UNIVERSITY OF PAVIA

Department of Brain and Behavioral Science

PhD program in Psychology, Neuroscience and Data Science

Use and misuse of P-values: a conditional approach to post-model-selection inference

Thesis by:

Mauro Gioè

Advisor:

Prof. Mario Comelli

Academic year 2019/2020, XXXIII Doctorate Cycle

Abstract

Adaptive generation of hypotheses is among the main culprits of the lack of replicability in science. Under conditions of uncertainty, the statements, or the process that generates them, can only be trusted if the reported error rates are reflected in the replication attempts. The discrepancy between the two is due to many factors, but interactive data analysis plays a major role in the inflation of type I error. In this regard, inference after model selection is of particular interest because its misuse can be analyzed through a Monte Carlo simulation. As the findings of this thesis show, inflation of type I error can be quite severe even in low dimensional scenarios, with up to 40% of false positives in the selected set of variables. Depending on the model selection strategy and the structure of the true data-generating mechanism, this percentage varies greatly. The results of the simulation show different performances between the Least Absolute Shrinkage and Selection Operator (LASSO) and the Forward Selection (FS). In particular, the LASSO yields a type I error lower than the FS when the structure of the true data-generating mechanism is additive and a higher one when the structure is multiplicative. The results also provide additional empirical evidence that given an extensive class of problems, most methods will provide on average comparable solutions. As shown in this thesis, the conditional probability approach to selective inference represents a viable solution to control type I error while avoiding any data loss due to data splitting. In the current research environment, incentives and funding policies need to be reshaped in order to bring about effective changes on the overall reliability of the published papers, but the tools to provide rigorous results, while meeting the needs of the researchers, are available for anyone conscientious enough.

Contents

Chapter 1

Introduction

1.1 Search for truth: determinism vs chaos in brief

Humankind has always tried to control the surrounding environment in order to improve its existence. To accomplish such task, we had to gain knowledge about the laws of the world, and the more we evolved the more we wanted to extend this control to the smallest of the details. Nowadays this is more true than ever, but even though we are able to create complex systems that make our everyday lives easier, things do not always go as planned. When we face discrepancies between what we planned and what actually happened, we often blame our misfortune, but was it simply our lack of knowledge that led to the unwanted result, or was it something we truly could not predict? In other words, is reality deterministic or stochastic?

In philosophy, the debate over the deterministic or stochastic nature of the world has been going on for centuries. Possibly the first to start the argument in modern literature was Descarte with the principle of causality in 1641 [1] "Every effect has an antecedent proximate cause". Eventually the concept of universal determinism of Laplace came into being in 1778 [2], he argued that through scientific causality we can exactly predict the state of things, we would "just" need to comprehend all the forces acting to cause any given event.

Although probability came into being into the 18th century, we had to wait until the 1892, for the birth of the chaotic universe theory. While studying the evolution of a physical system over time, Poincaré observed that even the slightest

difference in the initial conditions might result in large differences in the final phenomena, due to events that we may call random, which are unpredictable even if the laws of nature held no secret for us [3].

Many other arguments were brought in favor of the two views, but what remain certain is that with our actual tools we are not able to discern between our lack of knowledge and a truthfully random universe, actually at the moment these opposing views seem more like two sides of the same coin. A coin is also the perfect example to show how probability can be seen under both paradigms. Let us say that we are interested into knowing what the outcome of a fair coin toss will be. We know that in the long run we are going to observe half of the time tail and half of the time head. Unfortunately, that does not tell us anything at all about the exact result of the next coin toss, thus one could be led to think that it is in fact randomness that rules our world. Now, let us add some details to our previous case. Let us say that the coin toss is going to be performed inside a room and in total absence of wind. Moreover, imagine to know the details about the amount of strength that will be put in the coin toss, the exact spin that it will be given, when and how the spinning will be stopped, and all the rest of possible factors that may interfere with the outcome. Would it not be possible to exactly determine the outcome of the coin toss? If we believe the answer to be yes, and extend this reasoning to everything in our universe, then we are using statistics and probabilities just to deal with our lack of knowledge about a deterministic world by simply evaluating the chances of something happening in the long run. If instead we believe the answer to be no, then we are dealing with a stochastic world by trying to identify its stochastic rules. The debate is without doubt fascinating, but from a practical point of view, it does not matter which one reflects the underlying truth, since even if the world is in fact deterministic, it would simply be impossible to have perfect knowledge about all the factors influencing every outcome. So regardless of the true nature of reality, in order to put sense

into randomness, statistics has become our torch to peer into the unknown in the pursuit of knowledge.

Despite its focal role, we must bear in mind that statistics is not an exact science, and that we have to rely on the multiplicity of the evidences to draw a conclusion about the presence or absence of an effect. Unfortunately, replicability, which is the only source of real evidence of effects, is often hurt by researchers misconduct in particular by adaptive generation of hypotheses, that as expressed by Benjamini [4], may be referred to as the silent killer of replicability.

1.2 Replicability

Replicability is the ability of a scientific experiment or trial to be repeated by others and obtain a consistent result. The need for such a characteristic in a scientific experiment was firstly stated in the 17th century by Robert Boyle, who noted that a single experiment is not enough and one should continue to replicate it until we are convinced of its validity [5]. According to Fisher, replicability is also a characteristic of a good experiment; “A scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give ... significance.” [6]. Replicability cannot be assured but it can be enhanced through well and transparently designed experiments, and reproducible data analysis. It should be responsibility of journal editors to have only this kind of papers getting published. However, as of 2019, we are in the middle of a replicability crisis, both in academic ([7, 8]) and industry research ([9]), with replication rates often ranging from 40% to 10%, how did we get to this point? By summarizing what stated by Ioannindis in 2005 [10], and the main points of a 2016 workshop about this crisis [11], the main culprits are:

- The ratio of true to no relationships among the relationships probed in each scientific field;

- Suboptimal power due to small samples;
- Extreme multiplicity, with many factors being analyzed;
- Adaptive search of the results (i.e. selective inference, which will be discussed in detail in chapter 3);
- Financial and other interests, and prejudice;
- Chasing of statistical significance.

Some of these points are hardly improvable, due to the conflict of interest that revolves around a researcher career and his publications, and publishers' interest in the higher citation rate provided by positives findings. Having accounted for these problems, there are ways in which we may improve the scientific process, one way to do so, is to spread knowledge about the problem while hoping in researchers' good faith. One concrete solution has been proposed by Benjamini [4], the idea is to have a mandatory replication study, for the main outcomes of the studies, attached to the published article. In this way, we protect ourselves from the conflict of interest by having another researcher carrying out the study while having a probability of wrongly rejecting the null hypotheses in both studies, with effects in the same direction, equal to 0.00125 (for $\alpha = 0.05$). Of course, in order to make this possible, granting agencies would have to cooperate and allocate funds accordingly to the replication effort and its adequacy. However, inconclusive or negative replication results should not be a reason to cancel a grant. Another way to improve things is to have journals accept papers before studies are actually conducted, so that their decision would be based only on the starting hypothesis and the methodology, and not on the results, thus removing one of the incentives to follow statistical significance.

1.3 Evolution of statistics and statistical needs

Frequentist statistics taught in universities nowadays, tells us that statistical inference is the process through which we extend the results observed on a random sample to the underlying infinite population. In order to correctly infer results, the classical frequentist approach relies on a priori specification of hypotheses, in other words we fix a model, which is a set of assumptions, under which our results are valid. This singular view of the inferential process is actually a hybridization [12]. In fact, this topic created debate among whom we may call the father of Statistics, Ronald Fisher, and the duo formed by Neyman and Pearson. Where the first thought a statistician should be flexible and create ad-hoc models for every single case, the seconds thought a rigorous mathematical approach should be consistently followed each time, in order to obtain coherence in the formalization of statistical problems[12]. Several other differences can be seen in the two school of thoughts, the most of important of which is probably the role of the p-value. For Fisher it should have been used as a continuous index of evidence against the null hypothesis without any decision rule, which was later introduced by Neyman and Pearson, as it is necessary in order to express a judgment over the presence or absence of an effect. Despite their differences, what they could not foresee is that their approaches would eventually become known as one thing.

Statistics is still a new science, and since its purpose is to find answers to real problems, methods have constantly evolved with the emergence of new needs. Modern statistics was born between the late-19th and early-20th century, at that time computations had to be done by hand so methods were built to handle few variables and small samples. With the advent of the digital revolution, data availability and computation speed greatly grew, thus researchers tried to make the most out the data they had, giving rise to what we may call "adaptive statistics"

(i.e. model selection procedures) and to machine learning. Attention has been shifted to prediction (i.e. machine learning), which surely has its purpose, but the results stemming from prediction tools should not be lightly used to carry out frequentist inference, where the latter establishes objective rules in order to take decisions regarding the absence of an effect, the former does not establish any objective decision mechanism. At the boundary between statistics and machine learning lies model selection, which is embedded into the prediction mindset, but should find place in statistics only if applied with the proper care. Improper inference carried out on the selected model increases type I error [13], thus by treating the selected model as if it was pre-specified, we invalidate the procedure (see chapter 4 for a quantification of the extent of type I error). Users of any method should always be aware of its pros and cons in order to use it properly. Sadly, that is often not the case with statistics. Researchers often have so many data that it is almost impossible to make any sense out of them, thus leading to abuse of automatic procedures. These methods are used to quickly look through data in order to find best evidence of the presence of an effect, without realizing how this a posteriori formulation of hypotheses can easily lead to deceptive results, see chapter 4. Critics to this approach have been made for over 20 years[14] (see section 3.2), but up to the last 7 years, the only solution was to use data splitting [15], which never caught on in standard research, due to the loss of observations and thus statistical power. In the last half-decade, the so-called selective inference framework has started emerging. This framework allows us to take into account the model selection procedure used to identify a given model, thus allowing to carry out inference without sacrificing any data and while controlling type I error.

1.4 Aim of the thesis

The aim of this thesis is to show the threat posed by adaptive generation of hypotheses to the reliability of science. In order to fully understand the implications of the findings reported in chapter 4, it is necessary to get a grasp of the problems arising in knowledge creation and the role played by the current incentive system within the scientific environment. These problems are faced in chapter 2. In chapter 3, the selective inference framework is summarized in order to grant the reader a basic understanding of it.

Chapter 2

The role of science

In order to understand why it is important to address selectivity, we have to recall what is the goal of science and how can researchers be attuned to it while performing frequentist inference. As exemplified by Fig.2.1, science is a learning process that starts by observation of the surrounding world and ends with some acquired knowledge about it that eventually becomes integrating part of our shared knowledge.

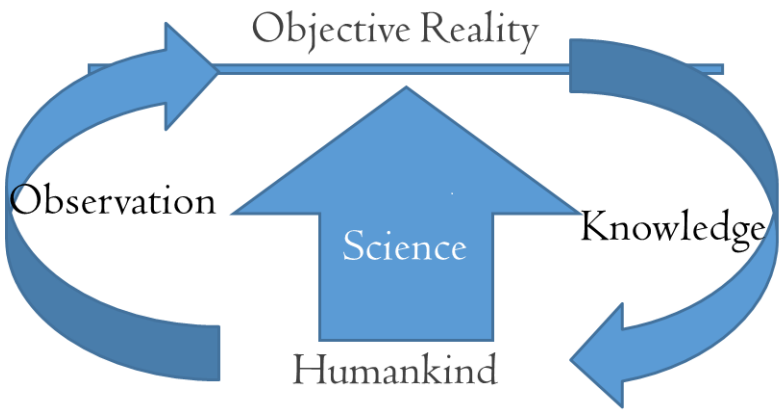

Figure 2.1: Exemplified scientific process. The solid bar represents the inaccessibility of an objective state of perception to humankind

We use science to build up over time a reliable source of knowledge, but since the only thing we are sure to perceive is our subjective sense of reality, how can we define what is reliable? Even if an objective reality were to exist, we as humans are bounded to never know for sure how far our perceived reality is from it. Epistemologists have long debated on how to define a reliable process [16], and despite the valid counter-arguments, an interesting definition is given by Goldman [17], which says that a process is reliable if its truth ratio (number of true

beliefs generated/number of beliefs generated) is close to 1.0. Unfortunately, the only way to use such definition would be to know what is true and what is not. Instead, what we humans do to evaluate the reliability of a process is to look at internal coherence of the generated beliefs, which does not imply at all that our beliefs are true or that the process is reliable, but this is as much as we can do. Given this abiding condition of uncertainty, on which basis can we define whether a learning process is scientific or not? This eternal dilemma goes by the name of demarcation problem and despite the long debate there is no complete consensus on a foolproof solution to it. Probably the best demarcation methodology is given by Lakatos [18], who properly explained Popper's point of view on the matter [19]. To summarize the concept, we could say that a theory is scientific if it is possible to produce data that could potentially falsify it. This leads inevitably to an ever-increasing degree of knowledge for humankind, with old views of the world being replaced by new ones. In other words, science can be seen as a collection of stochastic processes, one for any given set of hypotheses, H_0, which is evaluated through a model, M, so that over time the natural filtration of such processes can be written as

$$F_0(H_{0_0}, M_0) \subseteq F_t(H_{0_t}, M_t) \tag{2.1}$$

Hopefully reducing our uncertainty about the nature of H_0, while assuming that its true value stays fixed over time.

Up until now, I have glossed over the trickiest part of the process, how do we move from observation of phenomenons to acquisition of knowledge? Since the observation process occurs under lack of knowledge about the true data-generating mechanism, and thus under conditions of uncertainty about all the involved factors, we have to find ways to incorporate this uncertainty into our acquisition of knowledge. Inferential statistics was born to answer this need, bor-

rowing elements from both the deductive and the inductive reasoning in order to gauge the evidence brought by data against a given hypothesis. Now depending on the inferential framework we choose to follow, frequentist or Bayesian, the way in which we formalize the problem changes. Despite the long debate between ideologists of both sides upon which is the correct way to carry out inference, both are valid approaches to inference, as they are simply different ways of looking at the decision-making process. Where frequentist inference takes shelter under an objective approach, Bayesian inference chooses to rely on the strength of subjective views, so where the latter reflects a human approach to knowledge, the first tries to escape it. Both approaches have weaknesses and both are easily susceptible to exploitation through poor management of procedures. There is no doubt though, that suspicions about the reliability of a finding should arise when conclusions change systematically depending on the inferential framework used.

Having introduced the concept of inference under conditions of uncertainty we can now try to define what a reliable process should be like in this context. Following a frequentist framework, we know that we can commit type-I and type-II errors when carrying out inference, this means that a process which exploits such a methodology will be reliable (i.e. truth ratio close to 1) only if it sticks to the declared error rates. As shown in section 1.2, replication efforts will usually show very low replication rates, this is a strong evidence of a high presence of false positives in the published literature. If one stops to think about it though, this outcome should not be so surprising.

To start off, the publishing process acts as a discriminant between what is deemed to be worthy of being published and what is not, on the basis of several factors which are not related to the quality of the paper in question. The most relevant is the citation potential, which is higher for positive findings and which therefore makes them more sought after by journal publishers, thus leading to a biased

publication strategy. If you combine this behavior with the propensity of funding agencies to more easily fund researchers with a high number of citations, or those who promise novel findings, you can see how the positive feedback loop between a researcher's career and the publishing of such results continues to be fueled. This relationship has been increasingly exploited over the years, since as shown by [20], the number of positive findings in the literature kept rising from 70.2% in 1990 to 85.9% in 2007, with this last observation possibly not representing the final plateau. To sum it up, this means that when we talk about the reliability of the scientific process, we are actually also talking about the reliability of the publication process, which is the only evidence on which basis we can evaluate the underlying mechanism that generates knowledge.

Taking inspiration from Ioannidis [10], we can now think of a thought experiment. Let us assume that we happened to know the true state of several hypotheses being tested, if the ratio of false to true null hypotheses (R) was to be equal to 1, and if everybody was to run analysis with the same inferential error rates, we could think of some possible scenarios as the ones in Tab. 2.1. If this were the case, it would be easy to evaluate the reliability of researchers' actions by simply looking at the gap between the declared error rates and the ones observed through replications efforts.

	Hypotheses being tested		
	100 False H_0	**100 True H_0**	**Published false positives**
Scenario 1 ($\beta = 20\%, \alpha = 5\%$)	80 rejection	5 rejection	5/85=5.88%
Scenario 2 ($\beta = 80\%, \alpha = 20\%$)	20 rejection	20 rejection	20/40=50%

Table 2.1: Published literature scenarios with ratio of false to true null hypotheses equal to 1

If now we let R varies to 1/3, by looking at the higher degree of false positives in the literature, Tab.2.2, we can realize how dangerous it is to randomly test hypotheses. By trying to portray this thought experiment into reality, we know

that exploratory studies and phase I clinical trials test many hypotheses, so R is likely to be way lower than 1/3. Moreover, we know that before conducting a study researchers quite often run power analysis only for the main outcome, or they do not run it at all, and at the same time they often carry out several actions which lead to an inflation of type I error, see section 1.2. It becomes therefore easy to see why the literature is filled up with false positives.

	Hypotheses being tested		
	50 False H_0	**150 True H_0**	**Published false positives**
Scenario 1 ($\beta = 20\%, \alpha = 5\%$)	40 rejection	8 rejection	8/48=16.67%
Scenario 2 ($\beta = 80\%, \alpha = 20\%$)	10 rejection	30 rejection	30/40=75%

Table 2.2: Published literature scenarios with ratio of false to true null hypotheses equal to 1/3

In order to enhance the reliability of science, funding agencies and journal publishers should start acting differently. Either by removing the link between a researcher's career and the amount of positive findings he discovers or by introducing incentives to publish replicable results. Otherwise, as pointed out by Higginson and Munafò [21], within this ecosystem (i.e., incentive structures) researchers will keep maximizing their fitness (i.e., publication records), regardless of the scientific value of their discovery.

In this context the least one could do is to promote the proper way to carry out inference. In this thesis, I focus on the control of type I error, which is something that can be achieved even after data collection. Something that should be easily understood, but is often neglected, is the need to keep type I error under control when running several statistical tests or models. When performing multiple hypotheses tests the probability of making at least one false discovery (FWER), is equal to

$$\alpha_{FWER} = 1 - (1 - \alpha_{\text{(per comparison)}})^m \tag{2.2}$$

with m equal to the number of comparisons. Now, something more subtle occurs when out of all the possible tests we only select those passing a threshold, or in multipurpose surveys, when we ignore the total amount of tests being done while focusing only on one aspect of the whole survey. This selection leads to an inflation of type I error due to a reduction of the number of comparisons taken into account and thus to adaptive generation of hypotheses. In the next chapter the focus will slowly shift to inference after model selection, which is only one of the possible ways in which selectivity may occur, leading the way for the fourth chapter in which I will show through a simulation the extent of type I error inflation in post-model-selection inference.

Chapter 3

Selective inference background

When we talk about selective inference, we refer to a framework in which inference is carried out on a subset of parameters that turned out to be of interest after viewing the data, which therefore implies an adaptive generation of hypotheses. The selection can occur in different ways, we talk about out-of-study selection when it is not evident in the published works

- File drawer/publication bias: bias due to the tendency to publish only positive results.
- P-hacking: extensive "manipulation" of the analysis in order to obtain significant results.
- Interactive data analysis: selection of hypotheses after data snooping.

these actions will always lead to a bias into the pool of published papers. Instead, we talk about in-study selection when it is evident in the published work

- Selection by the abstract, table, figure.
- Selection by highlighting effects passing a threshold.
- Reported model selection.

Out-of-study selection cannot be recognized in most cases, but sometimes it is possible to highlight the presence of publication bias, through instruments like contour-enhanced funnel plots [22]. On the contrary, in-study selection is evident, but not easy to deal with. Often, papers have many results, so a researcher

has to focus on the most interesting ones in order to create an alluring paper, it is inevitable. As for inappropriate inference after model selection, this can either be reported in the published paper, due to unawareness of the inappropriateness of such procedure, or it can fall under the category of interactive data analysis when not reported.

3.1 Model selection procedures

A model is an abstract construct which we use to describe the relationship between different elements of reality. We usually use them to explain an asymmetrical relationship between some explanatory variables, X, and a response variable, $f(X)$. In order to understand which explanatory variables, among a given subset of candidate ones, are part of the true data-generating mechanism, we resort to selection procedures. The most basic form of model selection is the scientific inquiry itself, in fact over time, through evaluation of hypotheses, we are eventually able to build up enough evidence to add or remove variables from what we may call a standard model. When exploring the effects of new possible explanatory variables, we should always be careful in carrying out inference, in the words of Ronald Coase "If you torture the data enough, nature will always confess" [23]. When using frequentist inference to be careful means that the identification of the best fitting model should be done on observations that are different from the ones used to carry out inference (i.e., data splitting [15]), by doing this we are able to keep type I error under control in the long run. Sadly, this approach is not often applied due to the loss of statistical power it entails. To overcome this obstacle new methods which exploit conditional probabilities have been developed in the last 7 years. The main extensions regard the forward selection (FS) 3.1.1, and the Least Absolute Shrinkage and Selection Operator (LASSO) 3.1.2.

3.1.1 Forward selection

The forward selection is a kind of stepwise regression [24], it is an automatic procedure that allows to select the best model out of several ones by minimizing a given selection criterion. This procedure adds, or removes, a regressor at each step depending on the chosen direction (forward or backward). In the forward selection, the procedure starts from the null model (i.e., model with only intercept) and variables are added to it until there is no more improvement in the selection criterion. Originally, the selection criterion used was the drop in RSS (residual sum of squares),

$$RSS_{s-1}(y, X_{[p_1,\ldots,p_J]}) \leq RSS_s(y, X_{[p_1,\ldots,p_J,k_1]}) \tag{3.1}$$

with p representing the predictors at step s-1, and k the predictor added at the following step. The drop in RSS would be deemed to be significant by using the following F test

$$F = \frac{(RSS_p - RSS_{p+k})/k}{RSS_{p+k}/(n-(p+k)-1)} \tag{3.2}$$

where n is the number of observations. This procedure is actually invalid and several critics have been reported in the literature (see section 3.2). Eventually, in order to take into account model complexity in the selection process, other selection criteria such AIC or BIC were introduced. The idea is to control model complexity by simply adding a penalization factor, λ, to the log likelihood of a given model

$$\text{Information criterion} = \lambda - 2ln(\hat{L}) \tag{3.3}$$

with the procedure stopping when it is not possible to further minimize the information criterion. Regardless of the chosen criterion, these methods haven been

criticized for a long time, but as shown in chapter 4, their performance is similar to the usually more appreciated LASSO estimator.

3.1.2 LASSO

The LASSO [25] is an estimator which is obtained by simply adding a penalization, λ, for the magnitude of the regression coefficients to the classical OLS estimator

$$\arg\min_{\beta}(y - XB)^T(y - X\beta) + \lambda||\beta||_1 \tag{3.4}$$

where $||\cdot||_1$ represent the absolute value. For λ=0 the LASSO solution is equal to the OLS, while for $\lambda = \infty$, the solution is equal to $\overrightarrow{0}$. So for a given lambda the LASSO provides model selection by setting to 0 some regression coefficients. The idea is to introduce some bias in the estimation process in order to reduce the variance and thus obtain a lower mean squared error for the regression coefficients

$$MSE(\hat{\beta}) = (Bias(\hat{\beta}))^2 + Var(\hat{\beta}). \tag{3.5}$$

So the chosen value of λ will usually be either the one minimizing the MSE, or the one which is 1 standard error away from it, in order to avoid overfitting[26].

3.2 Critics to inference after model selection

In contrast to proper frequentist inference, which relies on pre-specified hypotheses, inference after model selection relies on hypotheses which are selected on the basis of the results; thus, these hypotheses have to be treated differently or the subsequent inference will be invalid due to inflation of type I error. Since the introduction of stepwise regression in 1960 [24], inappropriate inference after

model selection has begun to spread, despite having been recognized as a flawed procedure several times over the years. Lockart et al. [27] and Draper et al. [28], show that the F-tests used for selecting regressors do not have the claimed F distribution, thus a model selected in this way is not the best among the candidate ones, but even if the model was selected by other more performing means, such as the minimization of information criteria, the p-values of the select regressors could not be trusted as shown in chapter 4. Moreover, as stated by Copas and Long [29], "The choice of the variables to be included depends on estimated regression coefficients rather than their true values, and so X_j is more likely to be included if its regression coefficient is overestimated than if its regression coefficient is underestimated", in other words the use of such methods would lead to bias into the published effects and thus on the subsequent meta-analysis. The general framework to overcome these problems is summarized in the next section.

3.3 Post-model-selection inference

Adaptive generation of hypotheses, such as those formulated after model selection, violates the classical inference paradigm, resulting in inflation of type I error. In order to avoid such inflation we have to take into account that the model, M, was obtained through a model selection procedure, so that under the assumption of H_0 being true we can control for the selective type I error, α

$$\mathbb{P}_{M,H_0}(reject\ H_0 |\ (M, H_0)\ selected) \leq \alpha \tag{3.6}$$

As suggested by Fithian et al.[13], we can think about the scientific process as a random process, in which models and hypotheses are chosen (randomly) on the basis of the outcome of previous results, and it is implicitly assumed that the randomness in selecting M and H_0 is independent of the data used for infer-

ence. Data splitting tries to emulate this process by splitting the data Y= (Y_1, Y_2) with Y_1 independent of Y_2, so to then set aside Y_1 for selection and Y_2 for inference. Instead of sacrificing some of the data, which leads to a loss of statistical power, we could condition to the event that (M, H_0) is selected. As with data splitting, we treat data as if they were revealed in stages: in the first stage, we observe "just enough" data to resolve the decision of whether to test (M, H_0), after which we can treat the data $(Y|(M, H_0)\ selected)$ as "not yet observed" when the second stage commences. So we carry out inference only on the selected predictors in model M, which means that μ (i.e. $\mathbb{E}(Y)$), in the simple linear case $Y \sim N_n(\mu, \sigma^2 I_n)$, is equal to

$$\mu = X_M \beta^M \tag{3.7}$$

In the non selective case the OLS estimator would be $\hat{\beta} = (X_M^T X_M)^{-1} X_M^T Y$, instead after the model selection we can rewrite for a particular j and M the estimator as $\hat{\beta}_{M_j} = \eta_{M_j}^T Y$, with

$$\eta_{M_j} = \frac{X_{j \cdot M}}{||X_{j \cdot M}||^2}, \tag{3.8}$$

where $X_{j \cdot M} = P^{\perp}_{X_{M \setminus j}} X_j$, is the reminder after adjusting X_j for other columns of X_M, and $P^{\perp}_{X_{M \setminus j}}$ denotes the projection onto the column space of $X_{M \setminus j}$. In other words we project X_j in the space defined by the model selection procedure, with $X_{j \cdot M}$ being the adjusted value of X_j in the selected model, which is then normalized so to have a vector of length one.

For any distribution belonging to the exponential family, under the selected model we have

$$Y \sim exp\{\frac{1}{\sigma^2} \beta^T X_M^T y - \frac{1}{2\sigma^2} ||y||^2 - \psi(X_M \beta, \sigma^2)\} \tag{3.9}$$

If σ^2 is known, the sufficient statistics are $X_k^T Y$ for k$\in$ M. Letting A be the se-

lection event, inference for β_j is based on

$$\mathcal{L}_{\beta_{M_j}} = (X_{M_j}^T | X_{M\setminus j}^T, A) \tag{3.10}$$

Which is equivalent to carry out inference on the Z test statistic

$$Z = \frac{\eta_{M_j}^T Y}{\sigma ||\eta_{M_j}^T||} \tag{3.11}$$

with $\sigma^2 = ||P_{X_{M\setminus j}}^{\perp} Y||^2/(n - |M|)$. Although, Z is marginally independent of $X_{M\setminus j}^T Y$, it is not conditionally independent given A, which means that the distribution of Z depends on the model selection procedure, and thus on $X_{M\setminus j}^T Y$. In order to obtain a tractable likelihood, Lee et al.[30] have shown that conditioning to the selected model as in 3.10 is not enough and we need to condition also on the signs of the selected effects, thus obtaining a simpler selection event. In general, to obtain a tractable likelihood when conditioning to the model section procedure is not an easy task. Given that model M has been selected and being $X_j^T Y$ the sufficient statistic for $j \in M$, if σ^2 is known, by following [31], we can write the conditional post-selection likelihood as

$$\mathcal{L}(\beta_M) = \frac{P(M|X_j^T Y) f(X_j^T Y)}{P(M)} I_M \tag{3.12}$$

where P(M) is the unconditional probability of selecting M, and $I_M = I_{\{S(y)=M\}}$, is the indicator function for the selection event. The main problem of carrying out inference after model selection is to identify P(M). In this regard, probably the most interesting result in inference after model selection is related to the polyhedral condition set. Tibshirani et al. [32] shows that as long as it is possible to characterize the selection event as *y* falling into a polyhedral set, $Ay \leq b$ (see Fig. 3.1), it is possible to carry out valid inference.

In particular, [30], shows that for the LASSO with fixed λ, it is possible to rewrite the polyhedron in terms of $\eta^T y$ and Z, so that it is possible to explicitly write the

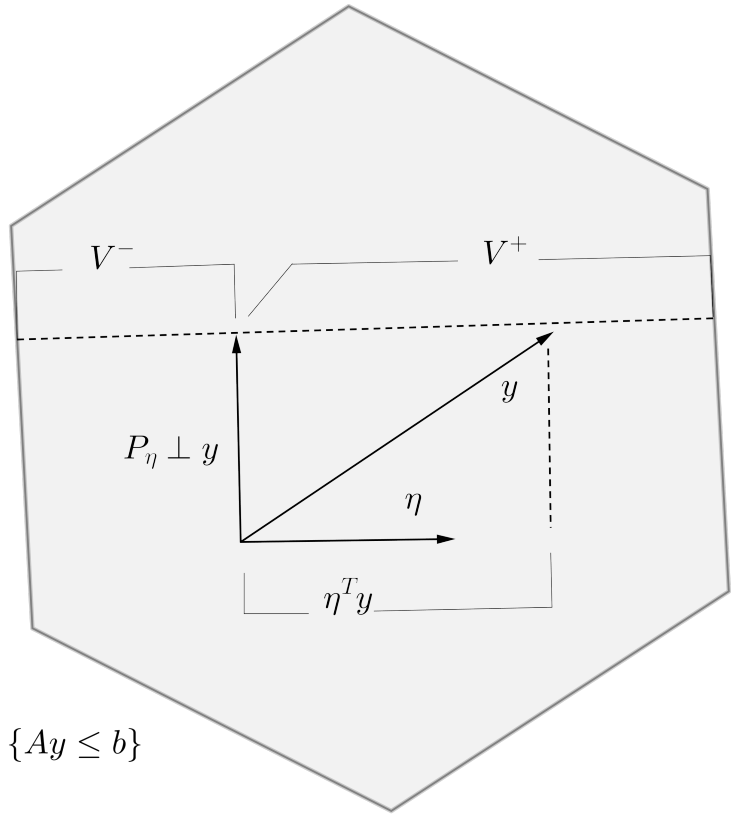

Figure 3.1: Reformulated representation by Ryan J Tibshirani et al.[32], that shows the quantities involved in the polyhedral selection

polyhedron boundaries

$$\{Ay \leq b\} = \{(Ac)_j(\eta^T y) \leq b_j - (AZ)_j \text{for all j}\} \tag{3.13}$$

$$\begin{cases} \eta^T y \leq \frac{b_j - (AZ)_j}{(Ac)_j}, & \text{for j:}(Ac)_j > 0\ , \\ \eta^T y \geq \frac{b_j - (AZ)_j}{(Ac)_j}, & \text{for j:}(Ac)_j < 0\ , \\ 0 \leq b_j - (AZ)_j, & \text{for j:}(Ac)_j = 0\ , \end{cases} \tag{3.14}$$

where A and b represent the constraints of the LASSO solution, and $c \equiv \Sigma\eta(\eta^T\Sigma\eta)^{-1}$.

By also conditioning on the observed value of Z, we have that

$$[\eta^T y | Ay \leq b, Z = Z_0] \sim TN(\eta^T\mu, \sigma^2||\eta||^2, V^-(Z_0), V^+(Z_0)) \tag{3.15}$$

The work by Tibshirani et al. [32], shows that similar results are also available for the FS, since it is possible to characterize the procedure as y falling into a polyhedral set. P-values for the FS can be computed in three different ways:

- Inference from sequential p-values: here we test sequentially the null hypothesis that a new predictor, which enters at step k into the active set A_k, is equal to 0.
- Inference at a fixed step k: here for every predictor in the active set A_k, at a fixed step k, we test the null hypothesis that the predictor is equal to 0.
- Inference at an adaptively selected step k: here for every predictor in the active set A_k, at a step k selected through a given criterion (e.g. AIC), we test the null hypothesis that the predictor is equal to 0.

This paradigm has been extended to accommodate quadratic inequalities which are needed in order to take into account the search for λ in the LASSO [33], or to allow inference in presence of categorical variables [34].
To summarize the rationale behind this section, we can say that to carry out inference under the selected model means to condition on the selection event A that identifies such model. This event affects the distribution of the Z-test statistic by making some values of Z impossible to observe, thus, in order to keep under control the percentage of false positives, we need to observe more extreme values of Z that if no model selection was carried out. In general, the selection event will reduce the range of possible observed values starting from values closer to the null hypothesis, in other words after the model selection procedure occurs, it is more likely to find significant predictors in the selected set than in the starting one.

Chapter 4

Adaptive generation of hypotheses in inference after model selection

4.1 Introduction

In this chapter the inferential performances of the FS and the LASSO are compared through a Monte Carlo simulation (see section 4.2). The comparison is carried out while keeping under control type I error, through the conditional probability approach, and while avoiding to do so. Some comparison between the two procedure have been carried out over the years, [35, 36], but none of these addresses an important question "What is the extent of type I error inflation in the selected models?". Awareness around this bad practice started to rise in the 2000s, but as any practitioner would surely recognize, model selection is still often practiced carelessly. The worst part is that models are treated as prespecified, thereby altering the perception about the quality of the work in question. Inflation of type I error, leads inevitably to adaptive generation of hypotheses, with researchers coming up with hypotheses in order to justify their results. While it is complicated to evaluate the extension of this behavior in general practice, the aforementioned severe lack of replicability in the literature is evidence of its presence. The idea of this work is to simulate results of bad practice in model selection usage. This means that for the LASSO, a variable recovered in the true set is treated as significant, which while being a misinterpretation it is also a useful way to show how the LASSO threshold fares in comparison with

the FS. Obviously, the standard LASSO should not be used to carry out inference, using the selection event as a threshold for significance, as much as the results stemming from the FS should not be used to carry out inference.

4.2 Simulation setup

4.2.1 Steps

The simulation can be described by the following steps which are repeated for each loop

(i) Data generation: data are generated from a multivariate normal distributions, $X \sim MVN(0, \sigma^2 I)$. The variance, σ^2, is fixed equal to 1. The response variable, Y, is generated as $Y = X\beta + \epsilon$, with $\epsilon \sim N(0, 1)$.

(ii) Model selection: the model selection procedure is applied to the dataset containing both true predictors and noise variables.

(iii) Outcomes: estimates of the outcomes reported in 4.2.3 are stored.

Afterwards, the average of these estimates is computed over 1000 loops, then this process is repeated again three times and results are averaged again. This is done in order to avoid any possible influence of the starting seed in the random generation process, to obtain more stable estimates and compute a measure of variability. The function to run the simulation is written in R [37] and is available in the supplementary material 5. The external packages required are the selectiveInference [38], for the conditional probability approach, the glmnet [39], for the LASSO, and their related dependencies. The FS is available in the standard R environment through the step function by simply setting the direction parameter as forward.

4.2.2 Parameters configuration

The parameters involved in the simulation are the following

Selective error:

Uncontrolled: inference carried out on the same data used for the model selection;

Controlled: inference carried out through the conditional probability approach.

Model selection procedure:

LASSO: the model is selected using the 1 SE rule, [26];

FS: the model selected is the one holding the smallest AIC, within the range of possible models. For the additive model this range spans from the null model (only intercept) to the maximum model (all regressors included additively). For the multiplicative model, the range spans from an additive model to the one with all possible pairwise interactions.

Structure of the fitted model:

Additive: $\mathbb{E}[Y] = X_1\hat{\beta}_1 + ... + X_k\hat{\beta}_k + X_{\text{noise}_1}\hat{\beta}_1 + ... + X_{\text{noise}_j}\hat{\beta}_j$; the true values of the regression coefficients associated to the noise variables are all equal to 0.

Multiplicative: it is created by adding to the additive model all the possible pairwise interactions between regressors; the true values of the regression coefficients associated to the interactions involving any noise variable are equal to 0.

Effect size: three predictors are used to generate Y. The effect sizes are fixed to 0.1, 0.3 and 0.5, in both the additive and the multiplicative scenario. In the

latter they are also the values for the regression coefficients of the interactions. The regressors respectively explain 0.74%, 6.67% and 18.52% of the variation of Y in the additive scenario. The focus is on small to medium effects, as they are the ones that are difficult to identify.

Number of noise variables: the amount of noise variables ranges from twice to four times the number of true effects. This values only reflect a low dimensional case, due to exponential increase of computational times as the number of variables increases.

Observation to variable ratio: this parameter is fixed either to the usual minimum recommended value of 10 observation per variable, [40], or to the half of it.

4.2.3 Outcomes of interest

In order to compare the procedures, both the power for each effect size and the selective alpha are computed for all the 32 combinations of the above parameters. The threshold used for rejection of the null hypothesis is a p-value smaller than 0.05 for the FS in the selected model or the inclusion of the variable in the recovered set for the LASSO. Additionally, in the uncontrolled scenario, the variation in inferential conclusions is computed as the difference in the amount of noise variables passing the threshold in the selected model with respect to the pre-specified full model. While for the controlled scenario, the percentage of times that the Truncated Gaussian (TG) constraints are not satisfied is evaluated, since when this occurs the procedure does not properly keep type I error under control.

All the results discussed in the next section are available in the supplementary material 5. High dimensional scenarios were not fully evaluated due to exponential increase of computational times. A constant variability was observed for all

comparisons, with the range of the average result usually being within $\pm$ 1-2%.

4.3 Results

4.3.1 Extent of type I error inflation

The observed extent of type I inflation in Fig. 4.1, shows that even in low dimensional scenarios, the inflation can be quite severe.

The minimum and the maximum observed values of the selective error for the LASSO are respectively 12% and 40%, while for the FS, values range from 20% to 33%. The LASSO performs better than FS in the additive scenario while it performs worse in the multiplicative scenario. Reducing the number of observations per variable leads to an increase in the selective error, but this inflation is stronger in the FS than in the LASSO. Observing this relationship in a more extreme additive scenario, using 40 noise variables and 80 observations, the selective error raises to 50% for the FS and to 21% for the LASSO, as for the multiplicative scenario, both procedures show a selective error equal to 50%.

When the number of noise variables increases and the number of observations decreases, by using the FS, variables which are not significant in the starting model that end up in the selected model tend to have smaller p-values and to become significant. In the aforementioned high dimensional case, on average 3 noise variables become significant in the selected model for the FS, while 1 noise variable which is significant in the starting model is removed on average by the LASSO. This variation in inferential conclusions tends to 0 as the number of observations per variable increases.

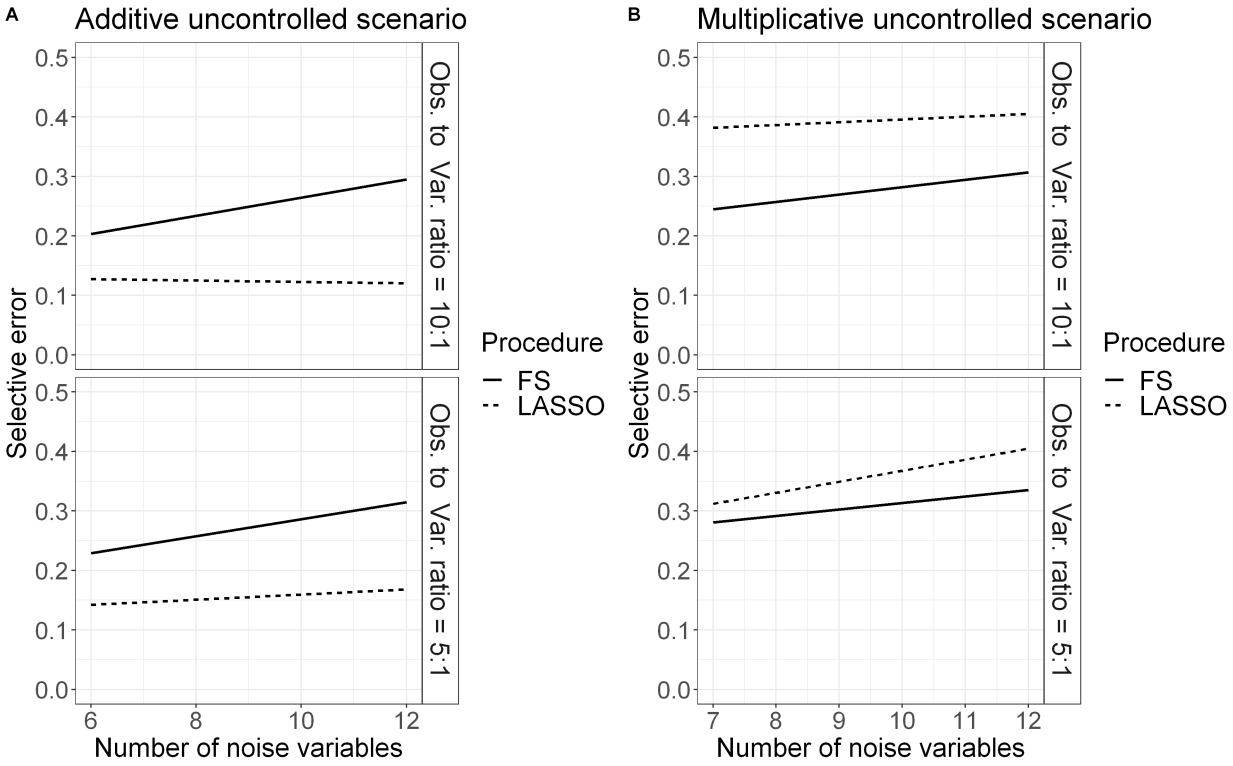

Figure 4.1: Inflation of type I error due to model selection

4.3.2 Trade-off between type I and type II error

It is straightforward to see that when we control for the selection, Fig.4.3 and Fig.4.4, the trade-off is way higher than when we do not control. The differences between the two procedures in the controlled approach are mostly due to the observation to variable ratio and the effect size. With the LASSO always improving its performances with respect to the FS when more observations are available. Also, The LASSO seems to perform slightly better than the FS for the two smaller effects but the result is reversed for the higher one. Another relevant difference can be seen in Fig. 4.2, with the LASSO failing to satisfy the TG constraints in some occurrences, leading to a small inflation of false positives in the long run with respect to the desired level α. This problem might be due to the search for lambda in the LASSO [33], which is not yet implemented in the package. As the number of observations per variable gets smaller and the

complexity of the model increases so does the percentage of times that the TG constraints are not satisfied. As for the uncontrolled scenario both procedures show the same results for the smallest effect considered, instead for the other two effects the LASSO performs better than the FS in the additive case and worse in the multiplicative case.

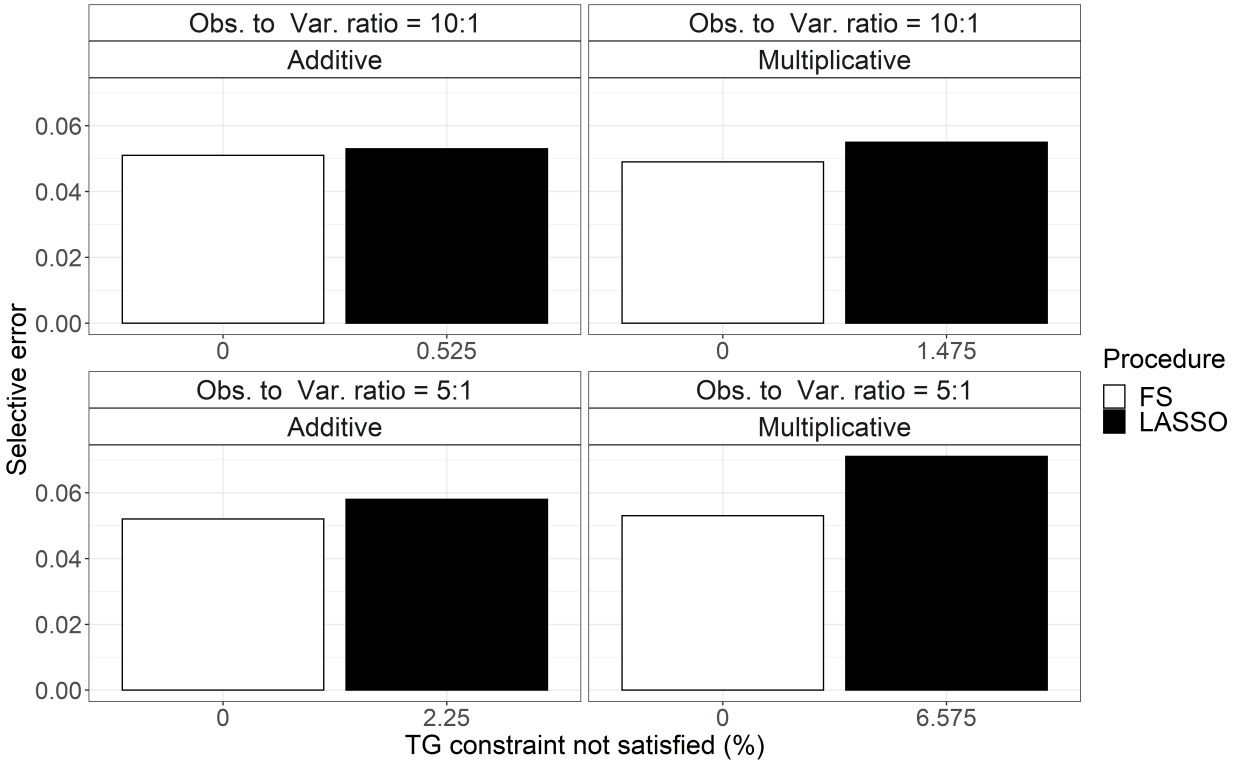

Figure 4.2: Relationship between the percentage of time that the TG constraints are not satisfied and the simulation parameters

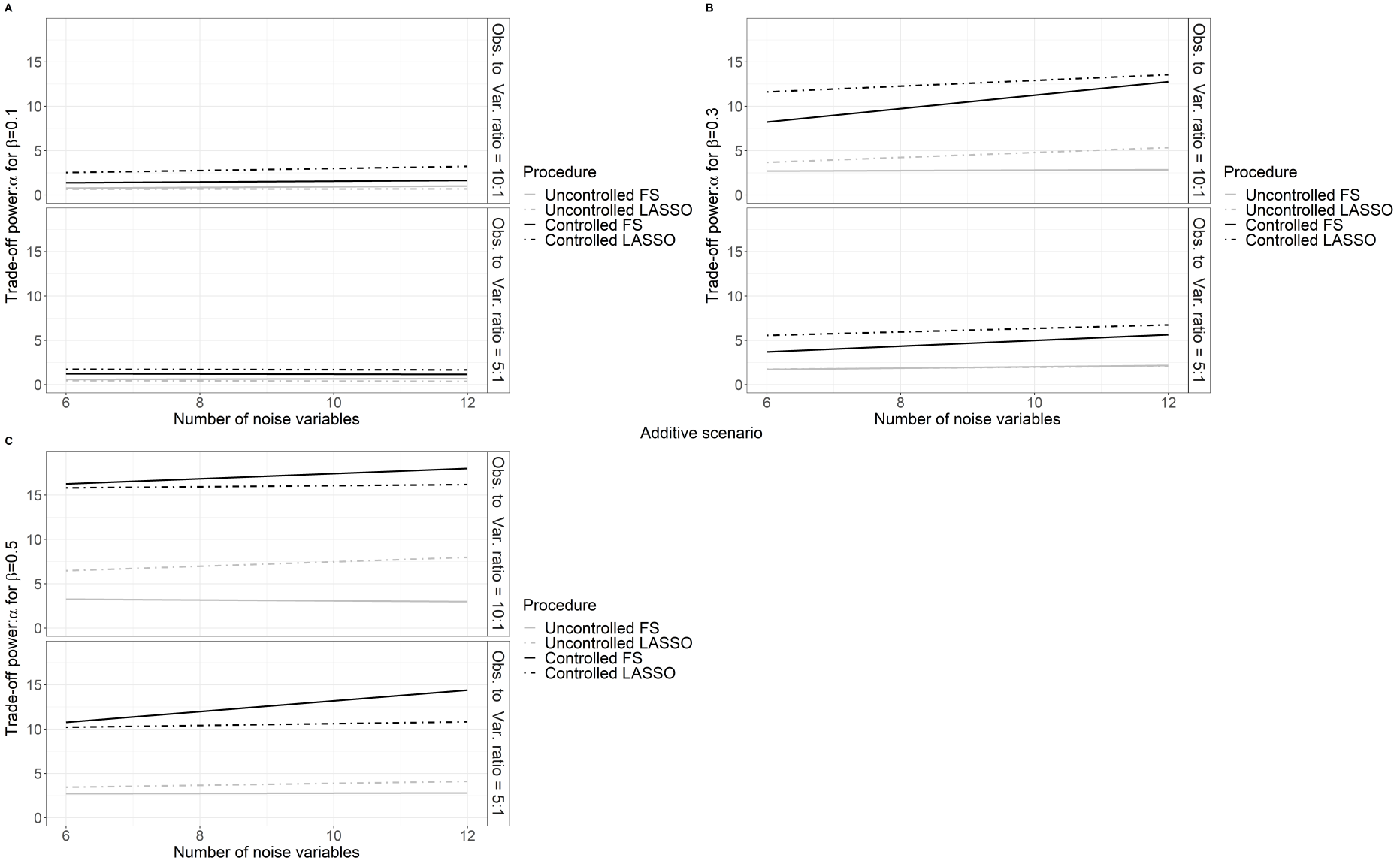

Figure 4.3: Relationship between the inferential errors trade-off and the simulation parameters in the additive scenario

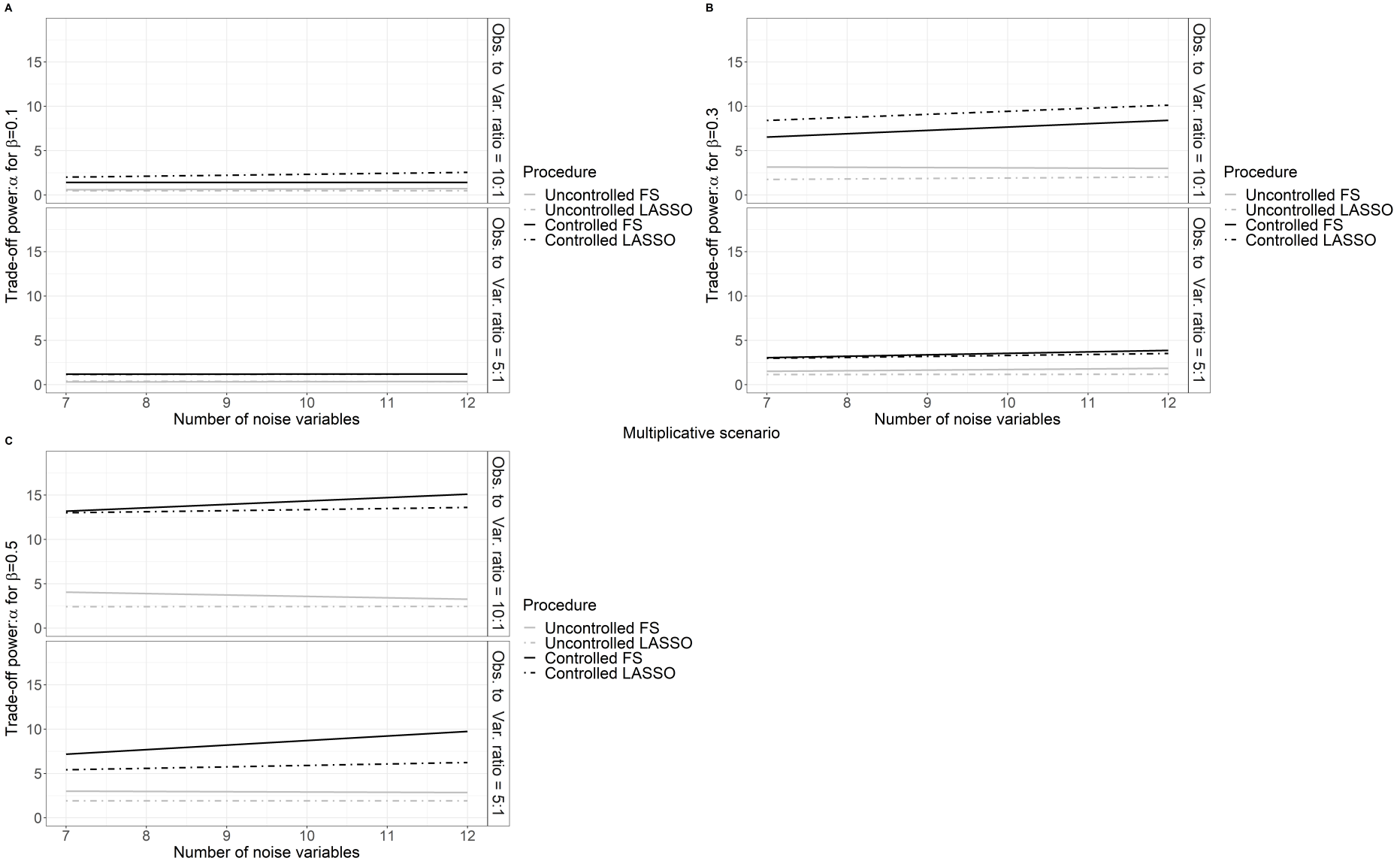

Figure 4.4: Relationship between the inferential errors trade-off and the simulation parameters in the multiplicative scenario

4.4 Conclusions

The results of this simulation clearly show that inflation of type I error can be extremely severe in inference after model selection. In order to keep type I error under control while carrying out inference, without using data-splitting, we can resort to the conditional probability approach to selective inference, which additionally leads to more powerful procedures for a fixed alpha. Differences between FS and the LASSO are negligible in the conditional probability approach, but when not accounting for the model selection in the inferential procedure, the LASSO seems to behave better in additive scenarios while the FS seems to behave better in multiplicative scenarios. This difference might represent a useful insight on the different behavior of the two procedures in the identification of the best fitting model. Despite our search for easy answers, these results represent additional evidence in favor of the no free lunch theorem [41], which shows how across all possible datasets the performances of machine learning algorithms are on average the same. Thus, without knowing for sure which is the structure of the underlying true data-generating mechanism we cannot know a priori which method will provide the best results.

Chapter 5

Discussion

This thesis rationale was to explore the reasoning behind the making of science, in order to understand why it is important to keep type I error under control when carrying out inference after model selection in a frequentist framework. To gather insight about the underlying data-generating mechanism is not easy by any means, and actually we may have already hit a dead end in the pursuit of the best methodology to carry out such task. But despite its lack of perfection, the scientific method is the best way to get a better understanding of what surround us and in order enhance its efficacy we should improve the environment in which research happens. The current publication and funding system allows and promotes the production of papers of low scientific value (i.e. little is done to promote replicable science). This means that a viable strategy to advance one's career is represented by the collection of high dimensional dataset with subsequent massive testing, which is equivalent to testing hypotheses randomly, in order to then select the most promising results. Something must be off with the system that allows such a practice, that is the farthest from something we could define scientific, to be productive. In this context, it becomes important to look at common bad practices in order to raise the awareness about such misbehaviors, since the first step to overcome problems is to realize that they exist. One of these misbehaviors, which could be easily corrected, is related to inference after model selection. It should be understood that inferential claims are trustworthy only if researchers carry out appropriate actions. Talking about statistical power makes sense only when the type I error of the procedure is fixed to the value de-

clared by the researcher. The conditional probability approach represents a solid solution to the problem of inference after model selection, which as shown can lead to a severe inflation of false positives even in low dimensional scenarios. It is also interesting to note how the structure of the true data-generating mechanism differently affects the two selection procedures in the uncontrolled scenario, which is something that has not been reported in the literature. Instead, when conditioning to the model selection event the differences between the FS and the LASSO become less relevant, and depend on the observation to variable ratio and the sizes of the effects rather than the structure of the true data-generating mechanism.

This thesis brings additional evidence to the threat posed by selective inference to the reliability of science. If nothing changes, we risk of spreading a sense of distrust in research, with loss of money and time that will become a common thing among researchers that start projects on the basis of the scientific literature. If we truly wish to improve the reliability of science, we need to start enacting policies which incentivize replicable findings rather than positive findings. Despite the difficulty of giving a truthful description of our reality, during the last three hundred years humankind has surely started making great leaps forward in this difficult task. As often happens in our daily lives it all comes to down to seeing the glass either half empty or half full, but in the making of science we should always be mindful about the importance of what we are doing, which means that we must be strict about all the steps leading to the production of knowledge. If we wish to further extend our control over nature, we first need to control how we produce science.

References

[1] René Descartes. *The philosophical writings of Descartes*. Vol. 2. Cambridge University Press, 1984.

[2] Pierre-Simon Laplace. *Pierre-Simon Laplace Philosophical Essay on Probabilities: Translated from the fifth French edition of 1825 With Notes by the Translator*. Vol. 13. Springer Science & Business Media, 1998.

[3] Henri Poincaré. *New methods of celestial mechanics*. Vol. 13. Springer Science & Business Media, 1992.

[4] Yoav Benjamini. *Selective Inference - the silent killer of replicability*. https://www.youtube.com/watch?v=6ZxIzVjV1DE, visited 07/05/2020.

[5] Steven Shapin. "Pump and Circumstance: Robert Boyle's Literary Technology". In: *Social Studies of Science* 14.4 (1984), pp. 488–490. ISSN: 03063127. URL: http://www.jstor.org/stable/284940.

[6] R.A. Fisher. "The arrangement of field experiments". In: *Journal of the Ministry of Agriculture of Great Britain* 33 (1926), p. 504.

[7] Richard A Klein et al. "Many Labs 2: Investigating variation in replicability across samples and settings". In: *Advances in Methods and Practices in Psychological Science* 1.4 (2018), pp. 443–490.

[8] Alexander A Aarts et al. "Estimating the reproducibility of psychological science". In: *Science* 349.6251 (2015), p. 943.

[9] CG Begley and LM Ellis. *Drug development: Raise standards for preclinical cancer research. Nature.[Online]. 483 (7391)*. 2012.

[10] John PA Ioannidis. "Why most published research findings are false". In: *PLoS medicine* 2.8 (2005), e124.

[11] Engineering National Academies of Sciences and Medicine. *Statistical Challenges in Assessing and Fostering the Reproducibility of Scientific Results: Summary of a Workshop*. Ed. by Michelle Schwalbe. Washington, DC: The National Academies Press, 2016. ISBN: 978-0-309-39202-0. DOI: 10.17226/21915. URL: https://www.nap.edu/catalog/21915/statistical-challenges-in-assessing-and-fostering-the-reproducibility-of-scientific-results.

[12] Johannes Lenhard. "Models and statistical inference: The controversy between Fisher and Neyman–Pearson". In: *The British journal for the philosophy of science* 57.1 (2006), pp. 69–91.

[13] William Fithian, Dennis Sun, and Jonathan Taylor. "Optimal inference after model selection". In: *arXiv preprint arXiv:1410.2597* (2014).

[14] Frank E. Harrell Jr. *Regression Modeling Strategies*. Berlin, Heidelberg: Springer-Verlag, 2006, pp. 67–72. ISBN: 0387952322.

[15] Julian J Faraway. *Data Splitting Strategies for Reducing the Effect of Model Selection on Inference*. Tech. rep. Citeseer, 1995.

[16] Alvin I Goldman. *Reliabilism and contemporary epistemology: essays*. Oxford University Press, 2012.

[17] Alvin I Goldman. "What is justified belief?" In: *Justification and knowledge*. Springer, 1979, pp. 1–23.

[18] Imre Lakatos. "Falsification and the methodology of scientific research programmes". In: *Can theories be refuted?* Springer, 1976, pp. 205–259.

[19] Karl Popper. *Conjectures and refutations: The growth of scientific knowledge*. routledge, 2014.

[20] Daniele Fanelli. “Negative results are disappearing from most disciplines and countries”. In: *Scientometrics* 90.3 (2012), pp. 891–904.

[21] Andrew D Higginson and Marcus R Munafò. “Current incentives for scientists lead to underpowered studies with erroneous conclusions”. In: *PLoS Biology* 14.11 (2016).

[22] Jaime L Peters et al. “Contour-enhanced meta-analysis funnel plots help distinguish publication bias from other causes of asymmetry”. In: *Journal of clinical epidemiology* 61.10 (2008), pp. 991–996.

[23] Ronald H Coase. “How should economists choose?” In: *Ideas, their origins, and their consequences: lectures to commemorate the life and work of G. Warren Nutter* (1988), pp. 63–79.

[24] Anthony Ralston and Herbert S Wilf. *Mathematical methods for digital computers*. Tech. rep. 1960.

[25] Robert Tibshirani. “Regression shrinkage and selection via the lasso”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 58.1 (1996), pp. 267–288.

[26] Friedman J. and Hastie T. and Tibshirani R. *The elements of statistical learning*. Vol. 1. 10. Springer series in statistics New York, 2001.

[27] Richard Lockhart et al. “A significance test for the lasso”. In: *Annals of statistics* 42.2 (2014), p. 413.

[28] NR Draper, Irwin Guttman, and H Kanemasu. “The distribution of certain regression statistics”. In: *Biometrika* 58.2 (1971), pp. 295–298.

[29] JB Copas and Tianyong Long. “Estimating the residual variance in orthogonal regression with variable selection”. In: *Journal of the Royal Statistical Society: Series D (The Statistician)* 40.1 (1991), pp. 51–59.

[30] Jason D Lee et al. “Exact post-selection inference, with application to the lasso”. In: *The Annals of Statistics* 44.3 (2016), pp. 907–927.

[31] Amit Meir and Mathias Drton. “Tractable post-selection maximum likelihood inference for the lasso”. In: *arXiv preprint arXiv:1705.09417* (2017).

[32] Ryan J Tibshirani et al. “Exact post-selection inference for sequential regression procedures”. In: *Journal of the American Statistical Association* 111.514 (2016), pp. 600–620.

[33] Joshua R Loftus. “Selective inference after cross-validation”. In: *arXiv preprint arXiv:1511.08866* (2015).

[34] Joshua R Loftus and Jonathan E Taylor. “Selective inference in regression models with groups of variables”. In: *arXiv preprint arXiv:1511.01478* (2015).

[35] Hastie T. and Taylor J. and Tibshirani R. and Walther G. “Forward stagewise regression and the monotone lasso”. In: *Electronic Journal of Statistics* 1 (2007), pp. 1–29.

[36] Hastie T. and Tibshirani R. and Tibshirani, R. J. “Extended comparisons of best subset selection, forward stepwise selection, and the lasso”. In: *arXiv preprint arXiv:1707.08692* (2017).

[37] R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria, 2017. URL: `https://www.R-project.org/`.

[38] Ryan Tibshirani et al. *selectiveInference: Tools for Post-Selection Inference*. R package version 1.2.5. 2019. URL: `https://CRAN.R-project.org/package=selectiveInference`.

[39] Jerome Friedman, Trevor Hastie, and Robert Tibshirani. “Regularization Paths for Generalized Linear Models via Coordinate Descent”. In: *Journal of Statistical Software* 33.1 (2010), pp. 1–22. URL: http://www.jstatsoft.org/v33/i01/.

[40] Harrell Jr F. E. and Lee K. L. and Califf R. M. and Pryor D. B. and Rosati R. A. “Regression modelling strategies for improved prognostic prediction”. In: *Statistics in medicine* 3.2 (1984), pp. 143–152.

[41] David H Wolpert. “The lack of a priori distinctions between learning algorithms”. In: *Neural computation* 8.7 (1996), pp. 1341–1390.

Supplementary material

- Results
- Simulation function
- Template for obtaining results