Nonparametric Variable Importance for Continuous Exposures with Applications in High-Dimensional Biology

Nima Hejazi, Ivana Malenica, Andre Kurepa Waschka, Alan E. Hubbard, and Mark J. van der Laan

Division of Biostatistics & Department of Statistics, University of California, Berkeley

School of Public Health
UNIVERSITY OF CALIFORNIA, BERKELEY

Overview & Motivations

1. We introduce a general nonparametric technique to assess the effects of continuous exposures, extending the methodology for use with data generated by modern experiments in high-dimensional biology.
2. Such nonparametric variable importance measures (NPVI) may be used to replace their parametric counterparts, which are standard practice in genomics.
3. NPVI captures the expected difference in outcome/response, based on differences in receiving a continuous exposure against a so-called "null range" of said exposure.
4. We apply the proposed method to analyzing DNA methylation data from a epidemiologic study, providing a sample of results for differentially methylated CpG sites.

Introduction & Data

- Data was generated by the **Infinium HumanMethylation450 BeadChip** platform, providing consistent measurements of nearly $450,000$ CpG sites.
- Data was made available on over 200 subjects, with relevant data on baseline characteristics, 450K methylation data collected from fetal blood, and outcome (IQ) scores assessed at several intervals across early years of development.
- The exposure of interest is a vector of 450K CpG methylation measures, with each array subjected to median normalization.
- The outcome of interest is a vector of IQ scores (from a particular type of test) assessed several years after DNA methylation was measured.

Methodology I

The general roadmap for using NPVI in problems of high-dimensional biology is as follows:

- We observe n independent copies $O_i = (W_i, X_i, Y_i)_{i=1}^n$ of the observed data structure $O \sim P_0 \in M$, where M is possibly a nonparametric (infinite-dimensional) model.
- The **NPVI** parameter is $\Psi(P) = \text{argmin}_{\beta \in \mathbb{R}} \mathbb{E}_P[(Y - \mathbb{E}_P(Y|X = x_0, W) - \beta(X - x_0))^2]$
  - $\Psi(P) = \frac{\mathbb{E}_P[X(\theta(P)(X,W) - \theta(P)(0,W))]}{\mathbb{E}_P(X^2)}$, a TMLE is available (**Proposition 1**). [4]
  - Neglecting W, we have a different mapping: $\mathcal{F}(P) = \text{argmin}_{\beta \in \mathbb{R}} \mathbb{E}_P[(Y - \beta X)^2] \equiv \frac{\mathbb{E}_P(XY)}{\mathbb{E}_P(X^2)}$
  - Both Ψ and $\mathcal{F}$ are pathwise differentiable, with known influence curves in semiparametics and in closed form.
- To estimate $\Psi(P)$, begin by initially estimating P with P_n^0 (setting $k = 0$), then iteratively:
  1. Construct a 1-dimensional model $P_n^k(\epsilon) : |\epsilon| < \|s\|_\infty^{-1} \subset \mathcal{M}$ by setting $\frac{dP_n^k(\epsilon)}{dP_n^k} = 1 + \epsilon s$ with $s = \nabla \Psi_{P_n^k}$.
  2. Compute the corresponding MLE ϵ_n^k for $P_n^{k+1} := P_n^k(\epsilon_n^k)$, updating $k \leftarrow k + 1$.
  3. By substitution, form the *targeted* minimum loss-based estimator $\psi_n = \Psi(P_n^K)$ where K is the last value of k (obtained when a stopping criterion is met).
- Apply the NPVI estimator to the observed data (with empirical measure P_n): $\hat{\Psi}_j(P_n) = \frac{\mathbb{E}_{P_n}[X(\theta(P_n)(X,W) - \theta(P_n)(0,W))]}{\mathbb{E}_{P_n}(X^2)}$, for j indexing genomic loci of interest.
- To estimate such a parameter at each genomic locus, we define a cutoff for the "null range" of NPVI as a particular percentile of the outcome at a given locus (e.g., 20^{th}), though a general procedure for defining a cutoff in a data adaptive manner is available [3].

Methodology II

To make this procedure feasible for use in high-dimensional biology, several adjustments are necessary:

- Use of screening procedures (e.g., simple linear models) helps to reduce the set of loci of interest to a computationally manageable number.
- A modified procedure for controlling the False Discovery Rate (FDR) with multi-stage analyses (FDR-MSA) [1] may be used to control the FDR at the same rate as if all genomic loci were tested.
- Methods to control the FDR with related hypotheses may be integrated into this framework.

Future extensions for combining multiple outcomes:

- The outcomes of interest are IQ scores measured by 8 different tests. As part of future work, we are interested in ways to combine several cognitive outcome measures into one outcome.
- Let $\psi_j(W) = \mathbb{E}(Y_j \mid X)$, where the outcome is defined as $Y = (Y_1, \ldots, Y_J)$.
- We define our target parameter as $\Psi_{\alpha_n}(P_0) = E_0(\sum_j \alpha_{nj} Y_j \mid W, X)$, with α_n corresponding to:

$$\alpha_n = \text{argmax}\left[1 - \frac{\frac{1}{V}\sum_v P_{n,v}^1(\sum_j \alpha_j(Y_j - \hat{\Psi}_j(P_{n,v}^0))^2}{P_n(\sum_j \alpha_j(Y_j - P_n Y_j))^2}\right]$$

- Honest cross-validated R^2 for estimator $\hat{\Psi}_{\alpha_n}(P_n)$ is:

$$R_{CV} = 1 - \frac{\sum_v P_{n,v}^1(\sum_j \hat{\alpha}_j(P_{n,v})Y_j - \sum_j \hat{\alpha}_j(P_{n,v})\hat{\Psi}_j(P_{n,v}))^2}{\sum_v \sum_j P_{n,v}^1(\sum_j \hat{\alpha}_j(P_{n,v})(Y_j - P_{n,v}Y_j))^2}$$

Results & Discussion

Table 1: Sample results for 10 CpG sites using the TMLE procedure to estimate the NPVI parameter.

	$\hat{\Psi}_j(P_n)$	raw p-value	adjusted p-value
1	**2.896E+00**	3.920E-09	4.263E-04
2	2.887E-01	6.850E-01	1.000E+00
3	**1.417E+01**	2.170E-141	3.147E-136
4	1.679E+00	1.730E-01	1.000E+00
5	1.132E+00	2.530E-02	1.000E+00
6	**-1.778E+00**	4.438E-292	1.931E-286
7	4.136E+00	5.903E-05	1.000E+00
8	**-8.347E-01**	1.838E-185	3.998E-180
9	2.589E-01	8.540E-01	1.000E+00
10	1.895E+00	3.740E-05	1.000E+00

- By applying the TMLE for estimating NPVI, the impact of DNA methylation on a phenotype-level outcome (e.g., IQ at age 7) may be nonparametrically estimated.
- The NPVI parameter is a complex (but convenient) substitute for estimating the relationships involving continuous exposures.
- Future work will make use of nonparametric methods for combining related outcomes into single scores.

Principal References

[1] Catherine Tuglus and Mark J. van der Laan. Modified FDR controlling procedure for multi-stage analyses. *Statistical Applications in Genetics and Molecular Biology*, 2009.

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

[3] Mark J. van der Laan, Alan E. Hubbard, and Sara Kherad-Pajouh. Statistical inference for data adaptive target parameters. *International Journal of Biostatistics*, 2016.

[4] Antoine Chambaz, Pierre Neuvial, and Mark J. van der Laan. Estimation of a non-parametric variable importance measure of a continuous exposure. *Electronic Journal of Statistics*, 2012.

Contact Information

N. Hejazi: NHEJAZI@BERKELEY.EDU
I. Malenica: IMALENICA@BERKELEY.EDU
A. K. Waschka: AKWASCHKA@BERKELEY.EDU
A. E. Hubbard: HUBBARD@BERKELEY.EDU
M. J. van der Laan: LAAN@BERKELEY.EDU