Generalized Application of Empirical Bayes Statistics to Asymptotically Linear Parameters

by

Nima S. Hejazi

A thesis submitted in partial satisfaction of the

requirements for the degree of

Master of Arts

in

Biostatistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor Alan E. Hubbard, Co-chair
Professor Mark J. van der Laan, Co-chair
Professor Martyn T. Smith

Spring 2017

Generalized Application of Empirical Bayes Statistics to Asymptotically Linear Parameters

Abstract

Generalized Application of Empirical Bayes Statistics to Asymptotically Linear Parameters

by

Nima S. Hejazi

Master of Arts in Biostatistics

University of California, Berkeley

Professor Alan E. Hubbard, Co-chair

Professor Mark J. van der Laan, Co-chair

The exploratory analysis of high-dimensional biological sequencing data has received much attention for its ability to allow the simultaneous screening of numerous biological characteristics at resolutions unimaginable just two decades ago. While there has been an increase in the dimensionality of such data sets in studies of environmental exposure and biomarkers, two important questions have received less attention than deserved: (1) how can individual estimates of independent associations be derived in the context of many competing causes while avoiding model misspecification, and (2) how can accurate small-sample inference be obtained when data-adaptive techniques are employed in such contexts. The central focus of this paper is on variable importance analysis in high-dimensional biological data sets with modest sample sizes, using semiparametric statistical models. We present a method that is robust in small samples, but does not rely on arbitrary parametric assumptions, in the context of studies of gene expression and environmental exposures. Such analyses are faced not only with issues of multiple testing, but also the problem of teasing out the associations of biological expression measures with exposure, among confounds such as age, race, and smoking. Specifically, we propose the use of targeted minimum loss-based estimation, along with a generalization of the moderated empirical Bayes statistics of Smyth, relying on the influence curve representation of a statistical target parameter to obtain estimates of variable importance measures. The result is a data-adaptive approach that can estimate individual associations in high-dimensional data, even in the presence of relatively small sample sizes.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis proposes a straightforward extension of an empirical Bayes inferential method — the moderated statistics of Smyth, as implemented in the popular "limma" software package [12] — for general use with asymptotically linear parameters [13, 16]. By way of this extension, estimators of complex target parameters can benefit from the inferential robustness that such moderated statistics provide, in the context of many comparisons. As a motivating example, consider a previous study of miRNA expression and occupational exposure to benzene [8]: The data consists of around $22,000$ genes (measured via the *Illumina Human Ref-8 BeadChips* platform) on 125 subjects in factories in China. In this study, the variable of interest was occupational exposure to benzene (measured in various ways), though information on confounding factors was also recorded (e.g., gender, smoking status). Taking benzene exposure to be binary, the quantity of interest can be framed as the adjusted association of each of the roughly $22,000$ expression values with exposure. One could easily use the approach based on moderated statistics by fitting a parametric linear model with, say, benzene as outcome and both exposure and confounders as predictors, performing a multiple comparison correction on the estimated coefficients associated with benzene. However, it is generally desirable to utilize a procedure that is less reliant on arbitrary assumptions, specifically one estimating a nonparametric estimand, where fitting the model of interest could be performed via automated, data-adaptive techniques (i.e., machine learning). We show that utilizing moderated statistics in such situations is possible if asymptotically linear parameters are used — that is, where the difference between the values taken by the estimator and the parameter may be approximated by a sum of i.i.d. random variables (i.e., the influence curve representation). Many complex parameters have representations that are asymptotically linear, and so with minor modifications, moderated statistics can be applied to a wide variety of settings. This is particularly valuable in smaller samples, as sampling distribution estimates for these complex estimators can be unstable, yielding false positives, a problem that moderated statistics are well-suited to ame-

liorate by borrowing estimates of the sampling variability across the variables of interest (in our case, gene expression measures). In this way, one can use data-adaptive methods to avoid the bias of arbitrary parametric assumptions, which are common in bioinformatic applications, while still adding a degree of robustness for these potentially unstable estimators.

In the following sections, we first detail a data-adaptive, machine learning-based estimator of a well-known estimand for deriving adjusted associations. We then show how the machinery of moderated statistics can be used to derive an empirical Bayes estimate of the standard error of this estimator — and, generally, for any asymptotically linear estimator. Finally, we apply the resulting procedure to the example of occupational benzene exposure previously described.

Chapter 2

Methodology

2.1 Data and Statistical Model

Others have proposed using estimators developed for low-dimensional causal inference problems to derive nonparametric estimators of association in the context of high-dimensional biomarker discovery studies [14]. In such cases, the goals of analyses are similar to those of more typical parametric approaches, but the approach is based on nonparametric estimands and can be estimated with data-adaptive techniques. Such data structures typically consist of large matrices of biological expression values as well as tables of phenotypic information on each subject. In particular, in later sections, we will illustrate the use of our technique on data generated by the *Illumina Human Ref-8 BeadChips* platform, from a study which included expression measures on about $22,000$ genes as well as phenotypic information, on a sample of 125 subjects. The analysis aims to evaluate the association of an environmental exposure (benzene) on the expression measures of the roughly $22,000$ genes simultaneously, controlling for the several aforementioned confounders. In our analysis, we consider three potential confounding factors on the relationship of exposure and expression: age, sex, and smoking status. This problem setup is easily generalizable to situations with greater numbers of potential biomarkers and confounders. Ultimately, the aim of analyzing such data sets is to rank the importance of a set of candidate biomarkers based on their independent associations with a treatment variable. In order to build a ranking of biomarkers, we start by defining a variable importance measure (VIM) [16].

Let $O = (W, A, Y)$ represent a random variable defined on the observed data, where W are the confounders, A the exposure of interest, and $Y = (Y_b, b = 1, \ldots, B)$ a vector of potential biomarkers. Note that we observe n i.i.d. copies of the random variable O, such that $O_i = \{O_1, \ldots, O_n\}$. Further, let $O \sim P_0 \in \mathcal{M}$, where P_0 is the unknown probability distribution of the full data, contained in an infinite-dimensional statistical model $\mathcal{M}$. For the specific data set described in above, $W =$

$(W_1, W_2, W_3, W_4, W_5)$, where age ($W_1$) is a continuous measure, gender (W_2) is binary, smoking status (W_3) is binary, BMI (W_4) is a continuous measure, and alcohol consumption (W_5) is binary; A is a binary exposure; and Y_b are miRNA expression measures.

2.2 The Target Parameter

To define the parameter of interest, generally, let $\Psi(P_0)$ be the target parameter based on a function Ψ that maps the probability distribution P_0 into the target feature of interest. Thus, the parameter $\Psi(P_0)$ is a function of the unknown probability distribution P_0, defined on the (unobserved) full data. Let P_n represent the empirical distribution of the observed data $O_1, O_2, \ldots, O_n$. Though we focus on cases where the O_i are i.i.d., the following is easily generalizable when the data are clustered (e.g., as repeated samples from the same biological unit). We are interested in substitution estimators of the form $\Psi(P_n^*)$ — that is, we apply the same mapping (Ψ) but to the empirical distribution P_n to derive our estimate (e.g., Ψ could merely be the expectation operator). In using this general definition, we expand the parameters of interest beyond coefficients in a misspecified parametric statistical model, by defining a parameter as a feature of the true probability distribution P_0 of the full data. Specifically, we propose here what is referred to as a targeted variable importance measure [2]:

$$\Psi_b \equiv \Psi_b(P_0) = \mathbb{E}_{W,0}[\mathbb{E}_0(Y_b \mid A = 1, W) - \mathbb{E}_0(Y_b \mid A = 0, W)]. \tag{2.1}$$

The parameter delineated in (2.1) above is generally referred to as the average treatment effect, often denoted simply as the ATE [10]. When the assumptions underlying the causal model, through which the target parameter is defined, do not hold, the estimand of the parameter of interest takes on a statistical interpretation: specifically, the difference of means within the strata W, averaged across levels of the treatment A. It has been shown that, under identifiability assumptions (e.g., no unmeasured confounding), this parameter can be statistically estimated via targeted maximum likelihood estimation [16]. Such parameters are significant in that they are not defined explicitly via parametric statistical models, leaving one free to fit the requisite models data-adaptively, minimizing assumptions wherever possible, and yet still estimating a relatively simple parameter with rich scientific interpretation.

2.3 Statistical Estimation

As noted previously in Section 2.2, the target parameter is defined as a feature of the unknown probability distribution P_0. While there are several general classes of estimators available for estimating Ψ, here we focus on a substitution estimator as noted above. Examining (2.1), one can anticipate that a substitution estimator will rely on estimates of two components of the data-generating mechanism, P_0: $\mathbb{E}_0(Y \mid A = a, W)$ and $P_0(W)$, or the true regression of Y on (A, W) and the marginal distribution of W. Let $Q_0^b(A, W) \equiv \mathbb{E}_0(Y_b \mid A, W)$, and $Q_n^b(A, W)$ an estimate of this regression. If we use the empirical distribution to estimate the joint marginal distribution of the W, then a substitution estimator is:

$$\Psi_b(P_n^*) = \frac{1}{n} \sum_{i=1}^{n} Q_n^b(A_i = 1, W_i) - Q_n^b(A_i = 0, W_i). \tag{2.2}$$

Below, we discuss recommendations for an initial estimate of Q_0, using the Super Learner algorithm, and a bias-reducing augmentation (targeted minimum loss-based estimation) with optimal properties for minimizing the error of estimation and deriving robust inference.

Using the Super Learner algorithm

The first step in the two-stage TMLE procedure is to derive an initial estimate of Q_0^b, referred to as $Q_n^{(b,0)}$. For instance, one may assume a parametric statistical model that results in (2.2) being equivalent to a regression coefficient (e.g., $Q_0^b(A, W) = \alpha^b + \beta_A^b A + \beta_W^b W$). By defining (2.2) in a nonparametric statistical model, using data-adaptive tools to estimate Q_0^b, we avoid settings wherein estimators based on parametric models would be inconsistent. Specifically, given that the true model Q_0^b is typically unknown, more accurate estimates may be derived by employing machine learning algorithms in the estimation procedure.

This reliance on machine learning algorithms leads naturally to the issue of choosing an optimal data-adaptive algorithm. To address this issue, we advocate use of the Super Learner algorithm, a generalized stacking algorithm for ensemble learning, implemented via cross-validation, which produces an optimally weighted combination of candidate estimators, minimizing the cross-validated risk. Using this procedure, the predictions from a set of candidate algorithms are combined, allowing for highly data-adaptive functional forms to be specified [15].

Though the set of candidate algorithms in the library may be arbitrary, the theoretical underpinnings of the Super Learner algorithm offer guidance as to the type and number of learning algorithms that ought to be considered in the fitting routine. In the rare case that one of the candidate learning algorithms captures the true model and, consequently, converges to the correct

estimate at a parametric rate, the Super Learner algorithm has been shown to converge to the same estimate at a near-parametric rate of $O\left(\frac{log(n)}{n}\right)$ [15]. As true relationships are rarely captured by single learning algorithms alone, Super Learner will, up to a first order term, do as well (in terms of risk) as an algorithm that chooses the particular candidate learner based on full knowledge of the true distribution — that is, an oracle selector — a result that holds as long as the number of candidate algorithms is polynomial in sample size. The principle implementation of the Super Learner algorithm is available as a software package [15] for the R language and environment for statistical computing [9].

Targeted minimum loss-based estimation

While the Super Learner estimate of Q_0 is performed to minimize the cross-validated risk based on an appropriate loss-function, Q_0 is not the target of our analysis, rather we seek to minimize the mean-squared error of an estimator of $\Psi_b(P_0)$, the target parameter of interest. There is no guarantee that, given a set of highly data-adaptive learning algorithms, the estimate of $\Psi_b(P_0)$ will have a normal sampling distribution, even in cases of fairly large sample size. Fortunately, an estimator of Q_0 that not only "targets" the estimate of the regression towards the particular parameter of interest but also "smooths" the estimator such that the sampling distribution converges reliably to a normal distribution is available [16]. This "targeting" step can be thought of as optimizing the bias-variance tradeoff, since the data-adaptive selection procedure of Super Learner results in an estimate of $\Psi_b(P_0)$ that suffers from residual confounding. This form of confounding can occur, for instance, if the variable selection step in the procedure estimating Q_0^b leaves out any regressors that are, in truth, confounders of the association of A and Y. In this case, bias in estimation of $\Psi_b(P_0)$ is caused by under-fitting. Thus, the resultant targeted minimum loss-based estimator (TMLE) is more robust to model misspecification than the initial substitution estimator, based on the initial fit of Super Learner, and is also, if one has consistent estimates of all relevant portions of P_0, semiparametrically locally efficient. For a detailed discussion of the theory of targeted minimum loss-based estimation and formal justifications of the efficiency of the resultant estimator, consult the appendix of van der Laan and Rose [16].

Algorithmically, the TMLE-based estimator in our case is a simple one-dimensional augmentation of the initial fit. Specifically, in the case of a continuous outcome, following the initial Super Learner fit, one proceeds by fitting a simple, one-dimensional regression:

$$Q_n^{(b,1)}(A,W) = Q_n^{(b,0)}(A,W) + \hat{\varepsilon} h_{g_n}(A,W)$$

where the initial fit, $Q_n^{(b,0)}(A,W)$ is treated as an offset, and $h_{g_n}(A,W)$ is a so-called "clever"

covariate:

$$h_{g_n}(A,W) = \frac{I(A=1)}{g_n(1 \mid W)} - \frac{I(A=0)}{g_n(0 \mid W)},$$

where $g_n(1 \mid W)$ is an estimate of $P(A = 1 \mid W)$, or the propensity score [10]; $\hat{\varepsilon}$ is the estimated coefficient from the regression of Y on $h_{g_n}(A,W)$, treating $Q_n^{(b,0)}(A,W)$ (or the logit of this quantity if regression is logistic) as the offset. The selection of g_n can be made via a process that minimizes the mean-squared error of the parameter of interest [5], but for application purposes, a simple main-terms logistic regression is usually sufficient. In the final step of this procedure, the targeted minimum loss-based estimate of Ψ_b is derived using the targeted estimate of Q:

$$\Psi_b(P_n^*) = \frac{1}{n}\sum_{i=1}^{n}[Q_n^{(b,1)}(A_i = 1, W_i) - Q_n^{(b,1)}(A_i = 0, W_i)], \tag{2.3}$$

where P_n^* is the estimate of the data-generating distribution based on TMLE, in this case, based on estimates of $g_n, Q_n^{(b,1)}$.

2.4 Statistical Inference

An influence curve-based approach

As shown in [16], $\Psi_b(P_n^*)$ is an asymptotically linear estimator of $\Psi_b(P_0)$, with influence curve $IC(O_i)$ if it satisfies

$$\sqrt{n}(\Psi_b(P_n^*) - \Psi_b(P_0)) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} IC(O_i) + o_p(1). \tag{2.4}$$

Note from (2.4) above that the variance of $\Psi_b(P_n^*)$ is well approximated by the sample variance of the influence curve divided by the sample size. When considering biomarkers, the estimated influence curve for the ATE is

$$\begin{aligned} IC_{b,n}(O_i) = &\left[\frac{I(A_i=1)}{g_n(1 \mid W_i)} - \frac{I(A_i=0)}{g_n(0 \mid W_i)}\right](Y_{b,i} - Q_n^{(b,1)}(A_i, W_i)) \\ &+ Q_n^{(b,1)}(1, W_i) - Q_n^{(b,1)}(0, W_i) - \Psi_b(P_n^*). \end{aligned} \tag{2.5}$$

With the above in hand, we easily derive asymptotic p-values and confidence intervals (CI) with a Wald-type approach:

$$\text{p-value} = 2\left[1 - \Phi\left(\frac{|\Psi_b(P_n^*)|}{\sigma_n^b/\sqrt{n}}\right)\right] \tag{2.6}$$

$$(1 - \alpha)\ \text{CI} = \Psi_b(P_n^*) \pm \frac{Z_{(1-\alpha)}\sigma_n^b}{\sqrt{n}} \tag{2.7}$$

where σ_n^b is the sample standard deviation of IC_b and $\Phi(\cdot)$ is the CDF of the standard normal distribution.

Moderated statistics for influence curve-based estimates

In high-dimensional settings, with small sample sizes, direct application of TMLE for obtaining joint inference for a targeted estimate of a variable importance measure can result in unstable standard error estimates, and thus potentially erroneous identification of biomarkers. This is particularly important if data-adaptive procedures are utilized, as these can add to finite-sample non-robustness. To address this problem, we apply moderated statistics [11], a technique that preserves accurate asymptotic inference, yet, provides robust inference in small sample settings by drawing on information across the many estimates of sampling variability (the σ_n^b) by invoking an empirical Bayes procedure. First developed for the analysis of data from microarray experiments, the moderated t-statistic is implemented in the immensely popular "limma" software package, which provides a suite of tools for analyzing differential expression of genes using linear models, borrowing information across all genes to provide stable and robust inference for microarray data [11]. Previously, we noted that a common way of making inference about the target parameter $\Psi_b(P_0)$ is to compute the influence curve-based values for $\Psi_b(P_n^*)$, which can then be used to calculate the corresponding standard errors of the influence curve of the target parameter. After obtaining these IC values, finding corresponding p-values and making inference about Ψ_b for each probe follow trivially.

The procedure for using moderated statistics on IC-based estimates of Ψ_b, using the approach of "limma" to impose variance shrinkage with an empirical Bayes procedure, is as follows:

1. Assume repeated tests, across all probes b, of null and alternative hypotheses:
   $H_0 : \Psi_b(P_0) = 0, H_A : \Psi_b(P_0) \neq 0$.

2. Find influence curve-based estimates for each probe, one at a time, using these to iteratively build a matrix of IC-based estimates of the target parameter across all subjects, for all probes.

3. Since the IC-based estimates have mean zero, add in the corresponding estimates of $\Psi_b(P_n)$ to each row (probe). This results in each row having an appropriate average ($\Psi_b(P_n)$) and sample variance equivalent to that of the influence curve for that probe (IC_b).

4. Using the implementation readily available in the "limma" R package, apply the moderated t-statistic ($\tilde{t}_b, b = 1, \ldots, B$) to the aforementioned matrix of IC-based estimates of the target parameter, resulting in individual estimates across each probe, relative to the null hypotheses above.

5. The resulting inference, based on the shrinkage estimate of the sampling standard deviation of the influence curve ($\tilde{\sigma}_n^b$) is a weighted average of σ_n^b and a value close to the average of all these sample standard deviation estimates across the biomarkers ($\bar{\sigma}_n^b \approx \frac{1}{B}\sum_{b=1}^{B}\sigma_n^b$, or $\widetilde{\sigma}_n^b = wt_b\sigma_n^b + (1-wt_b)\bar{\sigma}_n^b$, where $wt_b \in (0,1)$). See [11] for a rigorous and formal presentation. Asymptotically, as $n \to \infty$, $wt_b \to 1$, and thus $\widetilde{\sigma}_n^b \to \sigma_n^b$ as desired.

6. Use a multiple testing correction procedure to obtain accurate simultaneous inference for all probes (biomarkers) $b = 1,\ldots,B$. In standard practice, we recommend the well-known Benjamini-Hochberg procedure for controlling the False Discovery Rate (FDR) [3].

The procedure enumerated above will shrink aberrant estimates of variability towards the center of their joint distribution, with a particularly noticeable effect when the sample size is small. The practical effect is that application of this procedure reduces the number of significant biomarkers, largely false positives driven by potentially erroneous underestimates of the variation of the estimate of the parameter of interest, $\Psi_b(P_n)$. This approach is convenient in that it can handle any asymptotically linear estimator (has a representation as in (2.4)), which covers many estimators of parameters of scientific interest. An open source software package, "biotmle," implementing the described procedure, is publicly available [6].

Chapter 3

Data Analysis

For the gene expression data set previously described in Section 2.1, we applied the TMLE-based biomarker evaluation procedure to obtain separate, individual estimates of the association of each of the roughly $22,000$ biomarkers with benzene exposure, while controlling for potential confounding based on age, sex, and smoking status. The values obtained from applying this procedure on a biomarker-by-biomarker basis correspond to the contributions of each potential biomarker to changes measured by the ATE (the target parameter of interest, in this case), based on the influence curve decomposition of the ATE parameter. While having a direct interpretation in relation to the ATE, such transformed expression values hold little bearing on statistical inference.

Using the ATE, the moderated t-statistic for the test performed is as follows:

$$\tilde{t}_b = \frac{\sqrt{n}(\Psi_{b,n}(P_n^*))}{\tilde{s}_{b,n}},$$

where $\tilde{s}^2_{b,n} = \frac{d_0 s_0^2 + d_b(s_b^2(IC_{b,n}))}{d_0 + d_b}$, where d_b is the degrees of freedom for the b^{th} biomarker, d_0 is the degrees of freedom for the remaining biomarkers, $s_b(IC_{b,n})$ the standard deviation for the b^{th} biomarker, and s_0 the common standard deviation across all biomarkers towards which the empirical Bayes procedure performs shrinkage.

In order to isolate a set of differentially up-regulated or down-regulated biomarkers, we apply the moderated t-statistic [12] to test for group differences based on the observed benzene exposure status. This results in a table including the moderated t-statistic for each test of the ATE-transformed values between the exposed and unexposed groups (a coefficient corresponding to exposure in the gene-wise linear models fit via the approach of "limma"), standard errors of the coefficient, raw p-values, and adjusted p-values from application of the Benjamini-Hochberg procedure for controlling the FDR [3]. See table 3 below:

	Biomarker ID	ATE Change	p-value	adjusted p-value
1	198	1.69167E+01	1.04812E-54	2.90551E-51
2	1055	8.30585E+00	1.73105E-47	1.74498E-44
3	1764	-1.83308E+00	6.00103E-55	1.90121E-51
4	2469	1.70375E+02	2.87168E-47	2.76893E-44
5	3607	-4.36856E+00	6.07654E-47	5.39038E-44
6	4195	7.19651E+00	1.38153E-52	2.78529E-49
7	6207	-3.05520E+01	1.17986E-57	5.23316E-54
8	6262	-1.30293E+01	8.96437E-49	1.10446E-45
9	7481	-2.72348E+01	1.06992E-48	1.24883E-45
10	8664	-9.94950E+01	3.25553E-47	3.00824E-44
11	10255	1.07510E+01	9.07492E-54	2.01255E-50
12	11073	-2.88118E+01	7.45674E-54	1.83742E-50
13	12898	-2.50923E+01	1.34871E-58	7.47759E-55
14	14003	-1.84590E+01	5.86475E-59	4.33542E-55
15	14472	7.39674E-01	2.61339E-52	4.82976E-49
16	16255	-3.41521E+01	1.31512E-50	2.08324E-47
17	16454	-5.35507E+00	8.58888E-48	9.52378E-45
18	16608	-3.34112E+00	1.16964E-55	4.32320E-52
19	16658	-6.27276E+00	2.21905E-51	3.78552E-48
20	17537	-1.77342E+02	2.27910E-59	2.52718E-55
21	17982	-1.09417E+02	4.52028E-63	1.00246E-58
22	18337	1.49518E+00	1.87252E-49	2.44275E-46
23	19399	-1.06334E+02	3.36332E-50	4.97256E-47
24	20294	-1.16305E+02	1.64737E-47	1.73970E-44
25	22058	-1.13907E+01	6.07700E-50	8.42310E-47

Table 3.1: *The top 25 biomarkers isolated as a result of applying the moderated t-statistic to the ATE parameter. Applying empirical Bayes moderation to the variance of the ATE estimates produced by standard TMLE-based procedures identifies nearly* $5,000$ *biomarkers as significant in total.*

The analysis presented can be completely replicated by using the "biotmle" software package, which provides facilities for visualizing the results. Using this R package, a heatmap for visualizing differences in the ATE induced by benzene exposure, with all 125 subjects on the x-axis and the top 25 isolated biomarkers on the y-axis, may be produced. The heatmap, created using the "superheat" R package [1], is displayed as figure 3 below:

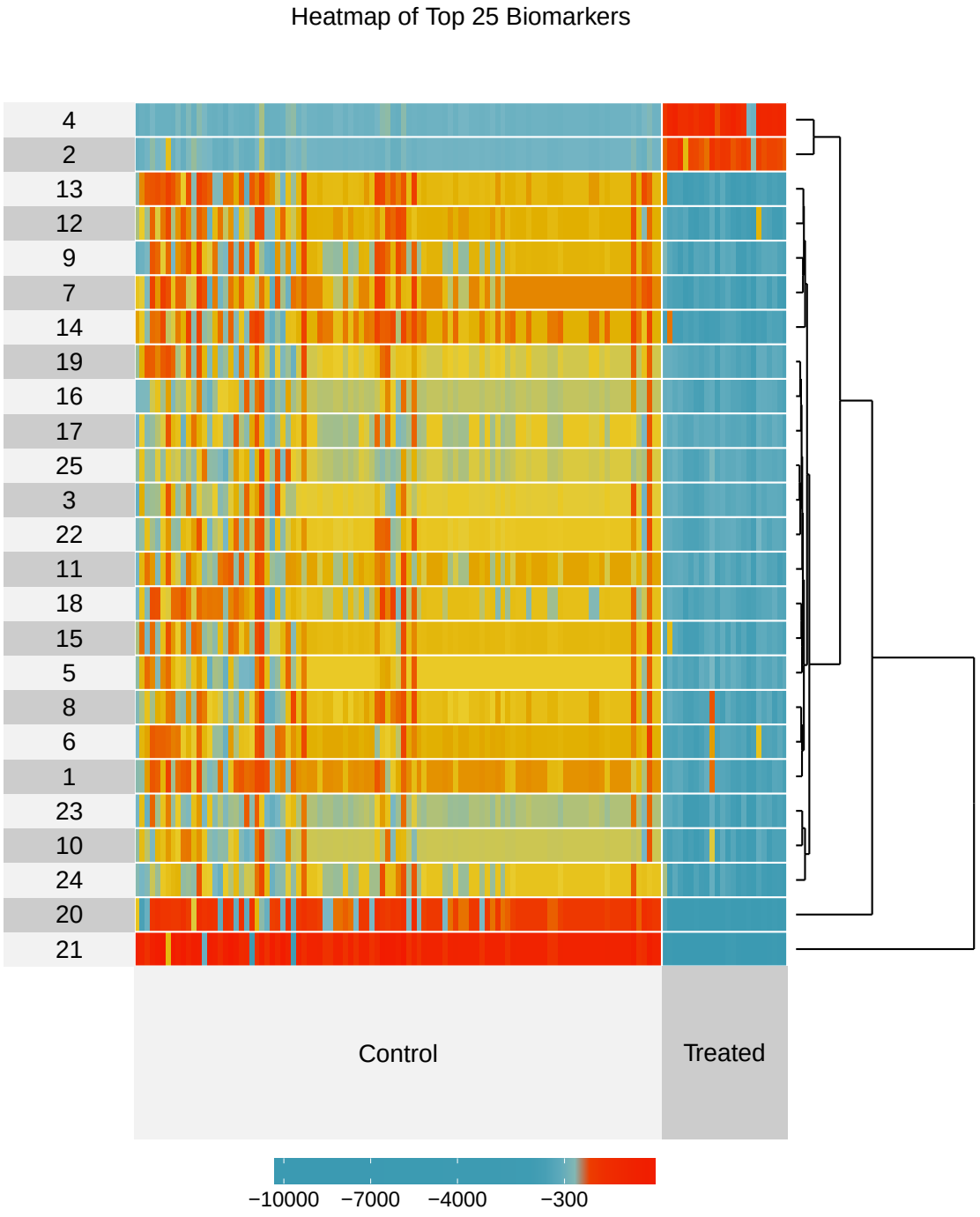

Figure 1: *Heatmap of the ATE estimates. Blue indicates a depression in the ATE, while red indicates elevation of the ATE, based on exposure to the maximal level of benzene as opposed to the minimal level. Hierarchical clustering is performed on the top 25 biomarkers identified by the proposed procedure.*

As expected, the use of moderated statistics reduces the spread of the standard deviation estimates of the influence curve by probe ($\tilde{s}_b^2$) across the approximately $20,000$ probes, and the corresponding Wald statistics for testing the target parameter, in comparison to using the original standard error. The results of our analysis indicate that application of moderated statistics to asymptotically linear parameters constitutes a powerful approach for assessing variable importance, based on treatment or exposure, in the context of high-dimensional investigations of biomarkers. We conclude that using this adaptation of TMLE, complimented by moderated statistics implemented in the "limma" R package, reduces the variability of standard errors and reduces the number of probes identified as significant, leading to more stable and robust inference, while providing the opportunity to evaluate biomarkers in the context of statistical parameters of scientific relevance, such as the average treatment effect focused on in the example discussed above.

Chapter 4

Discussion

This thesis has introduced an automated, robust method for analyzing high-dimensional biological sequencing data with relatively modest sample sizes. In the provided examples, the challenges presented were two-fold, including both obtaining simultaneous inference for a large number of comparisons and adjustment to account for potential confounders of the association of interest, in the context of a large statistical model and few biological replicates. Since the goal is estimation within a nonparametric (infinite-dimensional) statistical model, the techniques leveraged must involve data-adaptive estimation, while still providing trustworthy statistical inference, necessarily using estimators grounded in semiparametric efficiency theory. That is, given the parameter of interest and the nature of the statistical model, we maintain that the choice guiding the algorithm should not be *ad hoc*, but rather based on the relative efficiency of competing estimators. We have proposed methods that draw on existing work in statistical genomics and merge these with modern proposals for the analysis of variable importance, ultimately yielding a procedure that data-adaptively identifies promising biomarkers from a large set and that can be applied to data generated from experiments belonging to a large class of study designs.

We illustrated the method using an example miRNA data set (featuring benzene exposure) by applying, on a biomarker-by-biomarker basis, the outlined approach, combining TMLE with the moderated t-statistic to estimate the association of each potential biomarker with exposure. Thus, we present a flexible generalization of moderated statistics to the case of asymptotically linear parameters, obtaining robust small-sample inference, derived from influence curve-based estimation of the parameter of interest. The results suggest that instabilities inherent in small-sample inference can be ameliorated by combining this asymptotically efficient estimator of the ATE (based on TMLE) with the moderated t-statistic (implemented in the "limma" software package); in our example, this results in the isolation of fewer statistically significant biomarkers. Since application of the moderated statistics has no impact on asymptotic properties of TMLE-based estimation pro-

cedures — the adjustment for within-probe inference becomes negligible as sample size grows — we can readily use the asymptotic theory underlying TMLE.

This combination of existing methods offers many advantages: (1) it estimates target parameters relevant to specific scientific questions, in the presence of many confounders, without placing assumptions on the underlying statistical model; (2) it uses the theoretical optimality of loss-based estimation via the Super Learner algorithm, which optimally balances the bias-variance tradeoff in finite samples by appropriately choosing a level of parsimony to match the information available in the sample; (3) its reliance on TMLE-based estimators reduces residual bias and adds an appropriate degree of smoothing, making influence curve-based based inference available for the target parameters of interest; and (4) it robustifies inference by leveraging moderated statistics to derive joint inference with fewer false positives than would result from otherwise poor estimation of the sampling variability of the estimator. The result is a theoretically sound, data-adaptive estimation procedure, based on pre-specified, flexible learning algorithms, that guarantees robust statistical inference. While the continuing development of new biotechnologies promises new insights into the myriad relationships between biomarkers and health, procedures like the one presented here will surely be necessary to ameliorate the pitfalls of increasing dimensionality of the scientific problems of interest, by providing a rigorous and generalizable statistical framework for accurate, robust, and conservative biomarker discovery.

Chapter 5

Software Package

To support widespread use of this newly developed methodology, an R package, "biotmle," which provides a generalized implementation of this biomarker discovery procedure, has been made publicly available. As previously described in Section 2.3, the method, based on targeted minimum loss-based estimation [16] and a generalization of the moderated statistics of Smyth [12], is designed for use with both microarray and next-generation biological sequencing data. The statistical approach made available in this software package relies on the use of TMLE to rigorously evaluate the association between a set of potential biomarkers and another variable of interest while adjusting for potential confounding from another set of user-specified covariates. The implementation is in the form of a package for the R language for statistical computing [9].

There are two principal ways in which the biomarker discovery techniques in the "biotmle" R package may be used: to evaluate the association between (1) a phenotypic measure (say, environmental exposure) and a biomarker of interest, and (2) an outcome of interest (e.g., survival status at a given time) and a biomarker measurement, both while controlling for background covariates (e.g., BMI, age). By using a TMLE-based procedure to estimate the average treatment effect in a targeted manner, the package produces easily interpretable results in the form of a variable importance measure (see [16] for an extended discussion), making the "biotmle" package aptly suited for applications in bioinformatics, genomics, and molecular epidemiology.

While the principal results produced by this R package matches those produced by the immensely popular "limma" R package [11], "biotmle" provides several unique utilities, largely in the form of several expressive plotting methods — for example, a heatmap based on the recently developed "superheat" R package [1] — and a custom "biotmle" class, based on the popular "SummarizedExperiment" class [7]. While the R package is currently publicly available at `https://github.com/nhejazi/biotmle`, submission of the software package to the centralized repository maintained by the Bioconductor project [4] is underway.

Chapter 6

Future Work

The proposed procedure for applying empirical Bayes moderated statistics to asymptotically linear parameters currently suffers from several limitations, both in terms of pragmatic theoretical development and software implementation. As previously discussed, empirical Bayes moderation of estimates of parameters with asymptotically linear representations may be used to obtain robust standard deviation estimates when the estimated parameter is formulated in terms of a binary exposure variable. It is for this reason that we largely rely on the moderated t-statistic of Smyth [12] in our presentation of applied data analysis. The limitation imposed by assessing the difference between two levels of the exposure variable has several solutions: (1) the applied researcher could decide on a binarization scheme that fits their scientific question of interest, and (2) the range of target parameters currently supported by our software implementation [6] could be increased beyond just the average treatment effect.

Going further, this application of moderated statistics to the development of robust hypothesis testing procedures for asymptotically linear parameters may be further extended to a general notion of hypothesis testing for parameters that do not have straightforward asymptotically linear representations. Specifically, one might be interested in assessing more complex parameters (e.g., the causal dose-response curve) that allow a greater degree of flexibility in answering scientific questions of interest — that is, it is easy to conceive of investigations in which a dose-response relationship is of interest, rather than a mere difference in the levels of a binary exposure. Supporting the application of moderated statistics to such parameters would greatly increase the flexibility of the proposed procedure.

While the proposed method has been well-established theoretically, there are a number of computational improvements that can still be made, including both improvements of the software to meet the standards of the Bioconductor project and numerical simulations to confirm that the application of moderated statistics to asymptotically linear estimates of target parameters of interest

produces consistent results. Several ideas for simulation studies have been proposed, and work is underway to implement these ideas, to ensure that the proposed method and corresponding software implementation do hold to the theoretical properties presented in preceding sections.

Bibliography

[1] Rebecca L Barter and Bin Yu. *Superheat: An R package for creating beautiful and extendable heatmaps for visualizing complex data*. 2017.

[2] Oliver Bembom et al. "Biomarker discovery using targeted maximum-likelihood estimation: Application to the treatment of antiretroviral-resistant HIV infection". In: *Statistics in medicine* 28.1 (2009), pp. 152–172.

[3] Yoav Benjamini and Yosef Hochberg. "Controlling the false discovery rate: a practical and powerful approach to multiple testing". In: *Journal of the royal statistical society. Series B (Methodological)* (1995), pp. 289–300.

[4] Robert C Gentleman et al. "Bioconductor: open software development for computational biology and bioinformatics". In: *Genome Biology* 5.10 (2004), R80.

[5] Susan Gruber and Mark J van der Laan. "An application of collaborative targeted maximum likelihood estimation in causal inference and genomics". In: *The International Journal of Biostatistics* 6.1 (2010).

[6] Nima S Hejazi, Weixin Cai, and Alan E Hubbard. "biotmle: Targeted Learning for Biomarker Discovery". In: *The Journal of Open Source Software* submitted (2017).

[7] Wolfgang Huber et al. "Orchestrating high-throughput genomic analysis with Bioconductor". In: *Nature methods* 12.2 (2015), pp. 115–121.

[8] Cliona M McHale et al. "Global gene expression profiling of a population exposed to a range of benzene levels". In: *Environmental health perspectives* 119.5 (2011), p. 628.

[9] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2017. URL: `https://www.R-project.org/`.

[10] Paul R Rosenbaum and Donald B Rubin. "The central role of the propensity score in observational studies for causal effects". In: *Biometrika* 70.1 (1983), pp. 41–55.

[11] Gordon K Smyth. "Limma: linear models for microarray data". In: *Bioinformatics and computational biology solutions using R and Bioconductor*. Springer, 2005, pp. 397–420.

[12] Gordon K Smyth. "Linear Models and Empirical Bayes Methods for Assessing Differential Expression in Microarray Experiments". In: *Statistical Applications in Genetics and Molecular Biology* 3.1 (2004), pp. 1–25.

[13] Anastasios Tsiatis. *Semiparametric theory and missing data*. Springer Science & Business Media, 2007.

[14] Catherine Tuglus and Mark J van der Laan. "Targeted methods for biomarker discovery". In: *Targeted Learning*. Springer, 2011, pp. 367–382.

[15] Mark J van der Laan, Eric C Polley, and Alan E Hubbard. "Super learner". In: *Statistical applications in genetics and molecular biology* 6.1 (2007).

[16] Mark J van der Laan and Sherri Rose. *Targeted learning: causal inference for observational and experimental data*. Springer Science & Business Media, 2011.