# 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.
- While relevant estimation procedures are usually data-intensive, variance moderation stabilizes their small-sample properties, unlocking their use for differential expression analyses.
- Focusing on a targeted maximum likelihood estimator (TMLE) of the average treatment effect (ATE), we illustrate how the approach generalizes to asymptotically linear estimator.
- Supervised clustering is discussed in the context of influence functions, ...
- The `biotmle` R package [1] implements the inference and clustering approaches while leveraging state-of-the-art machine learning in the procedure.

## DATA: BENZENE BIOMARKERS

- With the growing number of methods for measuring biomarkers there arises a need for methodologies able to simultaneously analyze multiple kinds of exposome data.
- Data was generated by the **Illumina Human Ref-8 BeadChips** platform.
- There were 125 subjects, for which background characteristics and expression measures for $\sim 22,000$ genes were obtained.
- Covariates in W were age, sex, and smoking status; all were discretized.
- The treatment (A) is degree of Benzene exposure: none, <1ppm, and >5ppm.
- The outcome (Y) is a vector of gene expression measures, normalized by median.

## 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 $Y = (Y_b, b = 1, \ldots, B)$ 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 $\Psi_b$) [7]:

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

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

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

- A moderated test statistic [8, 9, 2] may be constructed for use wiht asymptotically linear estimators:

$$\tilde{t}_b = \frac{\sqrt{n}(\Psi_b(P_n) - \psi_0)}{S_b(D_{b,n})} \quad \text{where} \quad \tilde{S}_{b,n}^2 = \frac{d_0 S_0^2 + d_b S_b^2(D_{b,n})}{d_0 + d_b}, \tag{4}$$

  $\{S_b, d_b\}$: var. EIF and df for $b^{th}$ biomarker; $\{S_0, d_0\}$: 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 a 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.
- When $\widetilde{T}(b, b') = g(Z_b, Z_{b'})$ is small, biomarkers $b$ and $b'$ have similar associations between expression and the the target parameter $\Psi$, across the $n$ subjects.
- *Supervised clustering* may be performed by applying standard clustering algorithms to the matrix $\widetilde{T}$, 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 $\widetilde{T}$ of biomarkers is a group whose causal differential expression profiles varies similarly with the treatment $A \in \{0, 1\}$.

## RESULTS & DISCUSSION

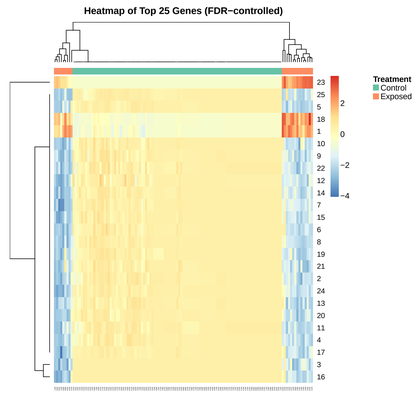


**Figure 1:** A *supervised heatmap* of biomarker association profiles with the ATE.

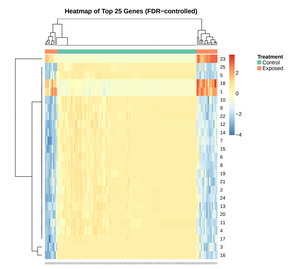


**Figure 2:** Improved error rate control of the variance moderated TMLE for the ATE.

- Concluding comment...
- Concluding comment...

## 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.
[9] ——, "Limma: linear models for microarray data," in *Bioinformatics and computational biology solutions using R and Bioconductor*. Springer, 2005, pp. 397–420.

## 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