Master's thesis

2019

Fredrik Nevjen

Master's thesis

NTNU
Norwegian University of
Science and Technology
Faculty of Information Technology and Electrical
Engineering
Department of Mathematical Sciences

NTNU
Norwegian University of
Science and Technology

Fredrik Nevjen

Properties of the closed testing procedure

With applications in pairwise comparisons and model selection

June 2019

NTNU
Norwegian University of
Science and Technology

NTNU

Norwegian University of
Science and Technology

Properties of the closed testing procedure

With applications in pairwise comparisons and model selection

Fredrik Nevjen

Applied Physics and Mathematics
Submission date: June 2019
Supervisor: Øyvind Bakke

Norwegian University of Science and Technology
Department of Mathematical Sciences

Abstract

In this thesis we discuss multiple hypothesis testing procedures and their properties in general, and the closed testing procedure introduced by Marcus et al. (1976) in particular. Various closed testing procedures were used to maintain familywise error rate (FWER) control for multiple pairwise comparisons of means. The specific case comparing three group means was explored, where the closed testing procedure provides FWER control with very little computational cost added. Simulation results using generated data show that the F-test commonly used in a one-way analysis of variance gives a powerful closed testing procedure in this scenario, confirming earlier results by Shaffer (1981).

Goeman et al. (2011) presented a way to use closed testing procedures for the purpose of making confidence statements about the false discovery proportion (FDP). This method was applied for the purpose of model selection in multiple linear regression, and was compared to conventional methods such as lasso regression and best subset selection based on the Akaike information criterion (AIC). False discovery rate control with the Benjamini & Hochberg procedure (Benjamini and Hochberg, 1995) was also tested. Simulation results using generated data from various randomly constructed linear models show that the performance of the multiple testing procedures was similar to that of conventional methods in many cases, and generally better with respect to identifying relevant covariates. The FDP based method appeared somewhat strict when it came to making predictions on unseen data, while the Benjamini & Hochberg procedure was comparable to conventional methods for this purpose.

I denne oppgaven diskuterer vi metoder for multippel hypotesetesting generelt, og fokuserer spesifikt på lukket testing, introdusert av Marcus et al. (1976). Ulike lukkede testmetoder ble brukt for å kontrollere *familywise error rate* (FWER) ved parvis sammenligning av forventningsverdier. Spesialtilfellet med sammenligning av tre grupper ble undersøkt, hvor lukket testing tilføyer svært lite ekstra beregningstid. Simuleringresultater basert på genererte data viser at F-testen fra en-veis variansanalyse gir en sterk lukket testmetode i dette tilfellet, noe som bekrefter tidligere resultater av Shaffer (1981).

Goeman et al. (2011) presenterte en metode som bruker lukket testing for å lage konfidensutsagn om *false discovery proportion* (FDP). Vi har anvendt denne metoden for å utføre modellseleksjon i multippel lineær regresjon, og sammenlignet den med konvensjonelle metoder som lassoregresjon og modellseleksjon basert på AIC, samt Benjamini & Hochbergs metode (Benjamini and Hochberg, 1995) for kontroll av *false discovery rate*. Resultater fra simuleringer med data fra ulike, tifeldig genererte lineære modeller tyder på at metodene basert på hypotesetesting var sammenlignbare med konvensjonelle metoder i mange tilfeller, og generelt bedre til å identifisere kovariater som påvirker responsvariabelen. Den FDP-baserte metoden synes allikevel å være for streng i forhold til å gjøre gode prediksjoner på usett data, mens Benjamini & Hochbergs metode ga like gode resultater som konvensjonelle metoder på dette området.

Preface

This master's thesis concludes my so far five years at the Norwegian University of Science and Technology (NTNU), and was carried out during the spring semester of 2019. The thesis is an extension of my master project, which also revolved around the closed testing procedure. Learning about multiple testing procedures in greater detail than what is covered by the university courses has been a joy, and the knowledge I have gained is definitely a benefit to an aspiring statistician.

I am grateful to my supervisor, Øyvind Bakke, who generally let me do what I wanted to do, and provided steady guidance when I needed it. He might be strict when it comes to details in latex, but the thesis looks better for it.

Trondheim, 04-06-2019

Fredrik Nevjen

Table of Contents

List of Tables

List of Figures

Abbreviations

FWER	=	Familywise error rate
FDP	=	False discovery proportion
FDR	=	False discovery rate
B&H	=	The Benjamini & Hochberg procedure
RSS	=	Residual sum of squares
MSE	=	Mean square error
AIC	=	Akaike information criterion
CV	=	Cross-validation
LOOCV	=	Leave-one-out cross-validation
ANOVA	=	Analysis of variance

Chapter 1

Introduction

Hypothesis testing is an important part of scientific research, as scientific discoveries are most commonly the results of rejected statistical hypotheses. It is therefore of great importance that the procedures used in hypothesis testing are mathematically and statistically sound, in order to have some control of potentially misleading scientific discoveries (Goeman and Solari, 2014).

A common occurrence is the testing of multiple hypotheses, which inflates the probability of committing false rejections. Our test methods must therefore be adjusted, resulting in multiple testing procedures. We discuss the control of the familywise error rate (FWER), the false discovery proportion (FDP), and the false discovery rate (FDR), and methods based on these criteria.

The closed testing procedure by Marcus et al. (1976) is most commonly known for its role in FWER control. It has played an important part in the development of FWER controlling methods, with specific examples being the development of methods by Holm (1979) and Hochberg (1988). These sequentially rejective methods are constructed as closed testing procedures, utilizing the well known Bonferroni method and the global test by Simes (1986), respectively. We define and discuss closed testing procedures. We further discuss their properties, explore options for how to construct them, and apply them to two different multiple hypothesis testing scenarios.

We discuss coherence and consonance, introduced by Gabriel (1969), which are properties a multiple testing procedure can have. The former is closely connected to the closed testing procedure, a connection which has been explored by Sonnemann (1988, 2008) and Finner (1988). Consonance was further explored by Romano et al. (2011) and is beneficial in methods that control the FWER.

The first application of closed testing procedures that we discuss is the pairwise comparisons of means. When multiple comparisons of means are made, the closed testing procedure is simplified, leading to strong and simple procedures. This is especially the case when three group means are compared. We present results from a study similar to that of Shaffer (1981). The goal was to investigate how well various closed testing procedures would perform in this scenario.

Goeman et al. (2011) also discussed consonance, and have presented an FDP based method for multiple testing using closed testing procedures. Their method takes advantage of the dissonant rejections that can occur if the closed testing procedure is not consonant. The information gained from these rejections is unused in FWER control, but still serves a purpose in FDP based methods.

We present results from a study of how well the FDP based method performs for the purpose of model selection in multiple linear regression, when compared to conventional methods of lasso regression and best subset selection with the Akaike information criterion (AIC). We additionally explore how useful the well known Benjamini & Hochberg procedure for FDR control is for the same purpose (Benjamini and Hochberg, 1995).

Chapter 2

Theory

This section contains basic theory regarding multiple hypothesis testing, specifically regarding the familywise error rate, false discovery proportion and false discovery rate, and the closed testing procedure. It also contains descriptions of properties a multiple testing procedure may have, namely coherence and consonance, and some results regarding these properties.

We define and discuss the closed testing procedure, and discuss a method proposed by Goeman et al. (2011) for constructing confidence sets for the false discovery proportion, methods for pairwise comparisons of means, and model selection methods in multiple linear regression. Finally we discuss various ways to test intersections of hypotheses, which we need to construct closed testing procedures.

2.1 Multiple testing

Let X be data from some distribution P_θ, where the parameter of interest, θ (potentially a vector), lies in some parameter space Ω. We consider hypotheses of the form $H\colon \theta \in \omega$, where $\omega \subset \Omega$ is some subset of the parameter space. We say that a hypothesis H is true if $\theta \in \omega$. Typical examples of hypotheses are $H\colon \theta = \theta_0$ and $H\colon \theta_1 = \theta_2$, corresponding to $\omega = \{\theta_0\}$ and $\omega = \{\theta \mid \theta_1 = \theta_2\}$, respectively, where θ_1 and θ_2 are components of the vector of parameters.

We consider a set or family $\mathcal{H} = \{H_1, H_2, \ldots, H_m\}$ of hypotheses of interest, that are to be tested simultaneously, where $H_i\colon \theta \in \omega_i \subset \Omega$ for $i \in \{1, 2, \ldots, m\} = M$. For a nonempty $I \subset M$ we use the notation H_I for the hypothesis $H_I\colon \theta \in \omega_I = \bigcap_{i \in I} \omega_i$. We somewhat misleadingly call H_I an intersection of hypotheses. Note that $H_{\{i\}} = H_i$ for all $i \in M$.

If $H_i \in \mathcal{H}$ for all $i \in I \subset M$ implies $H_I \in \mathcal{H}$ we say that $\mathcal{H}$ is closed (under intersection). Let $\overline{\mathcal{H}}$ denote the closure of $\mathcal{H}$, i.e.

$$\overline{\mathcal{H}} = \{H_I \mid I \neq \emptyset,\ I \subset M\}\,.$$

Thus a family $\mathcal{H}$ of hypotheses is closed if $\mathcal{H} = \overline{\mathcal{H}}$.

	Hypothesis is		
Hypothesis is	true	false	total
rejected	V	U	R
not rejected	$m_0 - V$	$m_1 - U$	$m - R$
total	m_0	m_1	m

Table 2.1: Table for multiple hypothesis testing, indicating the number of hypotheses involved in specific scenarios. R and m are known values, the rest are unknown.

Observe that a hypothesis H_i implies another H_j if $\omega_i \subset \omega_j$. In this case we say that H_j is a *component* of H_i, and a *proper* component if additionally $H_j \neq H_i$. We call a hypothesis H_i in a family $\mathcal{H}$ *elementary* if it implies no other hypothesis in $\mathcal{H}$, i.e. it has no proper components. If $\mathcal{H}$ is a family consisting only of (at least two) elementary hypotheses, it is easy to see that it is not closed.

If the family of hypotheses of interest is not closed, considering also the rest of its closure might provide useful information for testing. The additional hypotheses can be used in construction of methods that control the familywise error rate, and they have some more direct use in methods based on the false discovery proportion.

To obtain a framework for discussion of multiple testing procedures, we consider the different outcomes for a total of m hypothesis tests. Table 2.1 shows an overview of the number of hypotheses that are true and false, and the number of hypotheses that are or are not rejected. The number of true and false hypotheses are m_0 and m_1, respectively. V is the number of type I errors (rejections of true hypotheses) made and U is the number of true positives, adding up to R, the total number of rejected hypotheses. The number of true negatives is $m_0 - V$, and $m_1 - U$ is the number of type II errors (failures to reject false hypotheses) made. R and m are known, and V, U, m_0 and m_1 are unknown.

2.1.1 Familywise error rate

The familywise error rate (FWER) of a multiple testing procedure is the probability that at least one of the true hypotheses is rejected,

$$\text{FWER} = P(V > 0).$$

In other words, the FWER is the probability that at least one type I error is made. A multiple hypothesis testing procedure that guarantees that the FWER is at or below a threshold α is said to control the FWER at level α. Other types of control can be used when performing multiple hypothesis testing, but FWER plays a particularly important part as the main form of control used in confirmatory research.

If the procedure controls the FWER at level α only in the case where all hypotheses are true, it is said to control the FWER *weakly*. If the procedure controls the FWER at level α for any subset of the hypotheses, regardless of how many that are true, it is said to control the FWER *strongly*.

The most commonly known method for FWER control is that of Bonferroni, for which each hypothesis is tested at the adjusted level α/m, in the case where there are m hypotheses to test (Goeman and Solari, 2014).

2.1.2 False discovery proportion

The false discovery proportion (FDP) is the proportion of falsely rejected hypotheses, not to be confused with the false discovery rate. The FDP is defined as

$$\text{FDP} = \begin{cases} \frac{V}{R}, & \text{if } V > 0 \\ 0, & \text{otherwise.} \end{cases} \tag{2.1}$$

Methods based on the FDP aim to create confidence sets, confidence intervals or point estimates for the FDP. Methods that seek to estimate or create confidence statements for V, the number of true hypotheses in a given subset of hypotheses, are naturally equivalent to FDP based methods. In particular, the special case of estimating π_0, the number of true hypotheses in the complete set of hypotheses, has been explored by several researchers (Goeman et al., 2011).

FDP based methods let the user select the set of rejected hypotheses, and then confidence statements for the FDP of this selected set can be made. R is thus known because it is chosen directly. This contrasts methods that control the FWER or FDR at some predetermined level, where the methods themselves select which hypotheses to reject.

FWER control can be considered quite strict, which is beneficial in confirmatory research. FDP based methods, on the other hand, grant the user a lot more freedom, which is highly useful in exploratory research. The purpose of FDP based methods is not to produce final results, but to explore which hypotheses to look further into. This is particularly useful in for example genomics, as the initial number of hypotheses might be very large (Goeman and Solari, 2014).

2.1.3 False discovery rate

The false discovery rate (FDR) is the expected value of the FDP, and is thus defined as

$$\text{FDR} = \text{E}[\text{FDP}].$$

Similarly to how a procedure has FWER control at level α if it guarantees that the FWER is at most α, a procedure that has level α FDR control ensures that the FDR is at most α. In order words, it ensures that the expected value of the proportion of type I errors among all rejections is smaller than or equal to α.

The most common method for FDR control is the procedure by Benjamini and Hochberg (1995) (B&H). This is a step-up procedure with critical values $i\alpha/m,\ i = 1, 2, \ldots, m$. With $p_{(i)}$ meaning the ith smallest p-value, this means that the procedure finds the largest j such that $p_{(j)} \leq j\alpha/m$, and rejects all hypotheses corresponding to p-values $p_{(1)}, p_{(2)}, \ldots, p_{(j)}$. If no such j exists no hypotheses are rejected. The validity of the procedure is believed to be robust for the case with asymptotically normal, two-sided tests, which is what we will be mainly concerned with (Goeman and Solari, 2014).

2.2 The closed testing procedure

Closed testing was introduced by Marcus et al. (1976). We first describe the procedure in the context of a closed family of hypotheses $\mathcal{H} = \{H_i \mid i \in M\}$, with $M = \{1, 2, \ldots, m\}$.

A closed testing procedure controls the FWER at a predetermined level α for the hypotheses in $\mathcal{H}$. The rejection of hypotheses is discussed in two ways. The event that a hypothesis H_i is rejected by a level α test is denoted L_i, and the test is called a *local* test. The event that a hypothesis H_i is rejected by the closed testing procedure is denoted C_i. The procedure is then defined by

$$C_i = \bigcap_{\omega_j \subset \omega_i} L_j.$$

A hypothesis H_i is thus rejected by the closed testing procedure if all hypotheses that imply it (hypotheses of which H_i is a component) are rejected by the local tests.

Unless all hypotheses in $\mathcal{H}$ are false, there exists a unique true hypothesis in $\mathcal{H}$ such that all the other true hypotheses are components of it. To see this, let $T \subset M$ be the set of indices of true hypotheses in $\mathcal{H}$, and consider the hypothesis H_T. Since $\omega_T = \bigcap_{i \in T} \omega_i$, we must have $\omega_T \subset \omega_i$ for all true hypotheses H_i.

No true hypothesis is rejected by the closed testing procedure unless H_T is, and since the rejection of H_T by the closed testing procedure depends on a local level α test (as well as tests for any hypothesis $H_j \colon \theta \in \omega_j$ with $\omega_j \subset \omega_T$), this occurs with probability at most α. Thus the closed testing procedure ensures level α FWER control.

Consider now the case with a family of elementary hypotheses $\mathcal{H} = \{H_i \mid i \in M\}$. We know that this family is not closed, and that for any $i, j \in M,\ i \neq j$, we have $\omega_i \not\subset \omega_j$. Note that the set of true hypotheses in $\mathcal{H}$ is a subset of the set of true hypotheses in $\overline{\mathcal{H}}$, and thus at least one true hypothesis in $\mathcal{H}$ is rejected only if at least one true hypothesis in $\overline{\mathcal{H}}$ is rejected. Therefore FWER control of the hypotheses in $\overline{\mathcal{H}}$ implies FWER control of the hypotheses in $\mathcal{H}$, and we obtain FWER control of the hypotheses in $\mathcal{H}$ by applying the closed testing procedure on the hypotheses in $\overline{\mathcal{H}}$.

We denote the event that a hypothesis H_I is rejected by a level α test by L_I, and the event that a hypothesis H_I is rejected by the closed testing procedure by C_I. The definition of the closed testing procedure above now leads to the following

$$C_I = \bigcap_{J \supset I} L_I.$$

Regarding rejection of the elementary hypotheses we obtain

$$C_i = \bigcap_{J \ni i} L_J.$$

An elementary hypothesis H_i is rejected by the closed testing procedure if all intersections of hypotheses that have H_i as a component can be rejected by local level α tests.

2.3 Coherence

Gabriel (1969) discussed properties of a testing procedure. He argued that when a procedure rejects a hypothesis, it should also reject any hypothesis implying it. This property is called *coherence*. The rejection of a hypothesis means that we conclude that it is false. Thus, if a hypothesis has a component that we have rejected, we should conclude that the former is false as well.

A method is *coherent* if a hypothesis H_i is rejected only if H_j is rejected for every j such that $\omega_j \subset \omega_i$. The closed testing procedure is thus coherent by definition. If all the hypotheses in $\mathcal{H}$ are elementary, a method is coherent if a hypothesis $H_I \in \overline{\mathcal{H}}$ is rejected only if all hypotheses $H_J \in \overline{\mathcal{H}}$ such that $J \supset I$, are also rejected.

Romano et al. (2011) described the results by Sonnemann (1988, 2008) and Finner (1988), who showed both that any coherent multiple testing procedure is equivalent to a closed testing procedure, and that any incoherent procedure can be improved by a coherent one. This further underlines the role of coherence in regards to closed testing, and adds to the assertion by Gabriel (1969) that the property is beneficial for a testing procedure to have.

2.3.1 Any coherent multiple testing procedure is equivalent to a closed testing procedure

We now summarize briefly how any coherent multiple testing procedure can be expressed as a closed testing procedure. Let $\mathcal{H}$ be the family of hypotheses of interest, and $\mathcal{R} \subset \mathcal{H}$ be the set of hypotheses that the coherent procedure rejects. Now we express the procedure as a closed procedure as follows. The local test for $H_i \in \mathcal{H}$ rejects H_i if there exists any $H_j \in \mathcal{R}$ such that $\omega_j \supset \omega_i$. Thus the local test for H_i rejects it if any of its components (which could be the hypothesis itself) was rejected by the original procedure.

Now we can observe that if $H_i \in \mathcal{R}$, the local tests reject all $H_j \in \mathcal{H}$ such that $\omega_j \subset \omega_i$, which means that the closed testing procedure also rejects H_i, by the definition of the closed testing procedure.

If the closed testing procedure rejects $H_i \in \mathcal{H}$, it must be the case that the local test for H_i rejects it. This means that H_i has a component $H_j \in \mathcal{R}$ that was rejected by the coherent procedure. By coherence we therefore also have $H_i \in \mathcal{R}$, and we conclude that the original method and the closed testing procedure reject the exact same hypotheses.

2.3.2 Coherentization

Romano et al. (2011) described a method to construct a coherent multiple testing procedure that rejects the same hypotheses and possibly more than an incoherent one, while still maintaining FWER control at the same level. The method is appropriately named *coherentization*.

Suppose an incoherent multiple testing procedure controls the FWER for a closed family $\mathcal{H} = \{H_i \mid i \in M\}$ of hypotheses at level α, and rejects H_i when we observe data $X \in R_i$, for $i \in M$. R_i is called the *critical region* of H_i. The coherentized procedure is constructed by rejecting H_i when we observe data $X \in R_i'$, for $i \in M$, where

$$R_i' = \bigcup_{j\,:\,\omega_j \supset \omega_i} R_j.$$

Since $\omega_i \supset \omega_i$, we have $R_i \subset R_i'$, and so the coherentized procedure rejects at least as much as the incoherent one does. We add the rejections of all H_i for which there exists H_j with $\omega_j \supset \omega_i$ and where H_j is rejected by the incoherent procedure. This means that if a hypothesis H_i has a component that is rejected by the incoherent procedure, H_i will be rejected by the coherentized procedure. Thus the resulting procedure is coherent, since if a hypothesis is rejected, so will all that has it as a component.

If the coherentization adds the rejection of a true hypothesis H_i, it must have been the case that a component H_j of H_i, which thus means H_j is true, was already rejected by the incoherent procedure. Therefore the coherentization adds no rejection of a true hypothesis unless a true hypothesis was already rejected by the incoherent procedure. Thus the probability of rejecting at least one true hypothesis is not changed, and the FWER control is maintained at the same level.

2.4 Consonance

Another property discussed by Gabriel (1969) is *consonance*. A method is consonant if the rejection of a hypothesis H_i implies the rejection of at least one of its proper components, if such a hypothesis exists.

We argue that this general definition has some issues, for instance in a scenario with hypotheses $H_1 : \theta \in (0,2)$ and $H_2 : \theta \in (0,1)$. Note that H_2 is not elementary. In this case, rejecting H_2 and not rejecting H_1 causes a dissonance by the definition of Gabriel (1969), and we call it a *dissonant* rejection. However, it could be the case that $\theta \in (1,2)$, and so it would be nonsensical to reject H_1 for the sole reason that H_2 is rejected.

We will instead focus on the case where we test the hypotheses from a family $\mathcal{H}$ of elementary hypotheses, along with its closure $\overline{\mathcal{H}}$. We say that a closed testing procedure is consonant if the rejection of H_I implies the rejection of at least one of its elementary components, i.e. H_i for some $i \in I$. This is the definition used by Romano et al. (2011) and Goeman et al. (2011). If an intersection of hypotheses is false, at least one of the elementary components must also be false, which makes this definition of consonance seem like a natural property for a testing procedure to have.

In the context of using closed testing procedures to control the FWER, the rejection of an intersection of hypotheses, H_I, without the rejection of at least one of the involved elementary hypotheses, can be considered a wasted rejection (Goeman et al., 2011). Without it, the set of rejected elementary hypotheses remains the same.

Unlike the case for coherence, not all closed testing procedures are consonant. Marcus et al. (1976) discussed both consonant and non-consonant procedures. Goeman et al. (2011) discussed applications in exploratory research where the information gained from a dissonant rejection is used, in the context of creating confidence sets for the number of true hypotheses in any chosen subset of the elementary hypotheses.

Romano et al. (2011) showed results regarding consonance that were similar to the findings of Sonnemann and Finner regarding coherence. Specifically they showed that any non-consonant procedure can be replaced by a consonant one that rejects exactly the same elementary hypotheses, and thus still controls the FWER at the same level. They also showed that in specific cases, the procedure can even be improved to reject false hypotheses with greater probability. Thus consonant methods are preferable when only the rejections of elementary hypotheses are of interest, for example when the purpose is FWER control.

2.4.1 Consonantization

Romano et al. (2011) described a method of *consonatization*, which we summarize here. The method creates a consonant closed testing procedure from a non-consonant one, without altering which elementary hypotheses the procedure rejects, thus maintaining the same level of FWER control.

Suppose a non-consonant closed testing procedure controls the FWER for a family of elementary hypotheses $\mathcal{H} = \{H_i \mid i \in M\}$ at level α. Suppose further that the procedure rejects H_I when we observe data $X \in R_I$, for $I \subset M$. The consonantized procedure is constructed by rejecting H_I when we observe data $X \in R'_I$, for $I \subset M$, where

$$R'_I = \bigcup_{i \in I} \bigcap_{J \subset M, i \in J} R_J.$$

If any hypothesis H_i is rejected by the original procedure, we must have $X \in R_J$ for all $J \subset M$ such that $i \in J$, since the method is coherent. Thus $X \in \bigcap_{J \subset M, i \in J} R_J$, which means $X \in R'_I$ for all $I \subset M$ such that $i \in I$, including $I = \{i\}$. The new procedure thus rejects H_i as well.

If a hypothesis H_i is rejected by the new procedure, we must have $X \in R'_{\{i\}} = \bigcap_{J \subset M, i \in J} R_J$. Specifically we thus have $X \in R_{\{i\}}$, which means that the original procedure rejects H_i. Thus the two methods reach the exact same conclusions regarding the elementary hypotheses.

For any $I \subset M$ we have $R'_I = \bigcup_{i \in I} R'_{\{i\}}$, so that $X \in R'_I$ implies $X \in R'_{\{i\}}$ for at least one $i \in I$. Thus an intersection of hypotheses is rejected by the new procedure only if at least one of its elementary components is rejected, which makes the new procedure consonant.

Consonantization as described above does not impact which elementary hypotheses are rejected. Romano et al. (2011) did however describe how the method can be improved so that the consonant procedure created maintains the same level of FWER control, yet has increased power. The consonantization removes points from the critical regions of hypotheses with non-consonant local tests. This decreases the level of these local tests, which means other points may be added to the reduced critical regions without increasing the levels of the tests past their initial value. A simple, two-dimensional example was presented by Romano et al. (2011).

2.5 Confidence sets for number of false discoveries

Goeman et al. (2011) presented an FDP based method that takes advantage of the information gained from dissonant rejections caused by a non-consonant closed testing procedure. The resulting method grants the researcher a high degree of freedom in which hypotheses to investigate, as the method produces simultaneous confidence statements for the FDP of all possible subsets of the hypotheses. The setting is exploratory research, and the goal is to reduce a large number of hypotheses to a smaller number of promising hypotheses to further investigate with stricter testing procedures.

A set $\mathcal{H} = \{H_i \mid i \in M = \{1, 2, \ldots, m\}\}$ of elementary hypotheses is considered, along with its closure $\overline{\mathcal{H}}$. A closed testing procedure is applied at some level α. Let $\mathcal{U}$ denote the set of nonempty subsets $I \subset M$ for which H_I is rejected by a local test, and $\mathcal{X}$ denote the set of nonempty subsets $J \subset M$ for which H_J is rejected by the closed testing procedure.

A subset of the elementary hypotheses $\mathcal{R} \subset \mathcal{H}$ is selected by the user. Rather than creating confidence sets for the FDP of this set, Goeman et al. (2011) constructed confidence sets for $V(\mathcal{R})$, the number of true hypotheses in $\mathcal{R}$. Dividing by R, the number of hypotheses in $\mathcal{R}$, will result in a confidence set for the FDP.

Let $t_\alpha(\mathcal{R}) = \max\{|I| \mid \{H_i \mid i \in I\} \subset \mathcal{R},\ I \notin \mathcal{X}\}$, meaning $t_\alpha(\mathcal{R})$ is the size of the largest subset of $\mathcal{R}$ for which the intersection is not rejected by the closed testing procedure. If all such intersections are rejected, we set $t_\alpha(\mathcal{R}) = 0$. A $1 - \alpha$-confidence set for $V(\mathcal{R})$ is then

$$\{0, 1, \ldots, t_\alpha(\mathcal{R})\},$$

which means that with probability at least $1 - \alpha$ we have at most $t_\alpha(\mathcal{R})$ true hypotheses in $\mathcal{R}$, or that rejecting the hypotheses in $\mathcal{R}$ leads to at most $t_\alpha(\mathcal{R})$ false discoveries.

The reason behind the coverage probability ties into the proof that the closed testing procedure controls the FWER. The probability that no true hypothesis is rejected by the closed testing procedure is at least $1 - \alpha$. In the case that no true hypothesis is rejected, the number of true hypotheses in $\mathcal{R}$ can not be larger than $t_\alpha(\mathcal{R})$. If there were more than $t_\alpha(\mathcal{R})$ true hypotheses in $\mathcal{R}$, the intersection of these hypotheses would not have been rejected, which leads to a contradiction, since $t_\alpha(\mathcal{R})$ was the size of the largest subset of $\mathcal{R}$ for which the intersection is not rejected.

The confidence sets for all $\mathcal{R} \subset \mathcal{H}$ depend on the same event, that no true hypothesis is rejected by the closed testing procedure. Thus all of these confidence sets are simultaneous. This means that the user is free to consider the confidence sets for any subset, without compromising the coverage probability (Goeman et al., 2011).

2.5.1 Example of construction of confidence sets

Consider an example where we are interested in the set $\mathcal{H} = \{H_1, H_2, H_3, H_4\}$ of elementary hypotheses. Suppose all hypotheses with H_1 as a component, as well as the hypotheses $H_{\{2,3,4\}}$ and $H_{\{2,3\}}$ are rejected by the closed testing procedure, and the rest are not. See Figure 2.1 for an illustration.

$H_{\{2,4\}}$ and $H_{\{3,4\}}$ are the intersections involving the largest number of hypotheses in $\mathcal{H}$ that are not rejected. Thus $t_\alpha(\mathcal{H}) = 2$, and we conclude that $\{0, 1, 2\}$ is a $1 - \alpha$-confidence set for the number of true hypotheses in $\mathcal{H}$. Thus we observe that there are likely at least two false hypotheses among our elementary hypotheses, even though H_1 was the only elementary hypothesis rejected by the closed testing procedure.

Similarly, if we consider $\mathcal{R} = \{H_2, H_3\}$, H_2 or H_3 is the intersection involving the largest number (only one) of hypotheses in $\mathcal{R}$ that are not rejected. Thus $t_\alpha(\mathcal{R}) = 1$, and $\{0, 1\}$ is a $1 - \alpha$-confidence set for the number of true hypotheses in $\mathcal{R}$. This tells us that the second

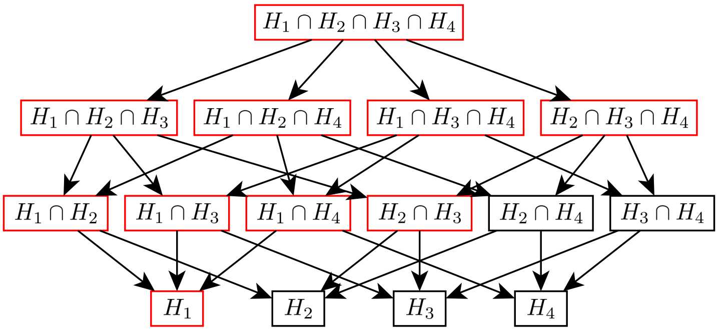

Figure 2.1: Intersections of elementary hypotheses H_1, H_2, H_3 and H_4. Hypotheses framed in red are rejected by a closed testing procedure. The rejections of $H_{\{2,3,4\}}$ and $H_{\{2,3\}}$ are dissonant rejections, since none of their elementary components are rejected. Note that the hypothesis H_I here is denoted $\bigcap_{i \in I} H_i$.

false hypothesis, the first being H_1, is likely either H_2 or H_3, and that investigating these hypotheses further could be useful.

Note that the information that lead to the conclusions in the previous paragraph is gained from the rejection of $H_{\{2,3\}}$, a dissonant rejection. A consonant method resulting in the same set of rejected elementary hypotheses would not have rejected $H_{\{2,3\}}$ (or $H_{\{2,3,4\}}$), which would result in a larger confidence set for the number of true hypotheses in $\mathcal{R}$. In fact, we would obtain $t_\alpha(\mathcal{R}) = 2$, resulting in the trivial confidence set $\{0, 1, 2\}$, and a complete loss of the information originally gained from the dissonant rejection.

2.5.2 Defining rejections

The *defining rejections* of the closed testing procedure are the rejected hypotheses $H_I \in \overline{\mathcal{H}}$ such that no H_J with $J \neq \emptyset, \;\; J \subset I$ is rejected (Goeman et al., 2011). In other words, a defining rejection is a rejected hypothesis with no rejected proper components. As an example, the defining rejections in Figure 2.1 are H_1 and $H_{\{2,3\}}$, since these are the only rejections with no rejections further down in the hierarchy.

If no true hypothesis is rejected, any rejected hypothesis must have at least one false elementary component. Since the defining rejections have no rejected proper components, the elementary hypotheses involved in defining rejections are the smallest subsets of elementary hypotheses of which at least one is false. For our example H_1 and $H_{\{2,3\}}$ are defining rejections, and so $\{H_1\}$ and $\{H_2, H_3\}$ are the smallest subsets that must contain at least one false hypothesis, conditioned on the event that no true hypothesis is rejected. This also means that at most all but one of the elementary components of a defining rejection are true.

Thus, if H_I is a defining rejection, and $\mathcal{R} = \{H_i \mid i \in I\}$, we have $t_\alpha(\mathcal{R}) = |I| - 1 = |\mathcal{R}| - 1$. If no incorrect rejections have been made, $\mathcal{R}$ contains at most $|I| - 1$ true hypotheses, and at least one false hypothesis. Note that the defining rejections with only one elementary component are the elementary hypotheses rejected by the closed testing procedure, and thus if only these are selected to be rejected we actually maintain FWER control at level α.

2.6 Pairwise comparisons of means

A common study is that of pairwise comparisons of means. This often appears in one-factor problems, such as a study of the effects of different treatments (Walpole et al., 2016, pp. 527, 543–544). An example can be for example testing to see if the expected time to finish a race is different for people applying different running techniques.

The random variables $Y_1, Y_2, \dots, Y_m$ corresponding to some response variable for m different groups, are investigated. The goal of the study is to determine whether or not the respective means, $\mu_1, \mu_2, \dots, \mu_m$, are equal, and which means that are. The parameter of interest is thus $\theta = (\mu_1, \mu_2, \dots, \mu_m)$, and the elementary hypotheses are

$$H_{ij} \colon \mu_i = \mu_j,$$

where $1 \leq i < j \leq m$, with alternative hypotheses

$$H'_{ij} \colon \mu_i \neq \mu_j.$$

Note that there are $m(m-1)/2$ elementary hypotheses. The *global* hypothesis is the intersection of all the elementary hypotheses,

$$H_{12\dots m} \colon \mu_1 = \mu_2 = \dots = \mu_m,$$

with alternative hypothesis

$$H'_{12\dots m} \colon \mu_i \neq \mu_j,$$

for some i and j.

In our discussion and later simulations, we assume $Y_i,\ i = 1, 2, \dots, m$, to be independent and to come from normal distributions with the same variance σ^2. For n independent realizations of each the m variables, we thus have $Y_{ij} \sim N(\mu_i, \sigma^2)$ for $i = 1, 2, \dots, m$, $j = 1, 2, \dots, n$.

The group sample mean, pooled sample mean, group sample variance and pooled sample variance are thus given by

$$\begin{aligned} \bar{Y}_i &= \frac{1}{n} \sum_{j=1}^{n} Y_{ij}, & \bar{Y} &= \frac{1}{m} \sum_{i=1}^{n} \bar{Y}_i, \\ S_i^2 &= \frac{1}{n-1} \sum_{j=1}^{n} \left(Y_{ij} - \bar{Y}_i\right)^2, & S_p^2 &= \frac{1}{m} \sum_{i=1}^{m} S_i^2, \end{aligned} \tag{2.2}$$

respectively (Casella and Berger, 2002, p. 528).

2.6.1 Testing the elementary hypotheses

The elementary hypothesis $H_{ij}: \mu_i = \mu_j$ is commonly tested with the two-sample t-test. The test statistic for this test is

$$T'_{ij} = \frac{\bar{Y}_i - \bar{Y}_j}{\sqrt{(S_i^2 + S_j^2)/n}},$$

which has a t-distribution with $2(n-1)$ degrees of freedom (Casella and Berger, 2002, p. 409). We will focus on an alternative test statistic, namely that of the pooled t-test

$$T_{ij} = \frac{\bar{Y}_i - \bar{Y}_j}{\sqrt{2S_p^2/n}}, \tag{2.3}$$

which has a t-distribution with $m(n-1)$ degrees of freedom (Casella and Berger, 2002, p. 529). Here the additional information from the samples of Y_k, $k \neq i, j$ is also used in the estimation of σ^2.

2.6.2 Familywise error rate control in pairwise comparisons

Since we are testing multiple elementary hypotheses, we should perform some correction. We consider how to achieve FWER control at level α. A simple approach is to use Bonferroni's method, and test each elementary hypothesis at level $\alpha/(m(m-1)/2)$.

Tukey's procedure simultaneously tests all pairwise comparisons while maintaining FWER control at a desired level α (Walpole et al., 2016, p. 546). The test is based on the studentized range distribution, which is the distribution of

$$Q = \frac{\bar{Y}_{\max} - \bar{Y}_{\min}}{\sqrt{S_p^2/n}}, \tag{2.4}$$

where $\bar{Y}_{\max}$ is the largest observed group mean and $\bar{Y}_{\min}$ is the smallest.

The test statistic used by Tukey's procedure for the elementary hypothesis $H_{ij}: \mu_i = \mu_j$ is

$$Q_{ij} = \frac{|\bar{Y}_i - \bar{Y}_j|}{\sqrt{S_p^2/n}},$$

which is tested using a studentized range distribution with m groups and $m(n-1)$ degrees of freedom (Walpole et al., 2016, p. 546). Note that this test statistic is similar to that of the pooled t-test, as $Q_{ij} = \sqrt{2}|T_{ij}|$. Each difference in observed means is tested by Tukey's procedure as though they had the distribution of the largest, and thus the tests for the non-largest observed differences will be conservative.

The test for the elementary hypotheses in Tukey's procedure is strictly more conservative than the pooled t-test when used only to test a single, arbitrary, elementary hypothesis, unless there are only two groups in total. This is illustrated for the case $m = 3$, $n = 30$ in Figure 2.2, which in black shows the density function of the studentized range distribution,

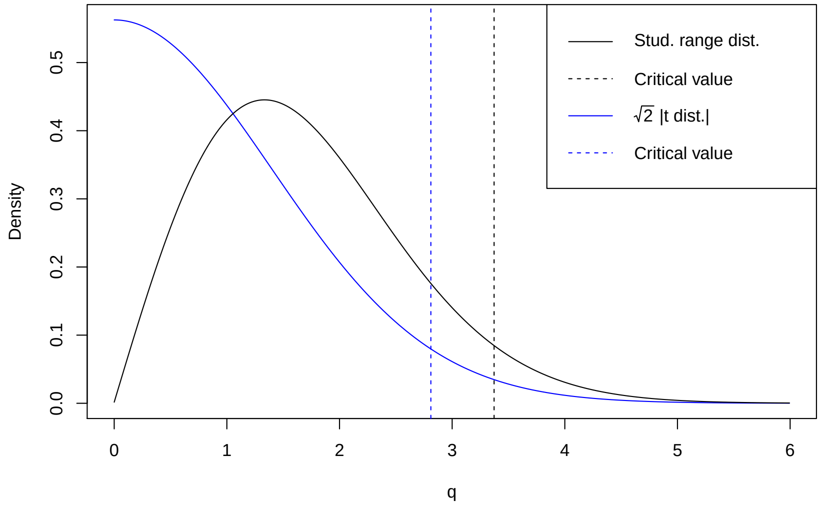

Figure 2.2: The density functions of Q and $\sqrt{2}|T|$, both with $3(30-1) = 87$ degrees of freedom, with three groups. The 0.95-quantile for each distribution is marked with a dotted line.

Q, and in blue the transformed t-distribution, $\sqrt{2}|T|$. The corresponding $1-\alpha$-quantiles are marked with dotted lines for $\alpha = 0.05$.

We see that if an observed test statistic q leads to a rejection by Tukey's procedure, meaning it is larger than the $1-\alpha$-quantile of the studentized range distribution, marked by the black dotted line, it must also be the case that it is rejected by the t-test, since its corresponding $1-\alpha$-quantile, marked with a blue dotted line, is smaller. The reason is that the studentized range distribution is based on the largest difference of means, and the t-distribution is based on an arbitrary difference, which explains that the probability mass for the former is shifted towards larger values compared to the latter.

The same reason that makes the tests in Tukey's procedure conservative also causes it to achieve level α FWER control, however. If all elementary hypotheses are true, no true hypothesis is rejected unless the one corresponding to the largest observed difference is, and the test for this has level α. If only a subset of the groups have the same mean, no true hypothesis is rejected unless we reject the one corresponding the the largest observed difference in mean between two of these groups. This observed difference is tested against a critical value which assumes that the number of groups with equal mean is larger, which naturally must be larger than the critical value corresponding to the actual number of groups with equal mean. Thus the probability of committing a type I error is always smaller than or equal to α.

Another alternative to achieve FWER control is to use a closed testing procedure. In pairwise comparisons this leads to some interesting simplifications, since some intersections of hypotheses coincide. Thus not all $2^{m(m-1)/2} - 1$ intersections have to be tested, reducing the computational cost. We consider the case $m = 3$ in particular in the next chapter and in our experiments.

2.7 Model selection in multiple linear regression

In multiple linear regression the goal is to model the relationship between a response variable Y and multiple covariates $x_1, x_2, \dots, x_m$. In the simplest case it is assumed that this relationship takes the form $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_m x_m + \epsilon$, where the βs are constant coefficients and ϵ is normally distributed noise with mean 0 and unknown variance σ^2. For n independent data points we thus have $Y_j \sim N(\beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \dots + \beta_m x_{mj}, \sigma^2)$, for $j = 1, 2, \dots, n$.

The coefficient estimates, $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \dots, \hat{\beta}_m$, are typically chosen by minimizing the residual sum of squares (RSS), defined as

$$\text{RSS} = \sum_{j=1}^{n} \left(y_j - \hat{y}_j\right)^2 = \sum_{j=1}^{n} \left(y_j - \left(\hat{\beta}_0 + \sum_{i=1}^{m} \hat{\beta}_i x_{ij}\right)\right)^2,$$

resulting in the *least squares coefficient estimates*. When comparing the predictions of models it is normal to report the mean of the RSS, the mean square error (MSE) (James et al., 2013, pp. 29, 62, 72).

An important part of regression is to determine which covariates that affect the response. The relevance of a particular covariate x_i is investigated through a hypothesis test of $H_i\colon \beta_i = 0$ versus its alternative $H_i'\colon \beta_i \neq 0$, and we call x_i a *significant* covariate if H_i is rejected.

The distribution of the least squares estimates for the coefficients is

$$\hat{\beta} \sim N(\beta, (X^T X)^{-1} \sigma^2),$$

where σ^2 is the unknown variance of ϵ and X is the design matrix, meaning row j of X is $(1, x_{1j}, x_{2j}, \dots, x_{mj})$, for observations $j = 1, 2, \dots, n$ (Hastie et al., 2001, p. 47). Thus a test statistic for the elementary hypothesis $H_i\colon \beta_i = 0$ is

$$T_i = \frac{\hat{\beta}_i}{\hat{\sigma}\sqrt{v_i}}, \tag{2.5}$$

where

$$\hat{\sigma}^2 = \frac{1}{n-m-1} \sum_{j=1}^{n} \left(y_j - \hat{y}_j\right)^2 = \frac{\text{RSS}}{n-m-1},$$

is the estimated variance of ϵ, and v_i is the ith diagonal element of $(X^T X)^{-1}$. T_i has a t-distribution with $n - m - 1$ degrees of freedom (Hastie et al., 2001, pp. 47–48).

Seldom will all the covariates be truly relevant for the response, and including the irrelevant ones in the model will add noise that increases the variance of its predictions. Thus it is beneficial to perform model selection to reduce the set of covariates in the model. Conventional methods for this are subset selection methods using some optimality criterion, and regularized regression (James et al., 2013, pp. 203–204).

The hypothesis tests for the significance of covariates are typically not used directly for model selection. F-tests can be used to compare the full model to a reduced one (described in the next section), but are not used extensively for model selection. Doing so would require some multiplicity correction, and we suspect that FWER control is too strict. Despite the fact that including irrelevant covariates in the model adds noise to the predictions, the exclusion of a relevant covariate may also have a large, negative impact on the model. Thus type II errors are also important to limit. We have explored if control milder than that of FWER has merit, specifically by applying the FDP based method by Goeman et al. (2011).

There is an issue with model selection in general when it comes to inference about the reduced model. The reported p-values for the covariates in a reduced model may not take the selection process into account, and may therefore be unreliable (Goeman et al., 2011). This is beyond the scope of what we explore here, although it is important to keep in mind when working with a reduced model.

2.7.1 Best subset selection

In best subset selection, all 2^m possible submodels of $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m + \epsilon$ are considered, and the one that optimizes some specified criterion is selected as the best model. We consider the Akaike information criterion (AIC), which for a model with k covariates is defined as

$$\mathrm{AIC} = -\frac{2}{n}\ell(\hat{\beta}) + \frac{2k}{n},$$

where $\ell(\hat{\beta})$ is the maximum log-likelihood for the model and n is the number of data points (Hastie et al., 2001, p. 231). For our linear regression model with normally distributed errors, this is equivalent to comparing

$$\mathrm{AIC}' = \frac{1}{n\hat{\sigma}^2}\left(\mathrm{RSS} + 2k\hat{\sigma}^2\right),$$

where $\hat{\sigma}^2$ is the estimated variance of ϵ, calculated by using the full model (James et al., 2013, pp. 211–212).

The AIC combines a measure of how well the model fits the data, the first term, with a penalty term for the complexity of the model. By selecting the model with the minimal AIC value, we thus end up with covariates that contribute to explain the response well, and exclude covariates that seem the least likely to affect the response.

2.7.2 Regularization

In regularized regression the coefficient estimates of the model are shrunk towards 0, in order to reduce the variance of the model's predictions. Lasso regression is a form of regularized regression, where instead of only minimizing the RSS, we restrict the coefficient space by the condition $\Sigma_{i=1}^{m}|\beta_i| \leq s$, where s is some tuning parameter. An equivalent formulation is to choose the coefficient estimates that minimize

$$\sum_{j=1}^{n}\left(y_j - \beta_0 - \sum_{i=1}^{m}\beta_i x_{ij}\right)^2 + \lambda\sum_{i=1}^{m}|\beta_i| = \text{RSS} + \lambda\sum_{i=1}^{m}|\beta_i|,$$

where λ is a tuning parameter (James et al., 2013, pp. 219–221).

If $\lambda = 0$, the minimization yields the regular least squares estimates, and if for $\lambda = \infty$ all estimated coefficients will be 0. For the first formulation with the restriction this corresponds to $s = \infty$ and $s = 0$, respectively.

The region defined by $\Sigma_{i=1}^{m}|\beta_i| \leq s$ has straight edges, and a consequence of this is that some coefficient estimates are forced to 0 for certain values of the tuning parameter. Figure 2.3 shows why this is the case. Because the edges of the region are straight, the contours of the error are likely to intersect the region at coefficient axes. Thus the corresponding coefficient estimates will be 0, and the method performs model selection in addition to the regularization.

The value of the tuning parameter is often chosen by cross-validation, where an estimate of the test MSE, the MSE the model would obtain when used on new data, is calculated for many values of λ. The data is partitioned, into what is called folds, and the data in each fold is treated as test data while the model is fitted on the remaining data. The average of the MSE values for each fold is then an estimate for how well the model fits new data, and the value of λ that minimizes this estimate is then used to fit the final model. If the number of folds is k the procedure is called k-fold cross-validation, and if it is equal to the number of data points, the procedure is called *leave-one-out* cross-validation (LOOCV) (James et al., 2013, pp. 176–182, 227).

2.8 Tests for intersections of hypotheses

In order to use a closed testing procedure, we need to have local tests for each intersection of hypotheses. There are many ways to select the local tests, where some depend only on the p-values for tests of the elementary hypotheses, while others depend on the joint probability distribution which the observed data comes from. A method that controls the FWER for a family $\mathcal{H}_\mathcal{I} = \{H_i \mid i \in I\}$ of hypotheses at level α can also be used to create a local level α test for H_I.

2.8.1 Constructing local tests using a procedure that controls the FWER

Suppose we have a multiple testing procedure for $\mathcal{H}_I$ that controls the FWER at level α. A level α test for H_I requires $P(L_I) \leq \alpha$ in the case that H_I is true, where L_I is the event that H_I is rejected.

To reject H_I whenever the multiple testing procedure with level α FWER control rejects H_i for at least one $i \in I$, is a level α test for H_I. To see this, let L_i be the event that H_i is rejected by the procedure, and note that in the case that H_I is true, the rejection of any elementary component of H_I is a type I error. Therefore

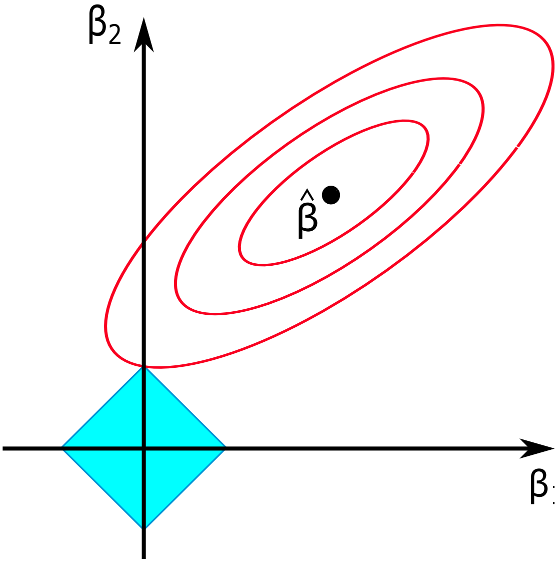

Figure 2.3: Two-dimensional example of lasso regression, with contours of the error as a function of the coefficients in red, the least squares estimate of β marked with a dot, and the restricted coefficient space in light blue. The figure is used with permission, and is made by Dag Johnsrud Kristiansen for the purpose of his own master's thesis, inspired by a similar figure by James et al. (2013, p. 222).

$$P(L_I) = P\left(\bigcup_{i\in I} L_i\right) = P(V > 0) = \text{FWER} \leq \alpha,$$

where V is the number of elementary components of H_I that are rejected. Thus any method that controls the FWER for a set of hypotheses can be used to construct a hypothesis test for the intersection of the same hypotheses.

2.8.2 Local tests for intersections based on the p-values from the elementary hypotheses

For the elementary hypotheses $H_1, H_2, \ldots, H_m$ with corresponding p-values $p_1, p_2, \ldots, p_m$, let $p_{(1)}, p_{(2)}, \ldots, p_{(m)}$ be the same p-values sorted in ascending order. For a non-empty subset $I \subset M$, let $p^I_{(1)}, p^I_{(2)}, \ldots, p^I_{(|I|)}$ be the sorted p-values for the hypotheses in $\mathcal{H}_I = \{H_i \mid i \in I\}$.

Bonferroni's method for FWER control can be used to create local test in the manner described above. For the hypothesis H_I, a level α test would thus be to reject H_I if, for some $i \in I$, H_i is rejected by the Bonferroni method at level α. H_i is rejected by the Bonferroni method if $p_i \leq \alpha/|I|$. Thus the resulting local test for H_I is to reject it whenever $p^I_{(1)} \leq \alpha/|I|$. This test is valid regardless of the distributions of the test statistics for the elementary hypotheses, as long as the tests of the elementary hypotheses themselves are valid.

The closed testing procedure obtained when using Bonferroni's method to construct local tests was found by Holm (1979). The procedure is a sequentially rejective multiple test procedure which rejects $H_{(j)}$ corresponding to $p_{(j)}$ for all j smaller than the smallest k such that $p_{(k)} > \alpha/(m-k+1)$, and all hypotheses if no such k exists. Note that using Holm's method to construct local tests for a closed testing procedure in the same manner also results in Holm's method, because it rejects at least one hypothesis in $\mathcal{H}_I$ in exactly the same case as Bonferroni's method, namely when $p^I_{(1)} \leq \alpha/|I|$.

Another test for an intersection of hypotheses based only on p-values is Simes' global test, introduced by Simes (1986). An intersection of hypotheses H_I can be rejected at level α by Simes' global test if for at least one $i \in I$ we have $p^I_{(i)} \leq i\alpha/|I|$. Simes' global test is valid when the test statistics are independent, but is also believed to be valid in other, specific cases; the two-sided t-tests are asymptotically normal, in which case it is believed that the validity of Simes' test is robust (Goeman and Solari, 2014), and simulations studies by Simes and others also indicate that Simes' global test is in valid for t-tests for pairwise comparisons (Shaffer, 1995).

Hommel (1988) considered the closed testing procedure obtained when using Simes' global test for the local tests. This procedure is slightly more complex than for example Holm's method, and rejects all H_j with $p_j \leq \alpha/k$, where $k = \max\{i \in M \mid p_{(m-i+k)} > k\alpha/i$ for $k = 1, 2, \ldots, i\}$. Hochberg (1988) presented a slightly weaker simplification of this procedure, which rejects $H_{(j)}$ corresponding to $p_{(j)}$ for all j smaller than or equal to the largest k such that $p_{(k)} \leq \alpha/(m-k+1)$, and no hypotheses if no such k exists.

2.8.3 Local tests for intersections based on the distribution of the data

We discuss how to test the intersections of elementary hypotheses from the pairwise comparisons and linear regression settings. In both cases, the elementary hypotheses are two-sided t-tests.

One-way analysis of variance

The one-way analysis of variance (ANOVA) is, despite its name, a common method for the analysis of means of random variables from different groups (Casella and Berger, 2002, pp. 521–534). The observations are assumed to come from the model $Y_{ij} = \mu_i + \epsilon_{ij}$, where $i = 1, 2, \ldots, m$ and $j = 1, 2, \ldots, n$, where ϵ_{ij} are independent and from a normal distribution with mean 0 and unknown variance σ^2, corresponding to the assumptions we made in Section 2.6.

A common way to test the global hypothesis $H_{12\ldots m}\colon \mu_1 = \mu_2 = \ldots = \mu_m$ is to use the F-statistic

$$F = \frac{\frac{n}{m-1}\sum_{i=1}^{m}\left(\bar{Y}_i - \bar{Y}\right)^2}{S_p^2}, \tag{2.6}$$

which under H is F-distributed with $m-1$ and $m(n-1)$ degrees of freedom (Casella and Berger, 2002, pp. 533–534).

The numerator estimates the variance of the group means, and the denominator estimates the overall variance. If the expected values are equal for each group, it is unlikely to observe an estimate of the variance of the group means that is large compared to the overall variance. Thus a large value of F suggests that the means are different, supporting the rejection of H. If H is rejected, we merely conclude that a difference in means exists, but we do not know which specific pair of groups that are different.

The ANOVA can also be used to construct tests for the intersections between the global hypothesis and the elementary hypotheses, but here we are only concerned with testing the global hypothesis, for reasons that will become clear in the next chapter.

The studentized range procedure

The test performed for the elementary hypotheses in Tukey's procedure can also be used to test the global hypothesis in the multiple comparisons setting. We will refer to this as the *range test*. We then only test the comparison of the largest and smallest observed mean, and use the test statistic

$$Q = \frac{\bar{Y}_{\max} - \bar{Y}_{\min}}{\sqrt{S_p^2/n}},$$

which has a studentized range distribution with m groups and $m(n-1)$ degrees of freedom (Shaffer, 1995).

Observe that this can be seen as the result of using Tukey's procedure for FWER control to construct a test for the global hypothesis. If at least one elementary hypothesis is rejected, it must be the case that the hypothesis corresponding to the largest observed difference in means is rejected, since all differences are compared to the same critical value.

F-tests for intersections of hypotheses in regression

F-tests can be used as local tests also in the regression setting. Assuming a hypothesis $H_I : \beta_i = 0$ for $i \in I$ is true can be seen as restricting the model from the full model $Y = \beta_0 + \sum_{i=1}^{m} \beta_i x_i + \epsilon$ to the reduced model $Y = \beta_0 + \sum_{i \in \{1,2,\dots,m\} \setminus I} \beta_i x_i + \epsilon$. If the full model fits the data much better than the reduced model, we have evidence supporting the rejection of H_I. The test statistic used is

$$F_I = \frac{(\mathrm{RSS}_I - \mathrm{RSS}_{\mathrm{tot}})/|I|}{\mathrm{RSS}_{\mathrm{tot}}/(n-m-1)}, \tag{2.7}$$

where RSS_I is the RSS of the model reduced by the restrictions of H_I, and $\mathrm{RSS}_{\mathrm{tot}}$ is the RSS of the full model. F_I is F-distributed with $|I|$ and $n-m-1$ degrees of freedom. When $I = \{i\}$, the test is equivalent to the previously mentioned t-test of H_i (James et al., 2013).

Chapter 3

Applications

In this section we discuss applications of the closed testing procedure. For the first application the closed testing procedure is used to control the FWER for pairwise comparisons of means. Here several intersections of hypotheses coincide, granting the benefit of a closed testing procedure with little added cost in the calculations. The second application is in model selection in multiple linear regression, and is discussed by Goeman et al. (2011). Here the FDP is of interest, and so we are interested in the information gained by dissonant rejections.

3.1 Pairwise comparisons of means

Pairwise comparisons of three means lead to some interesting simplifications in the closed testing procedure, since we end up with coinciding intersections of hypotheses.

Let Y_1, Y_2 and Y_3 be independently and normally distributed random variables with means μ_1, μ_2 and μ_3 and equal, unknown variance σ^2. We wish to compare their means, so that the parameter of interest is $\theta = (\mu_1, \mu_2, \mu_3)$. We have the elementary hypotheses $H_{12}\colon \mu_1 = \mu_2$, $H_{23}\colon \mu_2 = \mu_3$ and $H_{13}\colon \mu_1 = \mu_3$, with alternative hypotheses $H'_{12}\colon \mu_1 \neq \mu_2$, $H'_{23}\colon \mu_2 \neq \mu_3$ and $H'_{13}\colon \mu_1 \neq \mu_3$, respectively. Suppose n samples are collected of each of the three variables. A closed testing procedure will be used to maintain FWER control.

We realize that the closure of $\{H_{12}, H_{23}, H_{13}\}$ only introduces one additional hypothesis, namely the global hypothesis $H_{123}\colon \mu_1 = \mu_2 = \mu_3$, since $H_{\{12,23\}} = H_{\{23,13\}} = H_{\{12,13\}} = H_{123}$. Using the closed testing procedure, we thus first test the global hypothesis with a level α test, and if it is rejected, we test each of the elementary hypotheses also at level α. Hence we obtain level α FWER control without needing to adjust the level of the tests of the elementary hypotheses, so long as the global hypothesis is rejected by its local test.

We discuss options for the local test used for this closed testing procedure, and in the next chapter we discuss experiments we have performed to test these options.

3.1.1 Local tests for the elementary hypotheses

For the local tests of the elementary hypotheses we use the pooled t-tests. Thus we use the test statistic T_{ij} from (2.3), for $1 \leq i < j \leq 3$, which under H_{ij} has a t-distribution with $3(n-1)$ degrees of freedom.

Since we are using two-sided t-tests, we only consider the absolute value of T_{12},T_{23} and T_{13}. With observed test statistics t_{12}, t_{23} and t_{13} we thus get p-values $p_{ij} = 2(1 - F(|t_{ij}|))$, where F is the cumulative distribution function of a t-distribution with $3(n-1)$ degrees of freedom.

3.1.2 Local test for the global hypothesis based on the p-values for the elementary hypotheses

To test the global hypothesis, we may use Bonferroni's method, regardless of the dependence structure of the test statistics for the tests for the elementary hypotheses. This method allows us to reject the global hypothesis if $p_{(1)} \leq \alpha/3$. The resulting closed testing procedure lets us reject any elementary hypothesis H_{ij} with $p_{ij} \leq \alpha/3$, and if this is the case for at least one hypothesis, any remaining hypotheses with a p-value smaller than α.

Note that the closed testing procedure using Bonferroni's method to test the global hypothesis is uniformly stronger than a multiple testing procedure using only the t-tests with Bonferroni-correction. The latter rejects only elementary hypotheses with p-values smaller than or equal to $\alpha/3$, but the closed procedure might reject additional hypotheses if this is the case for at least one of them.

A second alternative is Simes' global test. In our case we can reject the global hypothesis if $p_{(1)} \leq \alpha/3, p_{(2)} \leq 2\alpha/3$, or $p_{(3)} \leq \alpha$. It is trivial to see that this test is stronger than that of Bonferroni.

For these tests we can observe a consequence of having coinciding intersections of hypotheses. In the general case, using Bonferroni's method to create tests for the intersections in a closed testing procedure results in Holm's method. For the case with three hypotheses, this method would reject the hypotheses in a step-down fashion with adjusted levels $\alpha/3, \alpha/2$ and α (Holm, 1979). For our case with coinciding intersections, we instead get adjusted levels $\alpha/3, \alpha$ and α, meaning the resulting method is slightly stronger.

Another consequence is that for our case of three elementary hypotheses, the method constructed with Simes' global test is consonant. In all three cases that the global hypothesis is rejected by Simes' global test, we must have at least one p-value smaller than α. Thus the corresponding elementary hypothesis is also rejected.

Note that the latter observation is not the case in general, when we also must consider $H_{\{12,13\}}$, $H_{\{12,23\}}$ and $H_{\{23,13\}}$ as separate hypotheses. The closed testing procedure will not reject any elementary hypothesis if for example $H_{\{12,13\}}$ and $H_{\{12,23\}}$ are not rejected by their local tests. If we get the p-values $p_{13} = p_{23} = 2\alpha/3$ and $p_{12} > \alpha$ for the elementary hypotheses, the global hypothesis is rejected, since $p_{(2)} \leq 2\alpha/3$. However, neither $H_{\{12,13\}}$ or $H_{\{12,23\}}$ are rejected, since $p_{13}, p_{23} > \alpha/2$ and $p_{12} > \alpha$. Thus none of the elementary hypotheses can be rejected by the closed testing procedure, and the rejection of the global hypothesis is a dissonant rejection.

3.1.3 Local test for the global hypothesis based on the distribution of the data

A one-way ANOVA can be used to test the global hypothesis. For our case with $m = 3$ the test statistic in (2.6) becomes

$$F = \frac{\frac{n}{2}\sum_{i=1}^{3}\left(\bar{Y}_i - \bar{Y}\right)^2}{S_p^2}, \tag{3.1}$$

which under H_{123} has an F-distribution with 2 and $3(n-1)$ degrees of freedom. Notably the resulting closed test procedure is equivalent to Fisher's LSD-test (Fisher, 1935), which for more than three groups does not provide FWER control (Sonnemann, 2008).

We may also use the range test for the global hypothesis, with the test statistic Q from (2.4). With $m = 3$ this has a studentized range distribution with 3 groups and $3(n-1)$ degrees of freedom.

The resulting closed testing procedure using the range test for the global hypothesis will be very similar to Tukey's procedure. The largest difference in means is tested using the studentized range distribution, as with Tukey's procedure, but the remaining two differences are tested with the t-distribution. Thus this resulting closed testing procedure is uniformly more powerful than Tukey's procedure. It is also consonant, since a hypothesis rejected using the studentized range distribution will also be rejected when using the t-distribution (see Section 2.6.2 and Figure 2.2). For our case this closed testing procedure is equivalent to the Newman–Keuls (Newman, 1939; Keuls, 1952) test (Sonnemann, 2008).

3.2 Multiple testing for model selection in regression

Goeman et al. (2011) discussed using their FDP based method to select covariates in multiple regression. When referencing the usage of the method by Goeman et al. (2011) in this manner, we will refer to it as the *confidence method*. Our family of elementary hypotheses is $\mathcal{H} = \{H_i \colon \beta_i = 0 \mid i \in \{1,2,\ldots,m\}\}$, where the βs are the constant coefficients in the model $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_m x_m + \epsilon$ and $\{x_1, x_2, \ldots, x_m\}$ are the covariates we wish to investigate. The elementary hypotheses are tested against their two-sided alternative hypotheses.

Differently from conventional model selection methods, described in the previous chapter, we now use the actual hypothesis tests for selection, instead of optimizing some loss function. The goal is to find some promising subset of the covariates for further research, and thus the confidence method thus seems appropriate. The method lets us choose a subset of covariates that the procedure tells us contains at least a certain number of relevant covariates with $1 - \alpha$ confidence.

We imagine a setting where we divide the selection process in two steps. The first step is exploratory, and the goal is merely to find candidates among the covariates that seem interesting to investigate further in the second step. We have tested how the confidence method performs in step one, before using conventional model selection methods on the resulting subset of covariates in step two. Additionally we have tested the Benjamini & Hochberg procedure for FDR control for the same purpose as the FDP based method.

Since we depend on a closed testing procedure, we must choose local tests for the hypotheses in the closure of $\mathcal{H}$. We benefit from dissonant rejections, since the information gained from these is of interest. Ideally, the local tests have some computational shortcuts, to combat the issue of having to compute the results of $2^m - 1$ tests.

3.2.1 Local tests for the elementary hypotheses

For the elementary hypotheses $H_i : \beta_i = 0, i = 1, 2, \ldots, m$, we use the test statistic T_i from (2.5), which under H_i has a t-distribution with $n-m-1$ degrees of freedom. With observed test statistics $t_1, t_2, \ldots, t_m$ we thus get p-values $p_i = 2(1 - F(|t_i|))$, for $i = 1, 2, \ldots, m$, where F is the cumulative distribution function of a t-distribution with $n - m - 1$ degrees of freedom.

3.2.2 Local tests for intersections based on the p-values for the elementary hypotheses

For the local tests of intersections of hypotheses we may consider the tests discussed for the application in pairwise comparisons, namely a test based on Bonferroni's method and Simes' global test. Note that none of the intersections of hypotheses coincide in this example, so saving time during the calculations is very beneficial.

The method based on Bonferroni's method is consonant in our case (see A.2 in the Appendix). Thus there will be no dissonant rejections, resulting in trivial confidence statements about the number of true hypotheses in a subset of all the hypotheses, and thus the only defining rejections are of elementary hypotheses. The calculations will be fast, but if we conclude to reject only hypotheses that were rejected by the closed testing method, we end up with FWER control. In this case no additional hypothesis we choose to reject can improve the number of false hypotheses we are confident to have rejected, and thus the strengths of the confidence method are not utilized.

Unlike Bonferroni's method, Simes' global test is not consonant, as shown by the example in section 3.1.2. The occurrence of such dissonant rejections does however appear to be rare, see Table 3.1, which shows the results of simulations where Simes' global test was used to test intersections of hypotheses in the regression setting. Simes' global test does however have a shortcut in its calculations, as described by Goeman et al. (2011), which is a benefit for the computations.

3.2.3 Local tests for intersections based on the distribution of the data

Similarly to the pairwise comparisons example, we can use the F-tests as local tests also in the regression setting. We use the test statistics F_I from (2.7), based on the restrictions of H_I, for all H_I in the closure of $\mathcal{H}$.

A disadvantage of the F-test is the computational cost, as it requires all $2^m - 1$ reduced models to be made in order to calculate their RSS values. There is no shortcut for the F-test, unlike the tests based on the p-values of the elementary hypotheses. An additional disadvantage is in the case that $m > n$, where there is no unique least squares estimates for the coefficients. In this case the confidence method with local F-tests cannot be used, since we are unable to fit the models we need. Model selection with for example lasso regression

Simes' global test			F-test			True
Bound	Defining	Rate	Bound	Defining	Rate	complexity
4.030	4.037	0.006	4.069	4.567	0.211	4.482

Table 3.1: Simulation results for comparing Simes' global test and the F-test in the regression setting. "Bound" is the lower confidence bound for the number of relevant covariates in the entire set of covariates, "Defining" is the number of hypotheses involved in any defining rejection, and "Rate" is the rate at which the method caused at least one dissonant rejection. 1000 simulations were run with randomly constructed linear models with 12 covariates to consider, of which a random number were in the true model.

can be used to avoid this problem (James et al., 2013). This disadvantage is present also for the previously mentioned local tests, since they depend on the p-values obtained from fitting the full model.

Table 3.1 shows results from simulations similar to those described in the next chapter, comparing the F-test to Simes' global test when used as local tests for the intersections. 1000 randomly constructed linear models were used to create data, and Simes' global test and the F-test were used in the confidence method. We kept track of the lower confidence bound for the number of relevant covariates that each method yielded, how large the subset of hypotheses involved in the defining rejections were (see the next subsection), and the rate at which the method resulted in at least one dissonant rejection. We also kept track of how many covariates that actually were relevant in each model.

The results indicate that the F-test more frequently yields dissonant rejections, and a very slightly larger confidence bound for the number of relevant covariates. To obtain the larger bound we must however include slightly more covariates in the model (reject slightly more hypotheses). With so few dissonant rejections the Simes' global test is close to FWER control, and we see that the average number of hypotheses kept is actually smaller than the average complexity of the true model. If we want to discover all relevant covariates the F-test thus seems to be a good choice, since even though it is less certain about which of the hypotheses that are most likely to be relevant, it informs us about a larger set of interesting covariates to select than Simes' global test does.

For our simulations we consider only the case where $n > m$, so that we are actually able to use a closed testing procedure in our experiments. Despite the fact that the F-tests come at an increased computational cost, we choose these as our local tests over Simes' global test.

3.2.4 Choosing which hypotheses to reject with the confidence method

To simplify the usage of the confidence method when running many simulations, and to avoid having to manually review every confidence statement each time, we use an algorithm to make the choice of which hypotheses to reject. The algorithm preferably avoids checking the confidence statements of every subset of hypotheses, to reduce computational cost. One natural choice is a union of non-overlapping subsets obtained by looking at the defining rejections. The lower confidence bound for the number of false hypotheses in this set grows by one each time we add the hypotheses involved in a defining rejection, so long as none of these hypotheses are in our set already.

Though simple, this is not the optimal choice. If for example the defining rejections are $H_{\{1,2\}}, H_{\{1,3\}}$ and $H_{\{2,3\}}$, the lower confidence bound for the number of false hypotheses in $\{H_1, H_2, H_3\}$ is 2. Choosing only non-overlapping subsets will however result in only rejecting two hypotheses, with a confidence bound of 1. A solution to this specific case is to iterate through all the sets of hypotheses from defining rejections, and add them to our set of rejections so long as this increases the lower confidence bound.

A problem can occur with this algorithm if for example the defining rejections are $H_{\{1,2\}}, H_{\{3,4\}}, H_{\{1,3\}}, H_{\{1,4\}}$ and $H_{\{2,3\}}$. Here we have a lower confidence bound of 2 for $\{H_1, H_2, H_3, H_4\}$ and 1 for $\{H_1, H_2\}$. This means that we would add H_3 and H_4 to our rejections if we iterate through the defining rejections in that order, although we only need to reject for example $\{H_1, H_2, H_3\}$ to obtain the same lower confidence bound of 2.

A very simple way to select covariates is to just reject all hypotheses involved in any defining rejection. The lower confidence bound for the number of false hypotheses in this resulting set is the same as the bound for the set of all elementary hypotheses (a proof of this is found in the Appendix, see A.1), so this choice ensures we select all the covariates that the closed testing procedure believes with $1 - \alpha$ confidence to be relevant.

Although this selection process can fail to fully use the information from the dissonant rejections, these rejections are not very frequent (seen in Table 3.1), and we remove the randomness that can occur with the previous algorithm proposed. If for example the defining rejections are $H_{\{1,2\}}$ and $H_{\{1,3\}}$, the previous algorithm would make an arbitrary choice. Perhaps more importantly, we use the confidence method as an exploratory first step, and therefore argue that it is more important to keep the potentially relevant covariate x_3, by rejecting H_1, H_2 and H_3, even though only rejecting the first two gives the same lower confidence bound.

Note that using an algorithm removes one of the method's strengths, namely that the user is normally able to choose freely which covariates to select. In our simulations the choice is done out of necessity, firstly because manually selecting a subset a large number of times is very time consuming, and secondly because even though we include some randomness in the construction of the true models, we have knowledge that may affect our choice of covariates. One could argue regarding the latter that this corresponds to using expert knowledge in a situation with real data, which is a scenario where the confidence method would benefit, though this argument is rather vague and subjective. When using the method on real data it may very well be the case that the algorithms described are not optimal, and a more informed choice of covariates should probably be made if time and knowledge is available for the researcher to do so.

Chapter 4

Experiments

In this section we describe the experiments that were performed to investigate the capabilities of the closed testing procedures in the applications described in the previous chapter. For both applications we ran experiments using artificially created datasets, in order to be able to know for certain which hypotheses were actually true, and that our model assumptions were fulfilled.

4.1 Pairwise comparisons of means

For each simulation we created $n = 30$ samples of each of Y_1, Y_2 and Y_3 from a normal distribution

$$\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{bmatrix}, \sigma^2 \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \right).$$

We calculated the sample means, $\bar{y}_1$, $\bar{y}_2$ and $\bar{y}_3$, sample variances, s_1^2, s_2^2 and s_3^2, pooled sample mean $\bar{y}$, and pooled sample variance s_p^2 according to (2.2). These were then used to calculate the corresponding test statistics t_1, t_2 and t_3, (2.3), and p-values p_1, p_2 and p_3 for the elementary hypotheses. The global hypothesis $H_{123}\colon \mu_1 = \mu_2 = \mu_3$ was tested with Bonferroni's method, Simes' global test, the range test, and the F-test from a one-way ANOVA.

H_1, H_2 and H_3 were thus tested through closed testing procedures with various global tests, to control the FWER at level $\alpha = 0.05$. Tukey's procedure for the elementary hypotheses was also used for comparison.

Shaffer (1981) conducted a similar experiment, and also tested the studentized range procedure and the F-test for the global hypothesis. Here we investigate some more tests for the global hypothesis. Additionally, we separate correct and incorrect rejections, and report the any-pair power (the rate at which at least one false hypothesis is rejected) and all-pairs power (the rate at which all false hypotheses are rejected), as suggested by Jaccard et al. (1984).

Two sets of experiments were run, one with $\mu_1 = \mu_2 = 0$ and with μ_3 taking 30 evenly spaced values from 0.1 to 1.0, and the other with $\mu_1 = 0$, μ_2 taking 30 evenly spaced values from 0.1 to 0.7, and $\mu_3 = 2\mu_2$. For each of these 60 experiments we ran $N = 10\,000$ simulations. Outside of these intervals, the methods gave trivial results, either rejecting close to all hypotheses or close to none. One experiment was also run with $\mu_1 = \mu_2 = \mu_3 = 0$ to confirm the validity of the methods also in this case, with $N_0 = 100\,000$ simulations. For all simulations, $\sigma^2 = 1$ was used.

The goal is to compare the different closed testing procedures obtained with the different local tests for the global hypothesis, with respect to their ability to reject false hypotheses and not reject true ones.

4.2 Multiple testing for model selection in regression

4.2.1 Data

We used $n = 1000$ data points in each simulation. We used $m = 12$ covariates for the regression, where $x_j = (x_{1j}, x_{2j}, \ldots, x_{mj})^T$, $j = 1, 2, \ldots, n$, were drawn either from $N(0, \Sigma)$ or $N(0, I)$. Σ was a semi-arbitrary covariance matrix with all variances equal to 1 (see Appendix B for more information), and I was the m by m identity matrix. This resulted in two design matrices X_{corr} and X_{uncorr}, respectively, which each were used for half of the experiments.

We wanted to test a varied set of linear models, with different number of relevant covariates and different sizes of coefficients. For each simulation we created n independent samples ϵ_j, $j = 1, 2, \ldots, n$, from a standard normal distribution. The number of covariates to be included in the model, m', was drawn uniformly from either $\{1, 2, \ldots, \lfloor m/3 \rfloor\}$ for one set of experiments and $\{\lfloor m/3 \rfloor + 1, \lfloor m/3 \rfloor + 2, \ldots, \lfloor 2m/3 \rfloor\}$ for the other. A set J of m' indices was then drawn without replacement from $\{1, 2, \ldots, m\}$, and m' coefficients $\{\beta_i \mid i \in J\}$ were drawn independently and uniformly from the interval $[-0.25, -0.05] \cup [0.05, 0.25]$ for one set of experiments and $[-0.5, -0.25] \cup [0.25, 0.5]$ for the other. Thus we had 8 experiments, with covariates that were uncorrelated or correlated, few or many relevant covariates, and small or large coefficients.

The response was calculated as $y_j = \beta_0 + \sum_{i=1}^{m} \beta_i x_{ij} + \epsilon_j$, for $j = 1, 2, \ldots, n$, where $\beta_i = 0$ for $i \notin J$. We focus only on the m parameters corresponding to covariates other than the intercept, and chose $\beta_0 = 5$, large enough that $H_0\colon \beta_0 = 0$ was rejected at any sensible level for all simulations.

4.2.2 Experiment design

The first step in our model selection is to use either the confidence method with local level $\alpha_{\text{FDP}} = 0.05$ F-tests, or FDR control with the Benjamini & Hochberg procedure to restrict the set of covariates. For the FDR control a mild level of $\alpha_{\text{FDR}} = 0.25$ was used, because further selection was to be performed.

Either lasso regression or best subset selection with AIC were then used as a second step, considering only the restricted set of covariates, to obtain a final set of covariates for the model. 5-fold cross-validation was used to choose the λ in the lasso regression. If one or zero covariates were kept during the first step, no further selection was performed.

Selection using only lasso regression or subset selection with AIC were also tested on the full set of covariates, without prior restrictions, and covariates selected by the methods in the first step were also used to fit models without any further selection performed. In each simulation the dataset was split equally in a training set and a test set. All selection and model fitting was performed using the training data, and the resulting models were tested on the test data.

In summary, the 8 resulting composite methods were tested in the 8 different experiments obtained by letting the covariates be uncorrelated or correlated, the number of relevant covariates be small or large, and the size of the coefficients be either small or large. In essence, we investigated if there were some benefit to restricting the set of covariates in a hypothesis testing setting before doing conventional model selection.

We ran one simulation to illustrate the process, and $N = 500$ simulations for each of the 8 experiments to evaluate the composite methods.

Chapter 5

Analysis

In this section we present and discuss the results from the simulation experiments described in the previous chapter.

5.1 Pairwise comparisons of means

5.1.1 Reported results

For all simulations we reported the rate at which no hypothesis was rejected (in the column named “None”), and how often only one elementary hypothesis was rejected (“Contr.”). The latter outcome is of interest because it is contradictory. If for example $\mu_1 \neq \mu_2$, it cannot be the case that $\mu_2 = \mu_3$ and $\mu_1 = \mu_3$.

For the simulations where not all means were different, in which case committing a type I error was impossible, we reported the rate at which at least one type I error among the elementary hypotheses occurred, which is an estimate of the FWER. Note that we only counted the occurrence of type I errors among the elementary hypotheses, meaning a dissonant rejection of the global hypothesis in the case that all means are equal was not counted. This was only a concern for the global F-test, as all the other methods are consonant.

For the simulations with at least one mean different from the others we also reported the rate at which at least one false elementary hypothesis was rejected (“Any-pair”), and the rate at which all false elementary hypotheses were rejected (“All-pairs”).

The names of the rows in the presented tables (Table 5.2 through 5.5) correspond to the different testing procedures tested. “Bonf.”, “Simes”, “F”, and “Range” refer to the closed testing procedures using Bonferroni’s method, Simes’ global test, the F-test from a one-way ANOVA, and the range test, respectively, as the local test for the global hypothesis. “Tukey” refers to Tukey’s procedure for FWER control, applied to the elementary hypotheses.

Note that for a probability p, the proportion of successes in N independent trials, $\hat{p}$, is the mean of N independent Bernoulli trials with probability p of success, and the limits of

$\hat{p}$	0.5	0.4 or 0.6	0.3 or 0.7	0.2 or 0.8	0.1 or 0.9	0.05 or 0.95
Width	0.0062	0.0061	0.0057	0.0050	0.0037	0.0027

Table 5.1: The width of approximate 95% confidence intervals for a probability p when estimated with $N = 100\,000$ simulations, for different values of the observed $\hat{p}$.

an approximate 95% confidence interval for the p are thus

$$\widehat{p} \pm 1.96\sqrt{\hat{p}(1-\hat{p})/N} \tag{5.1}$$

(Casella and Berger, 2002, p. 502).

For $N = 100\,000$ the widths of such intervals given for certain values of $\hat{p}$ in Table 5.1. Note that though many of the reported results for our analysis are probabilities, in many of the cases the probability is not the same across all experiments, contradicting the underlying assumptions for the confidence intervals. We are however not necessarily interested in the exact estimates themselves, but rather in comparing estimates for different methods, and believe the intervals will suffice for this purpose.

5.1.2 All means equal to zero

The simulation results from $N_0 = 100\,000$ simulations in the case where all the three means were equal to zero can be seen in Table 5.2. The limits of the corresponding approximate 95% confidence intervals are included for the FWER estimates. We see that for the first three methods the FWER estimates are smaller than $\alpha = 0.05$, and for the last two α is inside the approximate confidence intervals. Had we included the incorrect rejection of the true global hypothesis in our FWER estimate, α would also be inside the confidence interval for the F-test. A further explanation of this comes later.

Simes' global test rejects more than Bonferroni's method, which is expected since it is uniformly stronger. The results for the range test and Tukey's procedure are equal, since the local test the former uses for the global hypothesis is the same test used by the latter. Tukey's procedure yields contradicting results slightly more often than the other methods.

For all methods except the one using a local F-test, we see that the values in the two first columns add up to 1. The explanation is that the F-test is not consonant, and the missing 0.0008 is the rate at which there occurred a dissonant rejection. Rejecting the true global hypothesis is a type I error, but we have chosen to only report the estimated FWER among the elementary hypotheses, since only these are of interest with respect to the FWER control. The only resulting difference is for the closed testing procedure using the local F-test, and only in this particular case when all means are equal to zero.

5.1.3 One or two means different from zero

One mean different from zero

The results from simulations with $\mu_1 = \mu_2 = 0$ and $\mu_3 \in [0.1, 1.0]$ can be seen in Table 5.3. We see that all the methods maintain the desired level of FWER control. The closed testing procedures had smaller FWER estimates when the third mean was small. This is

	None	FWER	Contr.
Bonf.	0.9559	0.0441 ± 0.0013	0.0266
Simes	0.9534	0.0466 ± 0.0013	0.0266
F	0.9507	0.0485 ± 0.0013	0.0273
Range	0.9493	0.0507 ± 0.0014	0.0317
Tukey	0.9493	0.0507 ± 0.0014	0.0440

Table 5.2: The results from $N_0 = 100\,000$ simulations of pairwise comparisons of means with different methods when all means are equal to 0, with approximate 95% confidence limits for the FWER estimates.

because in this case the global hypothesis, and thus any elementary hypothesis, including the true one, was less likely to be rejected. We also observe that Tukey's procedure was conservative for all values of μ_3.

The any-pair power was quite even for all five methods, with a slight edge to the procedures using the F-test or the range test. Tukey's procedure also performed well according to this criterion, which was expected since the method is based on the largest observed difference. Because of this one could think that The range test and Tukey's procedure would have the same any-pair power. It is however possible that the largest observed difference in means is between group 1 and 2, and that the pooled t-test rejects the hypothesis corresponding to the second largest observed difference while Tukey's procedure does not.

When considering the all-pairs power we clearly see the weakness of Tukey's procedure compared to the closed testing procedures. The F-test provided the best results also for this criterion, and we observe that Simes' global test here beat the range test, while Bonferroni's method performed the worst of the closed testing procedures.

When it comes to making contradictory conclusions, Tukey's procedure once again provided the worst results. The closed testing procedures using Bonferroni's method or Simes' global test rejected exactly one elementary hypothesis under the same circumstances, namely when $p_{(1)} \leq \alpha/3$ and $p_{(2)}, p_{(3)} > \alpha$. The rate at which these two method rejected exactly one hypothesis was smaller than for the other methods, which corresponds to the rest of the results that suggest that their overall power was slightly lower.

Two means different from zero

The results from simulations with $\mu_1 = 0, \mu_2 \in [0.1, 0.7]$ and $\mu_3 = 2\mu_2$ can be seen in Table 5.4. In this case all the hypotheses were false, and thus no type I errors could be committed. Again we can observe that the closed procedure using the local F-test caused dissonant rejections. The sum of the overall rates at which no rejections at all, and at least one elementary hypothesis was rejected, is $0.3305 + 0.6684 = 0.9989$, meaning that the F-test caused a dissonant rejection 0.11% of the time.

When all three means were different from each other the any-pair power for Tukey's procedure and the closed testing procedure using the range test were equal, because the largest difference in means corresponded to a false hypothesis. Here we also see a benefit from using the range test for the global hypothesis, as this resulted in a larger any-pair

		None	FWER	Any-pair	All-pairs	Contr.
	Bonf.	0.8550	0.0274	0.1392	0.0551	0.0688
$\mu_1 = 0$	Simes	0.8485	0.0294	0.1457	0.0595	0.0688
$\mu_2 = 0$	F	0.8417	0.0305	0.1503	0.0605	0.0723
$\mu_3 \in [0.1, 0.4]$	Range	0.8413	0.0295	0.1520	0.0578	0.0786
	Tukey	0.8413	0.0193	0.1484	0.0292	0.1205
	Bonf.	0.4559	0.0430	0.5432	0.3659	0.1435
$\mu_1 = 0$	Simes	0.4423	0.0444	0.5568	0.3782	0.1435
$\mu_2 = 0$	F	0.4309	0.0453	0.5661	0.3816	0.1491
$\mu_3 \in [0.4, 0.7]$	Range	0.4331	0.0441	0.5658	0.3735	0.1578
	Tukey	0.4331	0.0197	0.5644	0.2532	0.2975
	Bonf.	0.0928	0.0493	0.9072	0.8136	0.0731
$\mu_1 = 0$	Simes	0.0859	0.0495	0.9141	0.8204	0.0731
$\mu_2 = 0$	F	0.0816	0.0496	0.9177	0.8218	0.0752
$\mu_3 \in [0.7, 1.0]$	Range	0.0839	0.0495	0.9160	0.8176	0.0779
	Tukey	0.0839	0.0192	0.9159	0.7094	0.1942
	Bonf.	0.4679	0.0399	0.5299	0.4115	0.0952
$\mu_1 = 0$	Simes	0.4589	0.0411	0.5389	0.4194	0.0952
$\mu_2 = 0$	F	0.4514	0.0418	0.5447	0.4213	0.0989
$\mu_3 \in [0.1, 1.0]$	Range	0.4528	0.0410	0.5446	0.4163	0.1048
	Tukey	0.4528	0.0194	0.5429	0.3306	0.2040

Table 5.3: Each of the three first parts of the table shows results from 100 000 simulations of pairwise comparisons of means with different methods when two means are equal to zero and the third is different. The last part shows the average of all 300 000 simulations.

		None	Any-pair	All-pairs	Contr.
	Bonf.	0.7556	0.2444	0.0024	0.1057
$\mu_1 = 0$	Simes	0.7472	0.2528	0.0024	0.1057
$\mu_2 \in [0.1, 0.3]$	F	0.7396	0.2584	0.0024	0.1086
$\mu_3 = 2\mu_2$	Range	0.7372	0.2628	0.0024	0.1192
	Tukey	0.7372	0.2628	0.0003	0.1890
	Bonf.	0.2496	0.7504	0.0605	0.1820
$\mu_1 = 0$	Simes	0.2408	0.7592	0.0605	0.1820
$\mu_2 \in [0.3, 0.5]$	F	0.2351	0.7637	0.0605	0.1840
$\mu_3 = 2\mu_2$	Range	0.2324	0.7676	0.0605	0.1935
	Tukey	0.2324	0.7676	0.0156	0.3839
	Bonf.	0.0186	0.9814	0.3513	0.0588
$\mu_1 = 0$	Simes	0.0173	0.9827	0.3513	0.0588
$\mu_2 \in [0.5, 0.7]$	F	0.0167	0.9832	0.3513	0.0589
$\mu_3 = 2\mu_2$	Range	0.0162	0.9838	0.3513	0.0603
	Tukey	0.0162	0.9838	0.1737	0.1815
	Bonf.	0.3413	0.6587	0.1381	0.1155
$\mu_1 = 0$	Simes	0.3351	0.6649	0.1381	0.1155
$\mu_2 \in [0.1, 0.7]$	F	0.3305	0.6684	0.1381	0.1172
$\mu_3 = 2\mu_2$	Range	0.3286	0.6714	0.1381	0.1243
	Tukey	0.3286	0.6714	0.0632	0.2514

Table 5.4: Each of the three first parts of the table shows results from 100 000 simulations of pairwise comparisons of means with different methods when all three means are different. The last part shows the average of all 300 000 simulations.

power than for the other closed testing procedures.

Similarly to the case where one mean was different from the others, the all-pairs power of Tukey's procedure was smaller also in the case where all means were different from each other. Here the difference is even larger, with Tukey's method having less than half the estimated probability of rejecting all false hypotheses, compared to the closed testing procedures.

Interestingly the estimates for the all-pairs powers were equal for all the closed testing procedures. Shaffer (1981) observed the same result in her simulations. This suggests that for all conducted simulations, when all the elementary hypotheses were rejected by their respective local t-tests, all the different local tests for the global hypothesis also lead to rejection. This is actually the case in general for all the closed testing procedures tested, at least when $\alpha = 0.05$ and $n = 30$.

All elementary hypotheses are rejected if the absolute value of the t-statistics from (2.3) are larger than or equal the $1 - \alpha/2$-quantile of the t-distribution, $t_{\alpha/2,87} = 1.99$. In other words we have $\bar{Y}_{(1)}/\sqrt{2S_p^2/n} + 2t_{\alpha/2,87} \leq \bar{Y}_{(2)}/\sqrt{2S_p^2/n} + t_{\alpha/2,87} \leq \bar{Y}_{(3)}/\sqrt{2S_p^2/n}$ This specifically implies that the largest observed absolute value of a t-statistic is at least $2t_{\alpha/2,87} = 3.98$.

Further we have that $3.98 > t_{\alpha/6,87} = 2.44$, which means that the p-value for the t-test corresponding to the largest observed difference is smaller than $\alpha/3$. We also have $3.98 > q_{\alpha,3,87} = 3.37$, where $q_{\alpha,3,87}$ is the $1-\alpha$-quantile of the studentized range distribution with 3 groups and 87 degrees of freedom. Thus the range test, Bonferroni's method and Simes' global test all reject the global hypothesis. The test statistic from (3.1) can be written as the mean of the squared t-statistics, and if each of these surpass their critical values, so does the F-statistic, meaning the F-test also rejects the global hypothesis.

The estimated probability of rejecting exactly one elementary hypothesis was very large for Tukey's method, approximately twice the value obtained for the other methods. For the range test this estimated probability was slightly larger than for the other closed testing procedures, for which the estimates were quite even.

Table 5.5 shows the averaged results from all the simulations with either one or two means different from zero, from a total of 600 000 simulations. Overall the closed testing procedures using the F-test or the range test performed better than the other closed testing procedures, and far better than Tukey's method. When it comes to the comparison between the two, the F-test generally had comparable or better all-pairs power and smaller probability of contradictory results, while the range test generally had comparable or better any-pair power. The differences were very slight, however.

5.2 Multiple testing for model selection in regression

5.2.1 Reported results

From our hypothesis testing point of view, we are interested in how many type I errors we make, which means the true hypotheses we reject, or the irrelevant covariates we include in our model. From the model selection point of view we are also interested in type II errors, the false hypotheses we fail to reject, or the relevant covariates that we fail to include in

	None	FWER	Any-pair	All-pairs	Contr.
Bonf.	0.4046	0.0200	0.5943	0.2748	0.1053
Simes	0.3970	0.0205	0.6019	0.2787	0.1053
F	0.3910	0.0209	0.6066	0.2797	0.1080
Range	0.3907	0.0205	0.6080	0.2772	0.1146
Tukey	0.3907	0.0097	0.6072	0.1969	0.2277

Table 5.5: The table shows the average of the results from 600 000 simulations of pairwise comparisons of means with different methods when either one or two means are different from zero.

our model. Additionally we care about the connection between the different types of errors and how well the model predicts the response for unseen data.

For our simulations, we thus reported the rate at which at least one type I error occurred, the estimated FWER, and the rate at which at least one type II error occurred, which we call FWER-II (familywise type II error rate). We also reported the average FDP, the proportion of true hypotheses among the ones that are rejected (irrelevant covariates included in the model), or V/R (see (2.1) and Table 2.1), which is an estimate of the FDR.

We additionally reported the rate at which exactly the correct covariates were chosen (in the column named “Perf.”), the complexity of the chosen model (“Comp.”), and the test MSE obtained by making predictions on the test set (“MSE”). The lower confidence bound for the number of relevant covariates chosen by the confidence method is mentioned in the caption of each table.

The names of the rows in the presented tables (Tables 5.9 through 5.12) correspond to the different composite model selection methods. “None” means no selection was done either prior to or after using the other listed method, with “None - None” meaning that the full model was used. “AIC” and “Lasso” refers to the conventional model selection methods, “Conf.” refers to the confidence method, and “B&H” refers to the Benjamini & Hochberg procedure for FDR control. “Truth” refers to the underlying model, which was used to create the data.

5.2.2 Illustrative example of the selection process

We include an example to illustrate the selection process. All subset models were fitted, AIC values and F-statistics were calculated, and the 8 composite methods described in the previous chapter were performed. The underlying true model was constructed with few relevant covariates, small coefficients and correlated covariates, as described in the previous chapter.

The summary of the full model is shown in Table 5.6. The reported p-values were used for the B&H procedure, and were used together with the F-statistics for the confidence method. The p-value of $0.0014 < \alpha_{\text{FDP}} = 0.05$ for the global hypothesis for the regression told us that there would be at least some defining rejection(s) when using the confidence method with local F-tests for the intersections of hypotheses. At level $\alpha_{\text{FDR}} = 0.25$, the B&H procedure selected the covariates x_5 and x_{12}.

The defining rejections obtained by the confidence method were $H_{\{5,12\}}$ and $H_{\{1,3,5,11\}}$, which meant our algorithm selected x_1, x_3, x_5, x_{11} and x_{12}. We could tell from the defining

	Estimate	Std. Error	t-statistic	p-value
β_0	4.9633	0.0467	106.29	0.0000
β_1	−0.0838	0.0508	−1.65	0.0994
β_2	0.0186	0.0504	0.37	0.7118
β_3	−0.0325	0.0560	−0.58	0.5616
β_4	−0.0101	0.0529	−0.19	0.8491
β_5	0.1466	0.0511	2.87	0.0043
β_6	−0.0289	0.0505	−0.57	0.5674
β_7	−0.0065	0.0559	−0.12	0.9070
β_8	−0.0675	0.0519	−1.30	0.1943
β_9	−0.0432	0.0545	−0.79	0.4283
β_{10}	0.0234	0.0531	0.44	0.6592
β_{11}	0.0516	0.0536	0.96	0.3356
β_{12}	−0.1731	0.0507	−3.41	0.0007
F-statistic: 2.727 on 12 and 487 DF			p-value: 0.0014	

Table 5.6: Summary of the multiple linear regression model fitted on the training data, using all 12 covariates.

rejections that $H_{\{1,3,11,12\}}$ was not rejected, as otherwise either it or one of its components would have been a defining rejection. This gave a lower confidence bound of only one relevant covariate for our selected set of covariates. Since none of the defining rejections had only one elementary component, FWER control with the closed testing procedure would in this case have lead to no rejected hypotheses, and thus included no covariates in the model.

Table 5.7 shows the 10 models with the smallest AIC values, and the one with the 19th smallest. The covariates are marked with different colors corresponding to whether or not they were in the subsets of covariates selected by the confidence method or the B&H procedure. The first model is the one that had the minimal AIC.

Note that the composite method of the confidence method and subset selection with AIC corresponds to selecting the model with the minimal AIC that only contains covariates selected by the confidence method. In Table 5.7 we see that the third model is the first containing only covariates selected by the confidence method. Similarly, the model selected by the composite method of the B&H procedure and subset selection with AIC is the one that had the 19th smallest AIC overall. These are marked with stars in Table 5.7.

Figure 5.1 shows the selection of the value for λ in the lasso regression with no prior restrictions. The value that minimized the estimation for the test MSE was chosen, and the corresponding model complexity can also be observed in the figure.

In Table 5.8 the resulting models are shown, including the true model. Covariates correctly included are marked with blue, incorrect inclusions with orange. In this case, the B&H procedure selected only the correct covariates, and naturally also achieved the best test MSE. Using only AIC, lasso or the confidence method resulted in too complex models, although the regularization of the lasso reduced the issue of including too many covariates. The composite method of the confidence method and AIC made only one type I error, and the resulting model was almost as good as the one achieved with the B&H procedure.

AIC-rankings	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}
1	🟦				🟩			⬛				🟩
2	🟦				🟩			⬛			🟦	🟩
3*	🟦				🟩							🟩
4	🟦				🟩						🟦	🟩
5					🟩			⬛				🟩
6	🟦				🟩			⬛	⬛			🟩
7	🟦		🟦		🟩			⬛				🟩
8	🟦				🟩	⬛		⬛				🟩
9	🟦				🟩				⬛			🟩
10	🟦			⬛	🟩			⬛				🟩
19*					🟩							🟩

Table 5.7: The best subset models judged by the AIC values, where the top model corresponds to the model with the minimal AIC. Covariates marked with green were in the subsets of covariates restricted by both the confidence method and the B&H procedure. Covariates marked in blue were only in the former, and covariates marked in black were in neither. The models marked with a star, the one with the 3rd smallest and the one with the 19th smallest AIC value, were the models selected by the AIC when the set of covariates was restricted by the confidence method or the B&H procedure, respectively.

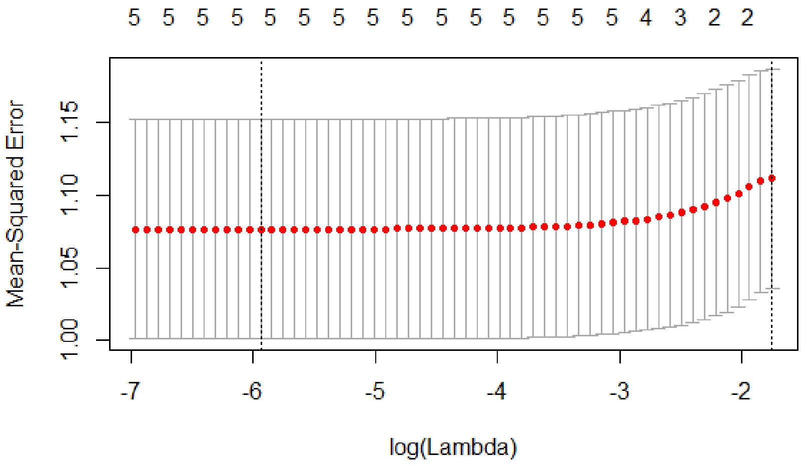

Figure 5.1: The selection of λ in the lasso regression. The x-axis shows the values of the tested values of λ (on a logarithmic scale), the y-axis shows the cross-validation estimate of the test MSE, and the top axis shows the model complexity the corresponding λ value yielded. The first dotted line shows the λ value that minimized the MSE estimate, and the second dotted line shows the largest λ that yielded an MSE estimate within one standard deviation of the minimum.

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	x_{10}	x_{11}	x_{12}	MSE
None - AIC	🟧				🟦			🟧				🟦	1.0047
None - Lasso	🟧		🟧		🟦			🟧				🟦	1.0008
Conf. - None	🟧		🟧		🟦						🟧	🟦	1.0049
Conf. - AIC	🟧				🟦							🟦	0.9997
Conf. - Lasso	🟧		🟧		🟦						🟧	🟦	1.0041
B&H - None					🟦							🟦	0.9914
B&H - AIC					🟦							🟦	0.9914
B&H - Lasso					🟦							🟦	0.9918
None - None	🟧	🟧	🟧	🟧	🟦	🟧	🟧	🟧	🟧	🟧	🟧	🟦	1.0091
Truth					🟦							🟦	0.989

Table 5.8: The resulting models from the various composite selection methods. Covariates that were correctly included are marked with blue, and erroneous inclusions are marked with orange.

5.2.3 Model selection for randomly generated linear models

Table 5.9 shows the mean values for all simulations when running $N = 500$ simulations for each of 8 experiments, for a total of 4000 simulated linear models. The best values for each category other than complexity are marked in **bold text**, the worst are underlined. Tables 5.10, 5.11 and 5.12 show the comparisons of mean values for all simulations for the experiments with few and many relevant covariates, small and large coefficients, and uncorrelated and correlated covariates, respectively.

Table C.1 in the Appendix shows the widths of approximate 95% confidence intervals for the probabilities or the expectations of the values in Table 5.9. The largest values of widths for the FWER estimates is 0.015, 0.014 for FWER-II estimates, 0.007 for FDR, 0.015 for rate of perfect model, 0.102 for complexity (this is for lasso, the other methods have slightly narrower confidence intervals), and 0.0021 for MSE. Note that these intervals do not take any multiplicity adjustment into account, as we are not concerned with any single result, but rather the overall performance of the procedures. The differences observed in the error rates, FDR and rate of perfect covariate selection between the results from procedures with different methods used in step one are all far greater than the widths of the confidence intervals. For the MSE the differences are less significant, although all procedures performed better than using the full model, and a lot worse than the true, underlying model.

One immediate observation, and something which might seem very natural, is that there is a balance going on between the type I and type II errors. Going to either extreme negatively impacts the test predictions. If we use the full model we make no type II errors, but the model complexity may cause overfitting and increase the test MSE by having a too large variance. If we use the empty model (predict only the sample mean, ignoring the covariates) we make no type I errors, but the model complexity may cause underfitting and increase the test MSE by having a too large bias. The optimal complexity thus must be somewhere between the extremes. We call a method *strict* if it rejects few hypotheses, and *mild* if it rejects many. Thus a strict method is likely to make fewer type I errors and more type II errors than a mild one.

The conventional selection methods appeared mild, and tended to select larger models

	FWER	FWER-II	FDR	Perf.	Comp.	MSE
None - AIC	0.712	0.224	0.250	0.220	5.386	1.0213
None - Lasso	0.904	**0.124**	0.400	0.062	7.492	1.0202
Conf. - None	0.242	0.299	0.088	0.550	4.556	1.0189
Conf. - AIC	**0.211**	0.323	**0.069**	**0.561**	4.318	1.0187
Conf. - Lasso	0.241	0.300	0.087	0.550	4.540	**1.0186**
B&H - None	0.504	0.269	0.154	0.337	4.912	1.0200
B&H - AIC	0.482	0.273	0.146	0.353	4.826	1.0198
B&H - Lasso	0.504	0.269	0.154	0.337	4.912	1.0196
None - None	1.000	0.000	0.626	0.000	12.000	1.0268
Truth					4.483	1.0001

Table 5.9: Model selection results for 4000 simulations, showing the mean across all simulations and all experiments. The best results within the different categories are marked in **bold text**, the worst results are underlined. The mean of the lower confidence bound of the number of relevant covariates chosen by the confidence method was 3.732.

than the hypothesis testing and composite methods, resulting in large FWER estimates as irrelevant covariates were included, and small FWER-II estimates as relevant covariates were rarely dropped. The B&H procedure yielded models of slightly smaller complexity, with a smaller FWER estimate and somewhat larger FWER-II estimate. The resulting test MSE values were very similar to those obtained by the conventional methods. The confidence method was stricter, with a much smaller FWER estimate and only slightly larger FWER-II estimate, compared to the B&H procedure. The balance obtained with the confidence method appeared optimal, as these methods achieved the best test predictions overall, and quite often selected exactly the covariates that were in the true model.

An interesting observation is thus that even though FDR and FDP based multiple testing procedures are normally considered to be mild in the context of testing hypotheses, they were stricter than the conventional model selection methods in the current context. The significance levels used in the multiple testing procedures, $\alpha_{\text{FDR}} = 0.25$ and $\alpha_{\text{FDP}} = 0.05$, can however be increased to obtain milder procedures.

Note that the estimated FDR and FWER are very closely connected. This is unsurprising, since the FDR is bounded by the FWER (Goeman and Solari, 2014). We see that the AIC and lasso methods yielded models where a relatively large proportion of the included covariates were irrelevant, corresponding to the large FWER and FDR estimates observed. Notice that even though the B&H procedure was used to control the FDR at level $\alpha_{\text{FDR}} = 0.25$, the estimated FDR seemed to indicate that the procedure was conservative. Table 5.10 does however show that when fewer covariates were actually relevant, the estimated FDR was closer to α_{FDR}. This is because the B&H procedure controls the FDR at level $\alpha_{\text{FDR}} \cdot m_0/m$, where m_0 is the number of true hypotheses, or irrelevant covariates (Goeman and Solari, 2014).

While inference on a reduced model is problematic, the lower confidence bound for the number of relevant covariates does give us a little piece of information about our resulting model in the case where we use only the confidence method for model selection. The lower

Few	FWER	FWER-II	FDR	Perf.	Comp.	MSE
None - AIC	0.804	0.150	0.376	0.167	3.836	1.0202
None - Lasso	0.841	**0.128**	0.471	0.104	5.348	1.0175
Conf. - None	0.210	0.265	0.107	0.574	2.701	1.0163
Conf. - AIC	**0.191**	0.279	**0.088**	**0.587**	2.494	**1.0159**
Conf. - Lasso	0.208	0.267	0.105	0.574	2.675	**1.0159**
B&H - None	0.463	0.218	0.196	0.393	3.015	1.0172
B&H - AIC	0.457	0.218	0.191	0.399	2.971	1.0171
B&H - Lasso	0.463	0.218	0.196	0.393	3.014	1.0168
None - None	1.000	0.000	0.793	0.000	12.000	1.0279
Truth					2.486	1.0014

Many	FWER	FWER-II	FDR	Perf.	Comp.	MSE
None - AIC	0.620	0.299	0.125	0.273	6.937	1.0224
None - Lasso	0.966	**0.120**	0.330	0.019	9.637	1.0230
Conf. - None	0.274	0.334	0.068	0.526	6.411	1.0216
Conf. - AIC	**0.232**	0.366	**0.050**	**0.536**	6.143	1.0216
Conf. - Lasso	0.274	0.334	0.068	0.526	6.407	**1.0213**
B&H - None	0.545	0.319	0.112	0.280	6.810	1.0228
B&H - AIC	0.508	0.328	0.101	0.308	6.682	1.0225
B&H - Lasso	0.545	0.319	0.112	0.280	6.810	1.0224
None - None	1.000	0.000	0.460	0.000	12.000	1.0257
Truth					6.481	0.9988

Table 5.10: Model selection results for 2000 simulations in each part, comparing the case with few covariates in the true model to the case with many. The best results within the different categories are marked in bold text, the worst are underlined. The mean of the lower confidence bound of the number of relevant covariates chosen by the confidence method was 2.034 for the experiments with few relevant covariates and 5.430 in the experiments with many.

bound given for the confidence method was on average a bit smaller than the complexity of the true model, 3.732 and 4.483, respectively. If we were to use FWER control, the number of rejected hypotheses would be smaller than or equal to the bound from the confidence method, since we would ignore any defining rejection with more than one elementary component. This seems to indicate that FWER control would be too strict, since we would almost surely ignore a lot of relevant covariates.

A good portion of the covariates selected by the confidence method seemed to be truly relevant ones, despite the pessimistic lower confidence bound, as we observed small values for the FDR. The small lower confidence bound did however indicate that the confidence method might have been a little strict, as the confidence method came with a cost of a higher FWER-II estimate compared to the other methods. Ultimately the method seemed to balance type I and type II errors well, since it achieved the best scores both for the test MSE and the rate at which the method selected the correct set of covariates.

In Table 5.10 one can observe the differences in the results when there were either few (1, 2, 3 or 4) or many (5, 6, 7 or 8) covariates in the true model. Table C.2 in the

Appendix shows the widths of approximate 95% confidence intervals for the probabilities or the expectations of the same values. The interval estimates are based on half as many observations as the mean across all simulations, and correspondingly have slightly larger widths (except for complexity, which naturally has a smaller variance when separated by size).

When there were few relevant covariates, all the methods performed significantly better than the full model with respect to the test MSE. The differences between the methods were also large, with the confidence method coming out on top, with respect to all but the FWER-II estimate.

When there were many relevant covariates, there were fewer type I errors to commit. As a consequence the FDR values were smaller for every method, compared to when there were few relevant covariates. As more covariates were relevant, the need for model selection was smaller. All methods performed worse than when few covariates were relevant, although the difference was largest for the hypothesis testing procedures. The AIC actually had a smaller FWER estimate when there were many relevant covariates, unlike the other procedures, although its FWER-II estimate was increased more. Naturally, the FWER-II estimate was larger for all methods when more relevant covariates existed that the methods could fail to include.

For the case with many relevant covariates, the confidence interval width of 0.003 for the MSE tells us that there were no significant differences between the different procedures with respect to this criterion, and barely any difference from using the full model.

The comparison of the experiments with small (absolute values between 0.05 and 0.25) and large (0.25 and 0.50) coefficients can be seen in Table 5.11. Table C.3 shows the widths of the corresponding approximate confidence intervals. It is very noticeable that when the coefficients were large enough, hardly any of the methods dropped any relevant covariates. Their effects were easily noticed by all methods, and thus the type I errors became more important to focus on. This gave the hypothesis testing methods an advantage, especially the confidence method. The confidence bound of 4.541 for the confidence method was very close to the complexity of its chosen models, 4.624, meaning there were few dissonant rejections, and the method was close to having FWER control. This can also be observed in the FWER estimate, which was very close to $\alpha_{\text{FDP}} = 0.05$.

When the coefficients were small, the FWER-II estimates were much larger for all methods, since the effects of the covariates were harder to detect. The confidence method gave a small lower bound, 2.924, while its average model complexity, 4.488, was similar to that of the true model. This means that the defining rejections generally had many components, and the large FWER-II estimate indicates that the selected covariates were not always the relevant ones. This suggests that the confidence method was too strict in this case. The B&H procedure, though only slightly milder, obtained significantly better results with respect to type II errors, rate of selecting the perfect model, and test MSE, though the lasso and AIC also performed well with respect to the latter.

A natural concern arises from the observations in Table 5.11, namely whether the models considered when the coefficients were large are representative of naturally occurring relationships between covariates and response variables. These models were used in half of the simulations, and thus impacted the overall results seen in Table 5.9 greatly. The apparent advantage the confidence method had over conventional methods in terms of test

Small	FWER	FWER-II	FDR	Perf.	Comp.	MSE
None - AIC	0.720	0.449	0.268	0.144	5.086	1.0195
None - Lasso	0.876	**0.248**	0.385	0.055	7.133	**1.0170**
Conf. - None	0.433	0.595	0.156	0.154	4.488	1.0218
Conf. - AIC	**0.372**	0.642	**0.119**	0.176	4.032	1.0214
Conf. - Lasso	0.431	0.598	0.153	0.155	4.457	1.0213
B&H - None	0.476	0.537	0.155	0.206	4.447	1.0187
B&H - AIC	0.459	0.546	0.148	**0.213**	4.369	1.0186
B&H - Lasso	0.476	0.537	0.155	0.206	4.446	1.0182
None - None	1.000	0.000	0.627	0.000	12.000	1.0237
Truth					4.473	0.9975

Large	FWER	FWER-II	FDR	Perf.	Comp.	MSE
None - AIC	0.705	**0.000**	0.233	0.295	5.687	1.0232
None - Lasso	0.931	**0.000**	0.416	0.068	7.851	1.0234
Conf. - None	0.051	0.004	0.020	**0.946**	4.624	1.0161
Conf. - AIC	**0.050**	0.004	**0.019**	**0.946**	4.604	**1.0160**
Conf. - Lasso	0.051	0.004	0.020	**0.946**	4.624	**1.0160**
B&H - None	0.532	**0.000**	0.153	0.468	5.378	1.0213
B&H - AIC	0.506	**0.000**	0.143	0.494	5.284	1.0210
B&H - Lasso	0.532	**0.000**	0.153	0.468	5.378	1.0209
None - None	1.000	0.000	0.626	0.000	12.000	1.0299
Truth					4.494	1.0028

Table 5.11: Model selection results for 2000 simulations in each part, comparing the case with small coefficients in the model to the case with large. The best results within the different categories are marked in bold text, the worst are underlined. The mean of the lower confidence bound of the number of relevant covariates chosen by the confidence method was 2.924 for the experiments with small coefficients and 4.541 in the experiments with large.

Uncorrelated	FWER	FWER-II	FDR	Perf.	Comp.	MSE
None - AIC	0.704	0.200	0.246	0.228	5.383	1.0197
None - Lasso	0.899	**0.107**	0.395	0.066	7.464	1.0188
Conf. - None	0.197	0.305	0.069	0.566	4.369	1.0178
Conf. - AIC	**0.191**	0.315	**0.063**	**0.567**	4.292	1.0177
Conf. - Lasso	0.197	0.305	0.069	0.566	4.369	**1.0176**
B&H - None	0.524	0.237	0.156	0.337	4.971	1.0185
B&H - AIC	0.512	0.240	0.152	0.347	4.921	1.0184
B&H - Lasso	0.524	0.237	0.156	0.337	4.971	1.0181
None - None	1.000	0.000	0.628	0.000	12.000	1.0252
Truth					4.470	0.9999
Correlated	FWER	FWER-II	FDR	Perf.	Comp.	MSE
None - AIC	0.721	0.249	0.255	0.211	5.390	1.0229
None - Lasso	0.908	**0.141**	0.406	0.057	7.521	1.0216
Conf. - None	0.287	0.293	0.107	0.533	4.743	1.0201
Conf. - AIC	**0.232**	0.330	**0.075**	**0.555**	4.345	1.0198
Conf. - Lasso	0.285	0.295	0.104	0.534	4.712	**1.0197**
B&H - None	0.484	0.300	0.152	0.337	4.854	1.0215
B&H - AIC	0.453	0.306	0.140	0.360	4.732	1.0212
B&H - Lasso	0.484	0.300	0.152	0.337	4.853	1.0211
None - None	1.000	0.000	0.625	0.000	12.000	1.0284
Truth					4.497	1.0003

Table 5.12: Model selection results for 2000 simulations in each part, comparing the case with uncorrelated covariates to the case with correlated ones. The best results within the different categories are marked in bold text, the worst are underlined. The mean of the lower confidence bound of the number of relevant covariates chosen by the confidence method was 3.756 for the experiments with uncorrelated covariates and 3.709 in the experiments with correlated.

MSE disappears if the simulations with large coefficients are ignored. It should however be noted that in the case where the coefficients were small the confidence method still selected fewer irrelevant covariates and overall selected exactly the correct ones slightly more often than the conventional methods.

What constitutes the optimal balance between type I and type II errors is evidently dependent on the underlying model. We see that while the balance greatly favoured the strictest method when the coefficients were large, the milder methods performed better when the coefficients were small. The B&H procedure appeared comparable to or better than subset selection with AIC and lasso in most of the experiments, and especially when there were few relevant covariates, as seen in Table 5.10.

The comparison of the results from the experiments with uncorrelated and correlated covariates is shown in Table 5.12, with the corresponding widths of approximate confidence intervals in Table C.4 in the Appendix. The differences between the two cases were generally small, except that all the methods performed worse with respect to the test MSE in the case where the covariates were correlated. The relative performance between the methods was the same in the two cases. With correlated covariates the confidence method resulted in slightly higher complexities and a slightly smaller confidence bound for the number of relevant covariates, suggesting that the defining rejections in general had more elementary components. Otherwise the methods seemed equally capable of identifying relevant covariates whether they were correlated or not.

Note that we have not looked into what specific effects the different correlations have in the case where the covariates are correlated. Goeman et al. (2011) mentioned that unlike other selection methods, the confidence method will not differentiate between two highly correlated covariates in its selection, but instead conclude that both are important. We have not concerned ourselves with this effect, and only looked at the overall effects having correlated covariates had on the measured results.

The largest disadvantage associated with the confidence method is its computational cost. When compared to subset selection with AIC this is not an issue, as all reduced models have to be fitted anyway. It is however possible to adjust the subset selection method to reduce computational time, for example by only considering models up to a certain complexity. This cannot be done with the F-test based confidence method, because we need to test every possible intersection of hypotheses. The lasso regression and B&H procedure do not share this issue. Restricting the original set of covariates with the B&H procedure can even be used to reduce the computational time for the AIC, as then only a reduced number of models have to be fitted.

The simulation results suggest that subset selection with AIC was quite robust to variations in the underlying model. Its performance was very similar whether there were few or many covariates, and was generally only slightly worse when the coefficients were large or the covariates were correlated, compared to when the coefficient were small or the covariates uncorrelated, respectively. With respect to identifying the relevant covariates it did however appear to be too mild, as it generally selected rather large portions of irrelevant covariates.

The lasso regression generally achieved good test MSE values, even though the error rate estimates seem suboptimal. The method often included a large number of irrelevant covariates, seen by the large FDR values. The shrinking of the coefficients did however seem to limit the negative impact of including these, since the test MSE was generally small. The considerations of error rates based on whether the coefficients are exactly zero or not might be unfair to the lasso regression, as some coefficients may be essentially ignored by having small, non-zero coefficients that contribute little to the predictions. Therefore, if the purpose of the model selection is not to make good predictions, but to accurately identify relevant covariates, lasso regression might not be perfectly suited for the task. Some adjustments can be done, however, such as choosing a slightly larger value for λ than the one that minimizes the estimated MSE, though this was not considered for this thesis.

The confidence method performed very well in all the experiments that used large coefficients for the covariates in the true model, and as such appeared to be a good method for specifically identifying covariates with large effects on the response. When the coefficients were small, the method appeared too strict. With real data expert knowledge can be applied to make a more informed selection of covariates, rather than relying on an algorithm. It does however appear from our simulations that the choice generally should not be stricter than ours. On the other hand a milder choice, selecting a larger subset than our algorithm, does not increase the lower confidence bound for the number of relevant covariates, and hence would not use all information gained from dissonant rejections.

We used a fixed value of $\alpha_{\text{FDR}} = 0.25$ for the B&H procedure, and cannot conclude whether or not this was optimal. Note that using for example cross-validation to select α_{FDR} will interfere with the validity of the FDR control. It could potentially improve the predictions, but the procedure must then be treated differently than as a form of correction for multiple hypothesis testing. The B&H procedure seemed generally promising for the purpose of model selection, both with respect to test MSE and selecting correct covariates. The procedure was milder than the confidence method, which was detrimental when the coefficients were large. It did however achieve better results with respect to test MSE when the coefficients were small. The B&H procedure was also more stable with respect to variations in the underlying model.

Another observation regarding the confidence method and the B&H procedure is that the resulting models very rarely changed when using lasso regression or subset selection with AIC afterwards. The former almost never lead to additional rejections, although the regularization of the coefficient estimates generally reduced the test MSE. The latter lead to a small amount of additional rejections, and also a general reduction of the test MSE. However, none of the results indicate that using the conventional methods after the initial application of the hypothesis testing procedures provides a significant advantage over using only the testing procedures.

Since the multiple testing procedures were stricter than the conventional methods in our simulations, this observation makes sense, and hints that a reversal of the roles could have been better. We have however already mentioned the downside to this, namely that if we use the hypothesis testing methods after the set of covariates has been reduced by for example AIC as part of one selection process, the p-values are not necessarily valid.

Chapter 6

Conclusions

6.1 Pairwise comparisons of means

The simulation results clearly indicate that in the case of pairwise comparisons of three means, a closed testing procedure should be used over both Bonferroni correction and Tukey's procedure. Bonferroni correction for the elementary hypotheses is a weaker multiple testing procedure than the closed testing procedure using Bonferroni's method to test the global hypothesis, and this was the closed testing procedure with the worst performance in the simulations. Tukey's procedure is weaker than the closed testing procedure using the range test for the global hypothesis, and performed worse than all the other procedures tested, especially with respect to the all-pairs power. Using Tukey's procedure did result in a large any-pair power, but since the tests for the remaining pairs were weaker than those used by the other methods, it more often lead to contradictory results. Since computational cost is a non-issue when there only is one additional hypothesis to test, closed testing procedures should be preferred in this scenario.

Regarding the different closed testing procedures, Simes is uniformly stronger than Bonferroni, and performed slightly better for all the experiments. The estimates for the any-pair and all-pairs powers of Simes were consistently smaller than those of the F-test, though the latter caused contradictory results slightly more often. Comparing the F-test to the range test, we see that the F-test achieved a slightly larger all-pairs power, and that the range test generally achieved a slightly larger any-pair power. In some cases the F-test was superior with respect to both kinds of power, though overall the differences were hardly significant. The range test caused contradictory results slightly more often. All in all the common F-test appears to be common for a reason, although there is little to lose from using the range test instead. This conclusion corresponds to the one made by Shaffer (1981).

Though the F-test is consonant, an adjustment can theoretically be made to increase the power of the corresponding closed testing procedure. Some of the critical region for the F-test can be removed through consonantization (2.4.1), and we saw in Table 5.2 that this would have resulted in a FWER estimate significantly smaller than $\alpha = 0.05$. This means that the consonantized procedure is conservative, and we can thus add points to the critical region without sacrificing FWER control. If we choose to add points that are also in the critical region of at least one of the local t-tests for the elementary hypotheses, the overall power of the procedure will increase. This example is three-dimensional, though very similar to the two-dimensional example presented by Romano et al. (2011).

The simplification of the closed testing procedure in the case $m = 3$ is well known, and, though interesting, only applies to a very small set of multiple comparisons studies. Closed testing procedures for larger m are also simplified, and it may very well be that the optimal local tests for these cases are different from when $m = 3$.

6.2 Multiple testing for model selection in regression

We experienced that the multiple testing procedures of FDR control with the Benjamini & Hochberg procedure and the FDP based method by Goeman et al. (2011) were stricter than conventional model selection methods of lasso regression and subset selection with AIC, resulting in sparser models. The confidence method was the strictest, and only performed better than conventional methods with respect to test MSE in a selection of the experiments. The test MSE of the B&H procedure was generally comparable to or better than that of the conventional methods. The overall optimal balance between type I and type II errors with respect to making good predictions thus appears to be somewhere around what the B&H procedure achieved. This seems to confirm our initial suspicion that FWER control is too strict for model selection in the general case.

Though it is questionable whether or not the models created with large coefficients (or any of the models, for that matter) are representable of naturally occurring covariate-response relationships, it shows that if the goal of the model selection is to indentify only the covariates which largely impact the response, a strict selection procedure is beneficial. Here the confidence method comes with the added benefit that the final selection is up to the user, who can make even stricter decisions than those our algorithm performed, if wanted or needed.

The confidence method in particular was unstable with regards to variation in the underlying model when it came to achieving a good test MSE. The B&H procedure and subset selection with AIC fared better at this regard, as their MSE values were generally more stable for the different experiments. Lasso regression resulted in good MSE values especially when the coefficients were small. Even though the method often had non-zero coefficients for irrelevant covariates, the shrinkage of the coefficients negated some of the negative impact these would otherwise have on the test MSE.

Overall the conclusion regarding the confidence method is that it was better than the conventional selection methods with respect to specifically identifying covariates that affect the response largely, but its predictions appeared less robust when it came to variations in the underlying model. The B&H procedure was generally better than the conventional methods at identifying relevant covariates. The resulting test MSE values for these methods were also comparable. The fact that the multiple testing procedures were stricter than the conventional methods meant that the second step in the selection often had very little effect. Performing the conventional methods before the testing procedures could be an interesting approach, but this requires the two steps to be separate experiments, in order to retain the validity of the testing procedures.

The fairly promising results should be more thoroughly investigated. The perhaps biggest unanswered question is whether or not the simulated models are representative of real covariate–response relationships, and thus if the same conclusions will apply when the methods are tested on actual datasets. The problem in that scenario is of course that there is no way to know what the true model actually is, though the ability to make predictions on an untouched dataset can still be tested. For this particular purpose it does however seem that the confidence method may be too strict, as the importance of avoiding type II errors appears equally important to or even more important than avoiding type I errors. Testing milder levels of α_{FDP}, and also other values of α_{FDR}, could be interesting. The effects of correlated covariates should also be investigated more specifically.

Bibliography

Benjamini, Y., Hochberg, Y., 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the Royal statistical society: series B (Methodological) 57, 289–300.

Casella, G., Berger, R.L., 2002. Statistical inference. 2 ed., Duxbury Pacific Grove, CA.

Fisher, R.A., 1935. The design of experiments, oliver and boyd ltd. London, UK .

Gabriel, K.R., 1969. Simultaneous test procedures–some theory of multiple comparisons. The Annals of Mathematical Statistics , 224–250.

Goeman, J.J., Solari, A., 2014. Multiple hypothesis testing in genomics. Statistics in medicine 33, 1946–1978.

Goeman, J.J., Solari, A., et al., 2011. Multiple testing for exploratory research. Statistical Science 26, 584–597.

Hastie, T., Tibshirani, R., Friedman, J., 2001. The elements of statistical learning. volume 1. Springer.

Hochberg, Y., 1988. A sharper bonferroni procedure for multiple tests of significance. Biometrika 75, 800–802.

Holm, S., 1979. A simple sequentially rejective multiple test procedure. Scandinavian journal of statistics , 65–70.

Hommel, G., 1988. A stagewise rejective multiple test procedure based on a modified bonferroni test. Biometrika 75, 383–386.

Jaccard, J., Becker, M.A., Wood, G., 1984. Pairwise multiple comparison procedures: A review. Psychological Bulletin 96, 589.

James, G., Witten, D., Hastie, T., Tibshirani, R., 2013. An introduction to statistical learning. volume 112. Springer.

Johnson, C., 1970. Positive definite matrices. The American Mathematical Monthly 77, 259–264.

Keuls, M., 1952. The use of the „studentized range" in connection with an analysis of variance. Euphytica 1, 112–122.

Kincaid, D., Kincaid, D.R., Cheney, E.W., 2009. Numerical analysis: mathematics of scientific computing. volume 2. American Mathematical Soc.

Marcus, R., Eric, P., Gabriel, K.R., 1976. On closed testing procedures with special reference to ordered analysis of variance. Biometrika 63, 655–660.

Newman, D., 1939. The distribution of range in samples from a normal population, expressed in terms of an independent estimate of standard deviation. Biometrika 31, 20–30.

Romano, J.P., Shaikh, A., Wolf, M., 2011. Consonance and the closure method in multiple testing. The International Journal of Biostatistics 7, 1–25.

Shaffer, J.P., 1981. Complexity: an interpretability criterion for multiple comparisons. Journal of the American Statistical Association 76, 395–401.

Shaffer, J.P., 1995. Multiple hypothesis testing. Annual review of psychology 46, 561–584.

Simes, R.J., 1986. An improved bonferroni procedure for multiple tests of significance. Biometrika 73, 751–754.

Sonnemann, E., 2008. General solutions to multiple testing problems. Biometrical Journal: Journal of Mathematical Methods in Biosciences 50, 641–656.

Sonnemann, E., Finner, H., 1988. Vollständigkeitssätze für multiple testprobleme, in: Multiple Hypothesenprüfung/Multiple Hypotheses Testing. Springer, pp. 121–135.

Walpole, R.E., Myers, R.H., Myers, S.L., Ye, K., 2016. Probability and Statistics for Engineers and Scientists. 10 ed., Pearson.

Appendix A

Proofs

A.1 Lower confidence bounds by the confidence method

The set $\mathcal{D}$, of all elementary hypotheses involved in a defining rejection of a closed testing procedure (see Section 3.2), has the same lower confidence bound for the number of false hypotheses as the set $\mathcal{H} = \{H_1, H_2, \ldots, H_m\}$ of all elementary hypotheses. Expressed with the notation used in Section 2.5:

$$|\mathcal{D}| - t_\alpha(\mathcal{D}) = |\mathcal{H}| - t_\alpha(\mathcal{H}),$$

where $t_\alpha(\mathcal{R})$ is the size of the largest subset of $\mathcal{R}$ for which the corresponding intersection of elementary hypotheses is not rejected (the upper confidence bound for the number of true hypotheses in $\mathcal{R}$).

Proof. Suppose we have a closed testing procedure for a family of elementary hypotheses $\mathcal{H}$ and its closure. Recall that a defining rejection is a rejected hypothesis with no rejected proper components. If H_J is a defining rejection and a component of H_I, we call it a *defining component* of H_I.

If H_I has a defining component, the coherence of the closed testing procedure tells us that H_I is also rejected. If H_I is rejected, it either is a defining rejection, or it has a proper component that is. Either way it has a defining component. Thus a hypothesis is rejected if and only if it has a defining component.

Let $M = \{1, 2, \ldots, m\}$, D be the set of indices of hypotheses in $\mathcal{D}$, and $E = M \setminus D$. Note that since E contains no index of a hypothesis involved in a defining rejection, H_E has no defining components, and is not rejected.

If there are no defining rejections, there are no rejections at all, and so $|\mathcal{D}| - t_\alpha(\mathcal{D}) = |\mathcal{H}| - t_\alpha(\mathcal{H}) = 0$. If $\mathcal{D} = \mathcal{H}$ the statement also obviously holds. Otherwise D and E are both proper, nonempty subsets of M.

Let J be the largest subset of D such that H_J is not rejected, and let K be the largest subset of M such that H_K is not rejected.

Note that $K \cap E = E$. Otherwise there would be an $e \in E \setminus K$, that could be added to K to create a larger set for which the corresponding hypothesis is not rejected, since H_K has no defining components and H_e is not involved in any defining rejection. Thus $|K| = |K \cap D| + |K \cap E| = |K \cap D| + |E|$. We want to show that

$$|K \cap D| = |J|.$$

Assume that $|K \cap D| > |J|$. By the definition of J, $H_{K\cap D}$ must be rejected. By coherence so is H_K, which is a contradiction to the definition of K. Note that the possibility of $J = \emptyset$ is covered by this case.

Assume that $|K \cap D| < |J|$. Consider $H_{J\cup E}$. This hypothesis must be rejected, otherwise it contradicts the definition of K. This means that $H_{J\cup E}$ has a defining component. Since E contains no indices of hypotheses involved in defining rejections, this means that the defining component must also be a component of H_J, which is a contradiction since H_J by definition is not rejected.

Thus we have $|\mathcal{H}| - t_\alpha(\mathcal{H}) = |M| - |K| = (|D| + |E|) - (|J| + |E|) = |D| - |J| = |\mathcal{D}| - t_\alpha(\mathcal{D})$. □

A.2 Closed testing with Bonferroni local tests in multiple linear regression is consonant

Here we present a short proof that in the case where no intersections of elementary hypotheses coincide, the closed testing procedure obtained by using the Bonferroni method to construct local tests is consonant.

Proof. Consider the elementary hypotheses $H_1, H_2, \ldots, H_m$, and let $M = \{1, 2, \ldots, m\}$. Suppose the closed testing procedure rejects H_I, where $I \neq \emptyset$ and $I \subset M$.

Since the closed testing procedure is coherent, H_J is rejected for all J such that $I \subset J \subset M$. Since H_I is rejected, it must be the case that for some $i \in I$, $p_i \leq \alpha/|I|$. Now let $\{i\} \subset K \subset I$. We have $|K| \leq |I|$, and so $p_i \leq \alpha/|I| \leq \alpha/|K|$. Thus H_K is rejected for all $\{i\} \subset K \subset I$ and H_J is rejected for all $I \subset J \subset M$, which means H_i is rejected by the closed testing procedure, since H_L is rejected for all $\{i\} \subset L \subset M$. □

Appendix B

Construction of a semi-arbitrary covariance matrix

To create an arbitrary covariance matrix, we needed to construct a positive definite matrix Σ. A symmetric, real matrix Σ is postive definite if and only if all its eigenvalues are positive, in which case $\Sigma = P^T DP$, where P is an orthogonal matrix and D is a diagonal matrix with the postive eigenvalues of A on its diagonal (Johnson, 1970).

We found an orthogonal matrix P by generating an m by m matrix M of m^2 samples from a standard normal distribution, and applying the Gram–Schmidt process (Kincaid et al., 2009). This works as long as M has full rank, which is likely. We then let $D = \text{diag}(\lambda_1, \lambda_2, \ldots, \lambda_m)$, where the λs were evenly spaced between $3/2$ and $2/3$ (not for any particular reason, except that we needed values different from all being equal to 1, since that would yield the identity matrix).

Thus $P^T DP$ was positive definite, and thus a covariance matrix. We chose to use the corresponding correlation matrix (scaling the matrix so that all diagonal elements are equal to 1), so that all the covariates would be on the same scale.

The resulting covariance matrix, which then was used in every experiment with correlated covariates in Chapter 5, can be seen in Table B.1. Most covariates were not particularly correlated, and the largest correlation coefficients were $\rho_{14} = -0.29$ and $\rho_{48} = 0.30$.

	1	2	3	4	5	6	7	8	9	10	11	12
1	1.00	0.18	0.13	−0.29	−0.03	−0.05	0.16	−0.02	0.00	−0.21	0.03	−0.17
2	0.18	1.00	0.18	−0.17	0.10	0.04	0.21	−0.01	0.05	−0.09	−0.24	−0.02
3	0.13	0.18	1.00	−0.02	−0.26	−0.11	−0.11	0.14	−0.08	−0.03	−0.06	0.15
4	−0.29	−0.17	−0.02	1.00	0.04	0.13	−0.09	0.30	0.06	−0.07	0.03	0.01
5	−0.03	0.10	−0.26	0.04	1.00	0.04	−0.06	−0.01	−0.12	−0.22	0.07	−0.07
6	−0.05	0.04	−0.11	0.13	0.04	1.00	0.08	−0.07	−0.12	0.10	−0.11	−0.15
7	0.16	0.21	−0.11	−0.09	−0.06	0.08	1.00	−0.13	−0.31	0.01	−0.16	−0.19
8	−0.02	−0.01	0.14	0.30	−0.01	−0.07	−0.13	1.00	0.22	−0.05	−0.06	−0.14
9	0.00	0.05	−0.08	0.06	−0.12	−0.12	−0.31	0.22	1.00	0.07	−0.14	−0.08
10	−0.21	−0.09	−0.03	−0.07	−0.22	0.10	0.01	−0.05	0.07	1.00	−0.06	0.13
11	0.03	−0.24	−0.06	0.03	0.07	−0.11	−0.16	−0.06	−0.14	−0.06	1.00	0.14
12	−0.17	−0.02	0.15	0.01	−0.07	−0.15	−0.19	−0.14	−0.08	0.13	0.14	1.00

Table B.1: Covariance matrix for the correlated covariates used in simulations in Chapter 5.

Appendix C

Additional simulation results in model selection

This section contains tables showing the width of the approximate confidence intervals for the values shown in Chapter 5. The estimates that are for probabilities, “FWER”, “FWER-II” and “Perf.” are calculated according to (5.1) (note that we assume that the probability of a given event is constant for all simulations, which is not likely to be the case, as for example type II errors are less likely when the coefficients are large), while the remaining values are assumed to be normally distributed. We also assume that the number of observations is large enough that the t-distribution can be approximated by a standard normal distribution, resulting in width

$$2 \cdot 1.96 \frac{\hat{\sigma}}{n},$$

where $\hat{\sigma}$ is the sample standard error and n is the number of observations used for the calculation of the estimated mean that the approximate 95% confidence interval corresponds to. The complexity is certainly not normally distributed, since it is discrete. We did however care more about the error and perfection rates, as well as the test MSE, and were thus not particularly concerned with the confidence interval estimate for the complexity.

	FWER	FWER-II	FDR	Perf.	Comp.	MSE
None - AIC	0.014	0.013	0.007	0.013	0.066	0.0020
None - Lasso	0.009	0.010	0.007	0.007	0.102	0.0020
Conf. - None	0.013	0.014	0.006	0.015	0.081	0.0021
Conf. - AIC	0.013	0.014	0.005	0.015	0.075	0.0021
Conf. - Lasso	0.013	0.014	0.006	0.015	0.080	0.0021
B&H - None	0.015	0.014	0.006	0.015	0.080	0.0020
B&H - AIC	0.015	0.014	0.006	0.015	0.077	0.0020
B&H - Lasso	0.015	0.014	0.006	0.015	0.080	0.0020
None - None	0.000	0.000	0.006	0.000	0.000	0.0021
Truth					0.071	0.0020

Table C.1: Widths of approximate 95% confidence intervals for the expectations or probabilities from all simulations shown in Table 5.9.

Few	FWER	FWER-II	FDR	Perf.	Comp.	MSE
None - AIC	0.017	0.016	0.011	0.016	0.066	0.0029
None - Lasso	0.016	0.015	0.011	0.013	0.131	0.0029
Conf. - None	0.018	0.019	0.010	0.022	0.085	0.0029
Conf. - AIC	0.017	0.020	0.009	0.022	0.071	0.0029
Conf. - Lasso	0.018	0.019	0.010	0.022	0.083	0.0029
B&H - None	0.022	0.018	0.011	0.021	0.077	0.0029
B&H - AIC	0.022	0.018	0.010	0.021	0.074	0.0029
B&H - Lasso	0.022	0.018	0.011	0.021	0.077	0.0029
None - None	0.000	0.000	0.004	0.000	0.000	0.0029
Truth					0.049	0.0028
Many	**FWER**	**FWER-II**	**FDR**	**Perf.**	**Comp.**	**MSE**
None - AIC	0.021	0.020	0.005	0.020	0.061	0.0029
None - Lasso	0.008	0.014	0.006	0.006	0.083	0.0029
Conf. - None	0.020	0.021	0.006	0.022	0.075	0.0029
Conf. - AIC	0.018	0.021	0.004	0.022	0.068	0.0029
Conf. - Lasso	0.020	0.021	0.006	0.022	0.075	0.0029
B&H - None	0.022	0.020	0.005	0.020	0.075	0.0029
B&H - AIC	0.022	0.021	0.005	0.020	0.072	0.0029
B&H - Lasso	0.022	0.020	0.005	0.020	0.075	0.0029
None - None	0.000	0.000	0.004	0.000	0.000	0.0029
Truth					0.049	0.0028

Table C.2: Widths of approximate 95% confidence intervals for the expectations or probabilities from all simulations where the true model contained few or many covariates, respectively, shown in Table 5.10.

Small	FWER	FWER-II	FDR	Perf.	Comp.	MSE
None - AIC	0.020	0.022	0.011	0.015	0.088	0.0028
None - Lasso	0.014	0.019	0.010	0.010	0.149	0.0028
Conf. - None	0.022	0.022	0.010	0.016	0.124	0.0029
Conf. - AIC	0.021	0.021	0.008	0.017	0.108	0.0029
Conf. - Lasso	0.022	0.021	0.009	0.016	0.124	0.0029
B&H - None	0.022	0.022	0.009	0.018	0.110	0.0029
B&H - AIC	0.022	0.022	0.009	0.018	0.106	0.0029
B&H - Lasso	0.022	0.022	0.009	0.018	0.110	0.0029
None - None	0.000	0.000	0.008	0.000	0.000	0.0029
Truth					0.100	0.0027

Large	FWER	FWER-II	FDR	Perf.	Comp.	MSE
None - AIC	0.020	0.000	0.010	0.020	0.096	0.0029
None - Lasso	0.011	0.000	0.009	0.011	0.138	0.0029
Conf. - None	0.010	0.003	0.004	0.010	0.103	0.0029
Conf. - AIC	0.010	0.003	0.004	0.010	0.102	0.0029
Conf. - Lasso	0.010	0.003	0.004	0.010	0.103	0.0029
B&H - None	0.022	0.000	0.008	0.022	0.112	0.0029
B&H - AIC	0.022	0.000	0.008	0.022	0.109	0.0029
B&H - Lasso	0.022	0.000	0.008	0.022	0.112	0.0029
None - None	0.000	0.000	0.008	0.000	0.000	0.0030
Truth					0.100	0.0028

Table C.3: Widths of approximate 95% confidence intervals for the expectations or probabilities from all simulations where the coefficients where small or large, respectively, shown in Table 5.11.

Uncorrelated	FWER	FWER-II	FDR	Perf.	Comp.	MSE
None - AIC	0.020	0.018	0.010	0.018	0.093	0.0029
None - Lasso	0.013	0.014	0.009	0.011	0.146	0.0029
Conf. - None	0.017	0.020	0.007	0.022	0.109	0.0029
Conf. - AIC	0.017	0.020	0.007	0.022	0.106	0.0029
Conf. - Lasso	0.017	0.020	0.007	0.022	0.109	0.0029
B&H - None	0.022	0.019	0.009	0.021	0.112	0.0029
B&H - AIC	0.022	0.019	0.008	0.021	0.109	0.0029
B&H - Lasso	0.022	0.019	0.009	0.021	0.112	0.0029
None - None	0.000	0.000	0.008	0.000	0.000	0.0029
Truth					0.101	0.0028

Correlated	FWER	FWER-II	FDR	Perf.	Comp.	MSE
None - AIC	0.020	0.019	0.010	0.018	0.093	0.0029
None - Lasso	0.013	0.015	0.009	0.010	0.143	0.0029
Conf. - None	0.020	0.020	0.009	0.022	0.119	0.0029
Conf. - AIC	0.018	0.021	0.007	0.022	0.105	0.0029
Conf. - Lasso	0.020	0.020	0.009	0.022	0.118	0.0029
B&H - None	0.022	0.020	0.009	0.021	0.114	0.0029
B&H - AIC	0.022	0.020	0.008	0.021	0.109	0.0029
B&H - Lasso	0.022	0.020	0.009	0.021	0.114	0.0029
None - None	0.000	0.000	0.008	0.000	0.000	0.0029
Truth					0.100	0.0028

Table C.4: Widths of approximate 95% confidence intervals for the expectations or probabilities from all simulations where the covariates where uncorrelated or correlated, respectively, shown in Table 5.12.

Appendix D

R Code used in simulations

D.1 Pairwise comparisons of means

```
#### Functions ----
library(mvtnorm)
Simulate_Y = function(mu,Sigma,m,n,N){...}
#Creates n*N realizations of Y from N(mu,Sigma)
Y_mean = function(Ys){...}
Y_var = function(Ys){...}
Y_pooled_mean = function(Y_means){...}
Y_pooled_var = function(Y_vars){...}
#Calculates the statistics in equation (2.2) in Chapter 2.6

ttops = function(Y_means,m,n,N){...}
#The numerator for the t-test and Tukey's procedure

t_test = function(Y_means,Y_pool_var,m,n,N){...}
#Calculates the p-values for the pooled t-tests

Tukeysrej = function(Y_means,Y_pool_var,m,n,N,alpha){...}
#Determines which hypotheses Tukey's procedure rejects

F_test=function(Y_means,Y_pool_mean,Y_pool_var,m,n,N){...}
Simes_test = function(ps,m,N){...}
Bonf_test = function(ps,m,N){...}
Range_test = function(tukeys){...}
#Performs the various global tests

make_ps = function(mu,Sigma,m,n,N,alpha){...}
#Calls the above functions, returns the rejections
```

```
#from the different methods

Run_simulation = function(mus,Facts ,Sigma ,m,n ,N, alpha){
ps = make_ps(mus,Sigma ,m,n ,N, alpha)
ttest = ps$t_ps
Ftest = ps$F_ps
Simes = ps$S_ps
Bonf = ps$B_ps
Range = ps$Range_rs
TukeyR = ps$T_rs
ResultArray = array(0,dim=c(5,5))
rownames(ResultArray) = c("Bonf.","Simes","F",
                          "Range","Tukey")
colnames(ResultArray) = c("None","FWER","Any␣power",
                          "Power␣all","Contr.")
globs = rbind(Bonf < alpha,Simes < alpha,Ftest < alpha)
#Rejections of the global hypothesis

None = c(1-apply(globs,1,mean),1-mean(Range),
         1-mean(TukeyR[1,] | TukeyR[2,] | TukeyR[3,]))
trejects = ttest < alpha
#Rejections of elementary hypotheses by the t-tests

BonfR = t(t(trejects) & globs[1,])
SimesR = t(t(trejects) & globs[2,])
FR = t(t(trejects) & globs[3,])
RangeR = t(t(trejects) & Range)
#Rejections by the closed testing procedures

if (Facts[1]*(!Facts[2])*(!Facts[3])){...
} else if ((!Facts[1])*(!Facts[2])*(!Facts[3])){...
} else if (Facts[1]*Facts[2]*Facts[3]){...
} else {...}
#Calculates "None", "FWER", "Any power", "Power all",
#depending on which hypotheses are true/false

Contra = c(mean(apply(BonfR,2,sum)==1),
         mean(apply(SimesR,2,sum)==1),
         mean(apply(FR,2,sum)==1),
         mean(apply(RangeR,2,sum)==1),
         mean(apply(TukeyR,2,sum)==1))

ResultArray[,1] = None
ResultArray[,2] = FWER
ResultArray[,3] = Power1
ResultArray[,4] = PowerAll
ResultArray[,5] = Contra

return(ResultArray)
}
#### Parameters ----
sigma2 = c(1,1,1)
Sigma = diag(3)*sigma2
m = 3
n = 30
N = 10000
N0 = 100000
alpha = 0.05
#### All zero ----
set.seed(0)
mu = c(0,0,0)
All_true = Run_simulation(mu,c(T,T,T),Sigma,m,n,N0,alpha)
All_true = All_true[,-c(3,4)] #Removes the power columns
All_true
#### One different ----
num_each = 10

mu1s = rep(0,3*num_each)
mu2s = rep(0,3*num_each)
mu3s = seq(0.1,1,length.out=3*num_each)
mus = rbind(mu1s,mu2s,mu3s)
Facts = c(T,F,F)
One_diff = array(0, dim=c(5,5,3*num_each))

for (i in 1:(3*num_each)){
One_diff[,,i] = Run_simulation(mus[,i],Facts,
        Sigma,m,n,N,alpha)
}
#### All different ----
mu1s = rep(0,3*num_each)
mu2s = seq(0.1,0.7,length.out=3*num_each)
mu3s = 2*mu2s
mus = rbind(mu1s,mu2s,mu3s)
Facts = c(F,F,F)
All_diff = array(0, dim=c(5,5,3*num_each))

for (i in 1:(3*num_each)){
All_diff[,,i] = Run_simulation(mus[,i],Facts,
        Sigma,m,n,N,alpha)
}
```

D.2 Model selection in multiple linear regression

```
#### Load functions and libraries ----
library(mvtnorm)
library(stats)
library(rje)
library(cherry) #Goeman (2011)
library(glmnet)
library(prodlim)
library(gtools)

make_data = function(X,n,m,sigma,traintestsplit,
        fewfalse=TRUE,smallcoef=TRUE,intercept=0){...}
#Creates data as described in Chapter 4.2.1

analysis = function(variables,data,subsets,traintestsplit,
        intrc=FALSE,silent=FALSE){...}
#Fits full model, then calls make_all

make_all = function(subsets,hypotheses,m,data,RSStot,sig2,
        intrc=FALSE,silent=FALSE){...}
#Creates all subset models, finds all AIC values, F stats

ct_FDP = function(hypotheses,m,pvals,global_pval,alpha){..}
#Performs closed testing with the Cherry package

select_FDP=function(defining,ct,FWER=FALSE,All=FALSE){...}
#Uses defining rejections to select a subset of covariates,
#described in Chapter 3.2.4

select_FDR = function(hypotheses,pvalues,alpha){...}
#Performs the Benjamini & Hochberg procedure

do_AIC = function(subsets,AICs,hypothesesfull,hypotheses,
        data,testdata,intrc=FALSE){...}
#Performs subset selection with AIC

do_Lasso = function(hypotheses,data,testdata,k=5,
        lambda_min=FALSE,stdz=FALSE,intrc=FALSE){...}
#Performs lasso regression with k-fold CV

do_Nothing = function(hypotheses,data,testdata,
        intrc=FALSE){...}
#Fits a model with the selected covariates (hypotheses)
```

```
few_left=function(var_left,data,testdata,intrc=FALSE){...}
#Handles the case when only 1 or 0 covariates
#are left after initial selection

#### Decide parameters ----
n = 1000
m = 12
alpha_FDP = 0.05
alpha_FDR = 0.25
sigma = 1
intercept = 5
traintestsplit = n/2
lassofolds = 5
lassomin = T
intrc = (intercept != 0)
parameters = data.frame(FewFalse=c(rep(T,4),rep(F,4)),
          SmallCoef=rep(c(T,F),4),IndX=rep(c(T,T,F,F),2))
NumComb = dim(parameters)[1]
NumRuns = 500

#### Create design matrices ----
#See Appendix B
set.seed(0)

p  =  qr.Q(qr(matrix(rnorm(m^2), m)))
#Arbitrary orthogonal matrix

Sigma = cov2cor(crossprod(p,p*seq(3/2,2/3,
          length.out = m)^2))
depX = array(data=t(rmvnorm(n,rep(0,m),Sigma)),
          dim=c(m,n))
Sigma2 = diag(m)
indX = array(data=t(rmvnorm(n,rep(0,m),Sigma2)),
          dim=c(m,n))

#### NumComb*NumRuns simulations ----
ResultArray = array(0,dim=c(10,7,NumComb,NumRuns))

hypothesesfull = sprintf("X%d",seq(1:m))
subsets = powerSetMat(m)[2:(2^m-1),]
subsets = subsets*col(subsets)

for (t in 1:NumRuns){
#Run simulation
for (j in 1:NumComb){
```

```
#for each experiment
fewfalse = parameters$FewFalse[j]
smallcoef = parameters$SmallCoef[j]
if (parameters$IndX[j]){X = indX}
else {X = depX}

d = make_data(X,n,m,sigma,traintestsplit,fewfalse,
        smallcoef,intercept)
Truth = d$Truth
TestFull = d$TestFull

traindata = d$data[1:traintestsplit,]
testdata = d$data[(traintestsplit+1):n,]
d = 0
#### Do the selection ---
#No restrictions
info = analysis(hypothesesfull,traindata,subsets,
        traintestsplit,intrc=intrc,silent=TRUE)
onlyAIC = do_AIC(subsets,info$AICs,hypothesesfull,
        hypothesesfull,traindata,testdata,intrc=intrc)
onlyLasso = do_Lasso(hypothesesfull,traindata,testdata,
        k=lassofolds,lambda_min=lassomin,intrc=intrc)

#Restricting with FDP/FDR
ct = ct_FDP(hypothesesfull,m,info$all_ps,info$global_p,
        alpha_FDP)
defining = defining(ct)
defining = defining[order(sapply(defining,length),
        decreasing=F)]
FDP_keeps = mixedsort(select_FDP(defining,ct,All=TRUE))
FDR_keeps = select_FDR(hypothesesfull,info$elementary_ps,
        alpha_FDR)

#Selecting with FDP-restricted data
m_FDP = length(FDP_keeps)

if (m_FDP == m){...
} else if (m_FDP > 1){...
} else {...}
#Performs AIC, lasso or nothing on with set of covariates
#restricted by the confidence method

#Selecting with FDR-restricted data
m_FDR = length(FDR_keeps)
```

```
if (m_FDR == m){...
} else if (m_FDR == m_FDP){
if (prod(FDR_keeps == FDP_keeps)){...
} else {...}
} else if (m_FDR > 1){...
} else {...}
#Performs AIC, lasso or nothing with set of covariates
#restricted by the Benjamini & Hochberg procedure

Results = list(AIC=onlyAIC,Lasso=onlyLasso,
    FDP_None=FDP_None,FDP_AIC=FDP_AIC,FDP_Lasso=FDP_Lasso,
    FDR_None=FDR_None,FDR_AIC=FDR_AIC,FDR_Lasso=FDR_Lasso,
    Full=TestFull,Truth=Truth)

K = length(Results)
for (k in 1:K){
result = Results[[k]]
m_res = length(result$Variables)

TP = length(intersect(Truth$Variables,result$Variables))
#In resvar, in truthvar (rejected, should be)
FP = m_res - TP
#In resvar, not truthvar (rejected, shouldn't be)
FN = length(Truth$Variables) - TP
#Not in resvar, in truthvar (not rejected, should be)
TN = m-TP-FN-FP
#Not in either (not rejected, shouldn't be)

FDR = 0
if ((TP + FP) !=0){FDR = FP/(TP+FP)}
ResultArray[k,1,j,t] = (FP != 0)
ResultArray[k,2,j,t] = (FN != 0)
ResultArray[k,3,j,t] = FDR
ResultArray[k,4,j,t] = ((FP+FN) == 0)
ResultArray[k,5,j,t] = m_res
ResultArray[k,6,j,t] = pick(ct,result$Variables,silent=T)
ResultArray[k,7,j,t] = result$MSE
}}}
colnames(ResultArray) = c("FWER","FWER-II","FDR","Perf.",
"Comp.","Bound","MSE")

rownames(ResultArray) = c("None␣-␣AIC","None␣-␣Lasso",
            "Conf.␣-␣None","Conf.␣-␣AIC","Conf.␣-␣Lasso",
            "B&H␣-␣None","B&H␣-␣AIC","B&H␣-␣Lasso",
            "None␣-␣None","Truth")
```