Maastricht University

Department of Advanced Computing Sciences

Christof Seiler

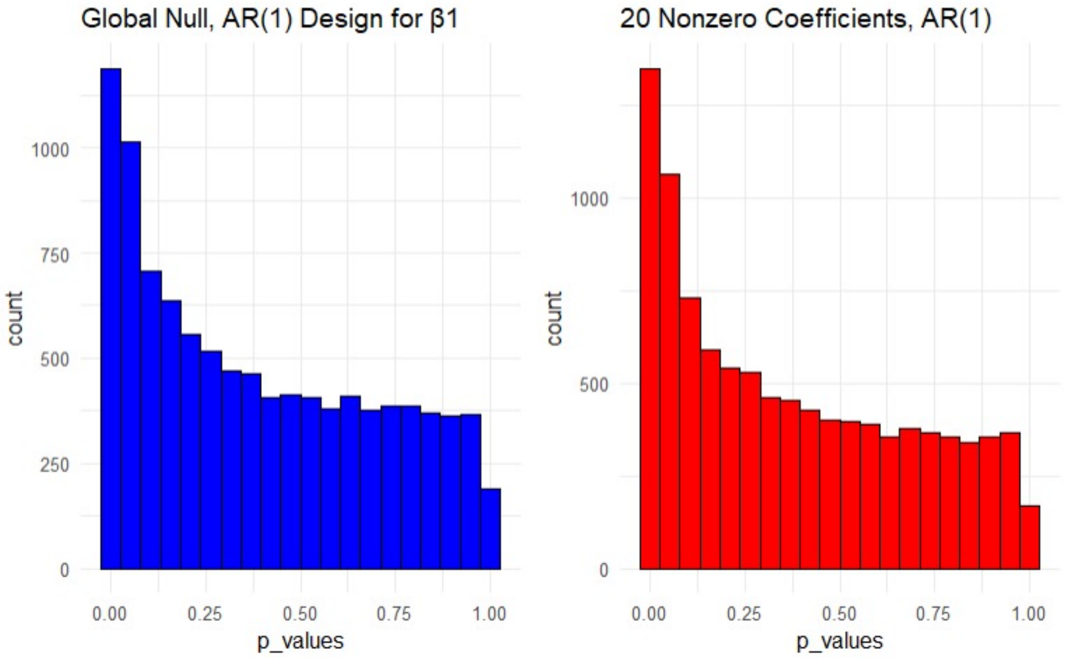
Assignment 2

Computational Statistics

Members:
Leonardo Ercolani
Lisa Incollingo
Chiara Magrone

March 22, 2024

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



The Figure was riproduced from (Candès et al. 2018) at the following link: `https://arxiv.org/abs/1610.02351`

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

In Figure 1 [1], we're examining the distribution of p-values for the coefficient β_1 under the null hypothesis, in the context of logistic regression models with a large number of predictors. Specifically, the figure reflects simulations with n=500 observations and p=200 predictors, which corresponds to a high-dimensional setting where $\frac{n}{p} \geq 0.1$ The two settings modeled are as follows:

- $(X_1, \ldots, X_p)$ is an AR(1) time series with AR coefficient 0.5 and

$$Y|X_1, \ldots, X_p \sim Bernulli(0.5)$$

.

- $(X_1, \ldots, X_p)$ again follows an AR(1) time series with an AR coefficient of 0.5, but in this setting, the response Y is modeled with a logistic regression on the first 20 predictors with non-zero coefficients:

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

The histograms for the p-values of β_1, which is null in all simulations, demonstrate that the null distribution of p-values deviates from the expected uniform distribution when the dimensionality

[1] `https://github.com/LisaIncollingo24/Computational-Statistics.git`

is high. Notably, we observe that the p-values for β_1, are not uniformly distributed; they show a higher concentration of small values, which suggests an inflated rate of type I errors (false positives). This inflation is especially pronounced in the second setting, where the true model includes non-zero coefficients for predictors X_2 through X_{21}, highlighting that the presence of actual effects in the model can influence the null distribution of p-values for other coefficients presumed to be null. This phenomenon arises because traditional asymptotic theory, which assures valid p-values only when n is much greater than p, does not hold in high-dimensional settings.

3 Propose a solution to address the problem.

To address this issue, we thought that the knockoff method as proposed by [1]. appeared to be the most suitable, offering an innovative, robust solution with better computational efficiency compared to, for example, the conditional randomization test. Specifically, we used the *knockoff.filter* function which effectively generates a set of knockoff variables $\tilde{X}$ that mimic the correlation structure of the original variables X while remaining independent from the response variable Y. This requires knowledge or estimation of the joint distribution of the predictor variables X. Subsequently, it calculates a statistic for each original variable and its corresponding knockoff that measures the association of the variable with the response Y. The statistics must possess the property that higher values indicate stronger evidence of the variable's association with the response. At this stage, we are able to select variables based on the comparison between the statistics of the original variables and their knockoffs. This is done by comparing the statistics of the original variables with those of their knockoffs and applying a criterion to control the FDR. The knockoff method aims to control the FDR, which is the expected proportion of type I errors (false positives) among all discoveries made. The FDR is then calculated based on the selected variables, adjusting the selection threshold to ensure that the FDR does not exceed a predetermined level.

4 Go back to Figure 1 and show that your solution fixes the problem.

We conducted experiments using the knockoff method, based on simulations designed to replicate Figure 1, then in the global null setting and with coefficients from x_2 to x_{21} being non-null. These experiments have demonstrated the effectiveness of the knockoff method in addressing the issues highlighted, providing a robust variable selection and keeping the false discovery rate (FDR) under control. Now we describe the two scenarios and our observations:

- **Global null setting**

  By applying the knockoff method to this scenario, with various FDR thresholds, we observed that, at each chosen threshold, no variable was selected as significant. At each chosen threshold, the method dynamically adjusted the significance threshold of the statistics used to determine which variables are genuinely important. This result matches our expectations and confirms the method's effectiveness, especially regarding the coefficient β_1. The latter, in contrast with the results shown in Figure 1, where a high frequency of low p-values for β_1 erroneously suggested its significance, is correctly identified by the knockoff method as non-significant.

- **Scenario with 20 Non-null Coefficients:**

Firstly, we implemented the knockoff method across various FDR threshold levels yielded significant insights, as detailed in 6 in the Appendix. Notably, the coefficient β_1 was never identified as significant, irrespective of the FDR threshold set. A particularly noteworthy outcome was observed at an FDR threshold of 0.10%, which achieved the optimal False Discovery Proportion (FDP $= \frac{\text{False Discoveries}}{\text{Discoveries}}$), correctly identifying all 17 selected variables as significant. Conversely, even at a more restrictive threshold of 0.06%, all variables from 2 to 21 were selected. However, at this threshold, additional variables were also deemed significant that, according to our simulation's design, should have been considered non-significant, resulting in an FDP of approximately 0.0952%. As the FDR threshold was raised, there was an observed increase in the misclassification of variables as significant. It's finally important to note that the observed phenomenon, where various variables are selected at an FDR threshold level of 0.06, none at 0.07, and then some again at 0.08, for example, may initially seem counter-intuitive. Indeed, we might expect that increasing the FDR threshold would always lead to an increase in the number of selected variables, as a higher threshold implies fewer restrictions in variable selection. However, this expectation does not account for a crucial aspect of the knockoff method: for each specified FDR level, a new set of knockoff variables is generated. This regeneration process implies that the selection of significant variables is influenced by the particular characteristics of the new set of knockoff variables created for that specific analysis.

Secondly, we have conducted an important experiment in order to have a comparison with Figure 1, we iteratively ran a loop with 1000 iterations and with a fixed FDR level of 0.25, focusing only on whether the first coefficient was ever detected. In this case, we recorded a count equal to 0, demonstrating that the method really works. The crucial point to underline is that the knockoff method, consistently, does not consider the coefficient β_0 as significant, unlike the experiment conducted in Figure 1.

5 Find a real dataset and apply your method.

The dataset chosen for this exercise was taken from Bioconductor, an open source software for bioinfromatics. Bionconductor has several packages accessible from R. We have chosen the ALL package which includes data of T- and B-cell Acute Lymphocytic Leukemia from the Ritz Laboratory at the DFC. The dataset is made of 128 samples and 12625 features. To ensure a smoother and shorter process we selected, filtered and transposed a subset after assigning the response variables. We then tried to perform a Logistic Regression, the estimates of the coefficients can be seen in Table 2. We notice that at a certain point the estimates started to be null. In fact, we notice that 90 of them weren't defined because of singularities. This implies that most likely there is perfect multicollinearity between variables. Additionally, we have a Fisher Scoring iterations number of 26, which might be a bit low. This may be related to the fact that in the process of likelihood maximization in Logistic Regression, the model risks to be stuck in a local optima instead of a global one, especially given the type of data presented. The method chosen in this assignment is however the knockoff, which we expect to significantly help in such a case. As we can see in Table 3, after applying the knockoff method it successfully selected 4 variables at a FDR of 0.5 (any lower FDR gave no output). In conclusion we can confirm that [1] indeed the knockoff method is effective when dealing with high-dimensional data.

6 Appendix

FDR Target	FDP	Selected Variables	# of Variables
0.05	0.0000	None	0
0.06	0.0952	2, 3, 4, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 112, 172	21
0.07	0.0000	None	0
0.08	0.1000	2, 3, 4, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 19, 20, 21, 95, 172	20
0.09	0.0000	2, 3, 4, 6, 8, 10, 11, 12, 13, 14, 15, 16, 17, 19, 20	15
0.10	0.0000	2, 3, 4, 6, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20	17
0.11	0.0000	None	0
0.12	0.0000	None	0
0.13	0.0000	4, 6, 9, 10, 12, 14, 16, 19, 20	9
0.14	0.0000	None	0
0.15	0.0000	None	0
0.16	0.1053	3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 18, 19, 20, 94, 172	19
0.17	0.0556	2, 3, 4, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 19, 20, 172	18
0.18	0.1579	4, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 18, 19, 20, 21, 94, 172, 182	19
0.19	0.0588	2, 3, 4, 6, 7, 8, 9, 10, 11, 12, 13, 14, 16, 19, 20, 21, 172	17
0.20	0.0000	2, 4, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 19, 20	16
0.21	0.1905	2, 4, 5, 6, 7, 8, 10, 11, 12, 13, 14, 15, 16, 18, 19, 20, 21, 94, 101, 112, 172	21
0.22	0.0000	2, 4, 6, 8, 9, 10, 11, 12, 14, 15, 16, 19, 20	13
0.23	0.1818	2, 4, 5, 6, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 94, 112, 141, 172	22
0.24	0.2692	2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 18, 19, 20, 21, 80, 94, 95, 112, 136, 172, 182	26
0.25	0.1111	2, 3, 4, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 19, 20, 112, 172	18

Table 1: Summary of Results from the Knockoff, Second Scenario

Variable	Estimate	Std. Error	z value	Pr(>—z—)
1000_at	1.878e+02	1.506e+07	0	1
1001_at	-2.625e+02	1.296e+07	0	1
1002_f_at	-3.961e+02	1.773e+07	0	1
1003_s_at	-6.108e+02	3.602e+07	0	1
1004_at	3.548e+02	1.887e+07	0	1
1005_at	1.670e+01	8.082e+05	0	1
1006_at	-1.387e+02	3.100e+06	0	1
1007_s_at	4.848e+01	2.844e+06	0	1
1008_f_at	1.191e+02	7.582e+06	0	1
1009_at	3.231e+01	5.833e+06	0	1
1010_at	5.321e+02	2.866e+07	0	1
1011_s_at	2.291e+02	2.071e+07	0	1
1012_at	2.412e+02	1.303e+07	0	1
1013_at	-6.981e+01	1.221e+07	0	1
1014_at	-2.870e+02	1.641e+07	0	1
1015_s_at	2.065e+02	1.465e+07	0	1
1016_s_at	-3.236e+02	1.954e+07	0	1
1017_at	4.076e+02	3.249e+07	0	1
1018_at	-5.921e+01	4.805e+06	0	1
1019_g_at	4.479e+01	6.259e+06	0	1
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
1100_at	NA	NA	NA	NA
1101_at	NA	NA	NA	NA
1102_s_at	NA	NA	NA	NA
1103_at	NA	NA	NA	NA
1104_s_at	NA	NA	NA	NA
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.

Table 2: Coefficients: (90 not defined because of singularities)

FDR Target	Selected Variables	# of Variables
0.5	8, 60, 118, 148	4

Table 3: Summary of Results from the Knockoff using the Bioconductor ALL package

References

[1] E. Candes, Y. Fan, L. Janson, and J. Lv. Panning for gold: Model-x knockoffs for high-dimensional controlled variable selection, 2017.