Variance Moderation of Locally Efficient Estimators in High-Dimensional Biology (and the `biotmle` R package)

Nima Hejazi

25 June 2019

Graduate Group in Biostatistics, and
Center for Computational Biology,
University of California, Berkeley

nshejazi
nhejazi
nimahejazi.org
bit.ly/2019_bioc_modtmle
Joint work with Alan Hubbard and Mark van der Laan

Preview

1. Model misspecification seriously undermines the utility of many common statistical modeling approaches.
2. Non/semi-parametric theory allows the construction of robust estimators that accommodate the use of machine learning.
3. Moderated variance estimators augment hypothesis testing strategies to reduce false positives in small-sample settings.
4. The moderation approach pioneered by the `limma` R package may easily be extended to non/semi-parametric estimators.

We'll go over this summary again at the end of the talk. Hopefully, it will all make more sense then.

Data structure and notation

- Consider a nonparametric structural equation model (NPSEM) to describe observed data O (Pearl 2000):

$$W = f_W(U_W); A = f_A(W, U_A); Y = f_Y(A, W, U_Y).$$

- f_W, f_A, f_Y are flexible but deterministic functions; U_W, U_A, U_Y are exogenous RVs specifying unobserved errors.
- Data on a single unit $O = (W, A, Y)$, where $O \sim P_0 \in \mathscr{M}$. Observe $O_1, \ldots, O_n$, i.e., n i.i.d. copies of O.
- $Y = (Y_b : b = 1, \ldots B)$ is a vector of biomarker outcomes.

Interventions and causal inference

- NPSEM: time-ordering and counterfactual RV distributions.
- *Static intervention* replaces f_A with an assigned value $A = a$.
- Generates a counterfactual RV $Y(a) = (Y_b^a, b : 1, \ldots B)$: expression of B biomarkers when A is set to a.
- Thus, we have potential outcomes $Y_b(1)$ (for $\text{do}(A = 1)$) and $Y_b(0)$ (for $\text{do}(A = 0)$) (Rubin 2005).
- We've now just about defined a canonical causal parameter, the ATE: $\psi_b = \mathbb{E}_W[Y_b(1) - Y_b(0)]$ (Pearl 2000).

Statistical parameter may be viewed as a simple adjusted difference in means even when identifiability conditions appear unsatisfiable.

A familiar workhorse: the linear model

- The linear model is *semiparametric* — linear in parameters!
- Flexible: accommodate transformations, interactions, etc.
- For each biomarker ($b = 1, \ldots, B$), fit a *working* linear model.
- Under the working model, the parameter β_b captures the ATE, allowing construction of estimators and inference.
- Test the coefficent of interest using a standard t-test:

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

There's nothing particularly wrong with this approach. It's exactly what we would come up with after a first-year statistics course. In practice, there are many issues: (1) we are forced to specify a functional form, the linear model; (2) we end up with unstable variance estimates that sharply increase the number of false positives detected, even after multiple testing corrections. In practice, the incredible flexibility of the linear mode is rarely taken advantage of — scientific guidance is usually lacking to justify the fitting of richer models.

Variance moderation robustifies inference

- When the sample size is small, s_b^2 may be so small that even small effects $(\hat{\beta}_b - \beta_{b,H_0})$ lead to large t_b.
- This results in false positives. Smyth proposes we get around this by an empirical Bayes shrinkage of the s_b^2.
- Test the coefficent of interest with a **moderated** t-test:

$$\tilde{t}_b = \frac{\hat{\beta}_b - \beta_{b,H_0}}{\tilde{s}_b} \quad \text{where} \quad \tilde{s}_b^2 = \frac{s_b^2 d_b + s_0^2 d_0}{d_b + d_0}$$

- Eliminates large t-statistics arising merely from very small s_b.

The substantive contribution here is the use of an empirical Bayes method to shrink the standard deviation across all of the biomarkers such that we obtain a larger (but accurate) estimate that reduces the number of test statistics that are marked as significant by low s_b^2 estimates alone.

Note that this is **not** the exact formulation of the moderated t-statistic as given by Smyth (his derivation assumes a hierarchical model; see original paper if interested). This formulation does a good enough job to help us see the bigger picture.

Variable importance measures as target parameters

- If the working model is incorrect, β_b does not correspond to the ATE — thus leading to biased results.
- The statistical functional identifying the ATE may be used as a variable importance measure (VIM):

$$\Psi_b(P_0) = \mathbb{E}_{W,0}[\mathbb{E}_0[Y_b \mid A = 1, W] - \mathbb{E}_0[Y_b \mid A = 0, W]]$$

- One-step and targeted minimum loss estimation build efficient, doubly robust estimators $\Psi_b(P_n^\star)$ of $\Psi_b(P_0)$.

By allowing scientific questions to inform the parameters that we choose to estimate, we can do a better job of actually answering the questions of interest to our collaborators. Further, we abandon the need to specify the functional relationship between our outcome and covariates; moreover, we can now make use of advances in machine learning.

Robust and locally efficient estimation

- Asymptotic linearity:

$$\Psi_b(P_n^\star) - \Psi_b(P_0) = \frac{1}{n}\sum_{i=1}^{n} D_b(O_i) + o_P\left(\frac{1}{\sqrt{n}}\right)$$

- The influence function D_b for the ATE takes the form

$$\begin{aligned} D_b(O_i) = &\left[\frac{2A_i - 1}{g_0(A_i \mid W_i)}\right](Y_{b,i} - Q_{0,b}(A_i, W_i)) \\ &+ Q_{0,b}(1, W_i) - Q_{0,b}(0, W_i) - \Psi_b, \end{aligned}$$

where $g_0(A \mid W) = \mathbb{P}_0(A = 1 \mid W)$ is the treatment mechanism and $Q_{0,b}(A, W) = \mathbb{E}_0[Y_b \mid A, W]$ is the outcome model.

Natural use of machine learning methods for the estimation of both Q_0 and g_0. Focuses effort to achieve minimal bias and asymptotic semiparametric efficiency bound for the variance, but still get inference (with some assumptions).

Robust and locally efficient estimation

- wrt the data $O = (W, A, Y)$, $D_b(O)$ admits an orthogonal decomposition — i.e.,$D_b(O) = D_b^Y(O) + D_b^A(O) + D_b^W(O)$.
- Under randomization, $D_b^A(O) = 0$ and need not be estimated, though estimation improves overall efficiency (Tsiatis 2007).
- No need to specify functional forms or assume we know the model underlying the true data-generating distribution P_0.
- Machine learning to estimate nuisance functions $g_0(A \mid W)$ and $Q_{0,b}(A, W)$, e.g., via stacked regression or cross-validation selectors (Breiman 1996, van der Laan et al. 2007).

Natural use of machine learning methods for the estimation of both Q_0 and g_0. Focuses effort to achieve minimal bias and asymptotic semiparametric efficiency bound for the variance, but still get inference (with some assumptions).

Robust inference via the influence function

- Suppose we have estimated $g_0(A \mid W)$ and $Q_{0,b}(A, W)$ via ML, yielding an estimate $D_{n,b}(O)$ of $D_b(O)$, for $b = 1, \ldots, B$.
- Conservative variance estimator for $\Psi_b(P_n^\star)$ based on $D_{n,b}(O)$:

$$se_b = \sqrt{\frac{s^2(D_{n,b})}{n}} \quad \text{where} \quad s^2(D_{n,b}) = \frac{1}{n}\sum_{i=1}^{n}\left(D_{n,b}(O_i)\right)^2$$

- Under $H_0 : \Psi_b(P_0) = 0$ (no treatment effect), test statistic:

$$t_b = \frac{\Psi_b(P_n^\star)}{se_b}$$

Using the influence curve representation, we can obtain all of the standard objects of statistical interest, but for more interesting parameters.

Moderated statistics based on influence functions

- Moderated t-statistic of Smyth (2004) naturally extends to locally efficient estimators:

$$\tilde{t}_b = \frac{\Psi_b(P_n^{\star})}{\widetilde{se}_b}$$

  where the posterior estimate of influence function variance is

$$\tilde{s}_b^2 = \frac{s_b^2(D_{n,b})d_b + s_0^2 d_0}{d_b + d_0}$$

- Preserves robust variance estimator but adds stability that smoothens its small-sample behavior.

- Consider this is repeated for $b = 1, \ldots, B$ different biomarkers, so that one has, for each b:

$$\Psi_b(Q^*_{b,n}), S^2_b(IC_{b,n}),$$

estimate of variable importance and standard error for all B.

- Propose an existing joint-inferential procedure that can add some finite-sample robustness to an estimator that can be highly variable.

That's nice and all but where's the proof?

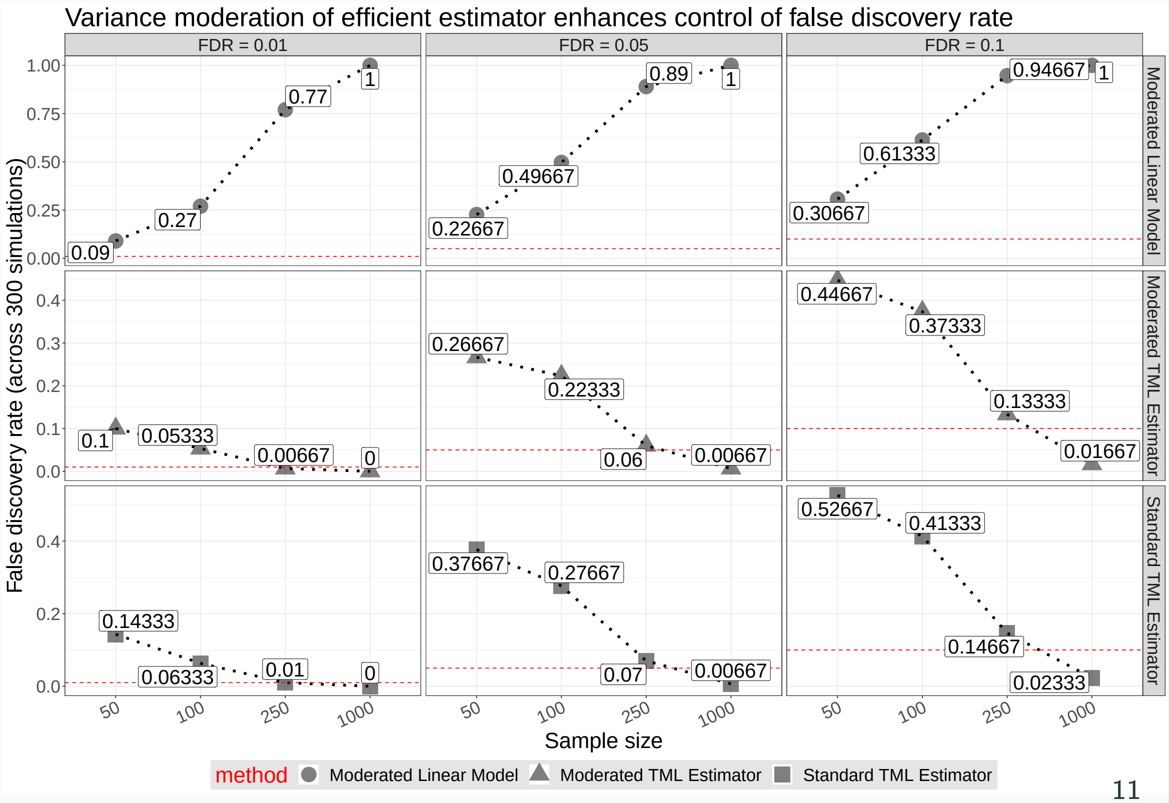
- Simulation study under two settings: (1) global null and (2) when half of probes respond to treatment.

$$\begin{aligned}
W_1 &\sim \text{Unif}(0,1); W_2 \sim \text{Unif}(0,1) \\
A &\sim \text{Bern}\left(\text{expit}(-1.2 - 2.5 \cdot W_1 + 3.5 \cdot W_2)\right) \\
Y_{\text{null}} &= 2 + 5 \cdot W_1 + 0.5 \cdot W_2 + W_1 \cdot W_2 + \varepsilon \\
Y_{\text{non-null}} &= 2 + 5 \cdot W_1 + 0.5 \cdot W_2 + W_1 \cdot W_2 + 5 \cdot A + \varepsilon,
\end{aligned}$$

- Data-adaptive estimation of relevant nuisance quantities.
- Compares TML estimator of ATE to working linear model, under moderated and standard variance estimates.

Essentially, we have the same concerns about using variable importance measures that we did about using the standard t-test — that is, non-robut estimates of the standard error of the estimator of the target parameter can cause erroneous identification of biomarkers (false positives). To reduce this, we can apply the same machinery that we did in the case of the standard t-test for our naive linear modeling approach.

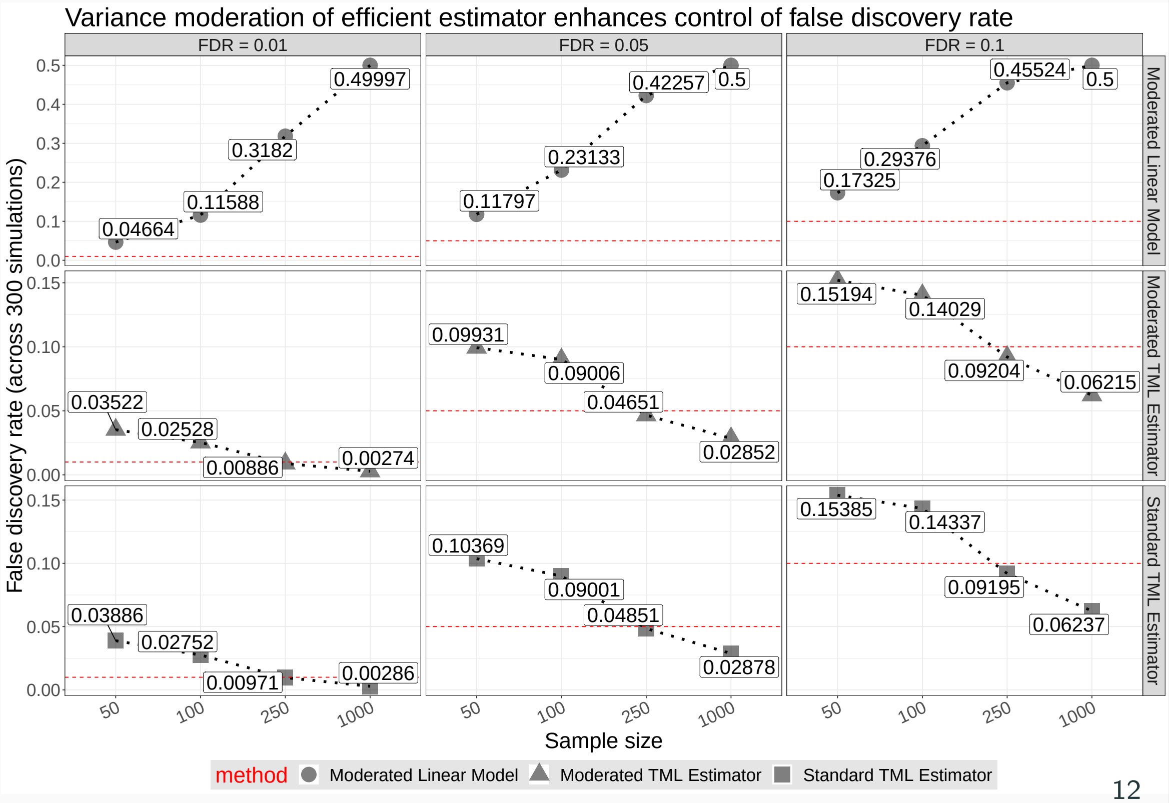
That's nice and all but where's the proof? Global null.



Variance moderation of efficient estimator enhances control of false discovery rate

- Control of the FDR using the Benjamini-Hochberg correction applied to the results of hypothesis tests based on `limma`, standard TML estimator without variance moderation, and TML estimator with variance moderation.
- TML estimators converges to the correct FDR asymptotically and achieves the nominal rate by $n = 250$ while the moderated linear model does not exhibit correct control of the FDR.

That’s nice and all but where’s the proof? Treatment effect.



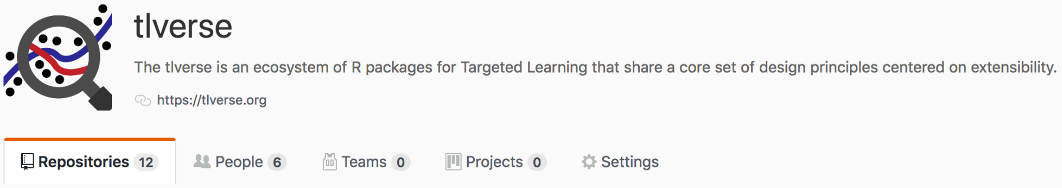
- Control of the FDR using the Benjamini-Hochberg correction applied to the results of hypothesis tests based on the moderated linear modeling approach of `limma`, the standard TML estimator without any variance moderation, and the TML estimator with variance moderation. Application of TML estimators converges to the correct FDR asymptotically and achieves the nominal rate by $n = 250$ while the moderated linear model does not exhibit correct control of the FDR.

Software implementation: `R/biotmle`

- `R` package for DE analysis based on TML estimators of the ATE that use machine learning for $g_0(A \mid W)$ and $Q_{0,b}(A, W)$.
- Statistical inference based on *moderated* variance estimator.
- Check out the package:
    - `https://github.com/nhejazi/biotmle`
    - `https://bioconductor.org/packages/biotmle`

Use it. File an issue. Help make it better!

The tlverse for Targeted Learning



tlverse

The tlverse is an ecosystem of R packages for Targeted Learning that share a core set of design principles centered on extensibility.

https://tlverse.org

Figure 1: `https://github.com/tlverse`

- An ecosystem of `R` packages for Targeted Learning, all sharing a core set of design principles centered on extensibility.
- *Draft phase* Targeted Learning handbook: `https://tlverse.org/tlverse-handbook`

Use it. File issues. Help make it better!

Review

1. Model misspecification seriously undermines the utility of many common statistical modeling approaches.
2. Non/semi-parametric theory allows the construction of robust estimators that accommodate the use of machine learning.
3. Moderated variance estimators augment hypothesis testing strategies to reduce false positives in small-sample settings.
4. The moderation approach pioneered by the `limma` `R` package may easily be extended to non/semi-parametric estimators.

It's always good to include a summary.

References

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

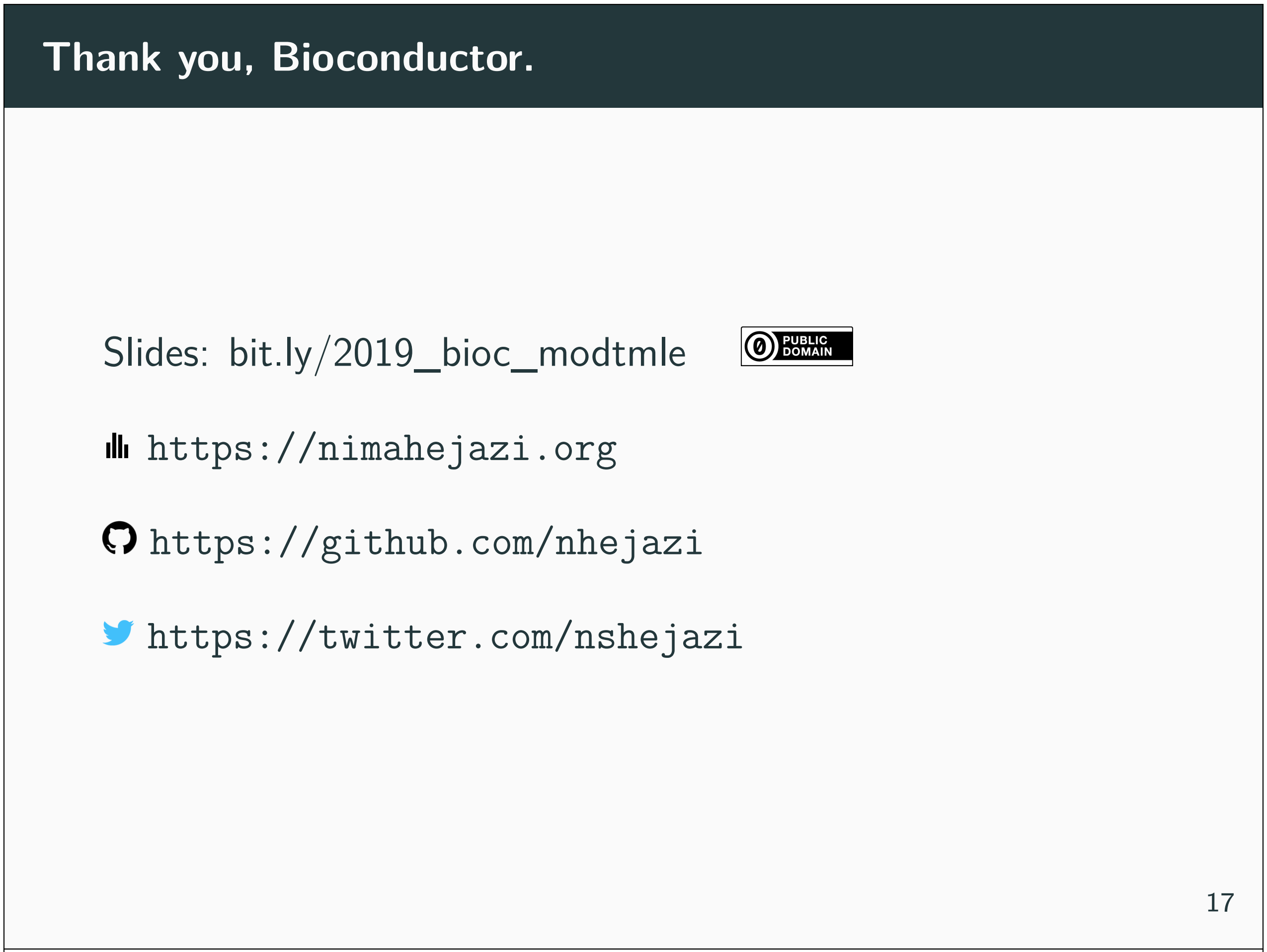
Pearl, J. (2000). *Causality*. Cambridge university press.

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.

Tsiatis, A. (2007). *Semiparametric theory and missing data*. Springer Science & Business Media.

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, Bioconductor.

Slides: bit.ly/2019_bioc_modtmle

https://nimahejazi.org

https://github.com/nhejazi

https://twitter.com/nshejazi

Here's where you can find me, as well as the slides for this talk.