

# Investigation into the second simulation question raised by the reviewers

September 19, 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  $Y|Z, X|Z$  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 Investigation into the second question

To address the second question, we first consider the case when  $\mathbf{Y}|\mathbf{Z}$  is a Poisson model whereas  $\mathbf{X}|\mathbf{Z}$  is a Gaussian model.

$$\mathbf{X}|\mathbf{Z} \sim \text{Pois}(-4 + \mathbf{Z}^\top \gamma), \mathbf{Y}|\mathbf{Z} \sim \text{Pois}(\mathbf{Z}^\top \beta), \beta = \gamma = \mathbf{1}_d.$$

We consider the simulation setting when  $n = 200$  whereas  $d = 10, 12, \dots, 18$ . The fitting methods applied are GLM both for  $\mathbf{X}|\mathbf{Z}$  and  $\mathbf{Y}|\mathbf{Z}$ . For methods comparison, we include both resampling strategies from  $\mathbf{X}|\mathbf{Z}, \mathbf{Y}|\mathbf{Z}$ . Also the unnormalized and normalized statistics are both considered. GCM test is added as well. The results are shown in Figure 1.

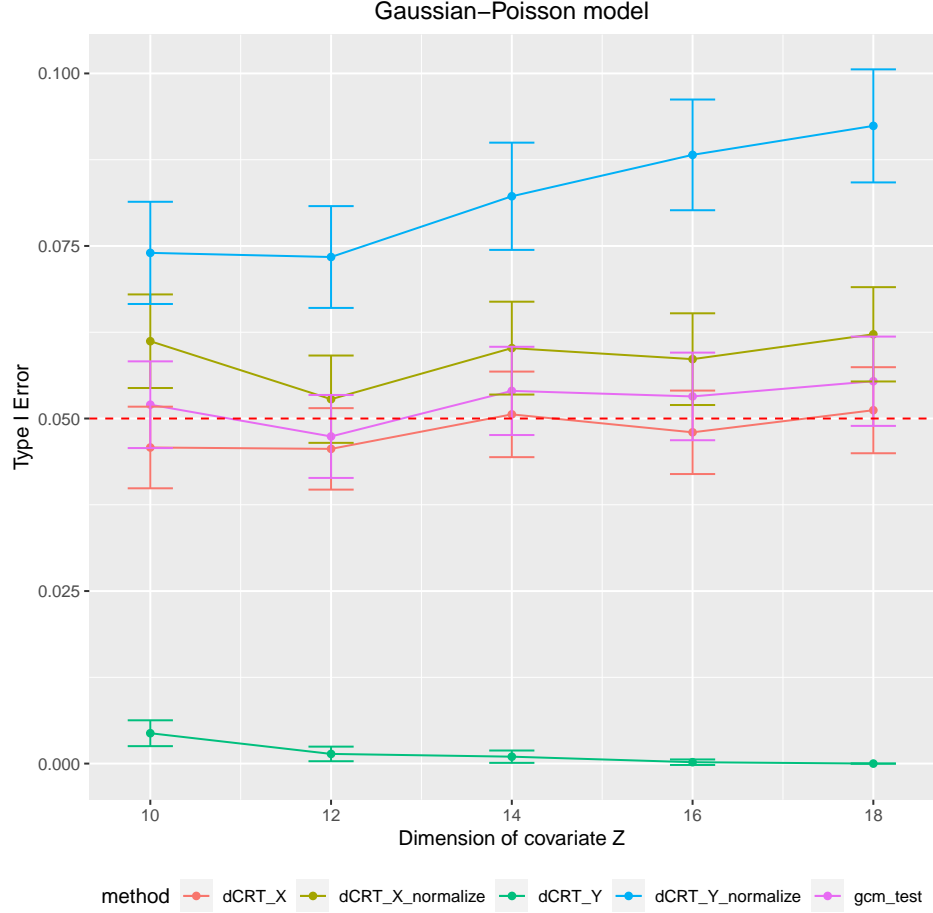


Figure 1: Type-I error comparison when  $\mathbf{Y}|\mathbf{Z}$  is a Poisson model. dCRT\_X and dCRT\_Y are tests based on unnormalized test statistics resampling from fitted model  $\mathbf{X}|\mathbf{Z}$  or  $\mathbf{Y}|\mathbf{Z}$ .

Clearly, there is a lot of information in this Figure. We summarise the main findings as follows:

1. we see the increasing trend of type-I error when applying dCRT\_Y\_normalize and slight increase for dCRT\_X, dCRT\_X\_normalize and gcm\_test;
2. the tests based on unnormalized statistics seem to have better type-I error control regardless of resampling from X or Y models;
3. type-I error are quite similar among dCRT\_X, dCRT\_X\_normalize and gcm\_test with slight type-I error inflation for dCRT\_X\_normalize;
4. dCRT\_Y tends to always accept the null hypothesis.

## 2.1 Unveil the mystery

We claim that the causes of the above phenomenon are due to the various behaviors of sampling variance and resampling variance. To see this, we focus on the setting where  $d = 16$ . We consider 1000 realizations of the data and compute the sampling variance. Then we fix the data generating mechanism to the first realization among the 1000 realizations and compute the resampling variance when resampling from model  $X$  or  $Y$ . The results are shown in Figure 2.

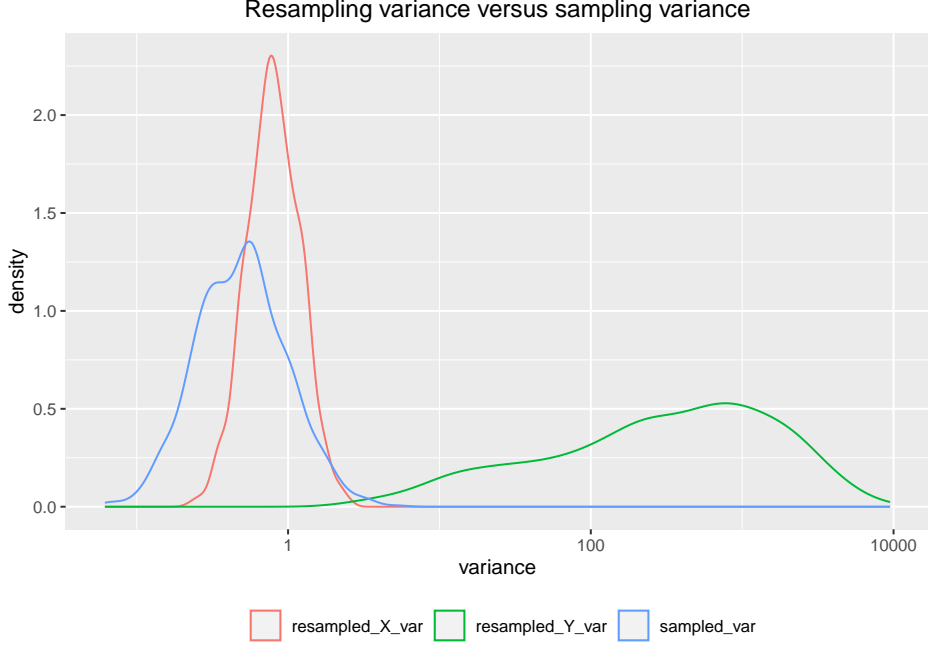


Figure 2: Comparison of sampling variance and resampling variance.

Clearly, the resampling variance of the unnormalized test statistics is slightly smaller than the sampling variance when resampling from model  $X$ . However, the difference between sampling variance and resampling variance is **huge** when resampling from model  $Y$ . This explains why `dCRT_Y` tends to always accept the null hypothesis. This is further confirmed when looking at the histogram of the unnormalized test statistics (Figure 3). The empirical CDF of the resampling unnormalized test statistics perfectly align with the empirical CDF of the sampling unnormalized test statistics when resampling from model  $X$ . However, the empirical CDF of the resampling unnormalized test statistics is much flatter than the empirical CDF of the sampling unnormalized test statistics when resampling from model  $Y$ .

Figure 2 is also informative to explain the behavior of normalized tests to some extent. When the sampling variance is much smaller than the resampling variance, the normalized resampling test statistics may tend to be smaller than the normalized sampling test statistics if the fluctuation of the numerator which is the unnormalized test statistics is smaller than their standard deviation. We did not see a huge difference in type-I error between `dCRT_Y_normalize` and `dCRT_X_normalize` indicates that the unnormalized resampling test statistics also fluctuate a lot during the resampling procedure. The empirical CDF of both normalized and unnormalized statistics are shown in Figure 3. Clearly, we see the behavior for resampling statistics from model  $Y$  deviates from the behavior of resampling statistics in both normalized and unnormalized cases.

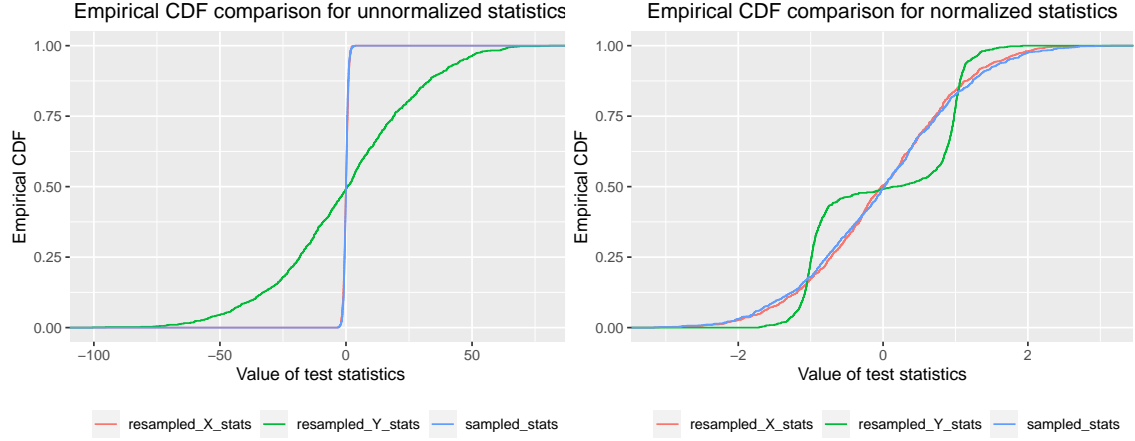


Figure 3: Histogram of the CDF of test statistics.

## 2.2 Think deeper into variance estimation

The deeper reason for causing such finite-sample phenomenon depends on the specific data generating mechanism. To better illustrate the mechanism of the sampling, we display a draw of  $\mathbf{Y}$  in Figure 4. A significant proportion of the draw are zeros but there are also a non-negligible proportion of huge values in the draw. This is because of we sometimes can get positive values with big magnitude in the link function of the Poisson model  $\mathbf{Y}|\mathbf{Z}$  while most of the conditional mean are very close to zero because of the  $-4$  intercept. However, the large conditional mean will make a big difference when performing the test especially when drawing resamples from the fitted  $Y$  model. The resampling variance is extremely variable causing the variance resampling variance estimation very inaccurate under finite sample.

On the contrary, the resampling variance estimation is somewhat easier when resampling from the fitted  $X$  model. This is due to the nice Gaussian model of the conditional model of  $\mathbf{X}|\mathbf{Z}$ .

```
> Y
[1] 0 63 0 0 0 0 0 0 0 0 0 0 0 0 17 50
[17] 0 0 0 2820 0 1 4 1 40 0 0 0 0 0 0 0
[33] 0 0 0 0 4661 0 0 0 0 0 0 1 0 0 0 37
[49] 0 0 0 2 1 0 0 0 0 0 0 0 0 2 1 13
[65] 0 1 0 0 0 112 0 0 0 0 0 4 0 0 0 9
[81] 0 0 206 0 1 2 0 5 0 0 0 0 0 0 1 78122
[97] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
[113] 1 0 0 0 0 1336 0 136 0 1 59 0 1 0 0 2
[129] 0 0 0 0 0 0 6 0 0 0 0 0 0 0 0 0
[145] 0 0 8 0 0 0 0 0 0 0 0 0 0 0 0 0
[161] 0 1 0 0 0 0 0 0 0 0 0 0 0 6 3 0
[177] 8 0 0 0 0 0 0 0 0 2 0 1 0 0 0 0
[193] 0 5 0 0 2 1 0 0
```

Figure 4: A draw of  $\mathbf{Y}$ .

## 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.