $\Psi_{b,n}$.
- $Z = \phi(W, A, Y)$ is then a $B \times N$ matrix, where each entry (b, i) may be interpreted as the degree to which subject i deviates from the target parameter $\Psi_{b,n}$ [2], and is thus an *association profile*.
- A *supervised distance matrix* [3] may be constructed by applying an appropriate distance metric of choice (e.g., Euclidean, correlation) to the transformed values Z.
- $\widetilde{T}(Z)$, the resultant $B \times B$ empirical distance matrix, encodes the dissimilarity between pairs of biomarker association profiles.
- *Supervised clustering* may be performed by applying standard unsupervised clustering algorithms to the matrix $\widetilde{T}$, thereby finding groups of biomarkers that share an association profile w.r.t. Ψ.
- In the case of the average treatment effect, a supervised cluster in $\widetilde{T}$ of biomarkers is a group whose causal differential expression profiles varies similarly with the treatment $A \in \{0, 1\}$.

NUMERICAL STUDY & RESULTS

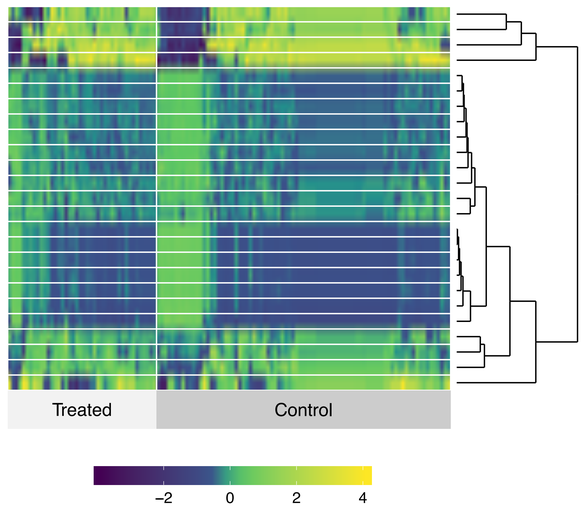

Figure 1: *Supervised heatmap* of the top 25 biomarkers visualizes groups with a shared exposure response.

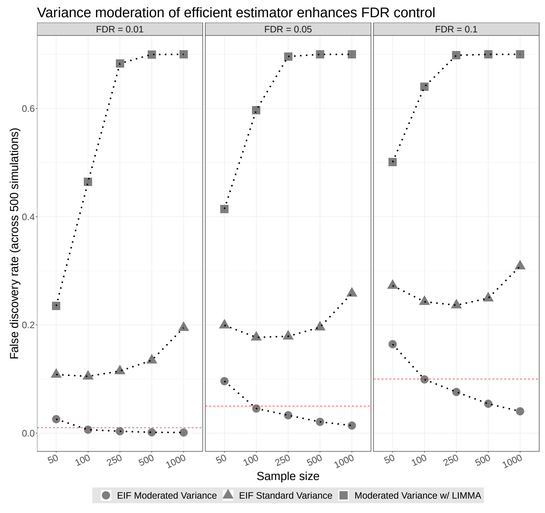

Figure 2: Enhanced control of the False Discovery Rate (FDR) with variance-moderated efficient estimator.

REFERENCES

[1] N. S. Hejazi, W. Cai, and A. E. Hubbard, "biotmle: Targeted learning for biomarker discovery," *The Journal of Open Source Software*, vol. 2, no. 15, July 2017.
[2] N. S. Hejazi, S. Kherad-Pajouh, M. J. van der Laan, and A. E. Hubbard, "Supervised variance moderation of locally efficient estimators in high-dimensional biology," 2018+.
[3] K. S. Pollard and M. J. van der Laan, "Supervised distance matrices," *Statistical applications in genetics and molecular biology*, vol. 7, no. 1, 2008.
[4] M. J. van der Laan, E. C. Polley, and A. E. Hubbard, "Super Learner," *Statistical applications in genetics and molecular biology*, vol. 6, no. 1, 2007.
[5] L. Breiman, "Stacked regressions," *Machine learning*, vol. 24, no. 1, pp. 49–64, 1996.
[6] D. H. Wolpert, "Stacked generalization," *Neural networks*, vol. 5, no. 2, pp. 241–259, 1992.
[7] M. J. van der Laan and S. Rose, *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media, 2011.
[8] G. K. Smyth, "Linear models and empirical bayes methods for assessing differential expression in microarray experiments," *Statistical Applications in Genetics and Molecular Biology*, vol. 3, no. 1, pp. 1–25, 2004.

CONTACT INFORMATION

- **N.S. Hejazi**: `nhejazi@berkeley.edu`; **M.J. van der Laan**: `laan@berkeley.edu`; **M.T. Smith**: `martynts@berkeley.edu`; **A.E. Hubbard**: `hubbard@berkeley.edu`
- `https://bioconductor.org/packages/biotmle`
- `https://arxiv.org/abs/1710.05451`