Robust and Efficient High-dimensional Inference With Surrogate Outcomes

Huiyuan Wang

University of Pennsylvania

Joint work with Jianmin Chen (Upenn), Yang Ning (Cornell) and Yong Chen (Upenn)

2024 ICSA Applied Statistics Symposium, Nashville

June 18, 2024

Background

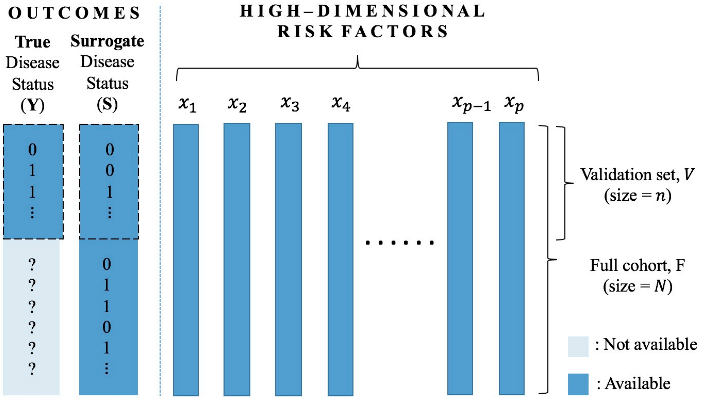
- One common use of EHR data is identification of novel risk factors for diseases
  - Y: binary phenotype of interest
  - $\mathbf{X} = (X_1, \ldots, X_p)^T$: the vector of p risk factors
  - The statistical association between $\mathbf{X}$ and Y is modeled by

$$\mathbb{P}(Y = 1 \mid \mathbf{X}) = \text{Expit}(X_1\beta_1^\star + \cdots + X_p\beta_p^\star)$$

- Identification of risk factors is equivalent to testing

$$H_{0,j} : \beta_j^\star = 0 \quad \text{versus} \quad H_{1,j} : \beta_j^\star \neq 0, \quad \text{for } j = 1, \ldots, p$$

Data Structure



- Data: $\{(\mathbf{X}_i, S_i)\}_{i \in F \setminus V} \cup \{(\mathbf{X}_i, S_i, Y_i)\}_{i \in V}$, where F denotes the full cohort and V the validation (chart-reviewed) set
- V is selected via random sampling (c.f. *Missing Completely at Random*)

Challenges

- **Small validated set**: The true phenotype Y is severely missing
  - Labeling Y relies on *manual chart review*, which is expensive often prohibitively

$$\frac{\text{\#chart-reviewed samples}}{\text{\#total samples}} \approx 0$$

  - Using only chart-reviewed samples for testing is often *inefficient*
- **High-dimensionality**:

$$\underbrace{\text{\#total samples}}_{N} \gg p \gg \underbrace{\text{\#chart-reviewed samples}}_{n}$$

Challenges

- Misclassified surrogates: S, a surrogate of Y, can be obtained for all samples from computational phenotyping algorithms
  - S is typically inaccurate; 28%–60% of patients are misclassified (Carroll et al. 2012)
  - Ignoring the misclassification and treating surrogates as true labels will *lead to substantial biased estimates and inflated Type I errors* (Duan et al. 2016)

Robert J. Carroll et al. Portability of an algorithm to identify rheumatoid arthritis in electronic health records. *Journal of the American Medical Informatics Association*, 19(e1):e162–e169, 2012.

Rui Duan et al. An empirical study for impacts of measurement errors on EHR based association studies. In AMIA Annual Symposium Proceedings, page 1764, 2016.

Score Test

- For a given j, consider

$$H_{0,j}:\beta_j^\star = 0 \quad \text{versus} \quad H_{1,j}:\beta_j^\star \neq 0$$

  - Let $\phi_j(\beta_j;\beta_{\setminus j},Y,\mathbf{X})$ be any score function of $\beta_j^\star$, where $\beta_{\setminus j} = (\beta_i, i \neq j)^T$
  - By properties of score function, at the truth $\beta_j = \beta_j^\star$ and $\beta_{\setminus j} = \beta_{\setminus j}^\star$

$$\frac{1}{\sqrt{n}}\sum_{i\in V}\phi_j(\beta_j^\star;\beta_{\setminus j}^\star,Y_i,\mathbf{X}_i) \to_d N(0,\text{Var}(\phi_j))$$

  - Replacing $\beta_{\setminus j}^\star$ and $\text{Var}(\phi_j)$ with sufficiently "good" estimators $\hat\beta_{\setminus j}$ and $\widehat{\text{Var}}(\phi_j)$, respectively, we can construct the score-based test statistic for the null

$$T_n^{(\alpha)}(\phi_j) = \begin{cases} 1, & \left|\sum_{i\in V}\phi_j(0;\hat\beta_{\setminus j},Y_i,\mathbf{X}_i)\right| \geq \sqrt{n\widehat{\text{Var}}(\phi_j)}z_{1-\alpha/2} \\ 0, & \text{otherwise} \end{cases}$$

  - *Smaller $\text{Var}(\phi_j)$ gives rise to more powerful $T_n^{(\alpha)}(\phi_j)$*

Decorrelated Score Test

- Viewing $\beta_j^\star$ as the target parameter, its score function is

$$\phi_j(\beta_j^\star; \beta_{\setminus j}^\star, \mathbf{X}, Y) = \frac{\partial \log\{\mathbb{P}(Y=1 \mid \mathbf{X})^Y \mathbb{P}(Y=0 \mid \mathbf{X})^{1-Y}\}}{\partial \beta_j}$$
$$= \{Y - \text{Expit}(\mathbf{X}^T \beta^\star)\} X_j$$

- The score function for the nuisance parameter $\beta_{\setminus j} = (\beta_i, i \neq j)^T$ is

$$\phi_{\setminus j}(\beta_{\setminus j}^\star; \beta_j^\star, \mathbf{X}, Y) = \{Y - \text{Expit}(\mathbf{X}^T \beta^\star)\} X_{\setminus j}$$

- The efficient score function for $\beta_j^\star$ (Tsiatis 2006, Ning and Liu 2017) is

$$\phi_j^{\text{val-eff}}(\beta_j^\star; \beta_{\setminus j}^\star, \mathbf{w}^\star, \mathbf{X}, Y) = \phi_j(\beta_j^\star; \beta_{\setminus j}^\star, \mathbf{X}, Y) - \mathbf{w}^{\star T} \phi_{\setminus j}(\beta_{\setminus j}^\star; \beta_j^\star, \mathbf{X}, Y),$$

where $\mathbf{w}^\star$ is chosen such that $\phi_j^{\text{val-eff}}$ is *not correlated* with $\phi_{\setminus j}$

Anastasios A. Tsiatis. Semiparametric theory and missing data. Vol. 4. New York: Springer, 2006.

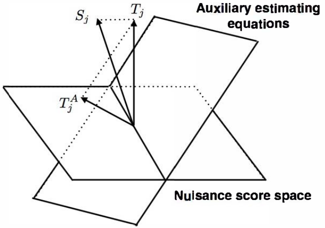
Yang Ning and Han Liu. A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *Annals of Statistics* 45.1:158-195, 2017.

Augmented Score Test for Variance Reduction

- Consider any function $h(S, \mathbf{X})$ with finite second moment Var$\{h(S, \mathbf{X})\} < \infty$
- $\mathbb{E}\{h(S, \mathbf{X})\}$ can be viewed as a nuisance parameter with score/influence function $h(S, \mathbf{X}) - \mathbb{E}\{h(S, \mathbf{X})\}$
- Note that $\mathbb{E}\{h(S, \mathbf{X})\}$ can be estimated by the whole sample

$$\mathbb{E}\{h(S, \mathbf{X})\} \approx \frac{1}{N} \sum_{i \in F} h(S_i, \mathbf{X}_i)$$

- Since $N \gg n$, we can view $\mathbb{E}\{h(S, \mathbf{X})\}$ as *known* asymptotically, which can offer us additional efficiency



Variance reduction by projection

Augmented Score Test for Variance Reduction

- **Proposition.** For any function $h(S,\mathbf{X})$ with finite second moment $\mathsf{Var}\{h(S,\mathbf{X})\} < \infty$ and any score function ϕ_j, if $\mathsf{Cov}\{\phi_j(Y,\mathbf{X}), h(S,\mathbf{X})\} \neq 0$, then the augmented score function

$$\phi_j^A(Y,\mathbf{X}) = \phi_j(Y,\mathbf{X}) - v^\star \underbrace{[h(S,\mathbf{X}) - \mathbb{E}\{h(S,\mathbf{X})\}]}_{\text{nuisance score}}$$

has a strictly smaller variance than ϕ_j, where $v^\star = \mathsf{Cov}\{\phi_j(Y,\mathbf{X}), h(S,\mathbf{X})\} / \mathsf{Var}\{h(S,\mathbf{X})\}$,

$$\mathsf{Var}(\phi_j) - \mathsf{Var}(\phi_j^A) = \frac{\{\mathsf{Cov}(\phi_j, h)\}^2}{\mathsf{Var}(h)} \leq \mathsf{Cov}\{\mathbb{E}\{\phi_j(Y,\mathbf{X}) \mid S,\mathbf{X}\}\},$$

and the equality holds when

$$h(S,\mathbf{X}) = h^\star(S,\mathbf{X}) \equiv \mathbb{E}\{\phi_j(Y,\mathbf{X}) \mid S,\mathbf{X}\}$$

Choice of h

- In practice, $h^\star$ is *unknown*
  - We can fit a regression model parametrized by γ on V: $\mathbb{E}(Y \mid S, \mathbf{X}) = f(S, \mathbf{X}; \gamma^\star)$ (e.g., *imputation*)
    - For the decorrelated score test,

$$h(S,\mathbf{X};\beta^\star,\mathbf{w}^\star,\gamma^\star) = \mathbb{E}\{\phi_j^{\text{val-eff}}(\beta^\star;\mathbf{w}^\star,\mathbf{X},Y) \mid S,\mathbf{X}\}$$
$$= \{f(S,\mathbf{X};\gamma^\star) - \text{Expit}(\mathbf{X}^T\beta^\star)\}\left(X_j - \mathbf{w}^{\star T}\mathbf{X}_{\setminus j}\right)$$

  - We can specify any other function $h(S,\mathbf{X})$ (*imputation-free*)
    - $h(S,\mathbf{X}) = (S,\mathbf{X}^T)^T$
    - $h(S,\mathbf{X};\gamma^\star) = \{S - \text{Expit}(\mathbf{X}^T\gamma^\star)\}g(\mathbf{X})$ for some weighting function $g(\cdot) \in \mathbb{R}^d$, where $\gamma^\star$ is the regression coefficient (Chen and Chen 2000)

$$\gamma^\star = \underset{\gamma}{\text{argmin}}\,\mathbb{E}\left\{-S\mathbf{X}^T\gamma + \log\left(1 + e^{\mathbf{X}^T\gamma}\right)\right\}$$

Chen, Yi-Hau, and Hung Chen."A unified approach to regression analysis under double-sampling designs." *Journal of the Royal Statistical Society Series B: Statistical Methodology* 62, no. 3 (2000): 449-460.

The Proposed Method for Hypothesis Testing

- Step 1: *Compute the decorrelated score function using validated samples* (under the null $H_{0,j} : \beta_j = 0$)

$$\phi_{ij}^{\text{val-eff}}(0, \hat{\beta}_{\setminus j}, \hat{\mathbf{w}}_j) = \left\{ Y_i - \text{expit}\left(\hat{\beta}_{\setminus j}{}^T \mathbf{X}_{i,\setminus j} \right) \right\} \left(X_{ij} - \hat{\mathbf{w}}_j{}^T \mathbf{X}_{i,\setminus j} \right),$$

where

$$\hat{\beta} = \underset{\beta}{\text{argmin}} \frac{1}{n} \sum_{i \in V} \left\{ -Y_i \mathbf{X}_i^T \beta + \log(1 + e^{\mathbf{X}_i^T \beta}) + \lambda \|\beta\|_1 \right\}$$

is the lasso estimator of $\beta^\star$, and

$$\hat{\mathbf{w}}_j = \left[\sum_{i \in F} \left\{ \hat{\mu}_{ij}(1 - \hat{\mu}_{ij}) \mathbf{X}_{i,\setminus j} \mathbf{X}_{i,\setminus j}^T \right\} \right]^{-1} \left[\sum_{i \in F} \left\{ \hat{\mu}_{ij}(1 - \hat{\mu}_{ij}) \mathbf{X}_{i,\setminus j} \mathbf{X}_{i,j} \right\} \right]$$

is the plug-in estimator of $\mathbf{w}^\star$ with $\hat{\mu}_{ij} = \text{Expit}(\mathbf{X}_{i,\setminus j}^T \hat{\beta}_{\setminus j})$

The Proposed Method for Hypothesis Testing

- Step 2: *Construct the augmented score function*:

$$\phi_{ij}^{A}(0, \hat{\beta}_{\setminus j}, \hat{\mathbf{w}}_j, h, \hat{\mathbf{v}}_j) = \phi_{ij}^{\text{val-eff}}(0, \hat{\beta}_{\setminus j}, \hat{\mathbf{w}}_j) - \hat{\mathbf{v}}_j^T \hbar(S_i, \mathbf{X}_i),$$

where $\hbar(S, \mathbf{X}) = h(S, \mathbf{X}) - (N-n)^{-1} \sum_{i \in F \setminus V} h(S_i, \mathbf{X}_i)$, and $\hat{\mathbf{v}}_j$ denotes the projection coefficient given by

$$\hat{\mathbf{v}}_j = \left[\frac{1}{N} \sum_{i \in F} \hbar(S_i, \mathbf{X}_i) \{\hbar(S_i, \mathbf{X}_i)\}^T \right]^{-1} \frac{1}{n} \sum_{i \in V} \left[\phi_{ij}^{\text{val-eff}}(0, \hat{\beta}_{\setminus j}, \hat{\mathbf{w}}_j) \hbar(S_i, \mathbf{X}_i) \right]$$

- Step 3: *Estimate the variance*

$$\widehat{\text{Var}}(\phi_j^A) = \widehat{\text{Var}}(\phi_j^{\text{val-eff}}) - \hat{\mathbf{v}}_j^T \left[\frac{1}{N} \sum_{i \in F} \hbar(S_i, \mathbf{X}_i) \{\hbar(S_i, \mathbf{X}_i)\}^T \right]^{-1} \hat{\mathbf{v}}_j$$

with $\widehat{\text{Var}}(\phi_j^{\text{val-eff}}) = n^{-1} \sum_{i \in V} \left\{ \phi_{ij}^{\text{val-eff}}(0, \hat{\beta}_{\setminus j}, \hat{\mathbf{w}}_j) \right\}^2$

- Step 4: Output the *test statistic*

$$T_n^{(\alpha)}(\phi_j^A) = \begin{cases} 1, & \left| \sum_{i \in V} \phi_{ij}^A(0, \hat{\beta}_{\setminus j}, \hat{\mathbf{w}}_j, h, \hat{\mathbf{v}}_j) \right| \geq \sqrt{n \widehat{\text{Var}}(\phi_j^A)} z_{1-\alpha/2} \\ 0, & \text{otherwise} \end{cases}$$

Theory

- Define $\phi_j^A(h)$ the augmented score function with $h(S, \mathbf{X})$
- Theorem. Under mild conditions
  - For any function h, the proposed test statistic is asymptotically valid

$$\lim_{n\to\infty} \mathbb{P}_{H_{0,j}}\{T_n^{(\alpha)}(\phi_j^A(h)) = 1\} = \alpha$$

  - $T_n^{(\alpha)}(\phi_j^A(h))$ is more powerful than $T_n^{(\alpha)}(\phi_j^{\text{val-eff}})$ in the sense that

$$\lim_{n\to\infty} \mathbb{P}_{H_{1,j}^{\text{loc}}}\{T_n^{(\alpha)}(\phi_j^A(h^\star)) = 1\} \geq \lim_{n\to\infty} \mathbb{P}_{H_{1,j}^{\text{loc}}}\{T_n^{(\alpha)}(\phi_j^A(h)) = 1\}$$
$$> \lim_{n\to\infty} \mathbb{P}_{H_{1,j}^{\text{loc}}}\{T_n^{(\alpha)}(\phi_j^{\text{val-eff}}) = 1\}$$

as long as Cov$\{\phi_j(Y, \mathbf{X}), h(S, \mathbf{X})\} \neq 0$, where

$$H_{1,j}^{\text{loc}} : \beta_j^* = Cn^{-1/2}$$

and *the first inequality is achieved* if $h = h_n$ and $\|\hat{h}_n(S, \mathbf{X}) - \mathbb{E}(Y \mid S, \mathbf{X})\| \to 0$ sufficiently fast, where $\hat{h}_n$ denotes the imputation model to learn $\mathbb{E}(Y \mid S, \mathbf{X})$ from the chart-reviewed sample V

Simulation

- Data generating process
  - $\mathbf{X}_i \sim N(\mathbf{0}_{50}, \Sigma)$ with $\sigma_{ij} = \rho^{|i-j|}$ for some $0 < \rho < 1$ and $i = 1, \ldots, 10^4$ ($N = 10^4, p = 50$)
  - $Y_i \mid \mathbf{X}_i \sim$ Bernoulli(Expit$(\mathbf{X}_i^T \beta^*)$) for $i = 1, \ldots, 100$ ($n = 10^2$)
  - $\mathbb{P}(S_i = s \mid Y_i = y, \mathbf{X}_i) = 0.8I(y = s) + 0.2I(y \neq s)$ for $y, s = 0, 1$
    - In this case,

$$\mathbb{P}(S = 1 \mid \mathbf{X}) = 0.6\mathbb{P}(Y = 1 \mid \mathbf{X}) + 0.2$$

and $\beta^\star$ can be purely identified by $(S, \mathbf{X})$ (Song et al. 2020):

$$\beta^\star = \underset{\beta}{\text{argmin}}\, \mathbb{E}\left\{ -\frac{S - 0.2}{0.6}\mathbf{X}^T\beta + \log\left(1 + e^{\mathbf{X}^T\beta}\right) \right\}$$

Song, Hyebin, Ran Dai, Garvesh Raskutti, and Rina Foygel Barber. "Convex and non-Convex approaches for statistical inference with class-conditional noisy labels." *Journal of Machine Learning Research*, no. 168 (2020): 1-58.

Simulation

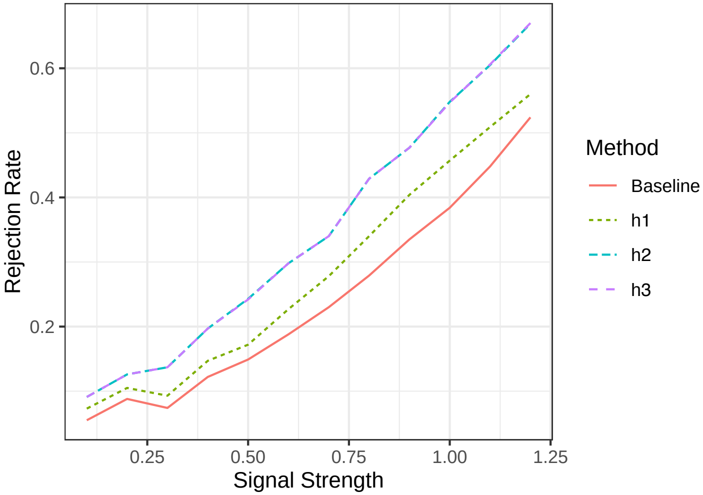
- We test $H_{0,6} : \beta_6^\star = 0$ versus $H_{1,6} : \beta_6^\star \neq 0$
  - Under $H_{0,6}$, we generate $\beta^* = (\beta_1^T, \mathbf{0}_{45}^T)^T \in \mathbb{R}^{50}$ with $\beta_{1:5} \sim N(\mathbf{0}_5, \mathbf{I}_5/\sqrt{5})$
- Power analysis
  - Under $H_{1,6}$, we generate $\beta^* = (\beta_{1:5}^T, \beta_6, \mathbf{0}_{44}^T)^T \in \mathbb{R}^{50}$ with $\beta_{1:5} \sim N(\mathbf{0}_5, \mathbf{I}_5/\sqrt{5})$, $\beta_6 = C/\sqrt{n}$ for $C = 0.5, 0.6, \ldots, 1.5$
- Choice of h
  - $h_1(S, \mathbf{X}) = (S, X_6)^T$
  - $h_2(S, \mathbf{X}; \hat{\gamma}_1) = \{S - \text{Expit}(\mathbf{X}^T \hat{\gamma}_1)\}(X_1, \ldots, X_6)^T$ with

$$\hat{\gamma}_1 = \underset{\gamma}{\text{argmin}} \sum_{i \in F} \left\{ -S_i \mathbf{X}_i^T \gamma + \log\left(1 + e^{\mathbf{X}_i^T \gamma}\right) \right\}$$

  - $h_3(S, \mathbf{X}; \hat{\gamma}_2) = \{(S - 0.2)/0.6 - \text{Expit}(\mathbf{X}^T \hat{\gamma}_2)\}(X_1, \ldots, X_6)^T$ with

$$\hat{\gamma}_2 = \underset{\gamma}{\text{argmin}} \sum_{i \in F} \{(S_i - 0.2)\mathbf{X}^T \gamma / 0.6 - \text{Expit}(\mathbf{X}^T \gamma)\}(X_1, \ldots, X_6)^T$$

Results



An improvement in power; *robust to the model for* $\mathbb{P}(S = 1 \mid \mathbf{X})$

Take-away Messages

- In the conventional literature of missing data, the theory regarding the semi-parametric efficiency is well-established but requires the *positivity* and *ignorability* (MAR) assumptions
- This work, by directly considering the problem of *variance reduction*, can be viewed as an extension of classic semiparametric theory in the sense of relaxing the *positivity* assumption
- Future work
    - Two-phase sampling, the optimal sampling rule, the *MAR* case
    - False discovery rate control
    - General high-dimensional M-estimation, time-to-event models
    - $\ldots$

Thanks!

Any Questions?