Combining Causal Inference and Machine Learning for Model-Agnostic Discovery in High-Dimensional Biology

Nima Hejazi

04 November 2022

Department of Biostatistics,
T.H. Chan School of Public Health,
Harvard University

nshejazi
nhejazi
nimahejazi.org
Biomedical AI Seminar,
University of Edinburgh
Joint work with A. Hubbard, M. van der Laan, P. Boileau

Preview

1. Modern computational biology research produces complex, heterogeneous data — innovative statistical inference still tied to simplistic and challenging-to-verify modeling assumptions.
2. *Model misspecification* seriously undermines the scientific utility of common, classical statistical modeling approaches.
3. Non/semi-parametric inference facilitates constructing robust estimators that easily bring machine learning into the fold.
4. Variance moderation strengthens hypothesis testing strategies, reducing false positives and preserving power under instability.

We'll go over this summary again at the end of the talk. Hopefully, it will all make more sense then.

Preview

1. Modern computational biology research produces complex, heterogeneous data — innovative statistical inference still tied to simplistic and challenging-to-verify modeling assumptions.
2. *Model misspecification* seriously undermines the scientific utility of common, classical statistical modeling approaches.
3. Non/semi-parametric inference facilitates constructing robust estimators that easily bring machine learning into the fold.
4. Variance moderation strengthens hypothesis testing strategies, reducing false positives and preserving power under instability.

We'll go over this summary again at the end of the talk. Hopefully, it will all make more sense then.

Preview

1. Modern computational biology research produces complex, heterogeneous data — innovative statistical inference still tied to simplistic and challenging-to-verify modeling assumptions.
2. *Model misspecification* seriously undermines the scientific utility of common, classical statistical modeling approaches.
3. Non/semi-parametric inference facilitates constructing robust estimators that easily bring machine learning into the fold.
4. Variance moderation strengthens hypothesis testing strategies, reducing false positives and preserving power under instability.

We'll go over this summary again at the end of the talk. Hopefully, it will all make more sense then.

Preview

1. Modern computational biology research produces complex, heterogeneous data — innovative statistical inference still tied to simplistic and challenging-to-verify modeling assumptions.
2. *Model misspecification* seriously undermines the scientific utility of common, classical statistical modeling approaches.
3. Non/semi-parametric inference facilitates constructing robust estimators that easily bring machine learning into the fold.
4. Variance moderation strengthens hypothesis testing strategies, reducing false positives and preserving power under instability.

We'll go over this summary again at the end of the talk. Hopefully, it will all make more sense then.

A common problem

- *Question*: What factors are associated ("causally" perhaps) with a health outcome of interest (e.g., cancer, death).
- *Experiment*: Assign[?] patients to novel therapy vs. standard of care (or exposure) and then evaluate outcome's occurrence.
- *Goal*: Deepen mechanistic insights — how does the therapy or exposure biologically operate? Identify intervention points.
- Combine tools from $\star$-omics and molecular biology, clinical trials, causal inference, (bio)statistics, epidemiology.

A common problem

- *Question*: What factors are associated ("causally" perhaps) with a health outcome of interest (e.g., cancer, death).
- *Experiment*: Assign[?] patients to novel therapy vs. standard of care (or exposure) and then evaluate outcome's occurrence.
- *Goal*: Deepen mechanistic insights — how does the therapy or exposure biologically operate? Identify intervention points.
- Combine tools from $\star$-omics and molecular biology, clinical trials, causal inference, (bio)statistics, epidemiology.

A common problem

- *Question*: What factors are associated ("causally" perhaps) with a health outcome of interest (e.g., cancer, death).
- *Experiment*: Assign[?] patients to novel therapy vs. standard of care (or exposure) and then evaluate outcome's occurrence.
- *Goal*: Deepen mechanistic insights — how does the therapy or exposure biologically operate? Identify intervention points.
- Combine tools from $\star$-omics and molecular biology, clinical trials, causal inference, (bio)statistics, epidemiology.

A common problem

- *Question*: What factors are associated ("causally" perhaps) with a health outcome of interest (e.g., cancer, death).
- *Experiment*: Assign[?] patients to novel therapy vs. standard of care (or exposure) and then evaluate outcome's occurrence.
- *Goal*: Deepen mechanistic insights — how does the therapy or exposure biologically operate? Identify intervention points.
- Combine tools from $\star$-omics and molecular biology, clinical trials, causal inference, (bio)statistics, epidemiology.

Let's meet the data: Benzene exposure and miRNA

- *Question*: Which miRNA (non-coding regulators) are affected by a target occupational exposure (benzene)?
- *Why?* Attempt to decipher how patterns of miRNA disregulation may impact subsequent disease states.
- *Study*: Cohort study of occupational exposure to benzene with 125 individuals and 22K candidate miRNA assayed.
- *Goal*: Characterize biological mechanisms or signatures derived from or attributable to exposure.

Let's meet the data: Benzene exposure and miRNA

- *Question*: Which miRNA (non-coding regulators) are affected by a target occupational exposure (benzene)?
- *Why?* Attempt to decipher how patterns of miRNA disregulation may impact subsequent disease states.
- *Study*: Cohort study of occupational exposure to benzene with 125 individuals and 22K candidate miRNA assayed.
- *Goal*: Characterize biological mechanisms or signatures derived from or attributable to exposure.

Let's meet the data: Benzene exposure and miRNA

- *Question*: Which miRNA (non-coding regulators) are affected by a target occupational exposure (benzene)?
- *Why?* Attempt to decipher how patterns of miRNA disregulation may impact subsequent disease states.
- *Study*: Cohort study of occupational exposure to benzene with 125 individuals and 22K candidate miRNA assayed.
- *Goal*: Characterize biological mechanisms or signatures derived from or attributable to exposure.

Let's meet the data: Benzene exposure and miRNA

- *Question*: Which miRNA (non-coding regulators) are affected by a target occupational exposure (benzene)?
- *Why?* Attempt to decipher how patterns of miRNA disregulation may impact subsequent disease states.
- *Study*: Cohort study of occupational exposure to benzene with 125 individuals and 22K candidate miRNA assayed.
- *Goal*: Characterize biological mechanisms or signatures derived from or attributable to exposure.

Let's meet the data: Smoking and DNA methylation

- *Question*: Which CpG sites, or larger functional units (e.g., "CpG islands"), are affected by long-term smoking?
- *Why?* Attempt to understand how smoking induces regulatory and functional changes that relate to disease (e.g., cancer).
- *Study*: Observational exposure study of 253 individuals (with 172 smokers, 81 non-smokers) and $\approx$ 450K CpG sites assayed.
- *Goal*: Characterize biological mechanisms or signatures derived from or attributable to exposure.

Let's meet the data: Smoking and DNA methylation

- *Question*: Which CpG sites, or larger functional units (e.g., "CpG islands"), are affected by long-term smoking?
- *Why?* Attempt to understand how smoking induces regulatory and functional changes that relate to disease (e.g., cancer).
- *Study*: Observational exposure study of 253 individuals (with 172 smokers, 81 non-smokers) and $\approx$ 450K CpG sites assayed.
- *Goal*: Characterize biological mechanisms or signatures derived from or attributable to exposure.

Let's meet the data: Smoking and DNA methylation

- *Question*: Which CpG sites, or larger functional units (e.g., "CpG islands"), are affected by long-term smoking?
- *Why?* Attempt to understand how smoking induces regulatory and functional changes that relate to disease (e.g., cancer).
- *Study*: Observational exposure study of 253 individuals (with 172 smokers, 81 non-smokers) and $\approx$ 450K CpG sites assayed.
- *Goal*: Characterize biological mechanisms or signatures derived from or attributable to exposure.

Let's meet the data: Smoking and DNA methylation

- *Question*: Which CpG sites, or larger functional units (e.g., "CpG islands"), are affected by long-term smoking?
- *Why?* Attempt to understand how smoking induces regulatory and functional changes that relate to disease (e.g., cancer).
- *Study*: Observational exposure study of 253 individuals (with 172 smokers, 81 non-smokers) and $\approx$ 450K CpG sites assayed.
- *Goal*: Characterize biological mechanisms or signatures derived from or attributable to exposure.

Data structure and notation

- Consider a structural causal model (SCM) (**?**) to describe how data on a single unit O was generated:

$$L = f_L(U_L); A = f_A(L, U_A); Y = f_Y(A, L, U_Y).$$

- f_L, f_A, f_Y are unknown but deterministic functions; U_L, U_A, U_Y are exogenous (unobserved) random errors.
- $Y = (Y_b : b = 1, \ldots B)$ is a vector of biomarker outcomes (e.g., $B = 22$K for miRNA, $B = 450$K for CpG sites).
- Temporal ordering between variables: L (sex-at-birth, age), A (smoking, benzene), Y_b (biomarker measurement for site b).
- Data on a single study unit $O = (L, A, Y)$, with $O \sim P_0 \in \mathscr{M}$, of which we observe n i.i.d. copies, $O_1, \ldots, O_n$.

Data structure and notation

- Consider a structural causal model (SCM) (**?**) to describe how data on a single unit O was generated:

$$L = f_L(U_L); A = f_A(L, U_A); Y = f_Y(A, L, U_Y).$$

- f_L, f_A, f_Y are unknown but deterministic functions; U_L, U_A, U_Y are exogenous (unobserved) random errors.
- $Y = (Y_b : b = 1, \ldots B)$ is a vector of biomarker outcomes (e.g., $B = 22$K for miRNA, $B = 450$K for CpG sites).
- Temporal ordering between variables: L (sex-at-birth, age), A (smoking, benzene), Y_b (biomarker measurement for site b).
- Data on a single study unit $O = (L, A, Y)$, with $O \sim P_0 \in \mathscr{M}$, of which we observe n i.i.d. copies, $O_1, \ldots, O_n$.

Data structure and notation

- Consider a structural causal model (SCM) (**?**) to describe how data on a single unit O was generated:

$$L = f_L(U_L); A = f_A(L, U_A); Y = f_Y(A, L, U_Y).$$

- f_L, f_A, f_Y are unknown but deterministic functions; U_L, U_A, U_Y are exogenous (unobserved) random errors.
- $Y = (Y_b : b = 1, \ldots B)$ is a vector of biomarker outcomes (e.g., $B = 22$K for miRNA, $B = 450$K for CpG sites).
- Temporal ordering between variables: L (sex-at-birth, age), A (smoking, benzene), Y_b (biomarker measurement for site b).
- Data on a single study unit $O = (L, A, Y)$, with $O \sim P_0 \in \mathscr{M}$, of which we observe n i.i.d. copies, $O_1, \ldots, O_n$.

Data structure and notation

- Consider a structural causal model (SCM) (**?**) to describe how data on a single unit O was generated:

$$L = f_L(U_L); A = f_A(L, U_A); Y = f_Y(A, L, U_Y).$$

- f_L, f_A, f_Y are unknown but deterministic functions; U_L, U_A, U_Y are exogenous (unobserved) random errors.
- $Y = (Y_b : b = 1, \ldots B)$ is a vector of biomarker outcomes (e.g., $B = 22$K for miRNA, $B = 450$K for CpG sites).
- Temporal ordering between variables: L (sex-at-birth, age), A (smoking, benzene), Y_b (biomarker measurement for site b).
- Data on a single study unit $O = (L, A, Y)$, with $O \sim P_0 \in \mathscr{M}$, of which we observe n i.i.d. copies, $O_1, \ldots, O_n$.

Data structure and notation

- Consider a structural causal model (SCM) (**?**) to describe how data on a single unit O was generated:

$$L = f_L(U_L); A = f_A(L, U_A); Y = f_Y(A, L, U_Y).$$

- f_L, f_A, f_Y are unknown but deterministic functions; U_L, U_A, U_Y are exogenous (unobserved) random errors.
- $Y = (Y_b : b = 1, \ldots B)$ is a vector of biomarker outcomes (e.g., $B = 22$K for miRNA, $B = 450$K for CpG sites).
- Temporal ordering between variables: L (sex-at-birth, age), A (smoking, benzene), Y_b (biomarker measurement for site b).
- Data on a single study unit $O = (L, A, Y)$, with $O \sim P_0 \in \mathscr{M}$, of which we observe n i.i.d. copies, $O_1, \ldots, O_n$.

Hypothetical interventions and causal inference

- *Static* interventions consider replacing f_A with an assigned value $a \in \mathscr{A}$ deterministically. "What if everyone smoked?"
- Generates "counterfactual" RV $Y(a) = (Y_b(a), b : 1, \ldots B)$: the expression of the B biomarkers if A had been set to a.
- Viewed as *potential outcomes* (POs) (**?**), $Y_b(1)$ when setting $A = 1$ and $Y_b(0)$ when setting $A = 0$.
- Note that $Y_b = AY_b(1) + (1 - A)Y_b(0)$ — only partially seeing the POs is the *fundamental problem of causal inference*.
- Causal inference yields interpretable, scientifically well-aligned estimands, e.g., the average treatment effect (ATE).

Statistical parameter may be viewed as a simple adjusted difference in means even when identifiability conditions appear unsatisfiable.

Hypothetical interventions and causal inference

- *Static* interventions consider replacing f_A with an assigned value $a \in \mathscr{A}$ deterministically. "What if everyone smoked?"
- Generates "counterfactual" RV $Y(a) = (Y_b(a), b : 1, \ldots B)$: the expression of the B biomarkers if A had been set to a.
- Viewed as *potential outcomes* (POs) (**?**), $Y_b(1)$ when setting $A = 1$ and $Y_b(0)$ when setting $A = 0$.
- Note that $Y_b = AY_b(1) + (1 - A)Y_b(0)$ — only partially seeing the POs is the *fundamental problem of causal inference*.
- Causal inference yields interpretable, scientifically well-aligned estimands, e.g., the average treatment effect (ATE).

Statistical parameter may be viewed as a simple adjusted difference in means even when identifiability conditions appear unsatisfiable.

Hypothetical interventions and causal inference

- *Static* interventions consider replacing f_A with an assigned value $a \in \mathscr{A}$ deterministically. "What if everyone smoked?"
- Generates "counterfactual" RV $Y(a) = (Y_b(a), b : 1, \ldots B)$: the expression of the B biomarkers if A had been set to a.
- Viewed as *potential outcomes* (POs) (**?**), $Y_b(1)$ when setting $A = 1$ and $Y_b(0)$ when setting $A = 0$.
- Note that $Y_b = AY_b(1) + (1 - A)Y_b(0)$ — only partially seeing the POs is the *fundamental problem of causal inference*.
- Causal inference yields interpretable, scientifically well-aligned estimands, e.g., the average treatment effect (ATE).

Statistical parameter may be viewed as a simple adjusted difference in means even when identifiability conditions appear unsatisfiable.

Hypothetical interventions and causal inference

- *Static* interventions consider replacing f_A with an assigned value $a \in \mathscr{A}$ deterministically. "What if everyone smoked?"
- Generates "counterfactual" RV $Y(a) = (Y_b(a), b : 1, \ldots B)$: the expression of the B biomarkers if A had been set to a.
- Viewed as *potential outcomes* (POs) (**?**), $Y_b(1)$ when setting $A = 1$ and $Y_b(0)$ when setting $A = 0$.
- Note that $Y_b = AY_b(1) + (1 - A)Y_b(0)$ — only partially seeing the POs is the *fundamental problem of causal inference*.
- Causal inference yields interpretable, scientifically well-aligned estimands, e.g., the average treatment effect (ATE).

Statistical parameter may be viewed as a simple adjusted difference in means even when identifiability conditions appear unsatisfiable.

Hypothetical interventions and causal inference

- *Static* interventions consider replacing f_A with an assigned value $a \in \mathscr{A}$ deterministically. "What if everyone smoked?"
- Generates "counterfactual" RV $Y(a) = (Y_b(a), b : 1, \ldots B)$: the expression of the B biomarkers if A had been set to a.
- Viewed as *potential outcomes* (POs) (**?**), $Y_b(1)$ when setting $A = 1$ and $Y_b(0)$ when setting $A = 0$.
- Note that $Y_b = AY_b(1) + (1 - A)Y_b(0)$ — only partially seeing the POs is the *fundamental problem of causal inference*.
- Causal inference yields interpretable, scientifically well-aligned estimands, e.g., the average treatment effect (ATE).

Statistical parameter may be viewed as a simple adjusted difference in means even when identifiability conditions appear unsatisfiable.

A familiar workhorse: the linear model

- The linear model is *semiparametric* — linear in parameters!
- Flexible: transformations (X_j^2), interactions ($X_j X_k$).
- For biomarker Y_b, fit *working* linear model, $\mathbb{E}_0[Y_b \mid X] = X\beta$, letting (wlog) $X_1 \equiv A$ be the exposure and β_1 its "effect".
- Under this working model, the parameter β_1 is the ATE, but *only if* there are no interactions (i.e., without flexibility).
- Test the contrast of interest with a standard t-test:

$$t_b = \frac{\hat{\beta}_b - \beta_{b,H_0}}{\hat{\sigma}_b}$$

There's nothing particularly wrong with this approach. It's exactly what we would come up with after a first-year statistics course. In practice, there are many issues: (1) we are forced to specify a functional form, the linear model; (2) we end up with unstable variance estimates that sharply increase the number of false positives detected, even after multiple testing corrections. In practice, the incredible flexibility of the linear mode is rarely taken advantage of — scientific guidance is usually lacking to justify the fitting of richer models.

A familiar workhorse: the linear model

- The linear model is *semiparametric* — linear in parameters!
- Flexible: transformations (X_j^2), interactions ($X_j X_k$).
- For biomarker Y_b, fit *working* linear model, $\mathbb{E}_0[Y_b \mid X] = X\beta$, letting (wlog) $X_1 \equiv A$ be the exposure and β_1 its "effect".
- Under this working model, the parameter β_1 is the ATE, but *only if* there are no interactions (i.e., without flexibility).
- Test the contrast of interest with a standard t-test:

$$t_b = \frac{\hat{\beta}_b - \beta_{b,H_0}}{\hat{\sigma}_b}$$

There's nothing particularly wrong with this approach. It's exactly what we would come up with after a first-year statistics course. In practice, there are many issues: (1) we are forced to specify a functional form, the linear model; (2) we end up with unstable variance estimates that sharply increase the number of false positives detected, even after multiple testing corrections. In practice, the incredible flexibility of the linear mode is rarely taken advantage of — scientific guidance is usually lacking to justify the fitting of richer models.

A familiar workhorse: the linear model

- The linear model is *semiparametric* — linear in parameters!
- Flexible: transformations (X_j^2), interactions ($X_j X_k$).
- For biomarker Y_b, fit *working* linear model, $\mathbb{E}_0[Y_b \mid X] = X\beta$, letting (wlog) $X_1 \equiv A$ be the exposure and β_1 its "effect".
- Under this working model, the parameter β_1 is the ATE, but *only if* there are no interactions (i.e., without flexibility).
- Test the contrast of interest with a standard t-test:

$$t_b = \frac{\hat{\beta}_b - \beta_{b,H_0}}{\hat{\sigma}_b}$$

There's nothing particularly wrong with this approach. It's exactly what we would come up with after a first-year statistics course. In practice, there are many issues: (1) we are forced to specify a functional form, the linear model; (2) we end up with unstable variance estimates that sharply increase the number of false positives detected, even after multiple testing corrections. In practice, the incredible flexibility of the linear mode is rarely taken advantage of — scientific guidance is usually lacking to justify the fitting of richer models.

A familiar workhorse: the linear model

- The linear model is *semiparametric* — linear in parameters!
- Flexible: transformations (X_j^2), interactions ($X_j X_k$).
- For biomarker Y_b, fit *working* linear model, $\mathbb{E}_0[Y_b \mid X] = X\beta$, letting (wlog) $X_1 \equiv A$ be the exposure and β_1 its "effect".
- Under this working model, the parameter β_1 is the ATE, but *only if* there are no interactions (i.e., without flexibility).
- Test the contrast of interest with a standard t-test:

$$t_b = \frac{\hat{\beta}_b - \beta_{b,H_0}}{\hat{\sigma}_b}$$

There's nothing particularly wrong with this approach. It's exactly what we would come up with after a first-year statistics course. In practice, there are many issues: (1) we are forced to specify a functional form, the linear model; (2) we end up with unstable variance estimates that sharply increase the number of false positives detected, even after multiple testing corrections. In practice, the incredible flexibility of the linear mode is rarely taken advantage of — scientific guidance is usually lacking to justify the fitting of richer models.

A familiar workhorse: the linear model

- The linear model is *semiparametric* — linear in parameters!
- Flexible: transformations (X_j^2), interactions ($X_j X_k$).
- For biomarker Y_b, fit *working* linear model, $\mathbb{E}_0[Y_b \mid X] = X\beta$, letting (wlog) $X_1 \equiv A$ be the exposure and β_1 its "effect".
- Under this working model, the parameter β_1 is the ATE, but *only if* there are no interactions (i.e., without flexibility).
- Test the contrast of interest with a standard t-test:

$$t_b = \frac{\hat{\beta}_b - \beta_{b,H_0}}{\hat{\sigma}_b}$$

There's nothing particularly wrong with this approach. It's exactly what we would come up with after a first-year statistics course. In practice, there are many issues: (1) we are forced to specify a functional form, the linear model; (2) we end up with unstable variance estimates that sharply increase the number of false positives detected, even after multiple testing corrections. In practice, the incredible flexibility of the linear mode is rarely taken advantage of — scientific guidance is usually lacking to justify the fitting of richer models.

Variance moderation to the rescue?!

- When sample size is small, σ_b^2 may be so small (by chance) that even small effects $(\hat{\beta}_b - \beta_{b,H_0})$ give large t_b.
- False positives! Many biomarkers flagged relevant despite small effect size, only since variance is even smaller still.
- Can we do better? A **moderated** t-test (**?**):

$$\tilde{t}_b = \frac{\hat{\beta}_b - \beta_{b,H_0}}{\tilde{\sigma}_b} \quad \text{where} \quad \tilde{\sigma}_b^2 = \frac{\sigma_b^2 d_b + \sigma_0^2 d_0}{d_b + d_0}$$

- Helps reduce erroneously large t_b by "averaging out" low variance across each of the many biomarkers.

The substantive contribution here is the use of an empirical Bayes method to shrink the standard deviation across all of the biomarkers such that we obtain a larger (but accurate) estimate that reduces the number of test statistics that are marked as significant by low s_b^2 estimates alone.

Note that this is **not** the exact formulation of the moderated t-statistic as given by Smyth (his derivation assumes a hierarchical model; see original paper if interested). This formulation does a good enough job to help us see the bigger picture.

Variance moderation to the rescue?!

- When sample size is small, σ_b^2 may be so small (by chance) that even small effects $(\hat{\beta}_b - \beta_{b,H_0})$ give large t_b.
- False positives! Many biomarkers flagged relevant despite small effect size, only since variance is even smaller still.
- Can we do better? A **moderated** t-test (**?**):

$$\tilde{t}_b = \frac{\hat{\beta}_b - \beta_{b,H_0}}{\tilde{\sigma}_b} \quad \text{where} \quad \tilde{\sigma}_b^2 = \frac{\sigma_b^2 d_b + \sigma_0^2 d_0}{d_b + d_0}$$

- Helps reduce erroneously large t_b by "averaging out" low variance across each of the many biomarkers.

The substantive contribution here is the use of an empirical Bayes method to shrink the standard deviation across all of the biomarkers such that we obtain a larger (but accurate) estimate that reduces the number of test statistics that are marked as significant by low s_b^2 estimates alone.

Note that this is **not** the exact formulation of the moderated t-statistic as given by Smyth (his derivation assumes a hierarchical model; see original paper if interested). This formulation does a good enough job to help us see the bigger picture.

Variance moderation to the rescue?!

- When sample size is small, σ_b^2 may be so small (by chance) that even small effects $(\hat{\beta}_b - \beta_{b,H_0})$ give large t_b.
- False positives! Many biomarkers flagged relevant despite small effect size, only since variance is even smaller still.
- Can we do better? A **moderated** t-test (**?**):

$$\tilde{t}_b = \frac{\hat{\beta}_b - \beta_{b,H_0}}{\tilde{\sigma}_b} \quad \text{where} \quad \tilde{\sigma}_b^2 = \frac{\sigma_b^2 d_b + \sigma_0^2 d_0}{d_b + d_0}$$

- Helps reduce erroneously large t_b by "averaging out" low variance across each of the many biomarkers.

The substantive contribution here is the use of an empirical Bayes method to shrink the standard deviation across all of the biomarkers such that we obtain a larger (but accurate) estimate that reduces the number of test statistics that are marked as significant by low s_b^2 estimates alone.

Note that this is **not** the exact formulation of the moderated t-statistic as given by Smyth (his derivation assumes a hierarchical model; see original paper if interested). This formulation does a good enough job to help us see the bigger picture.

Variance moderation to the rescue?!

- When sample size is small, σ_b^2 may be so small (by chance) that even small effects $(\hat{\beta}_b - \beta_{b,H_0})$ give large t_b.
- False positives! Many biomarkers flagged relevant despite small effect size, only since variance is even smaller still.
- Can we do better? A **moderated** t-test (**?**):

$$\tilde{t}_b = \frac{\hat{\beta}_b - \beta_{b,H_0}}{\tilde{\sigma}_b} \quad \text{where} \quad \tilde{\sigma}_b^2 = \frac{\sigma_b^2 d_b + \sigma_0^2 d_0}{d_b + d_0}$$

- Helps reduce erroneously large t_b by "averaging out" low variance across each of the many biomarkers.

The substantive contribution here is the use of an empirical Bayes method to shrink the standard deviation across all of the biomarkers such that we obtain a larger (but accurate) estimate that reduces the number of test statistics that are marked as significant by low s_b^2 estimates alone.

Note that this is **not** the exact formulation of the moderated t-statistic as given by Smyth (his derivation assumes a hierarchical model; see original paper if interested). This formulation does a good enough job to help us see the bigger picture.

Variable importance measures as target parameters!

- If the working model is incorrect, β_b does not correspond to the ATE — results polluted by *misspecification bias*.
- The statistical functional identifying the ATE may be used as an interpretable variable importance measure (VIM):

$$\psi_{b,0} \equiv \Psi_b(P_0) = \mathbb{E}_{L,0}[\mathbb{E}_0[Y_b \mid A = 1, L] - \mathbb{E}_0[Y_b \mid A = 0, L]]$$

- $\psi_{b,0}$ is a mapping ($\Psi_b(P_0)$) that depends on the underlying true (but unknown) distribution $P_0 \in \mathscr{M}$ — model-agnostic!
- The statistical functional *identifies* the ATE under untestable assumptions (no unmeasured confounding, positivity).

By allowing scientific questions to inform the parameters that we choose to estimate, we can do a better job of actually answering the questions of interest to our collaborators. Further, we abandon the need to specify the functional relationship between our outcome and covariates; moreover, we can now make use of advances in machine learning.

Variable importance measures as target parameters!

- If the working model is incorrect, β_b does not correspond to the ATE — results polluted by *misspecification bias*.
- The statistical functional identifying the ATE may be used as an interpretable variable importance measure (VIM):

$$\psi_{b,0} \equiv \Psi_b(P_0) = \mathbb{E}_{L,0}[\mathbb{E}_0[Y_b \mid A = 1, L] - \mathbb{E}_0[Y_b \mid A = 0, L]]$$

- $\psi_{b,0}$ is a mapping ($\Psi_b(P_0)$) that depends on the underlying true (but unknown) distribution $P_0 \in \mathscr{M}$ — model-agnostic!
- The statistical functional *identifies* the ATE under untestable assumptions (no unmeasured confounding, positivity).

By allowing scientific questions to inform the parameters that we choose to estimate, we can do a better job of actually answering the questions of interest to our collaborators. Further, we abandon the need to specify the functional relationship between our outcome and covariates; moreover, we can now make use of advances in machine learning.

Variable importance measures as target parameters!

- If the working model is incorrect, β_b does not correspond to the ATE — results polluted by *misspecification bias*.
- The statistical functional identifying the ATE may be used as an interpretable variable importance measure (VIM):

$$\psi_{b,0} \equiv \Psi_b(P_0) = \mathbb{E}_{L,0}[\mathbb{E}_0[Y_b \mid A = 1, L] - \mathbb{E}_0[Y_b \mid A = 0, L]]$$

- $\psi_{b,0}$ is a mapping ($\Psi_b(P_0)$) that depends on the underlying true (but unknown) distribution $P_0 \in \mathscr{M}$ — model-agnostic!
- The statistical functional *identifies* the ATE under untestable assumptions (no unmeasured confounding, positivity).

By allowing scientific questions to inform the parameters that we choose to estimate, we can do a better job of actually answering the questions of interest to our collaborators. Further, we abandon the need to specify the functional relationship between our outcome and covariates; moreover, we can now make use of advances in machine learning.

Variable importance measures as target parameters!

- If the working model is incorrect, β_b does not correspond to the ATE — results polluted by *misspecification bias*.
- The statistical functional identifying the ATE may be used as an interpretable variable importance measure (VIM):

$$\psi_{b,0} \equiv \Psi_b(P_0) = \mathbb{E}_{L,0}[\mathbb{E}_0[Y_b \mid A = 1, L] - \mathbb{E}_0[Y_b \mid A = 0, L]]$$

- $\psi_{b,0}$ is a mapping ($\Psi_b(P_0)$) that depends on the underlying true (but unknown) distribution $P_0 \in \mathscr{M}$ — model-agnostic!
- The statistical functional *identifies* the ATE under untestable assumptions (no unmeasured confounding, positivity).

By allowing scientific questions to inform the parameters that we choose to estimate, we can do a better job of actually answering the questions of interest to our collaborators. Further, we abandon the need to specify the functional relationship between our outcome and covariates; moreover, we can now make use of advances in machine learning.

Locally efficient estimation

- An estimator $\hat{\psi}_b$ is asymptotically linear if it admits the form

$$\hat{\psi}_b - \psi_{b,0} = \frac{1}{n}\sum_{i=1}^{n} D_b(O_i; P_0) + o_P\left(\frac{1}{\sqrt{n}}\right),$$

  where $D_b(O; P_0)$ is the efficient influence function (wrt $\mathscr{M}$), whose asymptotic variance is the *efficiency bound*.

- $D_b(O; P_0)$ helps construct efficient estimators. For ATE,

$$D_b(O_i; P_0) = \left[\frac{2A_i - 1}{g_0(L_i)}\right](Y_{b,i} - \overline{Q}_{0,b}(A_i, L_i)) + \overline{Q}_{0,b}(1, L_i) - \overline{Q}_{0,b}(0, L_i) - \psi_{b,0},$$

  where $g_0(L) = \mathbb{P}_0(A = 1 \mid L)$ is the "propensity score" and $\overline{Q}_{0,b}(A, L) = \mathbb{E}_0[Y_b \mid A, L]$ is conditional outcome mean.

Natural use of machine learning methods for the estimation of both Q_0 and g_0. Focuses effort to achieve minimal bias and asymptotic semiparametric efficiency bound for the variance, but still get inference (with some assumptions).

Locally efficient estimation

- An estimator $\hat{\psi}_b$ is asymptotically linear if it admits the form

$$\hat{\psi}_b - \psi_{b,0} = \frac{1}{n}\sum_{i=1}^{n} D_b(O_i; P_0) + o_P\left(\frac{1}{\sqrt{n}}\right),$$

  where $D_b(O; P_0)$ is the efficient influence function (wrt $\mathscr{M}$), whose asymptotic variance is the *efficiency bound*.

- $D_b(O; P_0)$ helps construct efficient estimators. For ATE,

$$\begin{aligned} D_b(O_i; P_0) = &\left[\frac{2A_i - 1}{g_0(L_i)}\right](Y_{b,i} - \overline{Q}_{0,b}(A_i, L_i)) \\ &+ \overline{Q}_{0,b}(1, L_i) - \overline{Q}_{0,b}(0, L_i) - \psi_{b,0}, \end{aligned}$$

  where $g_0(L) = \mathbb{P}_0(A = 1 \mid L)$ is the "propensity score" and $\overline{Q}_{0,b}(A, L) = \mathbb{E}_0[Y_b \mid A, L]$ is conditional outcome mean.

Natural use of machine learning methods for the estimation of both Q_0 and g_0. Focuses effort to achieve minimal bias and asymptotic semiparametric efficiency bound for the variance, but still get inference (with some assumptions).

Constructing locally efficient estimators

- Examining $D_b(O; P_0)$, we know we must estimate $g_0(L)$ and $\overline{Q}_{0,b}(A, L)$, but how exactly we do this is unspecified.
- No need to try to exactly specify functional forms or assume we know the underlying true data-generating distribution P_0.
- Instead, machine learning to estimate $g_0(L)$ and $\overline{Q}_{0,b}(A, L)$, e.g., by ensemble modeling (**?**).
- One-step estimator (**?**) "debiased" by an additive correction: $\hat{\psi}_b^+ = \hat{\psi}_b + n^{-1} \sum_{i=1}^n \hat{D}_b(O_i)$.
- A valid variance estimator: $\hat{\mathbb{V}}(\hat{\psi}_b^+) = n^{-1} \sum_{i=1}^n \hat{D}_b^2(O_i)$, but small-sample behavior may be erratic.

Natural use of machine learning methods for the estimation of both Q_0 and g_0. Focuses effort to achieve minimal bias and asymptotic semiparametric efficiency bound for the variance, but still get inference (with some assumptions).

Constructing locally efficient estimators

- Examining $D_b(O; P_0)$, we know we must estimate $g_0(L)$ and $\overline{Q}_{0,b}(A, L)$, but how exactly we do this is unspecified.
- No need to try to exactly specify functional forms or assume we know the underlying true data-generating distribution P_0.
- Instead, machine learning to estimate $g_0(L)$ and $\overline{Q}_{0,b}(A, L)$, e.g., by ensemble modeling (**?**).
- One-step estimator (**?**) "debiased" by an additive correction: $\hat{\psi}_b^+ = \hat{\psi}_b + n^{-1} \sum_{i=1}^n \hat{D}_b(O_i)$.
- A valid variance estimator: $\hat{\mathbb{V}}(\hat{\psi}_b^+) = n^{-1} \sum_{i=1}^n \hat{D}_b^2(O_i)$, but small-sample behavior may be erratic.

Natural use of machine learning methods for the estimation of both Q_0 and g_0. Focuses effort to achieve minimal bias and asymptotic semiparametric efficiency bound for the variance, but still get inference (with some assumptions).

Constructing locally efficient estimators

- Examining $D_b(O; P_0)$, we know we must estimate $g_0(L)$ and $\overline{Q}_{0,b}(A, L)$, but how exactly we do this is unspecified.
- No need to try to exactly specify functional forms or assume we know the underlying true data-generating distribution P_0.
- Instead, machine learning to estimate $g_0(L)$ and $\overline{Q}_{0,b}(A, L)$, e.g., by ensemble modeling (**?**).
- One-step estimator (**?**) "debiased" by an additive correction: $\hat{\psi}_b^+ = \hat{\psi}_b + n^{-1} \sum_{i=1}^n \hat{D}_b(O_i)$.
- A valid variance estimator: $\hat{\mathbb{V}}(\hat{\psi}_b^+) = n^{-1} \sum_{i=1}^n \hat{D}_b^2(O_i)$, but small-sample behavior may be erratic.

Natural use of machine learning methods for the estimation of both Q_0 and g_0. Focuses effort to achieve minimal bias and asymptotic semiparametric efficiency bound for the variance, but still get inference (with some assumptions).

Constructing locally efficient estimators

- Examining $D_b(O; P_0)$, we know we must estimate $g_0(L)$ and $\overline{Q}_{0,b}(A, L)$, but how exactly we do this is unspecified.
- No need to try to exactly specify functional forms or assume we know the underlying true data-generating distribution P_0.
- Instead, machine learning to estimate $g_0(L)$ and $\overline{Q}_{0,b}(A, L)$, e.g., by ensemble modeling (**?**).
- One-step estimator (**?**) "debiased" by an additive correction: $\hat{\psi}_b^+ = \hat{\psi}_b + n^{-1} \sum_{i=1}^n \hat{D}_b(O_i)$.
- A valid variance estimator: $\hat{\mathbb{V}}(\hat{\psi}_b^+) = n^{-1} \sum_{i=1}^n \hat{D}_b^2(O_i)$, but small-sample behavior may be erratic.

Natural use of machine learning methods for the estimation of both Q_0 and g_0. Focuses effort to achieve minimal bias and asymptotic semiparametric efficiency bound for the variance, but still get inference (with some assumptions).

Constructing locally efficient estimators

- Examining $D_b(O; P_0)$, we know we must estimate $g_0(L)$ and $\overline{Q}_{0,b}(A, L)$, but how exactly we do this is unspecified.
- No need to try to exactly specify functional forms or assume we know the underlying true data-generating distribution P_0.
- Instead, machine learning to estimate $g_0(L)$ and $\overline{Q}_{0,b}(A, L)$, e.g., by ensemble modeling (**?**).
- One-step estimator (**?**) "debiased" by an additive correction: $\hat{\psi}_b^+ = \hat{\psi}_b + n^{-1} \sum_{i=1}^n \hat{D}_b(O_i)$.
- A valid variance estimator: $\hat{\mathbb{V}}(\hat{\psi}_b^+) = n^{-1} \sum_{i=1}^n \hat{D}_b^2(O_i)$, but small-sample behavior may be erratic.

Natural use of machine learning methods for the estimation of both Q_0 and g_0. Focuses effort to achieve minimal bias and asymptotic semiparametric efficiency bound for the variance, but still get inference (with some assumptions).

Moderated test statistics with efficient influence functions

- Moderated t-statistic of **?** naturally extends to locally efficient estimators by noticing

$$\tilde{t}_b = \frac{\hat{\psi}_b^+ - \cancel{\psi_{b,0}}\, 0}{\tilde{\sigma}_b},$$

  where the *moderated* influence function variance is

$$\tilde{\sigma}_b^2 = \frac{\hat{\sigma}_b^2 d_b + \hat{\sigma}_0^2 d_0}{d_b + d_0}$$

- Preserves robust variance estimator while adding stability by "averaging out" potentially erratic variance across biomarkers.
- Avoid model misspecification while stabilizing inference.

Moderated test statistics with efficient influence functions

- Moderated t-statistic of **?** naturally extends to locally efficient estimators by noticing

$$\tilde{t}_b = \frac{\hat{\psi}_b^+ - \cancel{\psi_{b,0}}\ 0}{\tilde{\sigma}_b},$$

  where the *moderated* influence function variance is

$$\tilde{\sigma}_b^2 = \frac{\hat{\sigma}_b^2 d_b + \hat{\sigma}_0^2 d_0}{d_b + d_0}$$

- Preserves robust variance estimator while adding stability by "averaging out" potentially erratic variance across biomarkers.
- Avoid model misspecification while stabilizing inference.

Moderated test statistics with efficient influence functions

- Moderated t-statistic of **?** naturally extends to locally efficient estimators by noticing

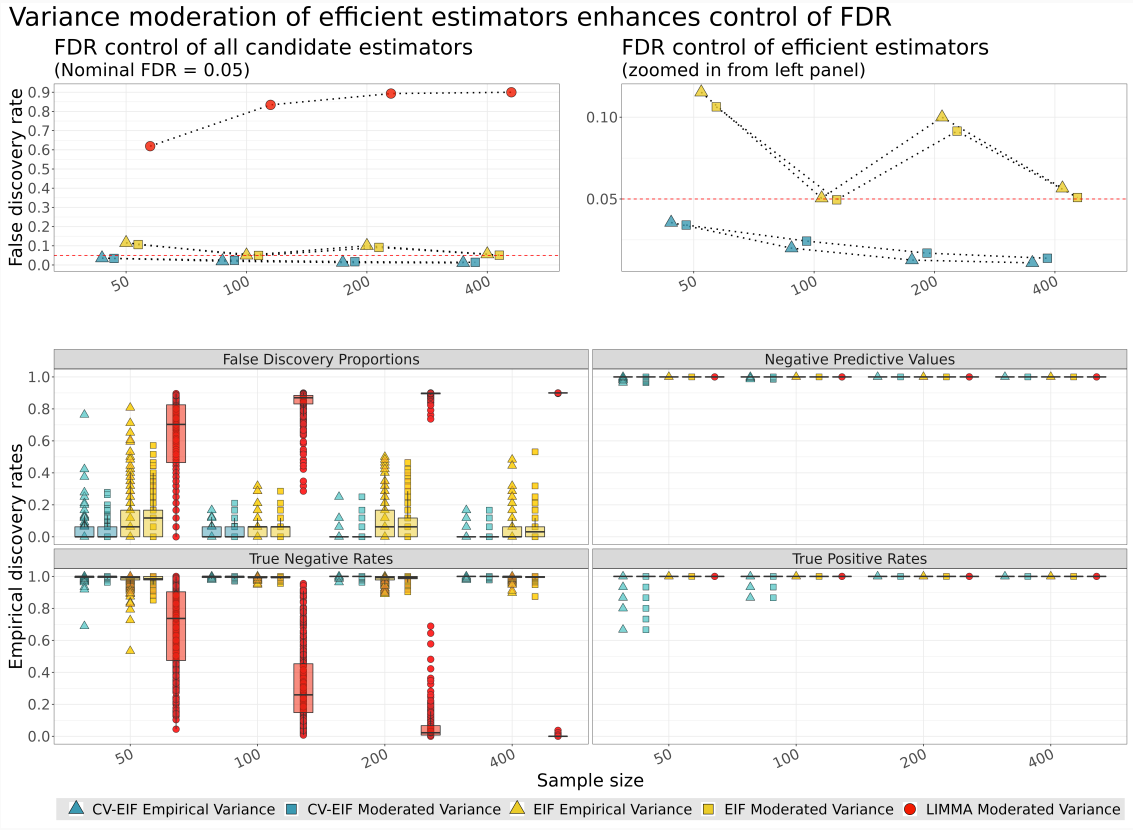
$$\tilde{t}_b = \frac{\hat{\psi}_b^+ - \cancel{\psi_{b,0}}\, 0}{\tilde{\sigma}_b},$$

  where the *moderated* influence function variance is

$$\tilde{\sigma}_b^2 = \frac{\hat{\sigma}_b^2 d_b + \hat{\sigma}_0^2 d_0}{d_b + d_0}$$

- Preserves robust variance estimator while adding stability by "averaging out" potentially erratic variance across biomarkers.
- Avoid model misspecification while stabilizing inference.

Let's take a look: Numerical study



Differential expression analysis algorithm

- Apply a filtering procedure to reduce the set of candidate biomarkers (**?**) (optional).
- For each biomarker, generate an efficient estimate of $\hat{\psi}_b$ of $\psi_{0,b}$ with EIF $\hat{D}_b(O_i)$ by estimating nuisances $(g_0, \overline{Q}_{0,b})$.
- Apply variance moderation across the EIF estimates, yielding *moderated* $\tilde{\sigma}_b^2$, to be used for hypothesis testing.
- Various techniques for inference are possible based on the moderated test statistics – taking advantage of near-normality, standardized logistic or concentration inequalities.
- Apply a multiple testing correction for accurate simultaneous inference inference across all B biomarkers, e.g., by controlling the False Discovery Rate (**?**).

Differential expression analysis algorithm

- Apply a filtering procedure to reduce the set of candidate biomarkers (**?**) (optional).
- For each biomarker, generate an efficient estimate of $\hat{\psi}_b$ of $\psi_{0,b}$ with EIF $\hat{D}_b(O_i)$ by estimating nuisances $(g_0, \overline{Q}_{0,b})$.
- Apply variance moderation across the EIF estimates, yielding *moderated* $\tilde{\sigma}_b^2$, to be used for hypothesis testing.
- Various techniques for inference are possible based on the moderated test statistics – taking advantage of near-normality, standardized logistic or concentration inequalities.
- Apply a multiple testing correction for accurate simultaneous inference inference across all B biomarkers, e.g., by controlling the False Discovery Rate (**?**).

Differential expression analysis algorithm

- Apply a filtering procedure to reduce the set of candidate biomarkers (**?**) (optional).
- For each biomarker, generate an efficient estimate of $\hat{\psi}_b$ of $\psi_{0,b}$ with EIF $\hat{D}_b(O_i)$ by estimating nuisances $(g_0, \overline{Q}_{0,b})$.
- Apply variance moderation across the EIF estimates, yielding *moderated* $\tilde{\sigma}_b^2$, to be used for hypothesis testing.
- Various techniques for inference are possible based on the moderated test statistics – taking advantage of near-normality, standardized logistic or concentration inequalities.
- Apply a multiple testing correction for accurate simultaneous inference inference across all B biomarkers, e.g., by controlling the False Discovery Rate (**?**).

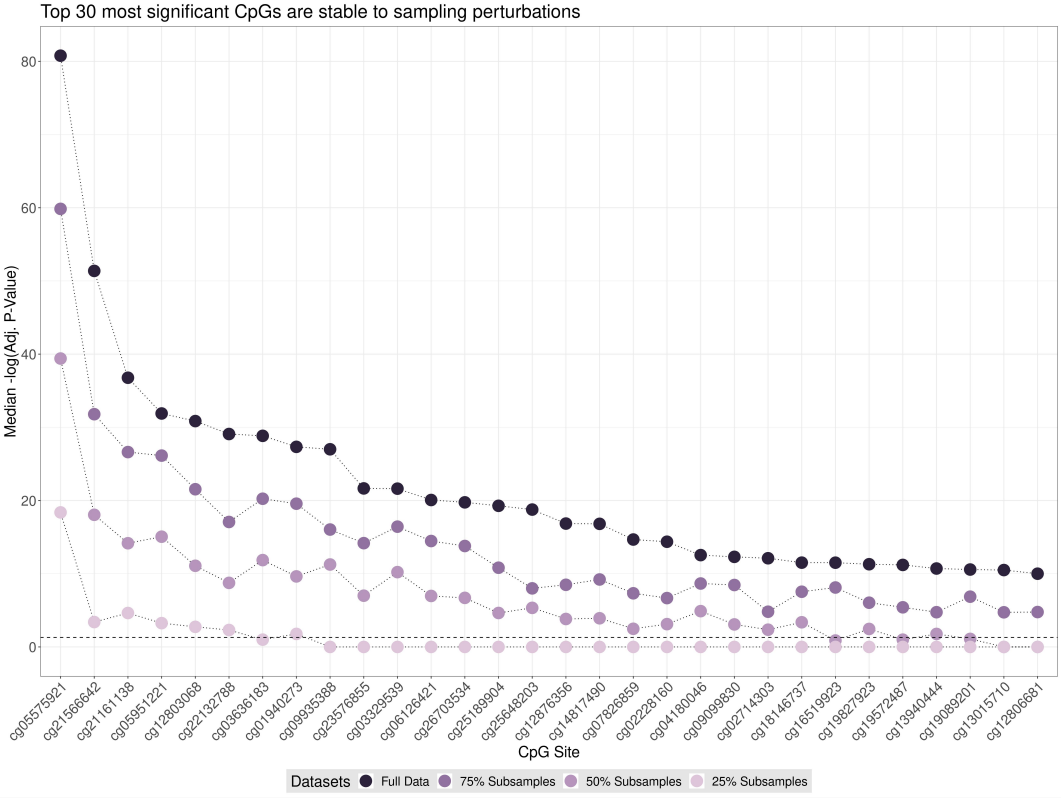
Differential expression analysis algorithm

- Apply a filtering procedure to reduce the set of candidate biomarkers (**?**) (optional).
- For each biomarker, generate an efficient estimate of $\hat{\psi}_b$ of $\psi_{0,b}$ with EIF $\hat{D}_b(O_i)$ by estimating nuisances $(g_0, \overline{Q}_{0,b})$.
- Apply variance moderation across the EIF estimates, yielding *moderated* $\tilde{\sigma}_b^2$, to be used for hypothesis testing.
- Various techniques for inference are possible based on the moderated test statistics – taking advantage of near-normality, standardized logistic or concentration inequalities.
- Apply a multiple testing correction for accurate simultaneous inference inference across all B biomarkers, e.g., by controlling the False Discovery Rate (**?**).

Differential expression analysis algorithm

- Apply a filtering procedure to reduce the set of candidate biomarkers (**?**) (optional).
- For each biomarker, generate an efficient estimate of $\hat{\psi}_b$ of $\psi_{0,b}$ with EIF $\hat{D}_b(O_i)$ by estimating nuisances $(g_0, \overline{Q}_{0,b})$.
- Apply variance moderation across the EIF estimates, yielding *moderated* $\tilde{\sigma}_b^2$, to be used for hypothesis testing.
- Various techniques for inference are possible based on the moderated test statistics – taking advantage of near-normality, standardized logistic or concentration inequalities.
- Apply a multiple testing correction for accurate simultaneous inference inference across all B biomarkers, e.g., by controlling the False Discovery Rate (**?**).

Ranking differentially methylated CpGs



Open-source software: `R/biotmle`!

- `R` package for differential expression or methylation analysis based on model-agnostic, efficient estimators of the ATE.
- Incorporates machine learning and allows cross-validation.
- Statistical inference based on variance *moderation*.
- Where can you find it?
  - `https://github.com/nhejazi/biotmle`
  - `https://bioconductor.org/packages/biotmle`

Use it. File an issue. Help make it better!

Open-source software: `R/biotmle`!

- `R` package for differential expression or methylation analysis based on model-agnostic, efficient estimators of the ATE.
- Incorporates machine learning and allows cross-validation.
- Statistical inference based on variance *moderation*.
- Where can you find it?
  - `https://github.com/nhejazi/biotmle`
  - `https://bioconductor.org/packages/biotmle`

Use it. File an issue. Help make it better!

Open-source software: `R/biotmle`!

- `R` package for differential expression or methylation analysis based on model-agnostic, efficient estimators of the ATE.
- Incorporates machine learning and allows cross-validation.
- Statistical inference based on variance *moderation*.
- Where can you find it?
  - `https://github.com/nhejazi/biotmle`
  - `https://bioconductor.org/packages/biotmle`

Use it. File an issue. Help make it better!

Open-source software: `R/biotmle`!

- `R` package for differential expression or methylation analysis based on model-agnostic, efficient estimators of the ATE.
- Incorporates machine learning and allows cross-validation.
- Statistical inference based on variance *moderation*.
- Where can you find it?
  - `https://github.com/nhejazi/biotmle`
  - `https://bioconductor.org/packages/biotmle`

Use it. File an issue. Help make it better!

Open-source software: `R/biotmle`!

- `R` package for differential expression or methylation analysis based on model-agnostic, efficient estimators of the ATE.
- Incorporates machine learning and allows cross-validation.
- Statistical inference based on variance *moderation*.
- Where can you find it?
  - `https://github.com/nhejazi/biotmle`
  - `https://bioconductor.org/packages/biotmle`

Use it. File an issue. Help make it better!

Open-source software: `R/biotmle`!

- `R` package for differential expression or methylation analysis based on model-agnostic, efficient estimators of the ATE.
- Incorporates machine learning and allows cross-validation.
- Statistical inference based on variance *moderation*.
- Where can you find it?
  - `https://github.com/nhejazi/biotmle`
  - `https://bioconductor.org/packages/biotmle`

Use it. File an issue. Help make it better!

Review

1. Modern computational biology research produces complex, heterogeneous data — innovative statistical inference still tied to simplistic and challenging-to-verify modeling assumptions.
2. *Model misspecification* seriously undermines the scientific utility of common, classical statistical modeling approaches.
3. Non/semi-parametric inference facilitates constructing robust estimators that easily bring machine learning into the fold.
4. Variance moderation strengthens hypothesis testing strategies, reducing false positives and preserving power under instability.

It's always good to include a summary.

Review

1. Modern computational biology research produces complex, heterogeneous data — innovative statistical inference still tied to simplistic and challenging-to-verify modeling assumptions.
2. *Model misspecification* seriously undermines the scientific utility of common, classical statistical modeling approaches.
3. Non/semi-parametric inference facilitates constructing robust estimators that easily bring machine learning into the fold.
4. Variance moderation strengthens hypothesis testing strategies, reducing false positives and preserving power under instability.

It's always good to include a summary.

Review

1. Modern computational biology research produces complex, heterogeneous data — innovative statistical inference still tied to simplistic and challenging-to-verify modeling assumptions.
2. *Model misspecification* seriously undermines the scientific utility of common, classical statistical modeling approaches.
3. Non/semi-parametric inference facilitates constructing robust estimators that easily bring machine learning into the fold.
4. Variance moderation strengthens hypothesis testing strategies, reducing false positives and preserving power under instability.

It's always good to include a summary.

Review

1. Modern computational biology research produces complex, heterogeneous data — innovative statistical inference still tied to simplistic and challenging-to-verify modeling assumptions.
2. *Model misspecification* seriously undermines the scientific utility of common, classical statistical modeling approaches.
3. Non/semi-parametric inference facilitates constructing robust estimators that easily bring machine learning into the fold.
4. Variance moderation strengthens hypothesis testing strategies, reducing false positives and preserving power under instability.

It's always good to include a summary.

References

Benjamini, Y. and 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)*, pages 289–300.

Bickel, P. J., Klaassen, C. A., Ritov, Y., and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press.

Pearl, J. (2000). *Causality*. Cambridge university press.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331.

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1):1–25.

Tuglus, C. and van der Laan, M. J. (2009). Modified FDR controlling procedure for multi-stage analyses. *Statistical Applications in Genetics and Molecular Biology*, 8(1).

van der Laan, M. J., Polley, E. C., and Hubbard, A. E. (2007). Super Learner. *Statistical applications in genetics and molecular biology*, 6(1).

Thank you

https://nimahejazi.org

https://github.com/nhejazi

https://twitter.com/nshejazi

Here's where you can find me, as well as the slides for this talk.