KEN4258: Computational Statistics

Assignment 2

Aurélien Bertrand Bart van Gool Gaspar Kuper
Ignacio Cadarso Quevedo Nikola Prianikov

March 2024

Assignment 2

Link to our GitHub repository: https://github.com/nprianikov/compstats

1) Reproduce Figure 1 from (Candès et al. 2018).

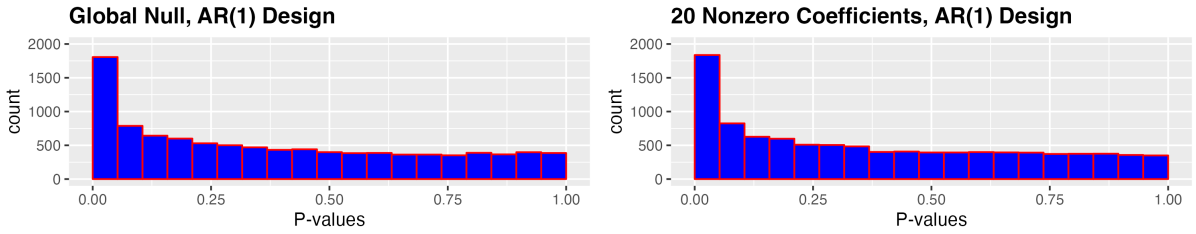

Figure 1: Replicated simulation from the source paper.

2) What is the problem that Figure 1 tries to illustrate?

The figure above illustrates the p-values for coefficient β_1 in two GLMs for two different systems. The figure shows, for each system, a histogram of the distribution of the β_1 p-value over $10,000$ replications of the experiment. The systems are given by the following functions:

(1) $Y|X_1, \ldots, X_p \sim \text{Bernoulli}(0.5)$

(2) $Y|X_1, \ldots, X_p \sim \text{Bernoulli}(\text{logit}(0.08(X_2 + \ldots + X_{21})))$

Where $(X_1, ..., X_p)$ are themselves random variables generated by an AR(1) time series with AR coefficient 0.5, and $p = 200$. In other words, the results of the first simulation are completely independent from the 200 values of X_i while the second simulation is only dependent on $(X_2, ..., X_{21})$. Neither system is influenced in any way by the predictor X_1 nor its coefficient β_1.

The figure shows that in both simulations, we see a significant amount of low p-values. In fact, for (1) we have 16.89% of p-values ≤ 0.05, while for (2) it is 19.17%. In GLMs, the p-values of a coefficient indicate the importance of a coefficient to the model. Low p-values are linked to coefficients that are significant to the model, and are therefore also significantly different from 0. This does not seem to be reflected in the figures, however, where we see that the β_1 coefficient is often having low p-values even though it has no impact at all on the underlying system that the model is trying to fit.

The problem can arise for multiple reasons in high-dimensional datasets, especially when the number of predictors is large relative to the amount of observations. One such reason is the multiple comparisons problem which occurs simply because so many hypotheses are being tested. In simulation (1) for example, where we have 200 predictors with no relation to the outcome, we can still expect to identify 10 important variables with a confidence of 95%. Another reason could be due to the model overfitting to the outcome variable. Even though predictors may not have any relationship with the outcome, they may still be able

to help the model fit to this specific set of outcomes. This problem is especially common when we have a relatively small amount of observations.

To correct this problem, we will need to implement a method to decrease the false discovery rate such that we are less likely to label a predictor as being significant.

3) Propose a solution to address the problem.

The Conditional Randomization Test (CRT) from the (Candès et al. 2018) presents a solution to our problem. CRT generates a distribution for the test statistic under the null hypothesis by conditioning on the observed values of the response Y and all covariates other than the one being tested. For each covariate Xj, the CRT samples a new Xj from its conditional distribution given the other covariates, computes the test statistic for this new X_j while Y and the other covariates remain the same, and then compares this statistic to the one computed on the actual data.

However, we introduce two modifications to the original CRT procedure. Firstly, the original procedure requires us to know the conditional distribution of covariates. Since the exact distribution of the data is unknown in practice we follow the approach of (Shaer et al. 2023), where conditional Gaussian distribution was used to sample the covariate X_j conditional on all the other covariates. Parameters of the mean and the variance are therefore estimated from all the other covariates. The second difference is that for the test statistic T_j we simply utilize absolute values of a coefficient fitted with Logistic regression without Lasso regularization. This is motivated by the fact that even with an extremely small penalty constant our experiments resulted in Lasso regularization to push the coefficients of the model to be too low which caused CRT to produce invalid p-values.

This process is repeated K times to generate an empirical distribution of the test statistic under the null hypothesis. The p-value is then calculated based on where the actual test statistic falls within this null distribution. By doing this, CRT accounts for the dependence between the covariates.

4) Show that your solution fixes the problem.

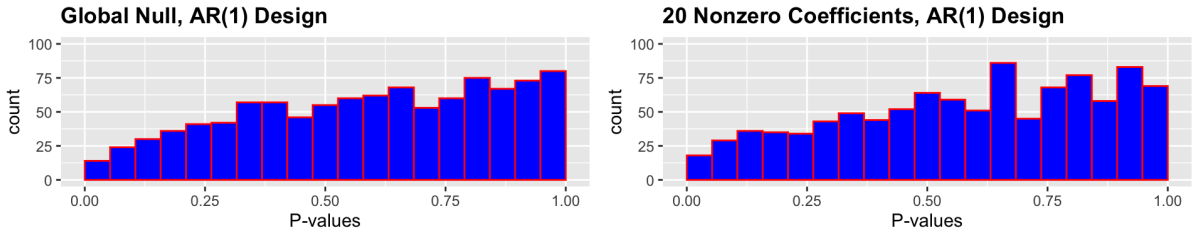

Figure 2: Demonstration of a fix through CRT. Simulation was 10^3 times with reduced dimensionality of p=80, n=200 and K=100

The histograms from CRT show that the distribution of p-values is more uniform. This uniform distribution is what we would expect if the null hypothesis is true. We can also see that there is no inflation of small p-values, which as explained in section 2 were identifying β_1 as significant. The Conditional Randomization Test has corrected for this inflation by better modeling the null distribution.

In conclusion, CRT has provided valid p-values that are a more reliable indication of the true significance of the predictors.

5) Find a real dataset and apply your method.

We found a real dataset with 92 variables and 6819 observations. We select 800 samples at random so that $p/n = 0.115 > 0.1$. We fitted a GLM and plotted the normal p-values and the ones obtained by CRT. The histogram of standard p-values shows most of the p-values are near 0, which means that all of the variables are considered statistically significant. However, this can be due to the high-dimensionality of the dataset, with many predictors, even by random chance, some will appear to be significant. On the other hand, the histogram of p-values obtained through CRT has a different distribution. The values are distributed more uniformly, with more variables showing high p-values which indicate no statistical significance. This means that after applying CRT, less predictors are considered significant, which aligns

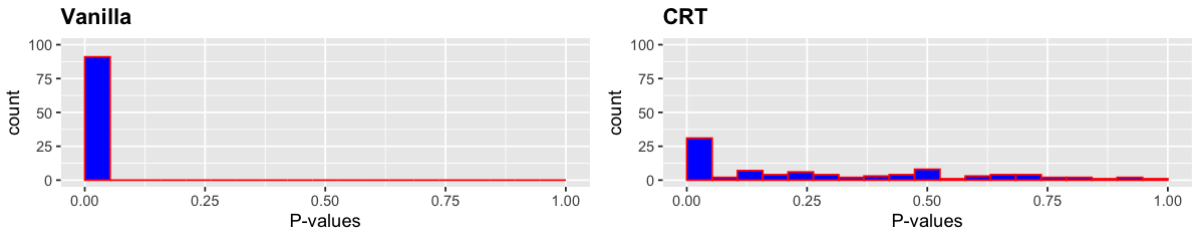

Figure 3: Comparision of p-values for all the model coefficients with a default procedure and CRT.

better with what we would expect under the null hypothesis, since it is unlikely that all predictors are significantly associated with the response.

The results of the CRT suggest that the original GLM identified many predictors as significant due to the high chance of type I errors in high-dimensional datasets. By providing a more stringent method for calculating p-values, CRT reduces the likelihood of these false discoveries, leading to valid p-values that identify the true predictors of the response variable.

NB: we removed some columns, because they had a high collinearity, leading to singular matrix exceptions when re-sampling.