Empirical Bayes Moderation of Asymptotically Linear Parameters

Nima Hejazi

Division of Biostatistics
University of California, Berkeley
stat.berkeley.edu/~nhejazi

nimahejazi.org
twitter/@nshejazi
github/nhejazi

slides: goo.gl/6ou8YR

These are slides for a talk given most recently at the Division of Biostatistics seminar series, at the University of California, Berkeley on 20 March 2017.

Source: `https://github.com/nhejazi/talk_biotmle`
Slides: `https://goo.gl/r3zsu6`
With notes: `https://goo.gl/6ou8YR`

Preview

1. Linear models are the standard approach for analyzing microarray and next-generation sequencing data (e.g., R package “limma”).
2. Moderated statistics help reduce false positives by using an empirical Bayes method to perform standard deviation shrinkage for test statistics.
3. *Beyond linear models:* we can assess evidence using parameters that are more scientifically interesting (e.g., ATE) by way of TMLE.
4. The approach of moderated statistics easily extends to the case of asymptotically linear parameters.

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

Motivation: Let's meet the data

- Observational study of the impact of occupational exposure (to benzene), with data collected on 125 subjects and roughly $22,000$ biomarkers.
- Biomarkers of interest are in the form of **miRNA**, assessed using the *Illumina Human Ref-8 BeadChips* platform.
- Occupational exposure to benzene reported as discrete values of interest (to epidemiologists): none, $<$ 1ppm, $>$ 5ppm.
- Background (phenotype-level) information available on each subject, including age, sex, smoking status.

This is not an atypical data set by modern standards in epidemiology, certainly not the standard for molecular biology. That is, sample sizes are usually much smaller in experiments examining biological processes.

Data analysis? Linear models!

- For each biomarker ($b = 1, \ldots, B$), fit a linear model:

$$\mathbb{E}[y_b] = X\beta_b$$

- Generally, we have a particular model coefficent in which we are interested (e.g., effect of benzene on biomarker expression).

- Controlling for baseline covariates, batch effects, and potential confounders happens by adding terms to the linear model.

- 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.

LIMMA: Linear Models for Microarray Data

- When the sample size is small, s_b^2 may be so small that small differences $(\hat{\beta}_b - \beta_{b,H_0})$ lead to large t_b.
- Uncertainty in the variance is an acute problem when the sample size is small.
- 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:

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

- Eliminates large t-statistics 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.

Beyond linear models

- It's not always desirable to specify a functional form: perhaps we can do better than linear models?
- Such models are a matter of convenience and not honest scientific practice: does $\hat{\beta}_b$ really answer our questions?
- We can do better by using parameters motivated by causal models (n.b., these will reduce to "variable importance measures" in our case).
- As long as the parameters we seek to estimate have asymptotically linear estimators, we can readily apply the approach of moderated statistics.

Linear models are convenient for communicating results — that is, all scientists are trained to understand them. This means they provide a basic way of easily communicating between statisticians and collaborators. That said, doesn't it seem a bit odd to use such elementary models to analyze complex biological sequencing data? We're using old statistical technology to analyze classes of data that have only recently become available.

Target parameters for complex questions

- Rather than being satisfied with $\hat{\beta}_b$ as an answer to our questions, let's consider a simple target parameter: the average treatment effect (ATE):

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

- No need to specify a functional form or assume that we know the true data-generating distribution P_0.

- Parameters like this can be estimated using *targeted minimum loss-based estimation* (TMLE).

- **Asymptotic linearity:**

$$\Psi_b(P_n^*) - \Psi_b(P_0) = \frac{1}{n}\sum_{i=1}^{n} IC(O_i) + o_P(\frac{1}{\sqrt{n}})$$

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.

Targeted Minimum Loss-Based Estimation

- TMLE produces a well-defined, unbiased, efficient substitution estimator of target parameters of a data-generating distribution.
- Iterative procedure (though there is a one-step now) that updates an initial estimate of the relevant part (Q_0) of the data generating distribution (P_0).
- Like corresponding A-IPTW estimators, removes asymptotic residual bias of initial estimator for the target parameter. If it uses a consistent estimator of g_0 (nuisance parameter), it is *doubly robust*.
- We can estimate the target parameter:

$$\Psi_b(P_n^*) = \frac{1}{n}\sum_{i=1}^{n}[Q_n^{(b,1)}(A_i = a_h, W_i) - Q_n^{(b,1)}(A_i = a_l, W_i)]$$

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).

Inference with influence curves

- The influence curve for the estimator is:

$$IC_{b,n}(O_i) = \left(\frac{\mathbb{1}(A_i = a_h)}{g_n(a_h \mid W_i)} - \frac{\mathbb{1}(A_i = a_l)}{g_n(a_l \mid W_i)} \right) \cdot (Y_{b,i} - \bar{Q}_n^{(b,1)}(A_i, W_i)) + \bar{Q}_n^{(b,1)}(a_h, W_i) - \bar{Q}_n^{(b,1)}(a_l, W_i) - \Psi_b(P_n^*) \tag{1}$$

- Sample variance of the influence curve:

$$s^2(IC_n) = \frac{1}{n} \sum_{i=1}^{n} (IC_n(O_i))^2$$

- Use sample variance to estimate the standard error:

$$se_n = \sqrt{\frac{s^2(IC_n)}{n}}$$

- Use this for inference — that is, to derive uncertainty measures (i.e., p-values, confidence intervals).

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

Moderated statistics for target parameters

- One can define a standard t-test statistic for an estimator of an asymptotically linear parameter (over $b = 1, \ldots, B$) as:

$$t_b = \frac{\sqrt{n}(\Psi_b(P_n^*) - \Psi_0)}{s_b(IC_{b,n})}$$

- This naturally extends to the moderated t-statistic of Smyth:

$$\tilde{t}_b = \frac{\sqrt{n}(\Psi_b(P_n^*) - \Psi_0)}{\tilde{s}_b}$$

where the posterior estimate of the variance of the influence curve is

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

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

$$\Psi_b(Q_{b,n}^*), S_b^2(IC_{b,n}),$$

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.

An influence curve transform

- Need the estimate for each biomarker (b) and the IC for every observation for that biomarker, repeating for all $b = 1, \ldots, B$.
- Essentially, transform original data matrix such that new entries are:

$$Y^*_{b,i} = IC_{b,n}(O_i; P_n) + \Psi_b(P^*_n)$$

- Since $\mathbb{E}[IC_{b,n}] = 0$ across the columns (units) for each b, the average will be the original estimate $\Psi_b(P^*_n)$.
- For simplicity, let's assume the null value is $\Psi_0 = 0$ for all b. Then, applying the moderated t-test to $Y^*_{b,i}$ will generate corrected, conservative test statistics $\tilde{t}_b$.

Just like the one-sample problem for estimation of parameter with associated standard error from the influence curve.

Why moderated statistics in this context?

- Often times, such data analyses are based on relatively small samples.
- To get a data-adaptive estimate, with standard implementation of these estimates, standard errors can be non-robust.
- Practically, "significant" estimates of variable importance measures may be driven by poorly and underestimated $s_b^2(IC_{b,n})$.
- Moderated statistics shrink these $s_b^2(IC_{b,n})$ (making them bigger), thus taking biomarkers with small parameter estimates but very small $s_b^2(IC_{b,n})$ out of statistical significance.

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.

Software implementation: “R/biotmle”

- An R package that “facilitates biomarker discovery by generalizing the moderated t-statistic of Smyth for use with asymptotically linear parameters.”
- Check it out on GitHub: nhejazi/biotmle

README.md

R/ `biotmle`

build passing | build passing | coverage 5% | repo status WIP | license MIT

> Targeted learning with the moderated t-statistic for biomarker discovery

Description

`biotmle` is an R package that facilitates biomarker discovery by generalizing the moderated t-statistic of Smyth for use with asymptotically linear target parameters. The set of methods implemented in this R package rely on the use of Targeted Minimum Loss-Based Estimation (TMLE) to transform biological sequencing data (e.g., microarray, RNA-seq) based on the influence curve representation of a particular statistical target parameter (e.g., the Average Treatment Effect). The transformed data is then used to test for group differences using the moderated t-statistic as implemented in the R package `limma`.

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

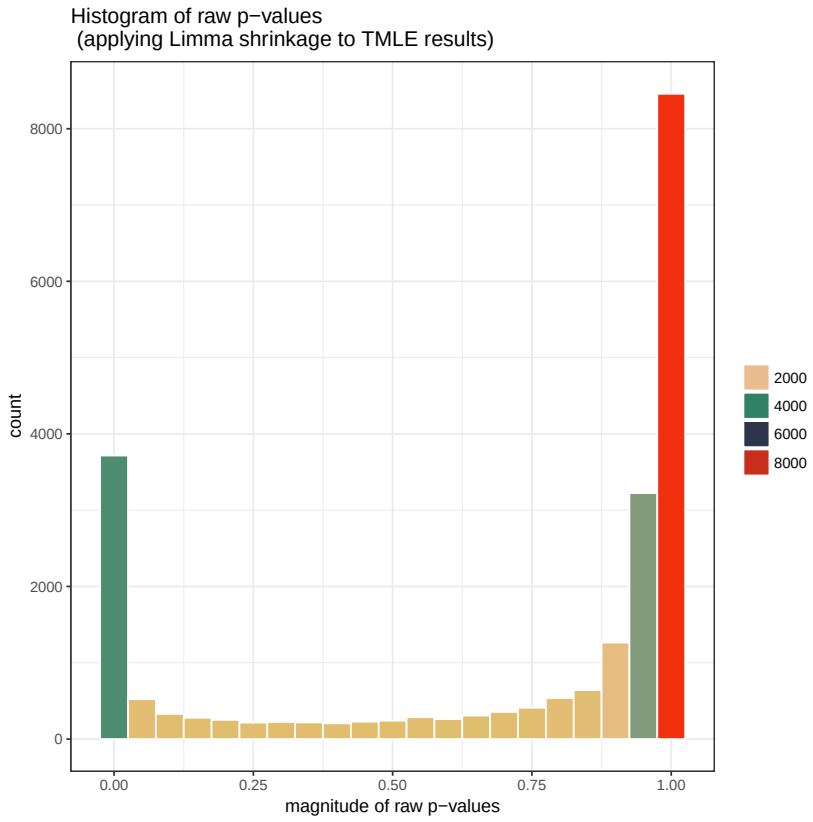
Data analysis with "R/biotmle"

- Observational study of the impact of occupational exposure (to benzene), with data collected on 125 subjects and roughly $22,000$ biomarkers.
- Baseline covariates W: age, sex, smoking status; all were discretized.
- Treatment A is degree of Benzene exposure: none, $<$ 1ppm, and $>$ 5ppm.
- Outcome Y is miRNA expression, median normalized.
- Estimate the parameter:

  $$\Psi_b(P_n^*) = \mathbb{E}[\mathbb{E}[Y_b \mid A = \max(A), W] - \mathbb{E}[Y_b \mid A = \min(A), W]]$$
- Apply moderated t-test as previously discussed.

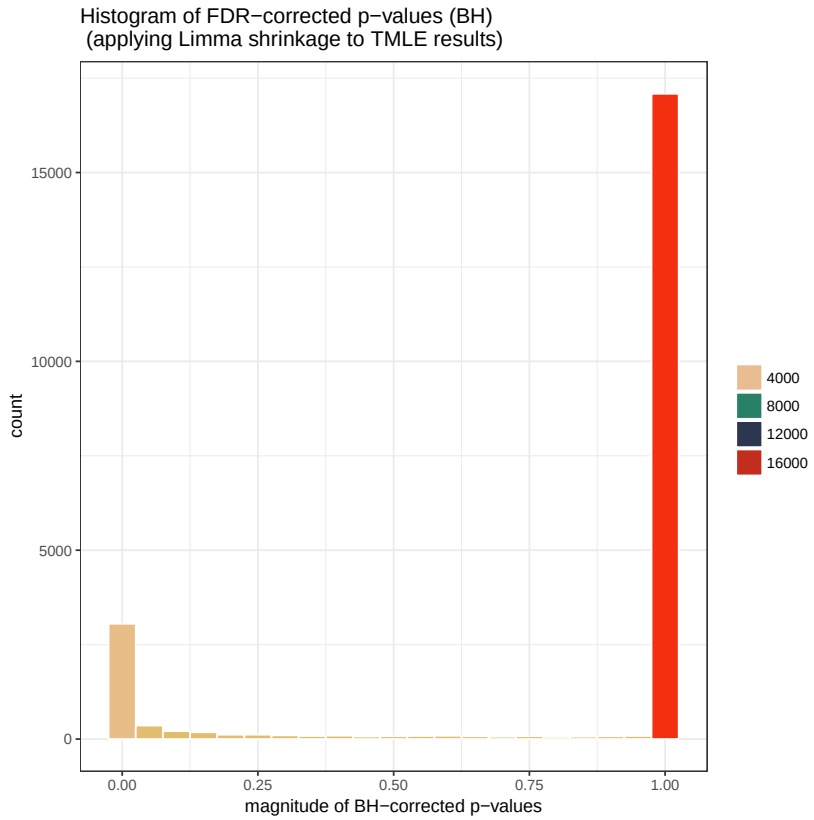
We are really just walking through the mechanistic procedure we outline, applying to the data set that served as our motivating example.

Analysis results I: Uncorrected tests



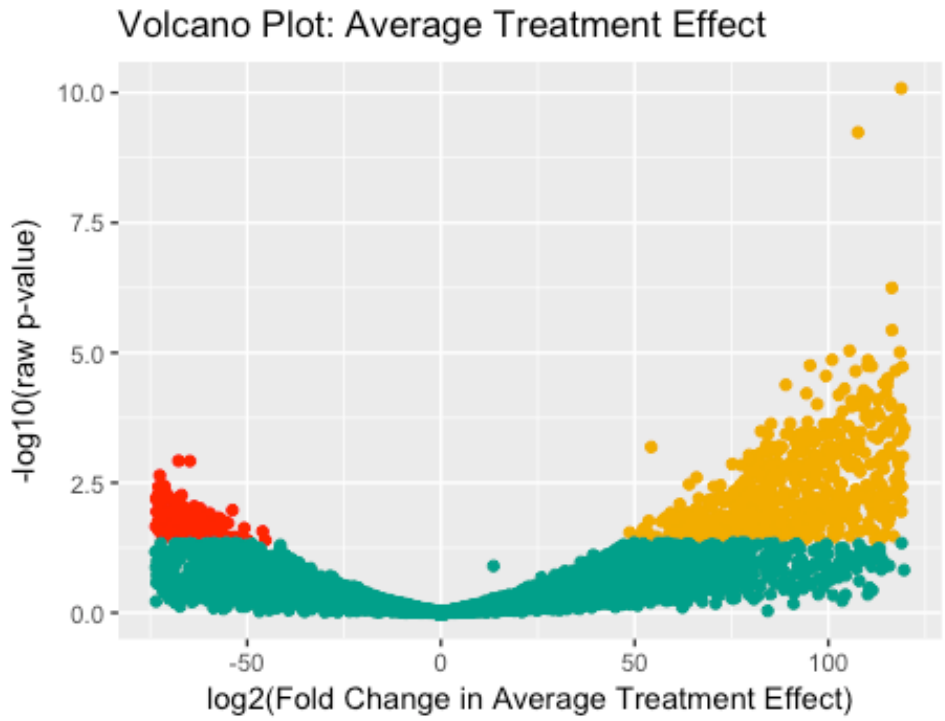
This is promising — we're not seeing too many biomarkers identified as "significant." But, we do have to correct for those $22,000$ tests that we just performed.

Analysis results II: Corrected tests



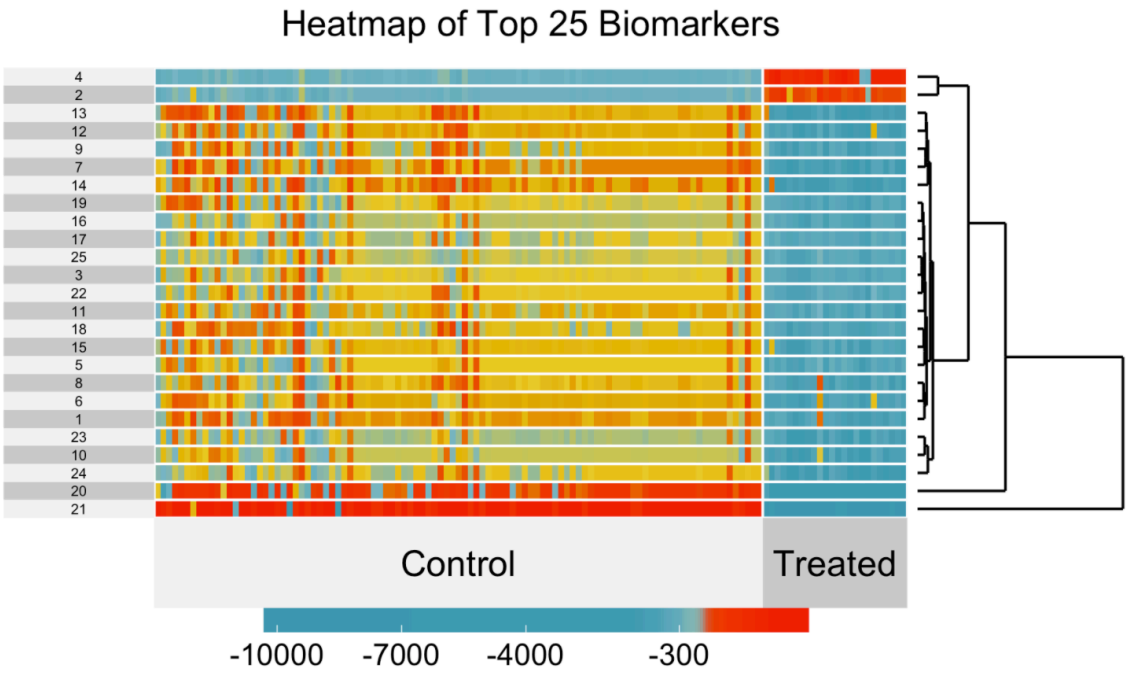
After application of the Benjamini-Hochberg procedure for controlling the False Discovery Rate (FDR).

Analysis results III: Volcano plot



Taking a look at a standard volcano plot adapted to the ATE quickly reveals that we really are not identifying any biomarkers with low fold change in the ATE as significant erroneously.

Analysis results IV: Heatmap of IC estimates



We can use our influence curve transform to identify biomarkers that are top contributers to the target parameter of interest — the ATE in this case.

Review

1. Linear models are the standard approach for analyzing microarray and next-generation sequencing data (e.g., R package “limma”).

2. Moderated statistics help reduce false positives by using an empirical Bayes method to perform standard deviation shrinkage for test statistics.

3. *Beyond linear models:* we can assess evidence using parameters that are more scientifically interesting (e.g., ATE) by way of TMLE.

4. The approach of moderated statistics easily extends to the case of asymptotically linear parameters.

It’s always good to include a summary.

References I

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.

Gruber, S. and van der Laan, M. J. (2010). An application of collaborative targeted maximum likelihood estimation in causal inference and genomics. *The International Journal of Biostatistics*, 6(1):1–31.

Hejazi, N. S., Cai, W., and Hubbard, A. E. (2017). biotmle: Targeted learning for biomarker discovery.

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.

References II

Smyth, G. K. (2005). Limma: linear models for microarray data. In *Bioinformatics and computational biology solutions using R and Bioconductor*, pages 397–420. Springer.

Tuglus, C. and van der Laan, M. J. (2011). Targeted methods for biomarker discovery. In *Targeted Learning*, pages 367–382. Springer.

van der Laan, M. J. and Rose, S. (2011). *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media.

Acknowledgments

Alan Hubbard	University of California, Berkeley
Mark van der Laan	University of California, Berkeley

This was all made possible by generous advising and collaboration.

Slides: goo.gl/6ou8YR

stat.berkeley.edu/~nhejazi

nimahejazi.org

twitter/@nshejazi

github/nhejazi

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