Revision: simulation study

September 4, 2023

Abstract

1 Questions in relation to simulation from reviewers

We can divide the questions reviewers have raised into two categories:

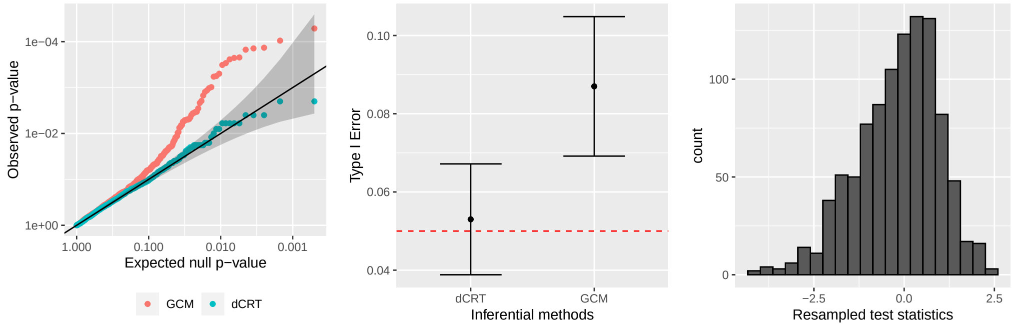
(1) **When should we use** dCRT **over** GCM**?**
In particular, the third question by the second reviewer is mostly relevant:
"When does the dCRT is preferred over the GCM test? This question is of most importance since the dCRT involves resampling while GCM does not. This implies that, from a computational perspective, the GCM is far more attractive."

(2) **The** dCRT **is an asymmetric test. Then when the conditional law $\boldsymbol{Y}|\boldsymbol{Z}, \boldsymbol{X}|\boldsymbol{Z}$ are of different difficulty for learning, how does the** dCRT **perform?**
This question is raised by the first reviewer:
"A distinction between the GCM test and the dCRT test is that the former is symmetric in X and Y but the latter is not (at least in the finite-sample sense). I wonder how this asymmetry affects the (finite-sample) behavior of dCRT. For example, would GCM perform differently than dCRT in situations where $Y|Z$ is more/less complex than $X|Z$. Or, what if different methods are used for fitting $\mu_{n,x}(\cdot)$ and $\mu_{n,y}(\cdot)$?"

(3) **What will be the performance when going beyond the linear Gaussian model?**
This is the fourth question given by the second reviewer:
"Numerical experiments are focused on a linear model. It will be great to further study the performance of the test in situations where the model for $Y|X$ and $X|Z$ is non-linear with heavy-tailed noise distribution. Here, it will be interesting to study the effect of a model estimating $Y|X$ and $X|Z$ is more flexible, such as random forests, neural networks, and kernel regression. Providing such a set of experiments is important given the careful design of experiments the authors offered, highlighting the current optimistic view on the robustness of MX knockoffs."

2 Simulation study for the first question

This is the most intriguing question raised by the reviewers. Conceptually, GCM is a asymptopia-based inferential method. The rejection region is determined by the asymptotic distribution of the test statistic which is the standard normal distribution. On the other hand, dCRT is a resampling-based inferential method. The rejection region is determined by the resampling



(a) (b) (c)

Figure 1: The illustration of finite-sample difference when the underlying distributions of X, Y are discrete. Left: Q-Q plot. Middle: Type-I error comparison with significance level 0.05. Right: histogram of the resampled test statistic.

distribution of the test statistic. Thus we will be very likely to find a discrepancy in terms of finite-sample performance when the distribution of test statistic is not close to the normal distribution while the resampling distribution can mimic the true sampling distribution of the test statistic for small sample size. Inspired by this idea, we consider the model

$$\boldsymbol{Y} \sim \text{Poisson}(\mu_y(\boldsymbol{Z})), \ \boldsymbol{X} \sim \text{Bern}(\mu_x(\boldsymbol{Z}))$$

where $\mu_y(\boldsymbol{Z}) = \exp(beta_0 + \boldsymbol{Z}\beta_1)$ and $\mu_x(\boldsymbol{Z})^{-1} = 1 + \exp(-(\gamma_0 + \boldsymbol{Z}\gamma_1))$, i.e. the $\boldsymbol{Y}|\boldsymbol{Z}$ is a Poisson GLM and the $\boldsymbol{X}|\boldsymbol{Z}$ follows a logistic GLM. We consider the following parameter settings:

$$\gamma_0 = -4, \ \gamma_1 = 1, \ \beta_0 = -3, \ \beta_1 = 1$$

and the sample size $n = 1000$. We employ the usual generalized linear regression to estimate $\mu_{n,x}(\cdot)$ and $\mu_{n,y}(\cdot)$ and use 1000 resamples for dCRT. The corresponding results are summarised in Figure 1. The simulation is repeated 1000 times. We can clearly see there is a significant amount of Type-I error inflation for GCM whereas dCRT controls the Type-I error very well. The histogram in the same Figure indicates the discrepancy between the the resampling distribution and normal distribution.

3 Simulation study for the second (and partly the third) question

To answer this question, we investigate into the following conditional models

$$\text{Poisson}(\exp[(-4, \boldsymbol{Z}^\top)\beta]), \ \text{Bern}(\text{expit}[(-3, \boldsymbol{Z}^\top)\beta]), \ N((0, \boldsymbol{Z}^\top)\beta, 1)$$

where $\beta = \mathbf{1}_{d+1} = (1, \ldots, 1)^\top \in \mathbb{R}^{d+1}$. It is well-acknowledged that the Poisson regression is more difficult to fit than the logistic regression and the Gaussian regression. This motivates us to vary the choice of $\boldsymbol{X}|\boldsymbol{Z}$ while fixing $\boldsymbol{Y}|\boldsymbol{Z} \sim \text{Poisson}(\boldsymbol{Z}^\top\beta)$. Consider

(1) $\boldsymbol{X}|\boldsymbol{Z} \sim N(\boldsymbol{Z}^\top\beta, 1), \boldsymbol{Y}|\boldsymbol{Z} \sim \text{Poisson}(\boldsymbol{Z}^\top\beta)$;

(2) $\boldsymbol{X}|\boldsymbol{Z} \sim \text{Bern}(\boldsymbol{Z}^\top\beta), \boldsymbol{Y}|\boldsymbol{Z} \sim \text{Poisson}(\boldsymbol{Z}^\top\beta)$.

To study how the strategy of resampling affect Type-I error, we include both the resampling from $\hat{\mathcal{L}}(\boldsymbol{X}|\boldsymbol{Z})$ and from $\hat{\mathcal{L}}(\boldsymbol{Y}|\boldsymbol{Z})$ and denoted the procedures as dCRT_X and dCRT_Y respectively. For comparison, we also incorporate the results of GCM. The parameter setting is $n = 100$ and d is varied within 11 to 15 with space 1. The simulation is repeated 10000 times and the results are summarised in Figure 2.

Clearly, we can see the discrepancy of the Type-I error between dCRT_X and dCRT_Y. Since it is a only a conjecture that Poisson model is hard to fit, we further verify it by comparing the mean squared error (MSE) of the estimated $\mu_{n,x}(\cdot)$ and $\mu_{n,y}(\cdot)$. The results are shown in the bottom row of Figure 2. The much larger MSE of $\mu_{n,y}(\cdot)$ indicates that the Poisson model is indeed harder to fit.

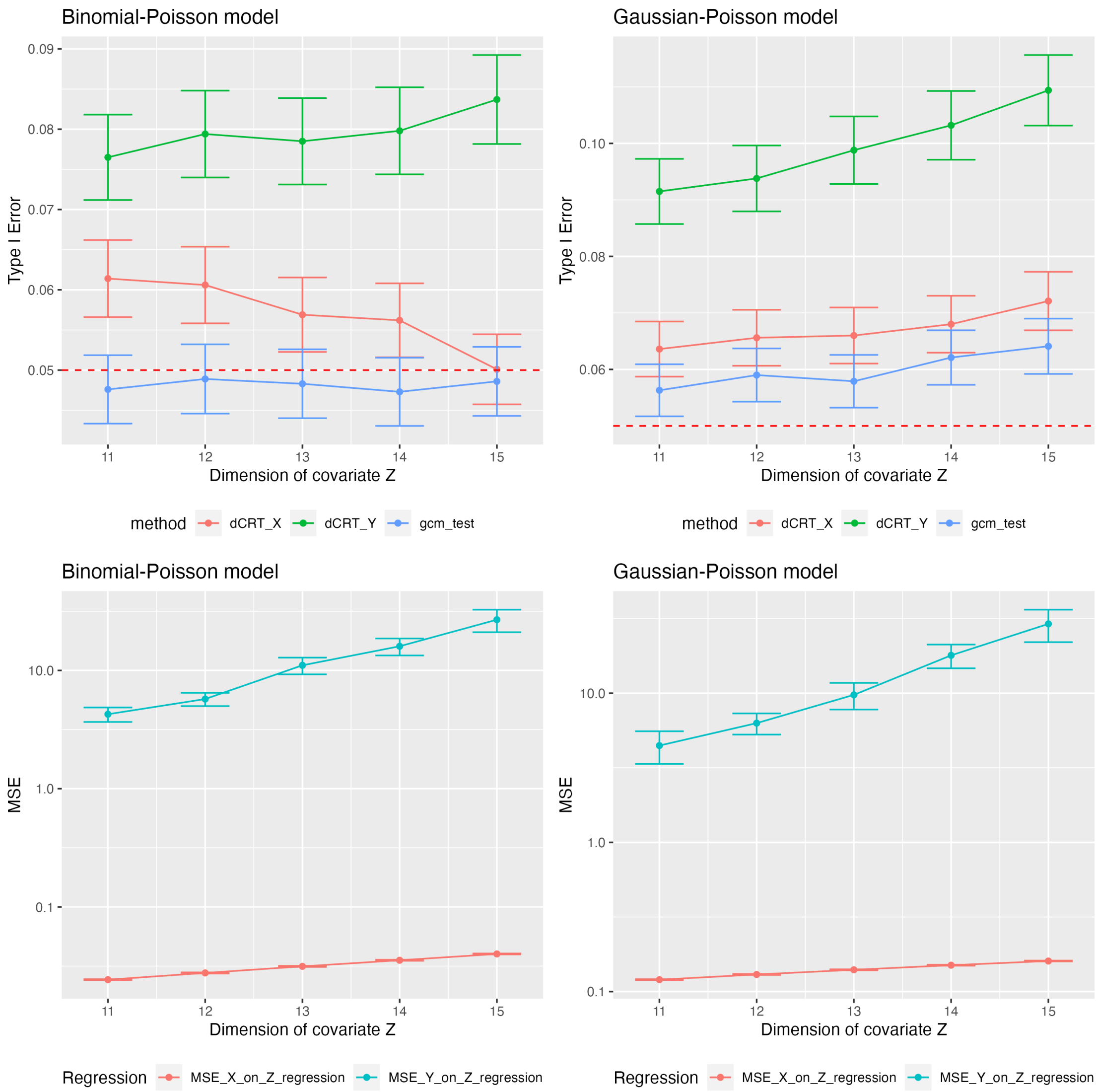


Figure 2: Type-I error and MSE comparison when $\boldsymbol{Y}|\boldsymbol{Z}$ is a Poisson model.

References

[1] E. Candes, Y. Fan, L. Janson, and J. Lv. “Panning for gold:‘model-X’knockoffs for high dimensional controlled variable selection”. In: *Journal of the Royal Statistical Society Series B: Statistical Methodology* 80.3 (2018), pp. 551–577.