Test of Significance for High-dimensional Thresholds with Application to Individualized Minimal Clinically Important Difference

Huijie Feng*[a], Jingyi Duan*[a], Yang Ning[a], Jiwei Zhao[b]

Department of Statistics and Data Science, Cornell University[a]; Department of Biostatistics and Medical Informatics, University of Wisconsin-Madison[b]

Introduction

Motivation A common and natural approach to assess clinical significance is to determine minimal clinically important difference(MCID). Since the magnitude of MCID would depend on patients' demographic variables, it is of scientific interest to estimate the individualized MCID (iMCID) based on each individual patient's clinical profile, which is typically complex and high-dimensional data.

Goal We aim at developing statistical methods that incorporate the high-dimensional data into both **magnitude determination and uncertainty quantification of iMCID**.

Problem Setup

$X \in \mathbb{R}$: a **continuous variable** representing the score change collected from the PRO

$Y = \pm 1$: a **binary variable** derived from the patient's response to the anchor question

$\boldsymbol{Z} \in \mathbb{R}^d$: patient's clinical profile including demographic variables

n i.i.d samples, $\left\{(x_i, y_i, z_i)\right\}_{i=1}^{n}$ of $(X, Y, \boldsymbol{Z})$, $d \gg n$

The objective function is formulated as

$$\boldsymbol{\beta}^* = \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\, R(\boldsymbol{\beta}), \text{ where } R(\boldsymbol{\beta}) = \mathbb{E}\left[w(Y) L_{01}\left\{Y\left(X - \boldsymbol{\beta}^T \boldsymbol{Z}\right)\right\}\right],$$

where $L_{01}(u) = \frac{1}{2}\{1 - \operatorname{sign}(u)\}$ is the 0-1 loss, $w(1) = 1/\pi$, $w(-1) = 1/(1-\pi)$ and $\pi = \mathbb{P}(Y = 1)$.

Denote $\boldsymbol{\beta}^* = \left(\theta^*, \boldsymbol{\gamma}^{*T}\right)^T$, where θ^* is an arbitrary one-dimensional component of β^*. We start from considering the hypothesis testing problem $H_0 : \theta^* = 0$ **versus** $H_1 : \theta^* \neq 0$ where we treat γ^* as a high-dimensional nuisance parameter.

Problem Overview

Estimation Under $d \gg n$, estimating $\boldsymbol{\beta}^*$ is hard:

- The non-smoothness of $L_{01}(u)$ would cause the estimator to have **a nonstandard convergence rate**, which happens even in the fixed low dimensional case *(Kim et al., 1990)*.
- Minimizing the empirical risk function based on the 0-1 loss is computationally **NP-hard** and is often very difficult to implement.
- There's no estimators of β^* with root-n convergence rate *(Feng et al. 2022)*.

Inference

- Non-regular model: **non-Gaussian limiting distribution** *(Kim et al., 1990)*.
- Under high dimensionality, there exists **bias** induced by the penalty term.

Methodology

Smoothed Surrogate loss To tackle the non-smoothness of $L_{01}(u)$, *Feng et al. 2022* considered the following **smoothed surrogate loss**:

$$R_\delta(\boldsymbol{\beta}) = \mathbb{E}\left[w(Y) L_{\delta,K}\left\{Y\left(X - \boldsymbol{\beta}^T \boldsymbol{Z}\right)\right\}\right],$$

where $L_{\delta,K}(u) = \int_{u/\delta}^{\infty} K(t)dt$ is a smoothed approximation of $L_{01}(u)$, K is a kernel function and $\delta > 0$ is a bandwidth parameter.

Penalized Smoothed Surrogate Estimator

$$\widehat{\boldsymbol{\beta}} := \underset{\boldsymbol{\beta}}{\operatorname{argmin}}\, R_\delta^n(\boldsymbol{\beta}) + P_\lambda(\boldsymbol{\beta}), \star$$

$P_\lambda(\boldsymbol{\beta})$: sparsity induced penalty; $R_\delta^n(\boldsymbol{\beta})$: empirical version of $R_\delta(\boldsymbol{\beta})$,

$$R_\delta^n(\boldsymbol{\beta}) = \frac{1}{n}\sum_{i=1}^{n} w\left(y_i\right) L_{\delta,K}\left(y_i\left(x_i - \boldsymbol{\beta}^T z_i\right)\right).$$

Bias Corrected Smoothed Decorrelated Score

the classical score test breaks down:

- When plugging in the estimate of the nuisance parameter $\boldsymbol{\gamma}$ ($\hat{\boldsymbol{\gamma}}$), the estimation error from $\hat{\boldsymbol{\gamma}}$ may become the leading term

Our solution: bias corrected score

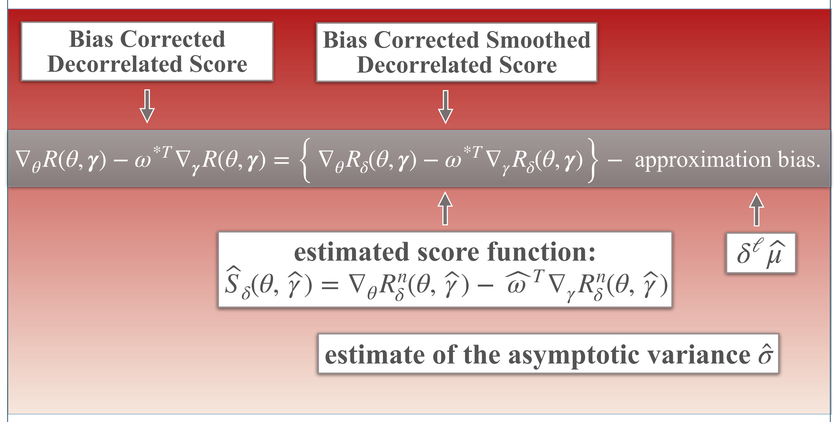
$$\nabla_\theta R(\theta, \boldsymbol{\gamma}) - \boldsymbol{\omega}^{*T} \nabla_\gamma R(\theta, \boldsymbol{\gamma}),$$

where the decorrelation vector is $\boldsymbol{\omega}^* = \left(\nabla^2_{\gamma,\gamma} R\left(\boldsymbol{\beta}^*\right)\right)^{-1} \nabla^2_{\gamma,\theta} R\left(\boldsymbol{\beta}^*\right)$.

- Sample version of $R(\theta, \boldsymbol{\gamma})$ is non-differentiable

Our solution: smoothed bias corrected score

We approximate $R(\theta, \boldsymbol{\gamma})$ by $R_\delta(\theta, \boldsymbol{\gamma})$, $\boldsymbol{\gamma}$ is estimated by $\hat{\boldsymbol{\gamma}}$ which is given by $\star$.



The test statistic is defined as

$$\widehat{U}_n = \sqrt{n\delta}\left(\frac{\widehat{S}_\delta(0,\hat{\gamma}) - \delta^\ell \hat{\mu}}{\hat{\sigma}}\right)$$

Theory

Theorem 1 Asymptotic Normality of The Decorrelated Score Under The Null Hypothesis

Under $H_0 : \theta^* = 0$, it holds that

$$\frac{\sqrt{n\delta}\,\widehat{S}_\delta(0,\hat{\gamma}) - \sqrt{n\delta^{2\ell+1}}\mu^*}{\sigma^*} \to N(0,1).$$

$\widehat{U}_n = \sqrt{n\delta}\left(\frac{\widehat{S}_\delta(0,\hat{\gamma}) - \delta^\ell \hat{\mu}}{\hat{\sigma}}\right)$: plug in the estimators $\hat{\mu}$ and $\hat{\sigma}$

$\left|\hat{\mu} - \mu^*\right| = o_p(1)$ and $\left|\hat{\sigma} - \sigma^*\right| = o_p(1) \rightarrow \widehat{U}_n \to N(0,1)$

Test function: $T_{DS} = I\left(\left|\widehat{U}_n\right| > \Phi^{-1}(1-\alpha/2)\right)$

Type I error $\mathbb{P}\left(T_{DS} = 1 \mid H_0\right) \to \alpha$

Theorem 2 Local Asymptotic Normality

- When the contiguous alternatives approach the null hypothesis at $n^{-\frac{\ell}{2\ell+1}}$, the proposed test → to a normal distribution
- If the alternatives deviate from the null hypothesis faster than $n^{-\frac{\ell}{2\ell+1}}$, the asymptotic power of our test is 1

Simulation

Model: $Y = \operatorname{sign}\left(X - \boldsymbol{\beta}^{*T}\boldsymbol{Z} + \epsilon\right)$

- **SDS**: the proposed smoothed decorrelated score test
- **DS**: the decorrelated score test method *(Ning and Liu, 2017)*
- **Honest**: Honest confidence region method *(Belloni et al., 2016)*

DS and Honest methods are devised for the logistic regression only

Table 6.1: The empirical Type I error rate of the tests under the Heteroskedastic Gaussian scenario from SDS, DS and Honest methods.

		$s = 3$			$s = 10$		
d	method	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0.7$	$\rho = 0.2$	$\rho = 0.5$	$\rho = 0.7$
100	SDS	5.6%	5.0%	6.4%	4.8%	4.8%	5.2%
	DS	1.2%	2.0%	2.0%	2.0%	1.8%	1.8%
	Honest	5.2%	5.6%	7.6%	5.4%	5.2%	6.8%
500	SDS	4.8%	4.4%	5.6%	5.6%	5.0%	4.8%
	DS	0.2%	0.4%	0.4%	0.2%	0.0%	0.4%
	Honest	7.0%	10.8%	7.6%	8.2%	6.8%	7.2%
1000	SDS	4.4%	6.0%	5.6%	5.0%	5.4%	5.0%
	DS	0.0%	0.4%	0.2%	0.0%	0.0%	0.4%
	Honest	10.0%	10.4%	12.6%	12.4%	6.4%	15.2%

References

1. Feng, H., Ning, Y., and Zhao, J. (2022), "Nonregular and minimax estimation of individu- alized thresholds in high dimension with binary responses," The Annals of Statistics, 50, 2284–2305.
2. Kim, J., Pollard, D., et al. (1990), "Cube root asymptotics," The Annals of Statistics, 18, 191–219.
3. Ning, Y. and Liu, H. (2017), "A general theory of hypothesis tests and confidence regions for sparse high dimensional models," The Annals of Statistics, 45, 158–195.
4. Belloni, A., Chernozhukov, V., and Wei, Y. (2016), "Post-selection inference for generalized linear models with many controls," Journal of Business & Economic Statistics, 34, 606– 619.