methyvim: Targeted, robust, and model-free differential methylation analysis in R

Nima S. Hejazi*[1,2], Rachael V. Phillips[1], Alan E. Hubbard[1], and Mark J. van der Laan[1,3]

[1]**Group in Biostatistics, University of California, Berkeley**
[2]**Center for Computational Biology, University of California, Berkeley**
[3]**Department of Statistics, University of California, Berkeley**

Supplementary File 1

Data Structure

We consider an observed data structure, on a single experimental subject (e.g., a patient), $O = (W, A, (Y(j) : j))$, where $(Y(j) : j = 1, \ldots J)$ is the set of CpG sites measured by the assay in question, $A \in \{0, 1\}$ represents a binary phenotype-level treatment, and W is a vector of the phenotype-level baseline covariates that are potential confounders (e.g., age, sex). We consider having access to measurements on a large number J of CpG sites (e.g., 850,000, as measured by the Illumina *MethylationEPIC* BeadChip arrays). Further, let $K : j \rightarrow S_j$ be a procedure that assigns to a given CpG site j a set of neighbors S_j – that is, S_j is a collection of indices of the neighbors of j just as j indexes the full set of CpG sites; moreover, since $Y(j)$ is the measured methylation value for a given CpG site j, $Y(S_j)$ is the measured methylation values at the neighbors S_j of j. Note that the definition of a neighborhood $\{j, S_j\}$ is left vague so as to facilitate the use of user-specified strategies in implementation. We consider the case of observing n iid copies of O (i.e., $O_1, \ldots, O_n$), where $O \sim P_0 \in \mathcal{M}$, which is to say that the random variable O is governed by an unknown probability distribution P_0, assumed only to reside in a nonparametric statistical model $\mathcal{M}$ that places no restrictions on the data-generating process.

Variable Importance Measure

With the data structure above in hand, we let the estimand of interest be a j-specific variable importance measure (VIM) $\Psi_j(P_0)$, which is defined through the *true* (and unknown) probability distribution P_0. As a motivating example, consider the case where $\Psi_j(P_0)$ is

$$\Psi_j(P_0) = \mathbb{E}_{P_0}(\mathbb{E}_{P_0}(Y(j) \mid A = 1, W, Y(S_j)) - \mathbb{E}_{P_0}\mathbb{E}_{P_0}(Y(j) \mid A = 0, W, Y(S_j)), \quad (1)$$

where $Y(S_j)$ is the subvector of Y that contains the measured methylation values of the *neighboring sites* of j (but not site j itself). This parameter of interest is a variant of the *average treatment effect*, which has been the subject of much attention in statistical causal inference [1, Hahn [2], Hirano et al. [3]]. As a measure of variable importance, we seek to estimate this target parameter ($\Psi_j(P_0)$), which quantifies the effect of changing a treatment A on the methylation $Y(j)$ of a CpG site j, accounting for any potential confounding from phenotype-level covariates W and observed methylation $Y(S_j)$ at the neighboring sites S_j of j. Importantly, the use of such well-defined parameters as variable importance measures allows the construction of inferential procedures (e.g., confidence intervals and hypothesis tests) for the estimate [4]. We propose a procedure that estimates this target parameter $\Psi_j(P_0)$ across all CpG sites of interest $j : 1, \ldots, J$. When data-adaptive regression procedures (i.e., machine learning) are employed in estimating $\Psi_j(P_0)$, this produces a nonparametric variable importance measure of differential methylation, allowing for the identification of differentially methylated positions (DMPs) while avoiding many assumptions common in the use of standard parametric regression procedures. We propose estimating $\Psi_j(P)$ via targeted minimum loss-based estimation, which allows for data-adaptive regression procedures to be employed in a straightforward manner using the Super Learner algorithm for the construction of ensemble models [5, Wolpert [6], Breiman [7], Gruber and van der Laan [8], van der Laan and Rose [9]].

Pre-Screening Procedure

As a matter of practicality, we consider only estimating the target VIM $\Psi_j(P_0)$ at a subset of CpG sites, so that $j : 1, \ldots J$ does not, in fact span the full set of assayed CpG sites. In order to determine this subset of CpG sites, we propose the use of a pre-screening procedure, which need be nothing more than a method for differential methylation analysis that is computationally less demanding than the method proposed here. Formally, let the pre-screening procedure $\theta(j) : j \rightarrow \{0, 1\}$, which is to say that θ takes as input a CpG site j and returns a binary decision rule of whether to include CpG site j in a subset to be evaluated further or not. As an example, one might consider employing a classical differential methylation analysis procedure, such as the linear modeling approach of the `limma` R package [10, Robinson et al. [11]], using only CpG sites that pass a reasonable cutoff based on the linear model (e.g., magnitude of t-statistics, p-values) for the subsequent analysis steps in the `methyvim` pipeline.

*`nhejazi@berkeley.edu`

Reducing Neighbors via Clustering

Due to the data-adaptive nature of the regression procedures employed in evaluating the target VIM $\Psi_j(P_0)$ via targeted minimum loss-based estimation, it is possible that a given CpG site j may have *too many* neighbors S_j to be controlled for in the estimation procedure. Heuristically, the inclusion of too many neighbors when controlling for potential confounders may lead to instability in the estimates produced. In such cases, we propose and implement the use of a clustering technique (e.g., partitioning around medoids) to select a *representative* subset of neighbors. Formally, there is likely no best choice of a specific clustering algorithm, so we leave this aspect of the proposed algorithm as flexible for the user [12]. Note that the goal of employing a clustering procedure in `methyvim` is to obtain a smaller but still highly representative set of neighbors $S(j)$, so as to allow for estimates of $\Psi_j(P_0)$ to account for as much confounding from neighboring sites as is allowed by the available data.

Targeted Minimum Loss-Based Estimation and Statistical Inference

Given the choice of target parameter $\Psi_j(P_0)$, we propose the use of targeted minimum loss-based estimation (TMLE) to construct and evaluate an estimator $\psi_{n,j}$ of $\Psi_j(P_0)$. In the case of our motivating example, where the target VIM is based on the average treatment effect, we make use of a TML estimator of this parameter, which has been implemented for a general case in the `tmle` R package [13]; users of `methyvim` may wish to consult the documentation of that software package as a supplement. Generally speaking, a TML estimator is constructed from a few simple components: (1) an estimator of the propensity score [14], often denoted $g(A \mid W)$; (2) an estimator of the outcome regression, often denoted $\bar{Q}(A, W)$; and (3) a targeting step that revises the initial estimators of the aforementioned components such that the resultant estimator satisfies a set of score-like equations [8]. While we describe a few of the key properties of TML estimators in the sequel, extended technical discussion is deferred to more comprehensive work [9, van der Laan and Rose [15]]. In order to construct an estimate $\psi_{n,j}$ of $\Psi_j(P_0)$, it is necessary to accurately estimate two nuisance parameters, these being the propensity score ($g(A \mid W)$) and the outcome regression ($\bar{Q}(A, W)$); moreover, data-adaptive regression procedures may be used to obtain consistent estimates of these quantities through the creation of a stacked regression model via the Super Learner algorithm [6, Breiman [7], van der Laan et al. [5]], which ensures that the resulting ensemble model satisfies important optimality properties with respect to cross-validation [16, Dudoit and van der Laan [17], van der Laan and Dudoit [18]]. To use our running example, one would construct a TML estimator of the average treatment effect (ATE) as a variable importance measure in the following short series of steps:

1. Construct initial estimates of the propensity score $g(A \mid W)$ and outcome regression $\bar{Q}(A, W)$. As noted previously, the Super Learner algorithm may be employed to construct such estimates.
2. Update these initial estimates (e.g., iteratively) so as to solve the efficient influence function estimation equation of the parameter of interest relative to a nonparametric statistical model.
3. In the efficient influence function for the ATE, $\text{EIF}(P_0)(O) = H_n(Y - \bar{Q}(A, W))$, the auxiliary term H_n may be denoted as $H_n = \frac{\mathbb{I}(A=1)}{g(1|W)} - \frac{\mathbb{I}(A=0)}{g(0|W)}$. Note that the exact form of the efficient influence function and auxiliary covariate varies with the target parameter of interest.
4. The TML estimator of the parameter of interest is constructed by a procedure that updates the initial estimators such that the empirical mean of the efficient influence function estimating equation(s) is nearly zero (with respect to an error proportional to the sample size, i.e., $\frac{1}{\sqrt{n}}$).

Importantly, TML estimators are well-suited for statistical inference, having an asymptotically normal limiting distribution [19, van der Laan and Rubin [20]], which allows for a closed-form expression of the variance to be derived:

$$\sqrt{n}(\psi_{n,j} - \Psi_j(P_0)) \sim N(0, \sigma^2_{\text{EIF}}), \tag{2}$$

where σ^2_{EIF} is the variance of the efficient influence function, with respect to a nonparametric statistical model, of the target parameter evaluated at the observed data. Such a convenience allows for confidence intervals and hypothesis tests to be constructed (i.e., $H_{0,j} : \Psi_j(P_0) = 0$) in a straightforward manner. What's more, under standard regularity conditions, and with consistent estimation of the propensity score and outcome regression, the TML estimator is asymptotically efficient and achieve the lowest possible variance among a large class of estimators [21].

Correction for Multiple Testing

Given that we seek to estimate $\Psi_j(P_0)$ for a possibly large number of CpG sites ($j : 1, \ldots J$), the need to perform corrections for multiple testing is clear. In order to curb the potential for false discoveries, we recommend the use of the Benjamini and Hochberg procedure for controlling the False Discovery Rate (FDR) [22]; however, as the proposed procedure involves a pre-screening step, naive application of the Benjamini and Hochberg procedure (BH) is invalid – instead, we rely on a modification of the procedure to control the FDR, with established theoretical guarantees when pre-screening is employed [23]. In brief, the modified marginal Benjamini and Hochberg procedure to control the FDR under pre-screening works by applying the standard BH procedure to a padded vector of p-values – that is, letting J be the number of CpG sites tested and K the number of CpG sites filtered out (so that $J + K = P$, where P is the original dimension of the genomic assay), the modified marginal BH procedure is the application of the original BH procedure to a vector of p-values, composed of the J p-values from performing J hypothesis tests (of the form $H_{0,j} : \Psi_j(P_0) = 0$) and K additional p-values automatically set to 1 (for the hypothesis tests not performed on account of pre-screening). This procedure is guaranteed to control the FDR at the same desired rate when pre-screening is performed whereas naive application of the BH procedure fails to do so.

References

[1] Paul W Holland. Statistics and causal inference. *Journal of the American statistical Association*, 81(396):945–960, 1986.

[2] Jinyong Hahn. On the role of the propensity score in efficient semiparametric estimation of average treatment effects. *Econometrica*, pages 315–331, 1998.

[3] Keisuke Hirano, Guido W Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

[4] Mark J van der Laan. Statistical inference for variable importance. *The International Journal of Biostatistics*, 2(1), 2006.

[5] Mark J van der Laan, Eric C Polley, and Alan E Hubbard. Super Learner. *Statistical applications in genetics and molecular biology*, 6(1), 2007.

[6] David H Wolpert. Stacked generalization. *Neural networks*, 5(2):241–259, 1992.

[7] Leo Breiman. Stacked regressions. *Machine learning*, 24(1):49–64, 1996.

[8] Susan Gruber and Mark J van der Laan. Targeted maximum likelihood estimation: A gentle introduction. 2009.

[9] Mark J van der Laan and Sherri Rose. *Targeted Learning: Causal Inference for Observational and Experimental Data*. Springer Science & Business Media, 2011.

[10] Gordon K Smyth. Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1):1–25, January 2004. doi: 10.2202/1544-6115.1027. URL `https://dx.doi.org/10.2202/1544-6115.1027`.

[11] Mark D Robinson, Abdullah Kahraman, Charity W Law, Helen Lindsay, Malgorzata Nowicka, Lukas M Weber, and Xiaobei Zhou. Statistical methods for detecting differentially methylated loci and regions. *Bioinformatics and Computational Biology*, 5:324, 2014.

[12] Jon M Kleinberg. An impossibility theorem for clustering. In *Advances in neural information processing systems*, pages 463–470, 2003.

[13] Susan Gruber and Mark J van der Laan. tmle: An R package for targeted maximum likelihood estimation. 2011.

[14] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1): 41–55, 1983.

[15] Mark J van der Laan and Sherri Rose. *Targeted Learning in Data Science: Causal Inference for Complex Longitudinal Studies*. Springer Science & Business Media, 2018.

[16] Mark J van der Laan, Sandrine Dudoit, and Sunduz Keles. Asymptotic optimality of likelihood-based cross-validation. *Statistical Applications in Genetics and Molecular Biology*, 3(1):1–23, 2004.

[17] Sandrine Dudoit and Mark J van der Laan. Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. *Statistical Methodology*, 2(2):131–154, 2005.

[18] Mark J van der Laan and Sandrine Dudoit. Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples. 2003.

[19] Anastasios Tsiatis. *Semiparametric Theory and Missing Data*. Springer Science & Business Media, 2007.

[20] Mark J van der Laan and Daniel Rubin. Targeted maximum likelihood learning. *The International Journal of Biostatistics*, 2(1), 2006.

[21] Peter J Bickel, Chris AJ Klaassen, Ya'acov Ritov, and Jon A Wellner. *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins University Press Baltimore, 1993.

[22] Yoav Benjamini and Yosef Hochberg. 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, 1995.

[23] Catherine Tuglus and Mark J. van der Laan. Modified FDR controlling procedure for multi-stage analyses. *Statistical Applications in Genetics and Molecular Biology*, 8(1):1–15, January 2009. doi: 10.2202/1544-6115.1397. URL `https://dx.doi.org/10.2202/1544-6115.1397`.