University of California, Berkeley

U.C. Berkeley Division of Biostatistics Working Paper Series

Year 2013 *Paper* 315

Adapting Data Adaptive Methods for Small, but High Dimensional Omic Data: Applications to GWAS/EWAS and More

Sara Kherad Pajouh[*] Alan E. Hubbard[†]

Martyn T. Smith[‡]

[*]UC Berkeley, kherad@berkeley.edu

[†]UC Berkeley, Division of Biostatistics, hubbard@berkeley.edu

[‡]University of California, Berkeley, martynts@berkeley.edu

Adapting Data Adaptive Methods for Small, but High Dimensional Omic Data: Applications to GWAS/EWAS and More

Sara Kherad Pajouh, Alan E. Hubbard, and Martyn T. Smith

Abstract

Exploratory analysis of high dimensional “omics” data has received much attention since the explosion of high-throughput technology allows simultaneous screening of tens of thousands of characteristics (genomics, metabolomics, proteomics, adducts, etc., etc.). Part of this trend has been an increase in the dimension of exposure data in studies of environmental exposure and associated biomarkers. Though some of the general approaches, such as GWAS, are transferable, what has received less focus is 1) how to derive estimation of independent associations in the context of many competing causes, without resorting to a misspecified model, and 2) how to derive accurate small-sample inference when data adaptive techniques are used in this context. This paper focuses on semi-parametric variable importance analysis of high dimensional data sets of modest sample size (e.g., gene expression, mRNA, etc). Though the methodology we propose is generally applicable to similar situations, we present the method in the context of a study of miRNA expression for an environmental exposure. Specifically, the analysis is faced with not just a large number of comparisons, but also trying to tease out of association of the expression of miRNA with an exposure apart from confounds such as age, race, smoking conditions, BMI, etc. Our goal is to propose a method that is reasonably robust in small samples, but does not rely on misspecified (arbitrary) parametric assumptions, and thus will be based on data-adaptive methods. The methodology proposed is we believe a powerful combination of existing semi-parametric statistical methods and theory, as well as a simple framework for use of commonly used empirical Bayes approaches to aid in small sample inference. Specifically, We propose using targeted maximum likelihood estimation (TMLE) for estimating variable importance measures along

with a general adaptation of the commonly used Limma approach, which relies on specification of the so-called influence curve of the proposed estimator. The result is a machine-based approach that can estimate independent associations in high dimensional data, but protects against the unreliability of small-sample inference that can result when using data adaptive estimation in relatively small samples.

1 Introduction

New methods are needed for evaluating both a rich variety of potential environmental exposures, along with the growing number of methods for producing biomarkers - so called more robust methods of Exposome or Environment Wide Association Studies (EWAS; Rappaport and Smith (2010) and Rappaport (2012)). There has been an effort to incorporate the types of methodologies used in similar data structures, for instance, Genome Wide Association Studies (GWAS; Manolio (2010)). For instance, Patel et al. (2010) has used fairly standard methods of parametric models with adjustments for multiple testing to rank potential environmental contributors to type 2 diabetes. However, as emphasized by Rappaport and Smith (2010), to thoroughly study the many facets of the exposome simultaneously, takes a complicated high dimensional situation and increases the dimensionality (more things to measure and associate with potential downstream effects of exposure). In this context, we will need data adaptive tools, that can be developed to target specific parameters related to the scientific questions posed by studying the exposome, avoiding bias automatically introduced by the use of parametric models. In this paper, we expand on an idea originally presented for an EWAS study of childhood exposures and later development, proposing a new combination of automated semiparametric estimation and inferential methods to assess the relative ”importance” of potential biomarkers in high dimensional, but relatively small data sets. The methodology can be adapted to almost any situation, in this case we discussed estimating variable importance in the context of a binary variable of interest (in our case exposure to a potential environmental toxin) and potentially many confounding variables. We present the method by analyzing miRNA expression array across subjects of different measured characteristics, but it obviously has the potential for use in for a very general data structure involving finding the impacts of specific factors among a large suite of candidates. However, one of the obvious dangers of data adaptive methods used in combination with very high dimensional data with relatively small samples sizes is the potential for spurious false positive findings, a potential epidemic which is occurring in the general field of high dimensional biology (e.g., see Ioannidis (2005)). However, we present a methodology that can at least avoid errors of analysis assuming a misspecified parametric model, and in a field of study where the dimension, and thus potential for false positives, will also grow without the use of more rigorous methods. Thus, we propose a semi-parametric (data-adaptive) method, that still can provide trustworthy robust inference, as well as the flexibility to target parameters that directly address the scientific question of interest.

Parametric or non-parametric (for simple comparisons across groups) methods have been widely studied in the literature by applying standard statistical tests perhaps with robust inference. When the estimation of the association of biomarkers involve adjustment of potential confounders, researchers have typically parametric regression models (e.g. Wang et al. (2012) and references therein). Though coefficients of linear model, multiple regression, or some other parametric methods have been considered natural candidates to measure the relationships of interest, these methods suffer from arbitrary assumptions (biased statistical models) of the data generating, since one typically knows very little in these circumstances about true probability distribution. In fact, though typically ignored, assuming an underlying model to make inference for the parameter may be misleading, claiming a certainty that does not exist. When the model is unknown, the dimension of the problem large, and the sample size small (often the case in studies of biomarkers), then data adaptive methods are warranted (nonparametric methods are impossible, and parametric methods unjustified). Such data adaptive methods, e.g., machine learning, avoids misspecification of the statistical model (since they are very large), but, at first blush, do not directly provide simple summary measures of the variable importance.

The methods of this paper are inspired by the goal of determining which measures (e.g., biomarkers) have the greatest *independent* association with some trait of interest, such as exposure to a potential toxin. In a simple motivating data set, we use differentially expressed miRNA's between two groups of exposed and non-exposed workers, by considering the all confounders of exposure in the model. As done in (Young et al. (2012)) the parameter of interest used is the average treatment (or exposure) effect (Holland (1986)) inspired by the causal inference literature, but can also be defined as a statistical parameter: the difference between conditional expectations of the outcome given different levels of treatment effect and covariates for each subject, averaged across all subjects (Bembom et al. (2008)). The particular estimator emphasized is a substitution estimator based on Targeted Maximum Likelihood Estimation (TMLE; Van der Laan and Rose (2011)).

Because the goal is to yield an algorithm that can be applied in automated fashion to find the relative importance of competing biomarkers in an unspecified statistical model, we concentrate on inferential methods for the TMLE estimates that are robust for small sample sizes, even when aggressive data adapted tools are used. As discussed below, one of the great advantages of TMLE is that the estimators can be so-called asymptotically linear (and thus normally distributed) with easy calculated standard errors (based on the influence curve). While these SE's are computationally inexpensive to derive, they can be unstable in small sample sizes. Thus, we introduce a general method of combining the empirical Bayesian method, (Smyth, 2005a), with influenced curve based inference, to derive more robust general methodology in finite sample sizes. The result is a computationally efficient, automated method for deriving the joint inference in a very large (semi parametric) statistical model of many related exposome measures.

In section 2, we introduce our propose methodology, by first motivating our methodology via a omic study of occupational exposure. We then outline the detailed steps of the method, from formally defining the data and the target parameter of interest, to estimating the statistical model using data-adaptive approaches (SuperLearner), to using the resulting prediction model to optimally estimate, via Targeted Maximum Likelihood, the variable importance measure.

We then present a generalization of Limma that can be used for the estimates of these type of variable importance measures. In section 3, we apply the proposed methodology to the example study, and we finish in section 4 with a discussion.

2 Methods

2.1 Data Structure

MicroRNAs (miRNA) are single-stranded RNA molecules of about 21-23 nucleotides in length, which regulate gene expression. In this study our aim is to find out the association of miRNA expression between an exposure in 18 subjects (20 samples, 2 technical replicates), from a study of occupational exposure in factories in Tianjin, China (McHale et al., 2011) and Lan et al. (2004). The miRNA expression was measured by Affymetrix GeneChip miRNA 2.0 microarrays based on 4592 probes specific to Homo Sapiens small RNAs, including miRNAs. Among the 18 subjects, 11 were exposed and the rest were unexposed controls. We considered 5 potential confounding factors: age, gender, BMI, cigarette smoking, and alcohol use.

In this example, our aim is to find the association of an environmental exposure on the gene expression of 4592 probes (miRNA) simultaneously and about some complicated situations and the challenging part which is due to the small sample sizes with existing of confounders in the model. This could be easily generalized to situations where one has a greater number of potential exposure biomarkers, as well as other confounding variables.

2.2 Defining the model and targeted parameter

The aim of analyzing these datasets is to rank the importance of a set of candidate biomarkers with regards to their independent association with the outcome (exposure in this case). In order to find a method for ranking biomarkers, we define a so-called variable importance measure (VIM) (Van der Laan and Rose, 2011). Let $O = (W, A, Y) \sim P_0$ represent a random variable regarding the observed data, where W are the corresponding confounders, A is the exposure of interest, Y is a vector of potential biomarkers ($Y = (Y_b, b = 1, .., B)$), and P_0 be the corresponding unknown probability distribution of the data. For our specific data set, $W = W_1, W_2, W_3, W_4, W_5$ where factor age (W_1) is continuous measure, gender (W_2) is binary, smoking situation (W_3) binary, BMI (W_4) is continuous measure and alcohol consumption (W_5) is a binary variable, A is binary exposure (yes=1, no=0), and Y_b are, one at a time, the miRNA expression values.

First, as a general definition of a parameter, consider $\Psi(P_0)$ be the target parameter according some function Ψ that maps the probability distribution P_0 in to the target feature of interest. The parameter $\Psi(P_0)$ is a function of the unknown probability distribution P_0. If P_n represents the empirical distribution of the $O_1, O_2, ..., O_n$, we are interested in the substitution estimators of the form $\Psi(P_n)$, that is we apply the same mapping but to the empirical distribution, to derive our estimate. Thus, we expand the parameters of interest beyond coefficients in a typically arbitrary (and certainly misspecified) parametric statistical model, to define a parameter as a feature of the true probability distribution P_0 of the data using true knowledge we have about P_0. Specifically, we propose here what we refer to as a targeted variable importance measure (VIM)(Bembom et al., 2008):

$$\Psi_b = \Psi_b(P_0) = \mathbb{E}_W[\mathbb{E}_0(Y_b|A=1,W) - \mathbb{E}_0(Y_b|A=0,W)]. \tag{1}$$

Note that under identifiability assumptions, such as no unmeasured confounding, that this parameter of the statistical model identifies a parameter of a causal model, specifically $\Psi(P_0)_b = E[Y_b(a=1) - Y_b(a=0)]$, or the differences in the mean of the biomarker in the population had everyone been exposed minus the mean had no one been exposed (Pearl, 2000). This is referred to as the average treatment effect (Rosenbaum and Rubin, 1983). The significance of this is that such parameters are not defined via a parametric model, and so one is free then to fit the model based on few if any assumptions; thus, one can use data-adaptive methods and still report the estimate of a relatively simple parameter.

2.3 Estimation

The targeted parameter is defined as a feature of the unknown probability distribution P_0 in Section 2.2, and the next step is to making inference about this parameter. There are general classes of estimators available for estimating Ψ. Here we focus on a substitution estimator (that is estimate 1 plugging estimates for Q and the empirical for $P_0(W)$), where the estimate of $Q_0^b(A,W) \equiv \mathbb{E}_0(Y_b|A,W)$ is based on TMLE (Targeted Maximum Likelihood Estimator) Gruber and Van Der Laan (2009). TMLE is a two stage estimation procedure, in which an initial estimate of Q_0 is updated (in the 2nd step) before deriving the substitution estimator,

$$\Psi(P_n)_b = \frac{1}{n}\sum_{i=1}^{n} Q_n^b(1,W_i) - Q_n^b(0,W_i) \tag{2}$$

where Q_n^b represents an estimate of Q_0^b.

2.3.1 SuperLearning

The first step in this two-stage estimator is to derive an initial estimate of Q_0^b, referred to as $Q_n^{(b,0)}$. One could, for instance, assume a parametric statistical model that results, for instance, 1 being equivalent to a coefficient (e.g., $Q_0^b(A, W) = \alpha^b + \beta_A^b A + \beta_W^b W$). However, given that 1 is defined in a nonparametric model, there's little sense in estimating with a misspecified regression model, when tools exists to data-adaptively estimate Q_0^b in a much bigger model. Specifically, given that the model Q_0^b is typically unknown, one should be able to derive less biased estimates by applying available machine learning algorithms, many of them imply a (near) nonparametric statistical model for Q^b.

However, this raises the issue of which data-adaptive algorithm is optimal, and no such theory exists for any one "learner". However, there is theory to support a procedure that examines a multitude of candidates. Specifically, the SuperLearner (SL) algorithm, which is a so-called stacking algorithm implemented as a cross-validated selector, that produces a estimate which is an optimally weighted (to minimize the cross-validated risk) the predictions from a set of candidate algorithms. These algorithms can very from very simple/smooth to highly data-adaptive (Van Der Laan et al., 2007).

Though the set of candidates is somewhat arbitrary, the theory developed behind SL (the Oracle Inequality) offers some guidance as to the type and number of candidates one should consider in the fitting routine. Specifically, if one of the candidate learners is actually the true model and converges at a parametric rate, the SL will converge at an close to parametric rate. In general, the SL will do, in first order, do as well (in terms of risk) as an algorithm that chooses the particular candidate learner based on knowing the true distribution (that is, the Oracle Selector), and this result holds as long as the number of candidates is on the order of a polynomial in sample size. Thus the SL theory encourages the use of a very large number of possible learning algorithms and ones that vary from very smooth, to aggressively data-adaptive. The SL algorithm is available as a statistical package (Polley and van der Laan, 2012) in the R programming language (Ihaka and Gentleman, 1996).

2.3.2 TMLE

The SL estimate of Q_0 is done to minimize the cross-validated risk, based on some loss-function (such as squared error, or -log likelihood), that is not the target of the analysis, which is to minimize the mean-squared error of Ψ_b. In addition, there is no guarantee that given a set of highly data-adaptive learners used by the SL, that the estimate of Ψ_b, particularly for a relatively small sample size, has a normal sampling distribution. Fortunately, Van der Laan and Rose (2011) introduced an estimator for Q_0, that "targets" the estimate of the regression, towards the particular parameter of interest, and also "smooths" the estimator such that the sampling distribution converges more quickly to a normal distribution. In addition, this targeting step can be thought of as a bias reduction step, given that data-adaptive selection represented by SL can result in an estimate of Ψ_b that suffers from residual confounding. This can occur, for instance, if selection of the variables in a procedure for estimating Q_0^b leaves out some regressors that are truly confounders of the association of A and Y. The resulting estimator is also more robust to model misspecification than the substitution estimator based on the initial SL fit, as well as being semi-parametrically locally efficient. For a full explanation of the theory behind TMLE and the formal justification of the estimator as most efficient among a class of estimators in a semiparametric model, see the appendix of Van der Laan and Rose (2011).

Algorithmically, the TMLE estimator in this context is a simple one-dimensional augmentation of the initial fit. Specifically, to the SL fit, if the outcome is continuous, one fits a simple, one-dimensional regression of type:

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

where the initial fit, $Q_n^{(b,0)}(A,W)$ is treated as an offset, and $h_{\hat{g}}(A,W)$ is a covariate:

$$h_{\hat{g}}(A,W) = \frac{I(A=1)}{\hat{g}(1|W)} - \frac{I(A=0)}{\hat{g}(0|W)}$$

where $\hat{g}(1|W)$ is an estimate of the $P(A=1 \mid W)$ or the propensity score (Rosenbaum and Rubin, 1983). The selection of $\hat{g}$ can also be made by a process that minimizes the mean-squared error of the parameter of interest (Gruber and Van

Der Laan, 2010), but in this case, we use simple main terms logistic regression. Finally, one derives the final TMLE estimate of the Ψ_b using the augmented estimate of Q, or

$$\hat{\Psi}_b(P_n) = \frac{1}{n}\sum_{i=1}^{n}[Q_n^{(b,1)}(1, W_i) - Q_n^{(b,1)}(0, W_i)]. \tag{3}$$

2.4 Inference

As shown in Van der Laan and Rose (2011), $\hat{\Psi}_b(P_n)$ of Ψ_b is asymptotically linear with influence curve $IC(O)$ if it satisfies

$$\sqrt{n}(\Psi_b(P_n) - \Psi_b(P_0)) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} IC(O_i) + o_p(1). \tag{4}$$

Thus, in other words, the variance of $\hat{\Psi}_b(P_n)$ is well approximated by sample variance of the influence curve divided by the sample size, n. In this case, the plug-in influence curve (IC) for the ATE is:

$$\begin{aligned} IC_{b,n}(O_i) = &\left(\frac{I(A_i = 1)}{g_n(1|W_i)} - \frac{I(A_i = 0)}{g_n(0|W_i)}\right)(Y_{b,i} - Q_n^{(b,1)}(A_i, W_i)) + Q_n^{(b,1)}(1, W_i) \\ &- Q_n^{(b,1)}(0, W_i) - \Psi_b(P_n). \end{aligned} \tag{5}$$

Finally, we can derive asymptotic p-values and confidence intervals using a Wald type approach, or

$$\text{pvalue} = 2\left[1 - \Phi(\frac{|\Psi_b(P_n)|}{\sigma_n^b/\sqrt{n}})\right]$$
$$95\ \%\ \text{CI} = \Psi_b(P_n) \pm 1.96\sigma_n^b/\sqrt{n},$$

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

2.4.1 Limma Applied to Influence Curve-Based Inference

By applying directly the TMLE estimator of targeted parameter, using data adaptive methods in small samples, one will often face high unstable standard error estimation, so that the joint inference of this targeted variable importance measure can result in erroneously significant biomarkers. Thus, we want a technique that asymptotically provides the correct inference, but in small samples, can "borrow" information across the many estimates of sampling variability (the σ_n^b), to provide more robust finite sample inference. Fortunately, an established empirical Bayes technique, developed for high dimensional situations (such as one involving 1000's of candidate biomarkers) accomplishes this twin goals. Specifically, we propose the *Limma* approach for improving the finite sample performance for the inference about targeted parameter. Limma is package in R which use linear model for microarray data by using empirical Bayes (Smyth et al., 2004). Limma borrows information across all genes and make more stable and robust inference for microarray data (Smyth, 2005a). In Section 2 we explained that a common way of making inference about the targeted parameter Ψ_b, (for the enough sample size), is to find the influence curve values for Ψ_b. Then by calculating the corresponding standard errors of the influence curve of Ψ_b and finding the corresponding p-values based on that and making inference about Ψ_b for each probe. We used Limma package for the applying empirical Bayes inference for all probes based on (Smyth, 2005b) as the following steps:

- Find influence curve for each probe, and find the corresponding matrix of influence curve for all subjects and all probes.

- Add the $\Psi_b(P_n)$ estimate to each row of the mean zero influence curve matrix, to get a row that has as simple sample average, $\Psi_b(P_n)$, but whose variance is the sample variance of the IC_b.

- Perform Limma on this matrix (assuming estimating the simple mean),

- Use the resulting inference, based on the shrinkage estimate of the sampling standard deviation of the influence curve, $\tilde{\sigma}_n^b$ which 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, or $\overline{\sigma_n^b} \approx \frac{1}{B}\sum_{b=1}^{B}\sigma_n^b$, or $\widetilde{\sigma}_n^b = wt_b\sigma_n^b + (1 - wt_b)\overline{\sigma_n^b}$, where the wt_b is between (0,1). see Smyth (2005a) for formal presentation. As $n \to \infty, wt_b \to 1$, and so $\widetilde{\sigma}_n^b \to \sigma_n^b$

- Use multiple testing corrections to make inference about all probes. Here we used False Discovery Rate (FDR) correction.

Thus, this procedure will shrink aberrant values variability estimates toward the center of the distribution, particularly if the sample sizes is relative small. The practical effect is that it tends to reduce the number of significant biomarkers, driven by (potentially) erroneous underestimate of variation of the parameter estimates of interest, $\Psi_b(P_n)$.

3 Data Analysis

3.1 miRNA

For the miRNA dataset in 2.1, we first applied a standard, linear model approach to get the VIM, using Limma as well. We show these results in Table 3.1.1, where for each method we calculated the parameter estimate (coefficient correspond to exposure in linear model), standard errors of the coefficient, original p-values and the corresponding adjusted p-values or q-values based on BH correction (Benjamini and Hochberg, 1995) for top 10 more important biomarkers.

Then we made inference about miRNA dataset based on the semiparametric method using TMLE 2.3.2 as well as adapting of empirical Bayes Limma method, that introduced in chapter 2.4.1. We used TMLE package in R which is introduced by Gruber and van der Laan (2012), and we used stepwise generalized linear models and choose the best model in a stepwise algorithm. If one had a relatively larger sample size, then a larger set of more flexible algorithms would be added.

As in the case for the simple linear model, we performed BH multiple testing procedure to find the significant probes. We show the results in Table 3.1.1 where we calculated the estimation of targeted parameter, standard error of parameter, original p-values and the corresponding q-values for both methods, after BH corrections for all probes. the results are shown for top 10 most significant probes.

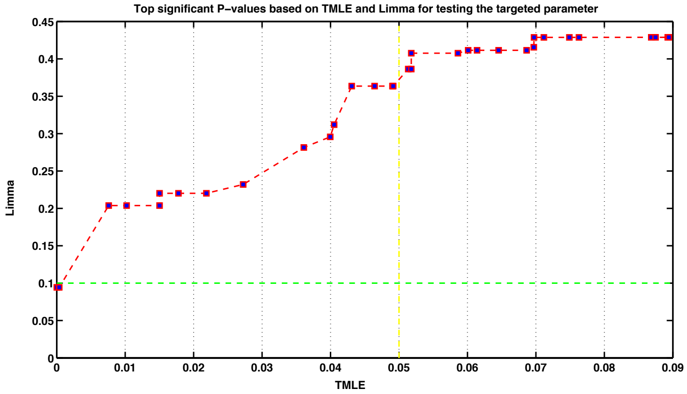

Figure 1: *P-values of top 10 significant probes according to TMLE and Limma for IC of* Ψ

3.1.1 Results

The results show that applying parametric linear regression model and also Limma for testing the corresponding parameter gives no significant Probes and inference based on TMLE estimation there are 23 significant probes at 5% of significant level and by applying Limma package on the matrix of influence curve added by estimation of Ψ_b effect, there are only 2 significant probes at 10% of significant level. In Figure 3.1.1 we compare the significant probes based on two methods. The results suggest that the shrinkage estimator using Limma for targeted parameter, provides more conservative inference for the targeted parameter, relative than directly using just σ_n^b based only on the IC within one biomarker, b.

As expected, Limma reduced the spread of the standard deviation estimates of the IC ($\widetilde{\sigma}_n^b$) across the 4952 probes, and the corresponding Wald-statistics for testing the targeted parameter, in comparison of using the original standard error, σ_n^b. In Figure 3.1.1 we compare the boxplots of Wald-statistics between two methods for top significant probes. Thus using directly TMLE for the small sample size of subjects leads many significant probes that may comes from variability of standard errors estimated by influence curve and adapting TMLE by Limma will reduce the variability of standard errors and reduce the number of significant probes which leads to the more robust inference.

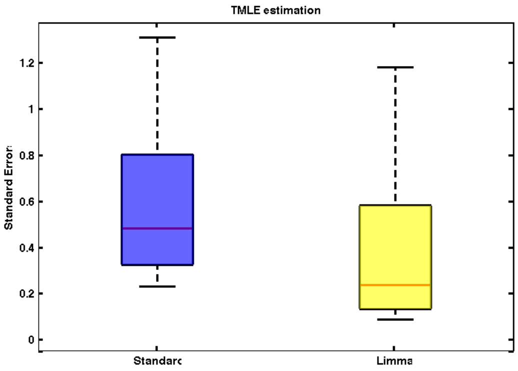

Figure 2: *Comparison of variability of standard errors for top 10 significant probes, according to TMLE estimation based on TMLE (Standard) and TMLE (Limma)*

3.2 Genomic Example

We performed simulations using mRNA data generated from subjects in the benzene occupational study, 59 shoe factory workers exposed to benzene (1<ppm) and 42 healthy unexposed control clothing factory workers McHale et al. (2011). First, we calculated statistical inference for all subjects using both methods. According to TMLE, differential expression of 2188 probes was found to be significant. According to Limma, 2086 significant probes were found after multiple test correction FDR (BH), treating these set as the set of "true" differentially expressed probes.We then randomly sampled from the list of arrays and re-did the analysis. This was repeated randomly as different sample for equal numbers of exposed vs. controls ($n = 8, 15, 20, 25, 30$) and, we report the performance (overlap) relative to the set found with all original samples (59 exposed, 42 controls).

Table 1: *Top 10 table of probes listing for testing the effect of exposure based on Linear model estimation according to standard Linear model and Limma inference. SE, p and q are corresponding to standard error of regression parameter (Ψ_{LM}) and corresponding p-values and q-values (adjusted p-values after BH correction)*

	Linear Model (standard)				Linear Model (Limma)			
	$\hat{\Psi}_{LM}$	SE	p	q	$\hat{\Psi}_{LM}$	SE	p	q
Prob 1	0.36	0.08	1×10^{-3}	>0.10	0.36	0.04	1×10^{-3}	>0.10
Probe 2	0.55	0.13	1×10^{-3}	>0.10	0.55	0.07	1×10^{-3}	>0.10
Probe 3	0.25	0.06	1×10^{-2}	>0.10	0.25	0.03	1×10^{-2}	>0.10
Probe 4	0.35	0.09	1×10^{-2}	>0.10	0.35	0.05	1×10^{-2}	>0.10
Probe 5	0.45	0.13	1×10^{-2}	>0.10	0.45	0.07	1×10^{-2}	>0.10
Probe 6	-1.14	0.33	1×10^{-2}	>0.10	-1.14	0.42	1×10^{-2}	>0.10
Probe 7	0.40	0.11	1×10^{-2}	>0.10	0.40	0.06	1×10^{-2}	>0.10
Probe 8	-0.68	0.20	1×10^{-2}	>0.10	-0.68	0.16	1×10^{-2}	>0.10
Probe 9	0.83	0.24	1×10^{-2}	>0.10	0.83	0.23	1×10^{-2}	>0.10
Probe 10	0.70	0.22	1×10^{-2}	>0.10	0.70	0.18	1×10^{-2}	>0.10

3.2.1 Results

In Table 3.2.1 the simulation is based on one draw for each sample size and shows that at small sample sizes the number of significantly differentially expressed probes decreases dramatically using the standard IC-based inference, versus adding the Limma approach, which also means a much lower false positive rate. This difference continues at larger sample sizes, but not as dramatically, as difference of Limma versus standard estimates of the IC's by probe diminishes. Though not a simulation from a known true distribution, this shows the potential danger of using IC-based inference with aggressive data-adaptive procedures at small sample sizes, and the benefit of Limma to make these inferences much more robust.

Table 2: *Top 10 table of probes listing for testing the effect of exposure based on Targeted Maximum Likelihood Estimation according to TMLE and TMLE adjusted by Limma inference . SE, p and q are corresponding to both the non-altered initial sampling standard deviation of the IC, σ_n^b, and that corresponding to the Limma estimate, $\widetilde{\sigma}_n^b$, p-values, and q-values (adjusted p-values after BH correction)*

	TMLE (σ_n^b)				TMLE (Limma; $\widetilde{\sigma}_n^b$)		
	$\hat{\Psi}_{TMLE}$	SE	p	q	SE	p	q
Probe 1	0.36	0.29	1×10^{-9}	<0.05	0.11	1×10^{-5}	<0.10
Probe 2	0.56	0.47	1×10^{-7}	<0.05	0.23	1×10^{-5}	<0.10
Probe 3	0.24	0.23	1×10^{-6}	<0.05	0.09	1×10^{-4}	>0.10
Probe 4	0.33	0.33	1×10^{-6}	<0.05	0.13	1×10^{-4}	>0.10
Probe 5	0.47	0.49	1×10^{-5}	<0.05	0.24	1×10^{-4}	>0.10
Probe 6	-1.23	1.31	1×10^{-5}	<0.05	1.28	1×10^{-4}	>0.10
Probe 7	0.40	0.43	1×10^{-5}	<0.05	0.20	1×10^{-4}	>0.10
Probe 8	-0.76	0.80	1×10^{-5}	<0.05	0.58	1×10^{-4}	>0.10
Probe 9	0.84	0.92	1×10^{-4}	<0.05	0.75	1×10^{-4}	>0.10
Probe 10	0.71	0.77	1×10^{-4}	<0.05	0.54	1×10^{-4}	>0.10

4 Discussion

The goal of this paper is to introduce an automated, robust method for analyzing high dimensional exposure and omic data with relatively modest sample sizes. In the example used, the challenge was not just from a large number of comparisons (potential biomarkers), but involve adjusting for potential confounders, all in the context of a large statistical model and small numbers of biological replicates. Given the goal is estimation within an unknown statistical model, the technique must involve data adaptive estimation, though the goal is to still provide trustworthy statistical inference and estimators that are based on semiparametric efficiency theory. That is, given the parameter of interest and the statistical model, the choices guiding the algorithm should not be ad hoc, but based on the relative efficiency of competing estimators. We have proposed methods that take the existing work on GWAS

Table 3: *Number of significants Probes after BH correction for the mRNA Data sets using TMLE and adapting by Limma. n is the number of sub-samples from each of exposed and non-exposed samples. FP TMLE shows the number of false positive Probes in TMLE method and FP Limma shows the number of false positive Probes in Limma.*

n	TMLE	FP TMLE	Limma	FP Limma
8	373	296	25	12
15	1136	312	771	159
20	1353	295	1047	168
25	1357	178	1159	120
30	2009	532	1741	403

and merge them with existing proposals for variable importance, and techniques for deriving empirical Bayes estimates of the sampling variance in the context of high dimensional data, for an automated procedure that can data adaptively list the most promising biomarkers among similar study designs.

We illustrated the method using an example miRNA versus benzene exposure by applying, on a probe by probe basis, TMLE/SuperLearning to estimate the association of each potential marker with exposure. The estimation was conducted using a combination of SuperLearning and TMLE. In addition, as has been proposed before, we present a simple way to generalize the Limma approach (Smyth, 2005a), to robustify small sample inference, that relies on deriving the influence curve of the estimator of the parameter of interest. The results suggest that one can ameliorate unstable small-sample inference by combining this asymptotically efficient estimator (TMLE) with Limma; in our data example, this results in fewer statistically significant biomarkers. In addition, because Limma has no impact on the asymptotics (the adjustment to the within probe inference becomes negligible as sample size grows), we can rely on the asymptotic theory developed for TMLE.

This combination of existing methods offers many advantages: 1) it estimates parameters relevant to scientific question, in context of confounders, making no assumptions about the statistical model, 2) it uses theoretical optimality of loss-based estimation (Oracle selector) via the SuperLearner algorithm, which does an optimal job balancing the variance-bias trade off in small samples by automatically choosing a level of parsimony to match the information available in the sample, 3) TMLE estimators reduce residual bias and add an appropriate amount of smoothing, along with the availability of influence curve based inference unavailable for the straight substitution estimator, 4) robustifies the inference by using an empirical Bayes approach (Limma) to derive joint inference with less false positives resulting from poor estimation of the sampling variability. The result is a theory-driven, data-adaptive procedure based on pre-specified algorithm with robust statistical inference. Though the continuing development of new technologies promises new insights into the relationship of biomarkers, and human health and disease, this procedure helps to ameliorate the pitfalls of increasing the dimension of the data (and thus avoiding a commiserate increase in false positive findings) by creating a rigorous statistical procedure for discovery.

References

Bembom, O., Petersen, M., Rhee, S., Fessel, W., Sinisi, S., Shafer, R., Van Der Laan, M., 2008. Biomarker discovery using targeted maximum-likelihood estimation: Application to the treatment of antiretroviral-resistant hiv infection. Statistics in medicine 28 (1), 152–172.

Benjamini, Y., 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), 289–300.

Gruber, S., Van Der Laan, M., 2009. Targeted maximum likelihood estimation: A gentle introduction.

Gruber, S., Van Der Laan, M. J., 2010. An application of collaborative tmle. The International Journal of Biostatistics.

Gruber, S., van der Laan, M. J., 2012. tmle: An R package for targeted maximum likelihood estimation. Journal of Statistical Software 51 (13), 1–35.
URL http://www.jstatsoft.org/v51/i13/

Holland, P. W., 1986. Statistics and causal inference. Journal of the American Statistical Association 81 (396), pp. 945–960.
URL http://www.jstor.org/stable/2289064

Ihaka, R., Gentleman, R., 1996. R: A language for data analysis and graphics. Journal of computational and graphical statistics 5 (3), 299–314.

Ioannidis, J. P. A., 08 2005. Why most published research findings are false. PLoS Med 2 (8), e124.
URL http://dx.doi.org/10.1371%2Fjournal.pmed.0020124

Lan, Q., Zhang, L., Li, G., Vermeulen, R., Weinberg, R., Dosemeci, M., Rappaport, S., Shen, M., Alter, B., Wu, Y., et al., 2004. Hematotoxicity in workers exposed to low levels of benzene. Science 306 (5702), 1774–1776.

Manolio, T. A., 2010. Genomewide association studies and assessment of the risk of disease. New England Journal of Medicine 363 (2), 166–176, pMID: 20647212.
URL http://www.nejm.org/doi/full/10.1056/NEJMra0905980

McHale, C., Zhang, L., Lan, Q., Vermeulen, R., Li, G., Hubbard, A., Porter, K., Thomas, R., Portier, C., Shen, M., et al., 2011. Global gene expression profiling of a population exposed to a range of benzene levels. Environmental health perspectives 119 (5), 628.

Patel, C. J., Bhattacharya, J., Butte, A. J., 2010. An environment-wide association study (ewas) on type 2 diabetes mellitus. PLoS One 5 (5), e10746.

Pearl, J., 2000. Causality: models, reasoning and inference. Vol. 29. Cambridge Univ Press.

Polley, E., van der Laan, M., 2012. SuperLearner: Super Learner Prediction. R package version 2.0-9.
URL http://CRAN.R-project.org/package=SuperLearner

Rappaport, S. M., 2012. Biomarkers intersect with the exposome. Biomarkers 17 (6), 483–489.

Rappaport, S. M., Smith, M. T., 2010. Environment and disease risks. Science(Washington) 330 (6003), 460–461.

Rosenbaum, P. R., Rubin, D. B., 1983. The central role of the propensity score in observational studies for causal effects. Biometrika 70 (1), 41–55.

Smyth, G., 2005a. Limma: linear models for microarray data. Bioinformatics and computational biology solutions using R and Bioconductor, 397–420.

Smyth, G., et al., 2004. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. Statistical applications in genetics and molecular biology 3 (1), 3.

Smyth, G. K., 2005b. Limma: linear models for microarray data. New York.

Van Der Laan, M., Polley, E., Hubbard, A., 2007. Super learner. Statistical Applications in Genetics and Molecular Biology 6 (1), 1–21.

Van der Laan, M., Rose, S., 2011. Targeted Learning: Causal Inference for Observational and Experimental Data. Springer.

Wang, X., Lin, Y., Song, C., Sibille, E., Tseng, G., et al., 2012. Detecting disease-associated genes with confounding variable adjustment and the impact on genomic meta-analysis: With application to major depressive disorder. BMC bioinformatics 13 (1), 52.

Young, J., Hubbard, A., Eskenazi, B., Jewell, N., 2012. Machine-learning algorithm for estimating and ranking the impact of environmental risk factors in exploratory epidemiological studies. In: Oakes, D., Hall, W. J., Almudevar, A. (Eds.), Modeling and Inference in Biomedical Sciences: In Memory of Andrei Yakovlev. IMS Collections Series, Institute of Mathematical Statistics, Beachwood, Ohio.