methyvim: Targeted, Robust, and Model-free Analysis of Differential Methylation 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**

Abstract We present `methyvim`, an R package implementing an algorithm for the nonparametric estimation of the effects of exposures on DNA methylation at CpG sites throughout the genome, complete with straightforward statistical inference for such estimates. The approach leverages variable importance measures derived from statistical parameters arising in causal inference, defined in such a manner that they may be used to obtain targeted estimates of the relative importance of individual CpG sites with respect to a binary treatment assigned at the phenotype level, thereby providing a new approach to identifying differentially methylated positions. The procedure implemented is computationally efficient, incorporating a preliminary screening step to isolate a subset of sites for which there is cursory evidence of differential methylation as well as a unique multiple testing correction to control the False Discovery Rate with the same rigor as would be available if all sites were subjected to testing. This novel technique for analysis of differentially methylated positions provides an avenue for incorporating flexible state-of-the-art data-adaptive regression procedures (i.e., machine learning) into the estimation of differential methylation effects without the loss of interpretable statistical inference for the estimated quantity.

Keywords

DNA methylation, differential methylation, epigenetics, causal inference, statistical variable importance, machine learning, targeted minimum loss-based estimation

[*]`nhejazi@berkeley.edu`

Introduction

DNA methylation is a fundamental epigenetic process known to play an important role in the regulation of gene expression. DNA methylation mostly commonly occurs at CpG sites and involves the addition of a methyl group (CH_3) to the fifth carbon of the cytosine ring structure to form 5-methylcytosine. Numerous biological and medical studies have implicated DNA methylation as playing a role in disease and development [1]. Perhaps unsurprisingly then, biotechnologies have been developed to rigorously probe the molecular mechanisms of this epigenetic process. Modern assays, like the Illumina *Infinium* HumanMethylation BeadChip assay, allow for quantitative interrogation of DNA methylation, at single-nucleotide resolution, across a comprehensive set of CpG sites scattered across the genome; moreover, the computational biology community has invested significant effort in the development of tools for properly removing technological effects that may contaminate biological signatures measured by such assays [2, Dedeurwaerder et al. [3]]. Despite these advances in both biological and bioninformatical techniques, most statistical methods available for differential analysis of data produced by such assays rely on over-simplified models that do not readily extend to such high-dimensional data structures without restrictive modeling assumptions and the use of inferentially costly hypothesis testing corrections. When these standard assumptions are violated, estimates of the population-level effect of an exposure or treatment may suffer from large bias. What's more, reliance on restrictive and misspecified statistical models naturally leads to biased effect estimates that are not only misleading in assessing effect sizes but also result in false discoveries as these biased estimates are subject to testing and inferential procedures. Such predictably unreliable methods serve only to produce findings that are later invalidated by replication studies and add still further complexity to discovering biological targets for potential therapeutics. Data-adaptive estimation procedures that utilize machine learning provide a way to overcome many of the problems common in classical methods, controlling for potential confounding even in high-dimensional settings; however, interpretable statistical inference (i.e., confidence intervals and hypothesis tests) from such data-adaptive estimates is challenging to obtain [4].

In this paper, we briefly present an alternative to such statistical analysis approaches in the form of a nonparametric estimation procedure that provides simple and readily interpretable statistical inference, discussing at length a recent implementation of the methodology in the `methyvim` R package. Inspired by recent advances in statistical causal inference and machine learning, we provide a computationally efficient technique for obtaining targeted estimates of nonparametric *variable importance measures* (VIMs) [5], estimated at a set of pre-screened CpG sites, controlling for the False Discovery Rate (FDR) as if all sites were tested. Under standard assumptions (e.g., identifiability, strong ignorability) [6], targeted minimum loss-based estimators of regular asymptotically linear estimators have sampling distributions that are asymptotically normal, allowing for reliable point estimation and the construction of Wald-style confidence intervals [7, van der Laan and Rose [8]]. In the context of DNA methylation studies, we define the counterfactual outcomes under a binary treatment as the observed methylation (whether Beta- or M-) values a CpG site would have if all subjects were administered the treatment and the methylation values a CpG site would have if treatment were withheld from all subjects. Although these counterfactual outcomes are, of course, impossible to observe, they do have statistical analogs that may be reliably estimated (i.e., identified) from observed data under a small number of untestable assumptions [6]. We describe an algorithm that incorporates, in its final step, the use *targeted minimum loss-based estimators* (**TMLE**) [9] of a given VIM of interest, though we defer rigorous and detailed descriptions of this aspect of the statistical methodology to work outside the scope of the present manuscript [9, van der Laan and Rose [7], van der Laan and Rose [8]]. The proposed methodology assesses the individual importance of a given CpG site, as a proposed measure of differential methylation, by utilizing state-of-the-art machine learning algorithms in deriving targeted estimates and robust inference of a VIM, as considered more broadly for biomarkers in Bembom et al. [10] and Tuglus and van der Laan [11]. In the present work, we focus on the `methyvim` software package, available through the Bioconductor project [12, Huber et al. [13]] for the R language and environment for statistical computing [14], which implements a particular realization of this methodology specifically tailored for the analysis and identification of differentially methylated positions (DMPs).

For an extended discussion of the general framework of targeted minimum loss-based estimation and detailed accounts of how this approach may be brought to bear in developing answers to complex scientific problems through tatistical and causal inference, the interested reader is invited to consult van der Laan and Rose [7] and van der Laan and Rose [8]. For a more general introduction to causal inference, Pearl [6] and Hernan and Robins [15] may be of interest.

Methods

Implementation

The core functionality of this package is made available via the eponymous `methyvim` function, which implements a statistical algorithm designed to compute targeted estimates of VIMs, defined in such a way that the VIMs represent parameters of scientific interest in computational biology experiments; moreover, these VIMs are defined such that they may be estimated in a manner that is very nearly assumption-free, that is, within

a *fully nonparametric statistical model*. The statistical algorithm consists in the several major steps summarized below. Additional methodological details on the use of targeted minimum loss-based estimation in this problem setting is provided in a brief appendix (see section **Appendix**).

1. *Pre-screening* of genomic sites is used to isolate a subset of sites for which there is cursory evidence of differential methylation. Currently, the available screening approach adapts core routines from the `limma` R package. Following the style of the function for performing screening via `limma`, users may write their own screening functions and are invited to contribute such functions to the core software package by opening pull requests at the GitHub repository: `https://github.com/nhejazi/methyvim`.
2. Nonparametric estimates of VIMs, for the specified target parameter, are computed at each of the CpG sites passing the screening step. The VIMs are defined in such a way that the estimated effects is of an binary treatment on the methylation status of a target CpG site, controlling for the observed methylation status of the neighbors of that site. Currently, routines are adapted from the `tmle` R package.
3. Since pre-screening is performed prior to estimating VIMs, we apply the modified marginal Benjamini and Hochberg step-up False Discovery Rate controlling procedure for multi-stage analyses (FDR-MSA), which is well-suited for avoiding false positive discoveries when testing is only performed on a subset of potential targets.

Parameters of Interest For CpG sites that pass the pre-screening step, a user-specified target parameter of interest is estimated independently at each site. *In all cases, an estimator of the parameter of interest is constructed via targeted minimum loss-based estimation*.

Two popular target causal parameters for discrete-valued treatments or exposures are

- The average treatment effect (ATE): The effect of a binary exposure or treatment on the observed methylation at a target CpG site is estimated, controlling for the observed methylation at all other CpG sites in the same neighborhood as the target site, based on an additive form. Often denoted $\psi_0 = \psi_0(1) - \psi_0(0)$, the parameter estimate represents the additive difference in methylation that would have been observed at the target site had all observations received the treatment versus the counterfactual under which none received the treatment.
- The relative risk (RR): The effect of a binary exposure or treatment on the observed methylation at a target CpG site is estimated, controlling for the observed methylation at all other CpG sites in the same neighborhood as the target site, based on a geometric form. Often denoted, $\psi_0 = \frac{\psi_0(1)}{\psi_0(0)}$, the parameter estimate represents the multiplicative difference in methylation that would have been observed at the target site had all observations received the treatment versus the counterfactual under which none received the treatment.

Estimating the VIM corresponding to the parameters above, for discrete-valued treatments or exposures, requires two separate regression steps: one for the treatment mechanism (propensity score) and one for the outcome regression. Technical details on the nature of these regressions are discussed in Hernan and Robins [15], and details for estimating these regressions in the framework of targeted minimum loss-based estimation are discussed in van der Laan and Rose [7].

Class `methytmle` We have adopted a class `methytmle` to help organize the functionality within this package. The `methytmle` class builds upon the `GenomicRatioSet` class provided by the `minfi` package so all of the slots of `GenomicRatioSet` are contained in a `methytmle` object. The new class introduced in the `methyvim` package includes several new slots:

- `call` - the form of the original call to the `methyvim` function.
- `screen_ind` - indices identifying CpG sites that pass the screening process.
- `clusters` - non-unique IDs corresponding to the manner in wich sites are treated as neighbors. These are assigned by genomic distance (bp) and respect chromosome boundaries (produced via a call to `bumphunter::clusterMaker`).
- `var_int` - the treatment/exposure status for each subject. Currently, these must be binary, due to the definition of the supported targeted parameters.
- `param` - the name of the target parameter from which the estimated VIMs are defined.
- `vim` - a table of statistical results obtained from estimating VIMs for each of the CpG sites that pass the screening procedure.

- `ic` - the measured array values for each of the CpG sites passing the screening, transformed into influence curve space based on the chosen target parameter.

The `show` method of the `methytmle` class summarizes a selection of the above information for the user while masking some of the wealth of information given when calling the same method for `GenomicRatioSet`. All information contained in `GenomicRatioSet` objects is preserved in `methytmle` objects, so as to easen interoperability with other differential methylation software for experienced users. We refer the reader to the package vignette, "`methyvim`: Targeted Data-Adaptive Estimation and Inference for Differential Methylation Analysis," included in any distribution of the software package, for further details.

Operation

A standard computer with the latest version of R and Bioconductor 3.6 installed will handle applications of the `methyvim` package.

Use Cases

To examine the practical applications and the full set of utilities of the `methyvim` package, we will use a publicly available example data set produced by the Illumina 450K array, from the `minfiData` R package.

Preliminaries: Setting up the Data We begin by loading the package and the data set. After loading the data, which comes in the form of a raw `MethylSet` object, we perform some further processing by mapping to the genome (with `mapToGenome`) and converting the values from the methylated and unmethylated channels to Beta-values (via `ratioConvert`). These two steps together produce an object of class `GenomicRatioSet`, provided by the `minfi` package.

```
suppressMessages(
  # numerous messages displayed at time of loading
  library(minfiData)
)
data(MsetEx)
mset <- mapToGenome(MsetEx)
grs <- ratioConvert(mset)
grs
```

```
## class: GenomicRatioSet
## dim: 485512 6
## metadata(0):
## assays(2): Beta CN
## rownames(485512): cg13869341 cg14008030 ... cg08265308 cg14273923
## rowData names(0):
## colnames(6): 5723646052_R02C02 5723646052_R04C01 ...
##   5723646053_R05C02 5723646053_R06C02
## colData names(13): Sample_Name Sample_Well ... Basename filenames
## Annotation
##   array: IlluminaHumanMethylation450k
##   annotation: ilmn12.hg19
## Preprocessing
##   Method: Raw (no normalization or bg correction)
##   minfi version: 1.21.2
##   Manifest version: 0.4.0
```

We can create an object of class `methytmle` from any `GenomicRatioSet` object simply invoking the S4 class constructor `.methytmle`:

```
library(methyvim)
```

```
## methyvim v1.3.1: Targeted Variable Importance for Differential Methylation Analysis
```

```
grs_mtmle <- .methytmle(grs)
grs_mtmle
```

```
## class: methytmle
## dim: 485512 6
## metadata(0):
## assays(2): Beta CN
## rownames(485512): cg13869341 cg14008030 ... cg08265308 cg14273923
## rowData names(0):
## colnames(6): 5723646052_R02C02 5723646052_R04C01 ...
##   5723646053_R05C02 5723646053_R06C02
## colData names(13): Sample_Name Sample_Well ... Basename filenames
## Annotation
##   array: IlluminaHumanMethylation450k
##   annotation: ilmn12.hg19
## Preprocessing
##   Method: Raw (no normalization or bg correction)
##   minfi version: 1.21.2
##   Manifest version: 0.4.0
## Target Parameter:
## Results:
## Object of class "data.frame"
## data frame with 0 columns and 0 rows
```

Additionally, a `GenomicRatioSet` can be created from a matrix with the function `makeGenomicRatioSetFromMatrix` provided by the `minfi` package.

Differential Methylation Analysis For this example analysis, we'll treat the condition of the patients as the exposure/treatment variable of interest. The `methyvim` function requires that this variable either be `numeric` or easily coercible to `numeric`. To facilitate this, we'll simply convert the covariate (currently a `character`):

```
var_int <- (as.numeric(as.factor(colData(grs)$status)) - 1)
```

n.b., the re-coding process results in "normal" patients being assigned a value of 1 and cancer patients a 0.

Now, we are ready to analyze the effects of cancer status on DNA methylation using this data set. We proceed as follows with a targeted minimum loss-based estimate of the Average Treatment Effect.

```
methyvim_cancer_ate <- methyvim(data_grs = grs, var_int = var_int,
                                vim = "ate", type = "Beta", filter = "limma",
                                filter_cutoff = 0.20, obs_per_covar = 2,
                                parallel = FALSE, sites_comp = 250,
                                tmle_type = "glm"
                               )
```

Note that we set the `obs_per_covar` argument to a relatively low value (just 2, even though the recommended value, and default, is 20) for the purposes of this example as the sample size is only 10. We do this only to exemplify the estimation procedure and it is important to point out that such low values for `obs_per_covar` will compromise the quality of inference obtained because this setting directly affects the definition of the target parameter.

Further, note that here we apply the `glm` flavor of the `tmle_type` argument, which produces faster results by fitting models for the propensity score and outcome regressions using a limited number of parametric models. By contrast, the `sl` (for "Super Learning") flavor fits these two regressions using highly nonparametric and data-adaptive procedures (i.e., via machine learning). Obtaining the estimates via GLMs results in each of the regression steps being less robust than if nonparametric regressions were used.

We can view a table of results by examining the `vim` slot of the produced object, most easily displayed by simply printing the resultant object:

```
methyvim_cancer_ate
```

```
## class: methytmle
## dim: 485512 6
## metadata(0):
## assays(2): Beta CN
## rownames(485512): cg13869341 cg14008030 ... cg08265308 cg14273923
## rowData names(0):
```

```
## colnames(6): 5723646052_R02C02 5723646052_R04C01 ...
##   5723646053_R05C02 5723646053_R06C02
## colData names(13): Sample_Name Sample_Well ... Basename filenames
## Annotation
##   array: IlluminaHumanMethylation450k
##   annotation: ilmn12.hg19
## Preprocessing
##   Method: Raw (no normalization or bg correction)
##   minfi version: 1.21.2
##   Manifest version: 0.4.0
## Target Parameter: Average Treatment Effect
## Results:
##                    lwr_ci       est_ate        upr_ci      var_ate
## cg14008030 -1.195660e-01 -0.0314159619  0.0567340489 2.022705e-03
## cg20253340 -8.850637e-02 -0.0588661418 -0.0292259165 2.286919e-04
## cg21870274 -9.499057e-02 -0.0291189817  0.0367526091 1.129495e-03
## cg17308840 -4.626018e-02 -0.0071524518  0.0319552773 3.981191e-04
## cg00645010 -2.677328e-02 -0.0134655543 -0.0001578256 4.609945e-05
## cg27534567  6.745648e-02  0.1157118536  0.1639672225 6.061486e-04
## cg08258224  1.365045e-01  0.3050884951  0.4736724770 7.398105e-03
## cg20275697 -3.021026e-01 -0.1231299608  0.0558427144 8.337989e-03
## cg24373735 -4.283533e-02  0.0076666975  0.0581687266 6.639043e-04
## cg12445832 -6.082411e-02  0.0150395574  0.0909032203 1.498151e-03
## cg01097950 -4.392159e-02  0.0658323796  0.1755863507 3.135656e-03
## cg01782097 -1.082072e-02  0.0010232901  0.0128672954 3.651615e-05
##                     pval n_neighbors n_neighbors_control max_cor_neighbors
## cg14008030 4.848468e-01           0                   0                NA
## cg20253340 9.917426e-05           0                   0                NA
## cg21870274 3.862537e-01           2                   1        0.94435796
## cg17308840 7.199943e-01           2                   1        0.94435796
## cg00645010 4.734007e-02           2                   2        0.52368097
## cg27534567 2.602936e-06           1                   0        0.93629683
## cg08258224 3.895914e-04           1                   0        0.93629683
## cg20275697 1.775155e-01           0                   0                NA
## cg24373735 7.660489e-01           0                   0                NA
## cg12445832 6.976022e-01           0                   0                NA
## cg01097950 2.397377e-01           0                   0                NA
## cg01782097 8.655302e-01           1                   1       -0.39410834
##  [ reached getOption("max.print") -- omitted 238 rows ]
```

Finally, we may compute FDR-corrected p-values, by applying a modified procedure for controlling the False Discovery Rate for multi-stage analyses (FDR-MSA) [16]. We do this by simply applying the `fdr_msa` function.

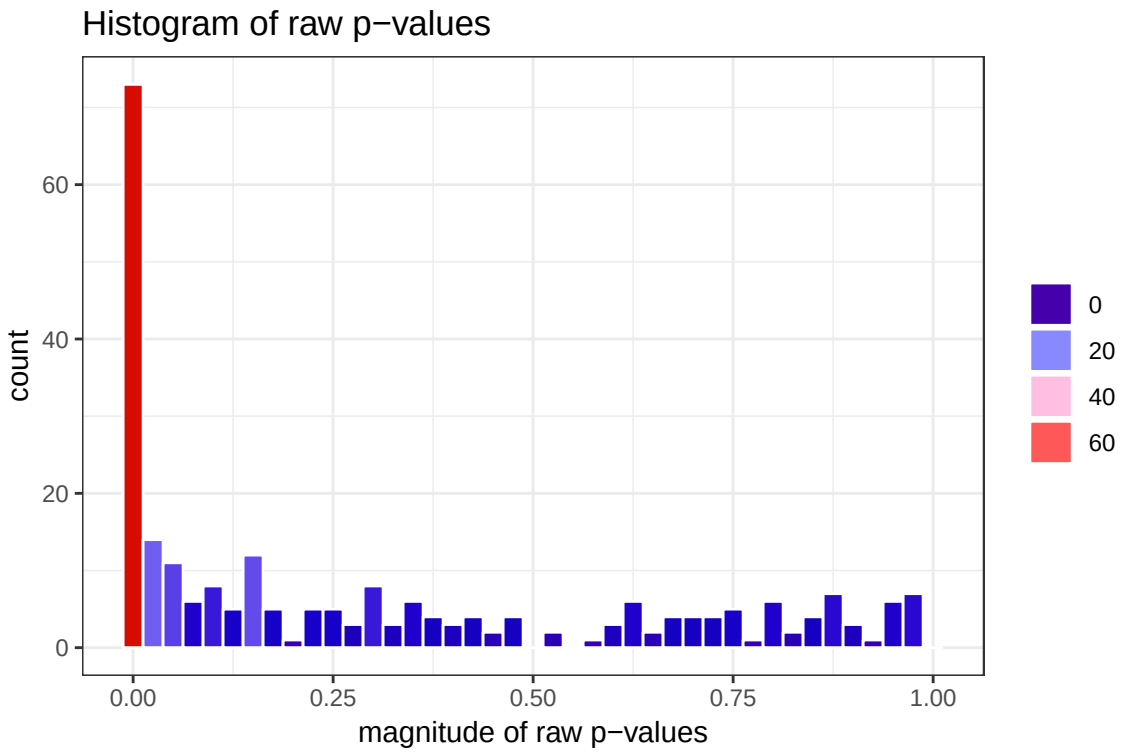
```
fdr_p <- fdr_msa(pvals = vim(methyvim_cancer_ate)$pval,
                 total_obs = nrow(methyvim_cancer_ate))
```

Having explored the results of our analysis numerically, we now proceed to use the visualization tools provided with the `methyvim` R package to further enhance our understanding of the results.

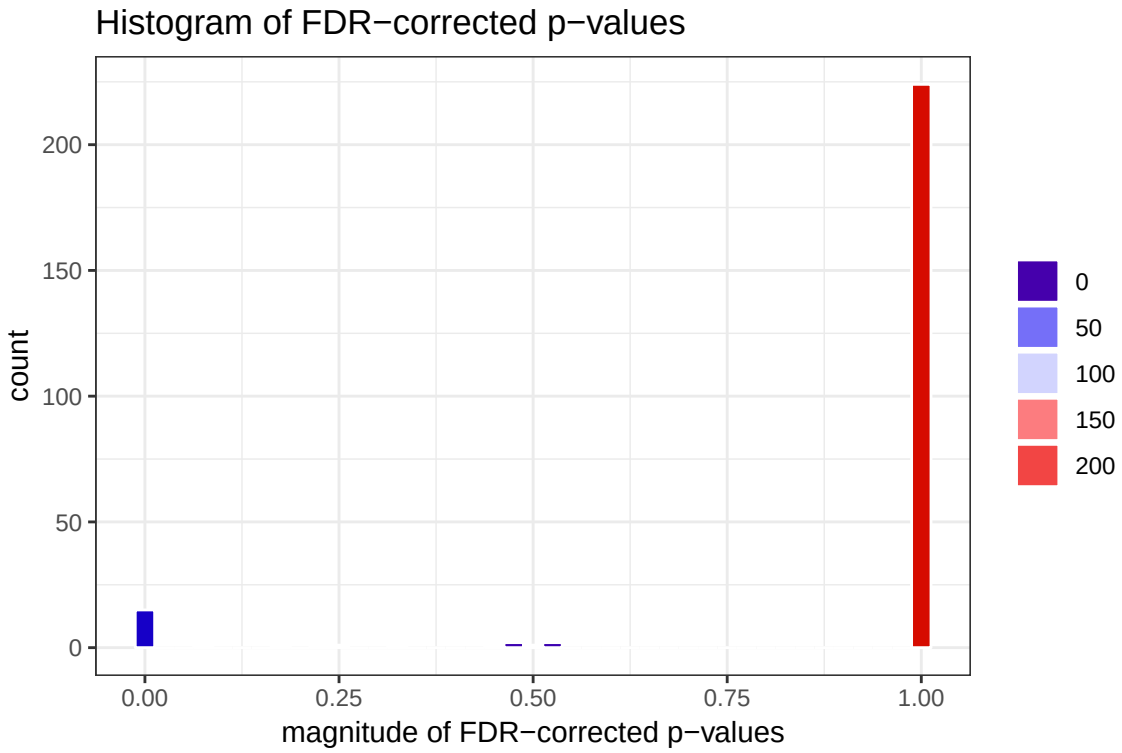
Visualization of Results While making allowance for users to explore the full set of results produced by the estimation procedure (by way of exposing these directly to the user), the `methyvim` package also provides *three* (3) visualization utilities that produce plots commonly used in examining the results of differential methylation analyses.

A simple call to `plot` produces side-by-side histograms of the raw p-values computed as part of the estimation process and the corrected p-values obtained from using the FDR-MSA procedure.

```
plot(methyvim_cancer_ate, type = "raw_pvals")
```



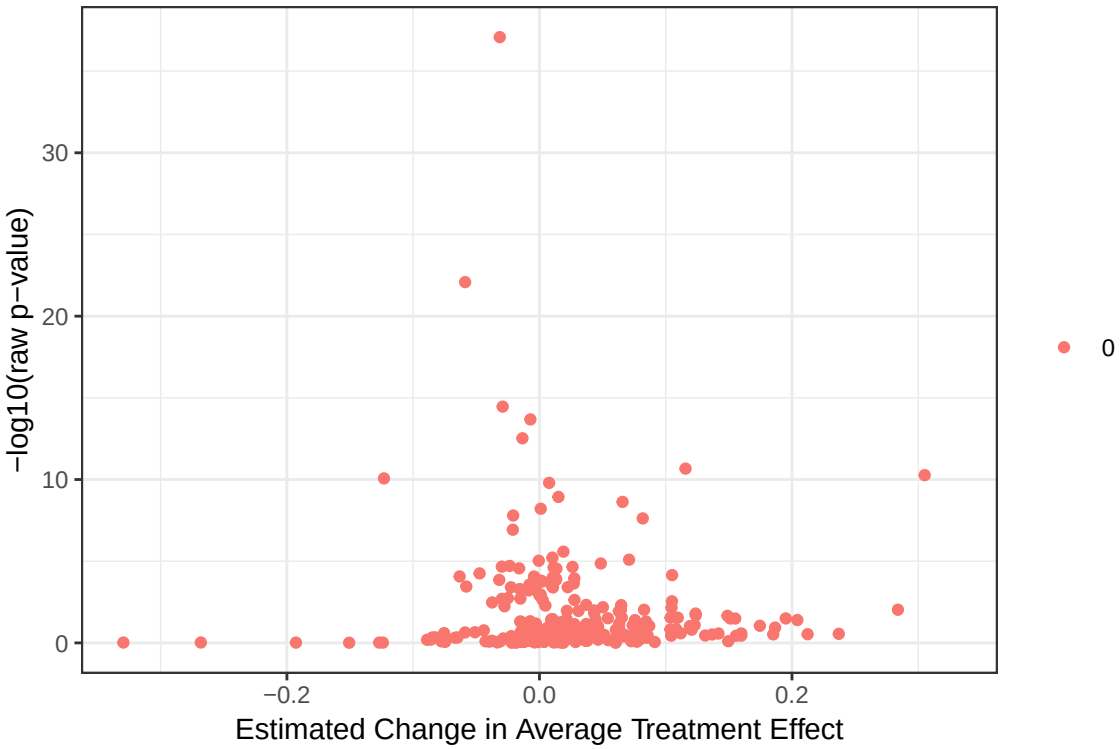

```
plot(methyvim_cancer_ate, type = "fdr_pvals")
```



Remark: The plots displayed above may also be generated as side-by-side histograms in a single plot object. This is the default for the `plot` method and may easily be invoked by specifying no additional arguments to the `plot` function, unlike in the above.

While histograms of the p-values may be generally useful in inspecting the results of the estimation procedure, a more common plot used in examining the results of differential methylation procedures is the volcano plot, which plots the parameter estimate along the x-axis and $-\log_{10}$(p-value) along the y-axis. We implement such a plot in the `methyvolc` function:

```
methyvolc(methyvim_cancer_ate)
```

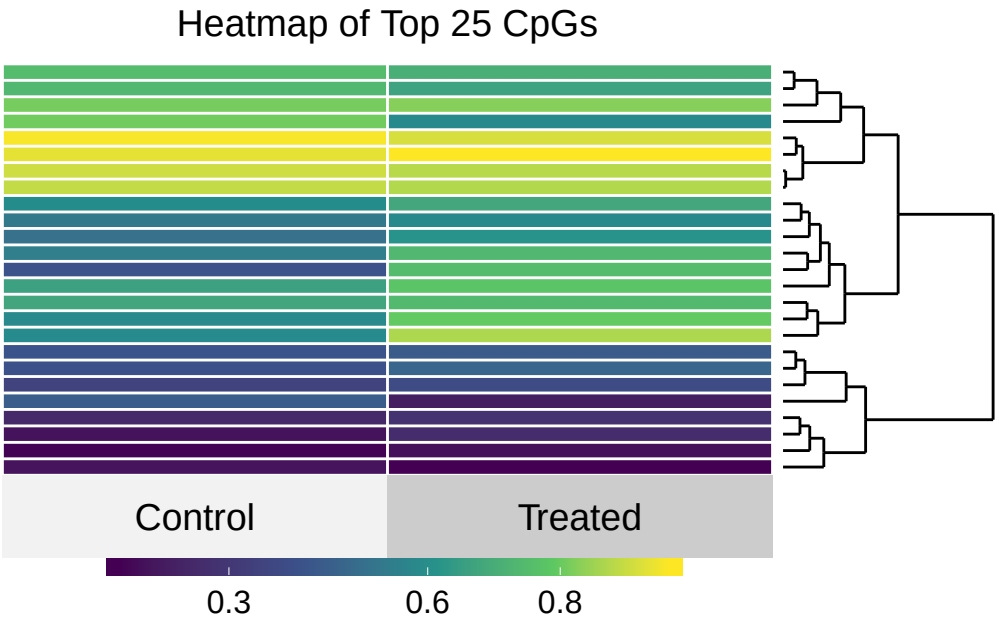


The purpose of such a plot is to ensure that very low (possibly statistically significant) p-values do not arise from cases of low variance. This appears to be the case in the plot above (notice that most parameter estimates are near zero, even in cases where the raw p-values are quite low).

Yet another popular plot for visualizing effects in such settings is the heatmap, which plots estimates of the raw methylation effects (as measured by the assay) across subjects using a heat gradient. We implement this in the `methyheat` function:

```
methyheat(methyvim_cancer_ate, smooth.heat = TRUE, left.label = "none")
```

```
## [1] 0.1 0.3 0.6 0.8 1.0
```



Remark: Invoking `methyheat` in this manner produces a plot of the top sites (25, by default) based on the raw p-value, using the raw methylation measures in the plot. This uses the exceptional `superheat` R package [17], to which we can easily pass additional parameters. In particulat, we hide the CpG site labels that would appear by default on the left of the heatmap (by setting `left.label = "none"`) to emphasize that this is only an example and *not* a scientific discovery.

Summary

Here we introduce the R package `methyvim`, an implementation of a general algorithm for differential methylation analysis that allows for recent advances in causal inference and machine learning to be leveraged in computational biology settings. The estimation procedure produces straightforward statistical inference and takes great care to ensure computationally efficiency of the technique for obtaining targeted estimates of nonparametric variable importance measures. The software package includes techniques for pre-screening a set of CpG sites, controlling for the False Discovery Rate as if all sites were tested, and for visualzing the results of the analyses in a variety of ways. The anatomy of the software package is dissected and the design described in detail. The `methyvim` R package is available via the Bioconductor project.

Software availability

Latest source code (development version): `https://github.com/nhejazi/methyvim`

Bioconductor (stable release): `https://bioconductor.org/packages/methyvim`

Archived source code as at time of publication: `https://github.com/nhejazi/methyvim/releases/tag/f1000`

Documentation (development version): `https://code.nimahejazi.org/methyvim`

Software license: The MIT License, copyright Nima S. Hejazi

Author contributions

NH designed and implemented the software package, applied the tool to the use cases presented, and co-drafted the present manuscript. RP helped in designing the software and co-drafted the present manuscript. AH and ML served as advisors for the development of this software and the general statistical algorithm it implements.

Competing interests

No competing interests were disclosed at the time of publication.

Grant information

NH was supported in part by the National Library of Medicine of the National Institutes of Health under Award Number T32-LM012417, by P42-ES004705, and by R01-ES021369. RP was supported by P42-ES004705. The content of this work is solely the responsibility of the authors and does not necessarily represent the official views of the various funding sources and agencies.

Appendix

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 $O \sim P \in \mathcal{M}$, which is to say that the random variable O is governed by an unknown probability distribution P, 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) $\Psi_j(P)$, which is defined through the *true* (and unknown) probability distribution P. As a motivating example, consider the case where $\Psi_j(P)$ is

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

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 [18, Hahn [19], Hirano et al. [20]]. As a measure of variable importance, we seek to estimate this target parameter ($\Psi_j(P)$), 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. We propose a procedure that estimates this target parameter $\Psi_j(P)$ across all CpG sites of interest $j : 1, \ldots, J$. When data-adaptive regression procedures (i.e., machine learning) are employed in estimating $\Psi_j(P)$, 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 ensemble modeling [21, Wolpert [22], Breiman [23], Gruber and van der Laan [24], van der Laan and Rose [7]].

Pre-Screening Procedure

As a matter of practicality, we consider only estimating the target VIM $\Psi_j(P)$ 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 [25, Robinson et al. [26]], 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. Details on the implementation of pre-screening approaches is provided in the sequel.

Reducing Neighbors via Clustering

Due to the data-adaptive nature of the regression procedures employed in evaluating the target VIM $\Psi_j(P)$ via targeted minimum loss-based estimation, it is possible the 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 [27]. 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 $\Psi_j(P)$ 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 $\Psi_j(P)$, we propose the use of targeted minimum loss-based estimation (TMLE) to construct and evaluate an estimator $\psi_{n,j}$ of $\Psi_j(P)$. 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 [28]; 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 [29], often denoted by $g(A \mid W)$; (2) an estimator of the outcome regression, often denoted $Q(A, W)$; and (3) a targeting step applied to the initial estimators constructed from the aforementioned components [24]. While we describe a few of the key properties of TML estimators in the sequel, extended technical discussion is deferred to more comprehensive work [7, van der Laan and Rose [8]]. In order to construct an estimate $\psi_{n,j}$ of $\Psi_j(P)$, it is necessary to accurately estimate to nuisance parameters, these being the propensity score ($g(A \mid W)$) and the outcome regression ($Q(A, W)$); 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 [22, Breiman [23], van der Laan et al. [21]], which ensures that the resulting ensemble learner satisfies important optimality properties with respect to cross-validation [30, Dudoit and van der Laan

[31], van der Laan and Dudoit [32]]. After constructing such estimates, an iterative procedure (i.e., the targeting step) may be used to combine these individual estimates into an estimate of the target parameter $\Psi_j(P)$.

Importantly, TML estimators are well-suited for statistical inference, having an asymptotically normal limiting distribution [33, van der Laan and Rubin [9]], which allows for a closed-form expression of the variance to be derived:

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

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., $H_{0,j} : \Psi_j(P) = 0$) in a straightforward manner.

Correction for Multiple Testing

Given that we seek to estimate $\Psi_j(P)$ 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) [34]; 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 [16]. 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 $H_{0,j} : \Psi_j(P) = 0$) 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] Keith D Robertson. DNA methylation and human disease. *Nature reviews. Genetics*, 6(8):597, 2005.

[2] Jean-Philippe Fortin, Aurelie Labbe, Mathieu Lemire, Brent W Zanke, Thomas J Hudson, Elana J Fertig, Celia MT Greenwood, and Kasper D Hansen. Functional normalization of 450k methylation array data improves replication in large cancer studies. *bioRxiv*, 2014.

[3] Sarah Dedeurwaerder, Matthieu Defrance, Martin Bizet, Emilie Calonne, Gianluca Bontempi, and François Fuks. A comprehensive overview of Infinium HumanMethylation450 data processing. *Briefings in bioinformatics*, page bbt054, 2013.

[4] Maxwell W Libbrecht and William Stafford Noble. Machine learning in genetics and genomics. *Nature Reviews. Genetics*, 16(6):321, 2015.

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

[6] Judea Pearl. *Causality: Models, Reasoning, and Inference*. Cambridge University Press, 2009.

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

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

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

[10] Oliver Bembom, Maya L Petersen, Soo-Yon Rhee, W Jeffrey Fessel, Sandra E Sinisi, Robert W Shafer, and Mark J van der Laan. Biomarker discovery using targeted maximum-likelihood estimation: Application to the treatment of antiretroviral-resistant HIV infection. *Statistics in medicine*, 28(1):152–172, 2009.

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

[12] Robert C Gentleman, Vincent J Carey, Douglas M Bates, Ben Bolstad, Marcel Dettling, Sandrine Dudoit, Byron Ellis, Laurent Gautier, Yongchao Ge, Jeff Gentry, Kurt Hornik, Torsten Hothorn, Wolfgang Huber, Stefano Iacus, Rafael Irizarry, Friedrich Leisch, Cheng Li, Martin Maechler, Anthony J Rossini, Gunther Sawitzki, Colin Smith, Gordon Smyth, Luke Tierney, Jean YH Yang, and Jianhua Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, 5(10):R80, 2004.

[13] Wolfgang Huber, Vincent J Carey, Robert Gentleman, Simon Anders, Marc Carlson, Benilton S Carvalho, Hector Corrada Bravo, Sean Davis, Laurent Gatto, Thomas Girke, et al. Orchestrating high-throughput genomic analysis with Bioconductor. *Nature methods*, 12(2):115–121, 2015.

[14] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2018. URL `https://www.R-project.org/`.

[15] Miguel A Hernan and James M Robins. *Causal Inference*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis, 2018, forthcoming.

[16] 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`.

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

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

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

[20] 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.

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

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

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

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

[25] 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`.

[26] 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.

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

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

[29] 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.

[30] 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.

[31] 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.

[32] 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.

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

[34] 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.