



# Variance Moderation of Locally Efficient Estimators and Supervised Clustering with Applications in High-Dimensional Biology

NIMA S. HEJAZI, MARK J. VAN DER LAAN, & ALAN E. HUBBARD *Graduate Group in Biostatistics*



School of  
Public Health

UNIVERSITY OF CALIFORNIA, BERKELEY

## OVERVIEW & MOTIVATIONS

- A general approach for applying variance moderation to locally efficient estimators in nonparametric models is introduced.
- Variance moderation stabilizes small-sample properties of semiparametric-efficient estimators,
  - curbing the error rate of tests relative to classical approaches
  - and facilitating *supervised clustering* from derived association profiles.
- Focusing on a targeted maximum likelihood estimator (TMLE), we illustrate how the proposed approach generalizes readily to any asymptotically linear estimator.
- The `biotmle` R package [1] implements the inference and clustering methods, and leverages 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 a data from an occupational exposure study, generated by the *Illumina Human Ref-8 BeadChips* platform.
- Baseline (phenotypic) confounders and exposure status were collected for  $n = 125$  participants, alongside expression measures for  $\sim 22,000$  genes.
- Baseline covariates ( $W$ ): age, sex, and smoking status.
- Exposure ( $A$ ): degree of Benzene exposure (none,  $< 1\text{ppm}$ ,  $> 5\text{ppm}$ ).
- Outcome ( $Y = (Y_b : b = 1, \dots, B)$ ): vector of gene expression measures, after full quantile normalization.

## 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  $Z$ .
- $\tilde{T}(Z)$ , the resultant  $B \times B$  empirical distance matrix, encodes the dissimilarity between pairs of biomarker association profiles.
- When  $\tilde{T}(b, b')$  is small, the biomarkers  $b$  and  $b'$  have similar contributions to the target parameter  $\Psi$ , across the  $n$  subjects.
- Supervised clustering* may be performed by applying standard unsupervised clustering algorithms to the matrix  $\tilde{T}$ , thereby finding groups of biomarkers that share an association profile w.r.t.  $\Psi$ .
- In the case of the average treatment effect, a supervised cluster in  $\tilde{T}$  of biomarkers is a group whose causal differential expression profiles varies similarly with the treatment  $A \in \{0, 1\}$ .

## METHODOLOGY I: VARIANCE MODERATION & LOCAL EFFICIENCY

- 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  $Y = (Y_b, b = 1, \dots, B)$  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:

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

where  $Q_0^b(A, W) \equiv \mathbb{E}_{P_0}(Y_b | A, W)$  and may be estimated via *ensemble machine learning* [4, 5, 6].

- Like the estimator  $\hat{\beta}$  in a linear model  $m(A, W | \beta)$ ,  $\Psi_b(P_n)$  is **asymptotically linear** (for  $\Psi_b$ ) [7]:

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

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

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

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

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

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

## RESULTS

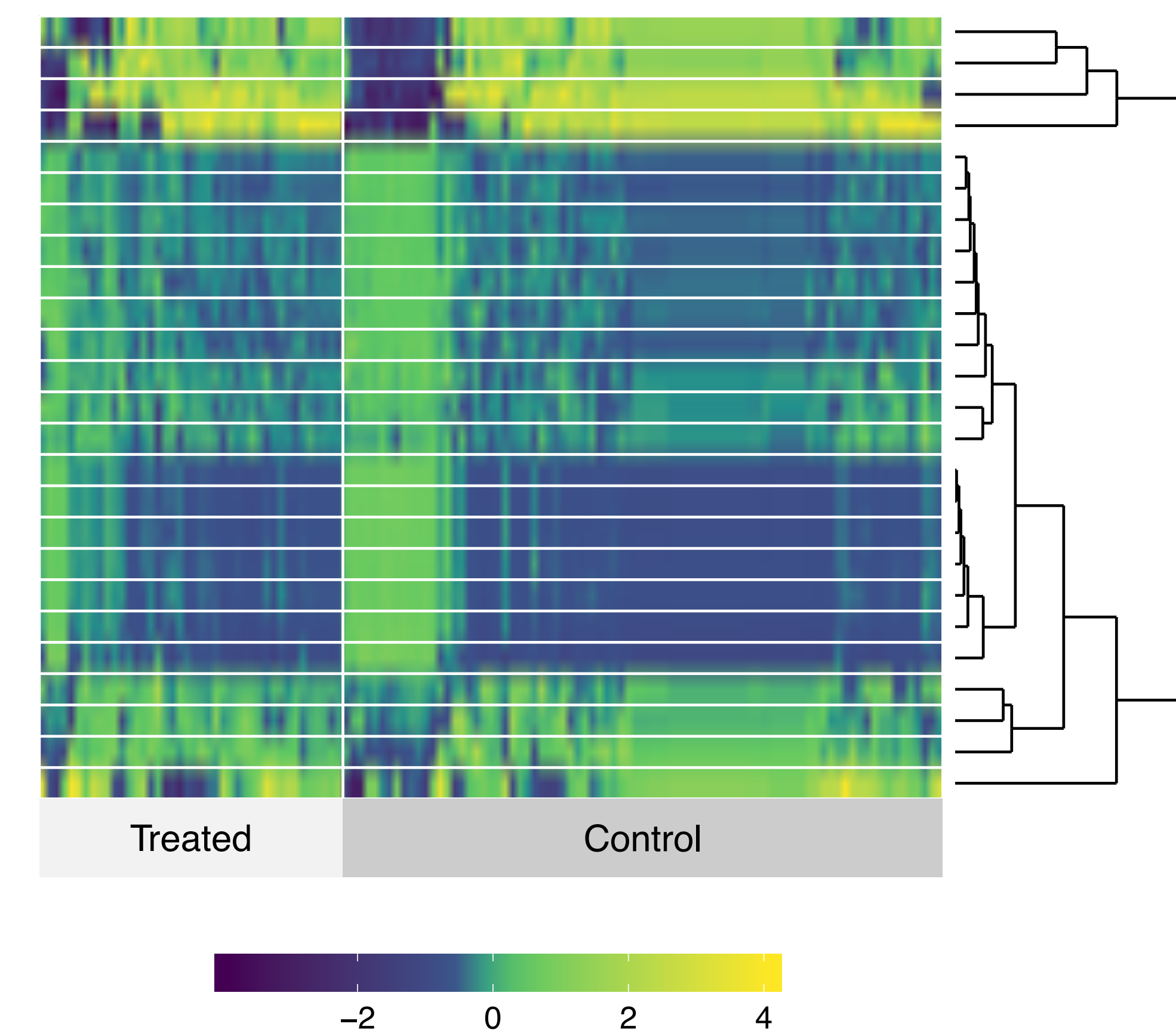


Figure 1: Supervised heatmap of top 25 biomarkers.

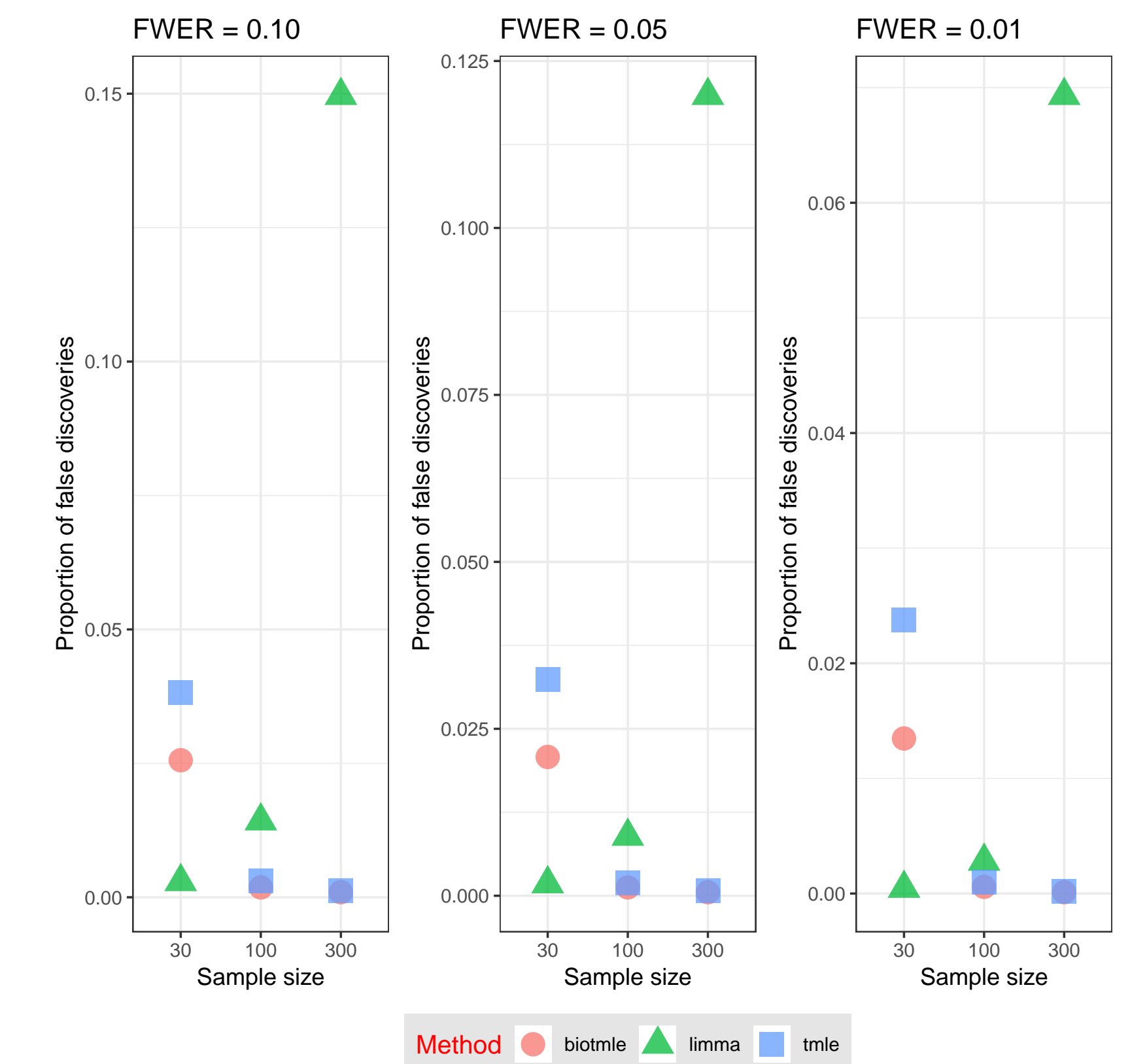


Figure 2: Enhanced error rate control with variance-moderated efficient estimator (*n.b.*, y-axis scales differ).

## 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**, Ph.D. candidate, Group in Biostatistics, NHEJAZI@BERKELEY.EDU
- M.J. van der Laan**, Professor of Biostatistics & Statistics, LAAN@BERKELEY.EDU
- A.E. Hubbard**: Professor of Biostatistics, HUBBARD@BERKELEY.EDU