Estimation and Testing Following Model Selection

Amit Meir

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2018

Reading Committee:

Mathias Drton , Chair

Noah Simon

Yoav Benjamini

Program Authorized to Offer Degree:
Statistics

University of Washington

Abstract

Estimation and Testing Following Model Selection

Amit Meir

Chair of the Supervisory Committee:
Mathias Drton

The field of post-selection inference focuses on developing solutions for problems in which a researcher uses a single dataset to both identify a promising set of hypotheses and conduct statistical inference. One promising heuristic for adjusting for model/hypothesis selection in inference is that of *conditioning on the selection event* (conditional inference), where the data is constrained to a subset of the sample space that guarantees the selection of a specific model. Two major obstacles to conducting valid and tractable conditional inference are that the conditional distribution of the data does not converge to a normal distribution asymptotically, and that the likelihood itself is often intractable in multivariate problems. A key idea underlying most recent works on conditional inference in regression is the polyhedral lemma which overcomes these difficulties by conditioning on information beyond the selection of a model to obtain a tractable inference procedure with finite sample guarantees. However, this extra conditioning comes at a hefty price, as it results in oversized confidence intervals and tests with less power. Our goal in this dissertation is to propose alternative approaches to conditional inference which do not rely on any extra conditioning.

First we tackle the problem of estimation following model selection. To overcome the intractable conditional likelihood, we generate noisy unbiased estimates of the post-selection score function and use them in a stochastic ascent algorithm that yields correct post-selection maximum likelihood estimates. We apply the proposed technique to the problem of estimating linear models selected by the lasso. In an asymptotic analysis the resulting estimates

are shown to be consistent for the selected parameters, and in a simulation study they are shown to offer better estimation accuracy compared to the lasso estimator in most of the simulation settings considered.

In Chapter 3 we consider the problem of inference following aggregate tests in regression. There, we formulate the polyhedral lemma for inference following model selection with aggregate tests, but also propose two alternative approaches for conducting valid post-selection inference. The first is based on conducting inference under a conservative parametrization, and the other a regime switching method which yields point-wise consistent confidence intervals by estimating the post-selection distribution of the data. In a simulation study, we show that the proposed methods control the selective type-I error rate while offering improved power.

In Chapter 4 we generalize the regime switching approach to a more general setting of conducting inference after model selection in regression. We propose a modified bootstrap approach in which we seek to consistently estimate the post-selection distribution of the data by thresholding small coefficients to zero and taking parametric bootstrap samples from the estimated conditional distribution. In an asymptotic analysis we show that the resulting confidence intervals are point-wise consistent. In a simulation study we show that our modified bootstrap procedure obtains the desired coverage rate in all simulation settings considered while producing much shorter confidence intervals with improved power to detect true signals in the selected model.

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGMENTS

In chronological order, I would like to start by thanking my parents who have supported me throughout my life, through some rough times in high-school (where my grades were not at all predictive of any academic success), very rough times in the military, and when the time came encouraged me pursue advanced degrees in academia, culminating in a move far away from home to pursue a PhD degree at the University of Washington.

Early on at the Hebrew University, I was fortunate to have many great teachers. I would like to thank Micha Mendel who's engaging teaching style lead me to change my course of studies and switch my majors from business to statistics in the first year of my academic studies. Pavel Chigansky who through his uncompromising rigorous approach to teaching statistical theory has provided me with my first major academic challenges and introduced me to the beauty of mathematical and statistical theory. Zvi Gilula who gave me the last nudge I needed in order to pursue a Master's degree, and Yosi Rinott who apart from mentoring me in writing my Master's dissertation, also shared his vast knowledge in architecture, music and philosophy.

At the University of Washington, I was fortunate to have a great teacher in Mathias Drton, and the degree of gratitude I owe him cannot be expressed in a single paragraph. Mathias was always open to discussing any and all research problems, and giving me much needed guidance and advise regarding next steps. Coming into the program, my writing and mathematical skills were somewhat raw, and getting them up to par took much time and patience on Mathias' part. Finally, I am grateful to Mathias for going out of his way to assist me in other aspects of graduate school on numerous occasions.

Special thanks go to Yoav Benjamini who in time of need gave me a home at Tel-Aviv

university, introduced me to the post-selection inference problem, and has kindly agreed to be a part of my reading committee.

I am grateful for the opportunity I had to work with Raphael Gottardo and Greg Finak at the Hutchinson Cancer Research Center, who apart from giving me a challenging and interesting problem to work on in Flow-Cytometry modeling and instructing me in the ways of clinical research, also provided the financial support that allowed me to dedicate myself to research full-time.

I was fortunate to have many great collaborators during my PhD studies. I would like to thank Ruth Heller, Nilanjan Chatterjee, Asaf Weinstein, and Yuval Benjamini, for the excellent discussions and fruitful collaborations.

I also thank my other committee members, Noah Simon, Yen-Chi Chen, Jing Tao, and Kevin Jamieson for their time and for the helpful discussions.

During five years in Seattle, I not only obtained an education, but also built a life, and I am grateful for my Seattle area friends who provided much needed distractions, companionship and opportunities to commiserate over the hardships inherent in pursing a PhD degree. Last but not least, I would like to thank my partner Ilana, who made the last few years far easier and much more enjoyable.

DEDICATION

To my parents Zipi and Yehuda, who throughout my life have actively refused to tell me what to do, and who gave me their unconditional support in whichever path I chose.

Chapter 1

INTRODUCTION

1.1 Inference after model selection

Consider the linear regression model

$$y = \boldsymbol{X}\,\beta + \varepsilon,$$

where $y \in \mathbb{R}^n$ is a response vector, $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ is a matrix of covariate values and $\varepsilon \in \mathbb{R}^n$ is a noise vector. When the number of available covariates p is large, it is often desirable or even necessary to specify a more succinct model for the data. This is commonly done by selecting a subset of the columns of $\boldsymbol{X}$ to serve as predictors for y.

Model selection in regression is a well studied problem. Frequently used methods include exhaustive search based on information criteria such as AIC (Akaike, 1973) or BIC (Schwarz, 1978), stepwise regression (Hocking, 1976), univariate screening (Fan and Lv, 2008), and regularization methods that induce sparsity via a variety of penalty functions (Hastie et al., 2015).

The topic of this dissertation is inference after model selection, a well known yet not as well understood problem. In particular, it is known that confidence intervals for parameters in selected models often do not achieve target nominal coverage rates, hypothesis tests tend to suffer from an inflated type-I error rate and point estimates are often biased. A simple Gaussian example serves well to illustrate the issues that may arise when using the same data for selection and inference.

Example 1.1. Let $Y_1, \ldots, Y_n \sim f$ i.i.d., with $\mathbb{E}_f(Y_i) = \mu$ and $\mathrm{Var}_f(Y_i) = 1$. Furthermore, suppose that estimation of μ is of interest only if a statistical test provides evidence that it

is nonzero. Specifically, suppose that at a 5%-level, we reject $H_0 : \mu = 0$ if $|\bar{y}| > 1.96/\sqrt{n}$. In this setting, if $|\mu| < 1.96/\sqrt{n}$, the uncorrected estimator $\hat{\mu} = \bar{y}$ will overestimate the magnitude of μ whenever we choose to estimate it.

An example of early work emphasizing the fact that data-driven model selection may invalidate standard inferential methods is the article by Cureton (1950), with its aptly chosen title 'validity, reliability and baloney'. Subsequently, this problem has been studied in the context of regression modeling. In particular, it has been shown that it is impossible to uniformly approximate the post-selection distribution of linear regression coefficient estimates (Leeb and Pötscher, 2005, 2006; Pötscher, 1991).

The field of post-selection inference is concerned with developing statistical methods that account for model selection in inference. *Conditional post-selection inference* is one promising approach for conducting post-selection inference, and it can be defined as:

> ***The practice of conditioning on a specific and well defined selection event in order to adjust for model selection.***

To introduce the premise of conditional post-selection inference, suppose that we observe a p dimensional random vector $y \sim f$ taking values in Ω and that we have a pre-determined model selection function $S : \Omega \to \mathcal{M}$. The model selection function selects a model M from a set of candidate models $\mathcal{M}$ based on the observed vector y, where each model comes paired with a corresponding parameter of interest θ^M. Then, were we able to construct a level q confidence interval $\text{CI}(M)$ that satisfies

$$P(\theta^M \in \text{CI}(M) | S(Y) = M) \geq q,$$

we would be able to guarantee that for any realization of y, our confidence intervals will have the correct coverage rate for the selected parameter

$$P\left(\theta^{S(y)} \in \text{CI}(S(y))\right) = \sum_{M \in \mathcal{M}} P\left(\theta^M \in \text{CI}(M) \middle| S(Y) = M\right) P(M) \geq q \sum_{M \in \mathcal{M}} P(M) = q.$$

One of the first published work utilizing conditional techniques in the context of selective inference is the seminal paper by Weinstein et al. (2013) who construct confidence intervals for selected univariate normal means. The majority of the subsequent developments in the field of conditional post-selection have been based on the *polyhedral lemma* or *affine framework* first introduced by Lee et al. (2016). In the next section we will give an overview of the conditional inference techniques in the relatively simple univariate case, and give a brief description of the polyhedral lemma. There, we will also mention some important generalizations of the polyhedral lemma as well as highlight some of its drawbacks. In Section 1.3 we will give an overview of the novel material presented in this dissertation.

1.1.1 Notation

Throughout the dissertation we will use the following set of notations. We denote a random variable by a capital letter Y and its realization in lower case y. Matrices are denoted in boldface $\boldsymbol{X}$. We use the notation $\boldsymbol{X}_{i,j}$ to denote the entry in the ith row and jth column of $\boldsymbol{X}$. Similarly, we use the notations $\boldsymbol{X}_{i,\cdot}$ and $\boldsymbol{X}_{\cdot,j}$ to denote the ith row and jth column of $\boldsymbol{X}$. We denote the jth coordinate of a vector x by x_j and by x_{-j} the vector x with the jth coordinate removed. We denote by e_j the unit vector with 1 at the jth coordinate and zeros everywhere else.

The model selection function $S : \Omega \to \mathcal{M}$ maps from the sample space of an observation Y, Ω, to a countable set of candidate models $\mathcal{M}$. We denote a specific selected model by $M \in \mathcal{M}$. When conditioning on a selection event $S(Y) = M$, we will often shorten notation by writing $P(A|M)$ in place of $P(A|S(Y) = M)$. In the context of regression, M will usually denote the set of columns of $\boldsymbol{X}$ that are included in the selected model, and in such cases we denote by $\boldsymbol{X}_M$ the sub-matrix of $\boldsymbol{X}$ consisting of the selected columns, and by $\boldsymbol{X}_{-M}$ the matrix consisting of the columns which are not included in M.

We denote the parameters of probability distributions in the usual manner using greek characters (e.g., μ, θ) and their estimates using the usual hat notation (e.g., $\hat{\mu}$, $\hat{\theta}$). In some cases it will be useful to explicitly differentiate between an arbitrary parameter value and

the true value of the parameter, in such cases we will denote the true parameter value with an asterisk (e.g. μ^*).

We denote the pdf of a random variable by $f(y)$ and its CDF by $F(y) = P(Y < y)$, regardless of whether a random variable is univariate or multivariate. We denote the pdf and CDF of a normal distribution with parameter μ and covariance $\mathbf{\Sigma}$ by $\varphi(y; \mu, \mathbf{\Sigma})$ and $\Phi(y; \mu, \mathbf{\Sigma})$, respectively.

1.2 Conditional post-selection inference

1.2.1 Conditional inference in univariate problems

Before jumping off the deep end and describe complex solutions for complex post-selection problems it may be worth asking, why do standard inference techniques fail post-selection? The problem of inferring on a single selected normal mean, while relatively simple, is sufficient to exemplify most of the problems we will run into when conducting post-selection inference in more complex settings.

Recall Example 1.1 where we observed $Y_1, \ldots, Y_n \sim N(\mu, \sigma^2)$ and estimated μ iff $|\bar{Y}| = |n^{-1} \sum_{i=1}^n Y_i| > c/\sqrt{n} > 0$ for some predetermined critical value c. The standard unbiased estimator for μ in the normal model is $\bar{Y}$. However, post-selection the distribution of $\bar{Y}$ is truncated normal and not normal, because we only observe extreme values of $\bar{Y}$. This is exemplified in Figure 1.1 where we plot the densities of some normal distribution, as well as truncated normal distributions where the observed value $\bar{y}$ is forced to fall above a threshold $c = 1$ in absolute value. So, post-model selection inference can be viewed as a model misspecification problem where we attempt to estimate a parameter of a truncated distribution based on a misspecified likelihood.

In many other statistical problems model misspecification can be resolved to some extent by relying on M-estimation theory and Sandwich type variance estimates. For such techniques to work, we require that our estimates are approximately normal for large enough sample sizes. This condition is not necessarily satisfied post-selection. For example, suppose

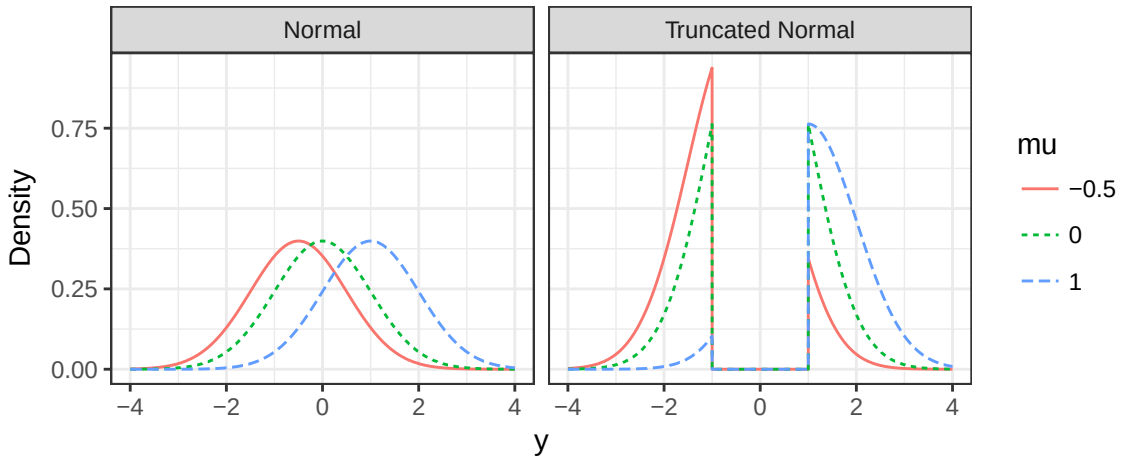

Figure 1.1: Univariate normal and truncated normal densities. In the lefthand side panel we draw normal densities with mean values $\mu = (-0.5, 0, 1)$. These densities are symmetric about the mean and are location invariant. In the righthand side panel we plot the same normal densities except that the observed values are truncated to fall above the threshold of $c = 1$ in absolute value. These distributions are no longer symmetric or location invariant.

that $Y_1, \ldots, Y_n \sim N(0, \sigma^2)$ and that, as before, we estimate μ if $|\bar{Y}| > c/\sqrt{n}$. Then, it is easy to see that $\sqrt{n}\bar{Y}$ remains truncated regardless of the sample size n.

Since we are unable to rely on standard large-sample theory when conducting post-selection inference, it appears necessary to work with the truncated (conditional) distribution directly. This is the approach taken by Weinstein et al. (2013) for constructing post-selection confidence intervals which we describe next. The CDF of the truncated normal distribution of Y is given by (for $n = 1$):

$$F_\mu(y \mid |Y| > c) = \frac{\Phi(\min(y, -c); \mu, \sigma^2) + [\Phi(y; \mu, \sigma^2) - \Phi(c; \mu, \sigma^2)]\, I\{y > c\}}{\Phi(-c; \mu, \sigma^2) + 1 - \Phi(c; \mu, \sigma^2)}. \tag{1.1}$$

The CDF described in (1.1) is pivotal in the sense that $F_{\mu^*}(Y \mid |Y| > c) \sim U(0, 1)$ post-selection. Thus, we can construct a level q confidence interval for μ

$$\text{CI}(|Y| > c) := \{\mu : 1 - q/2 \leq F_\mu(y \mid |Y| > c) \leq q/2\}$$

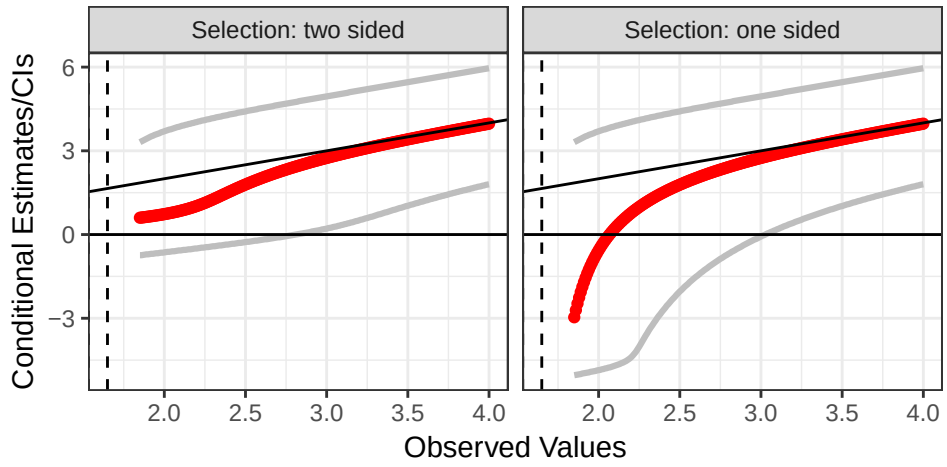

Figure 1.2: Conditional estimates and confidence intervals for a univariate normal mean under selection. The conditional estimates (red line) and confidence intervals (grey lines) are plotted as a function of the (truncated) normal observation. The diagonal line is the $x = y$ line and the dashed vertical line is the $c = 1.65$ threshold used for selection. In the lefthand panel we plot the estimates for two-sided selection $|y| > 1.65$ and in the righthand panel we plot the estimates for one-sided selection $y > 1.65$.

which has the correct post-selection coverage rate for μ by definition. In Figure 1.2 we plot the conditional confidence intervals for one-sided selection $y > 1.65$ and two-sided selection $|y| > 1.65$. The conditional confidence intervals are adaptive to selection in the sense that they tend to be wider when the observed value is close to the threshold, and narrower when the observed value is far away from the threshold and the distribution of the estimate is less likely to be affected by selection.

It is also possible to adjust for model selection in point-estimation by maximizing the conditional likelihood instead of the normal density. The concept of estimation has been introduced in the context of post-selection inference independently and in close succession by Reid et al. (2014), Benjamini and Meir (2014), and Routtenberg and Tong (2015). However, the conditional MLE was proposed in other contexts in much earlier works such as those by Hedges and Olkin (1985) and Iyengar and Greenhouse (1988). Define the univariate

conditional maximum likelihood estimator of μ:

$$\hat{\mu} = \arg\max_{\mu} \frac{\varphi(y; \mu, \sigma^2)}{P_\mu(|Y| > c)} I\{|y| > c\}.$$

The conditional MLE can in most cases be found via line search on a bounded set, as the following theorem shows.

Theorem 1.1. Suppose that $Y \sim N(\mu, \sigma^2)$ and that we estimate μ iff $y < c_1 \leq 0$ or $y > c_2 \geq 0$. Assume w.l.o.g that we observed $y > c_2$. Then if $-\infty < c_1 \leq c_2 < \infty$,

$$\hat{\mu} \in \left[\frac{c_1 + c_2}{2}, y\right]$$

Otherwise, if $c_1 = -\infty$ then $\lim_{y \to c_2} \hat{\mu} = -\infty$.

We defer the proof of the theorem to the appendix. We plot the conditional MLE for truncated normal observations in Figure 1.2. The conditional MLE acts as an adaptive shrinkage estimator, shrinking observed values that are close to the threshold, and leaving large observe values where they are.

Conditional post-selection inference is simple enough to implement in univariate problems. However, things become difficult once we attempt to extend the inference techniques described here to multivariate problems. The following example serves well to demonstrate the difficulties that may arise.

Example 1.2. Suppose that a data vector $Y \in \mathbb{R}^p$ is generated from $N_p(\mu, \mathbf{\Sigma})$ and that the model selection function selects a set of means for estimation M based on the rule $S(y) = \{j : |y_j| > c_j\}$, so $\mathcal{M}$ is the power set of $\{1, \ldots, p\}$. The conditional likelihood is now given by:

$$\mathcal{L}(\mu | M) = \frac{\varphi(y; \mu, \mathbf{\Sigma})}{P_\mu(M)} I\{S(y) = M\}, \tag{1.2}$$

where $P_\mu(M)$ is a p dimensional integral over $2^{|M|}$ disjoint regions. Because $P_\mu(S(Y) = M)$ is difficult to evaluate, maximizing the conditional likelihood becomes a non-trivial task for

large p. Similarly, computing confidence intervals becomes difficult because quantities such as $P_\mu(Y_j < y_j|M)$ are also intractable. To make things worse, the truncated multivariate normal distribution is not location invariant, meaning that even if we are only interested in constructing a confidence intervals for a single coordinate μ_j, the unknown values of the other coordinates μ_{-j} will often have a large influence on the conditional distribution of Y_j.

In the next section we describe the Polyhedral Lemma which was introduced by Lee et al. (2016) to resolve the aforementioned computational and non-invariance problems.

1.2.2 The affine framework and the polyhedral lemma

To describe the Polyhedral Lemma, we will again consider a normal vector $Y \sim N(\mu, \mathbf{\Sigma})$ where $\mathbf{\Sigma}$ is arbitrary and known, and μ is unknown. For the affine framework to be applicable we must be able to write the event of selecting a specific model as an affine constraint:

$$\boldsymbol{V}(M)y \leq b(M)$$

where $\boldsymbol{V}(M)$ and $b(M)$ are a matrix and a vector that are determined by the selected model M. For each model $M \in \mathcal{M}$ we assume that there exists a set linear functions of the form $\theta^M = \eta^T\mu$, that we are interested in estimating $\eta, \mu \in \mathbb{R}^p$. Next, we make these quantities explicit for the normal means problem.

Example 1.2. Continued. We select a coordinate j of μ for estimation if $|y_j| > c_j$. This event is impossible to encode as an affine constraint, but by conditioning on the signs of the selected coordinates $s = \text{sign}(y)$ in addition to M, we can define the jth row of a matrix $\boldsymbol{V}^1$ and the jth coordinate of a vector b^1 which encode the affine constraints for the selected coordinates:

$$\boldsymbol{V}^1(M, s)_{j,\cdot} = -s_j e_j, \qquad b^1_j = -c_j,$$

where e_j is the unit vector with 1 at the jth coordinate and zeros everywhere else. To encode the event $S(y) = M$ for the coordinates which were not selected we require two sets

of matrices and vectors:

$$\boldsymbol{V}_1^0(M,s)_{j,\cdot} = e_j, \qquad b_{1j}^0 = c_j,$$

$$\boldsymbol{V}_2^0(M,s)_{j,\cdot} = -e_j, \qquad b_{2j}^0 = c_j.$$

We can now write the selection event as an affine constraint:

$$\{S(Y) = M, \text{sign}(Y) = s\} = \left\{\boldsymbol{V}^1(M,s)Y < b^1, \boldsymbol{V}_1^0(M,s)Y < b_1^0, \boldsymbol{V}_2^0(M,s)Y < b_2^0\right\}.$$

We estimate the linear functions $\theta_j = \eta_j^T \mu = e_j^T \mu$, for all $j \in M$.

For the rest of this section we will fix a model M and suppress the dependence of $\boldsymbol{V}$ and b on M in our notation. As we already mentioned, the selection event constrains the data to a subspace of the original sample space giving rise to an intractable truncated likelihood. The Polyhedral Lemma resolves this issue by conditioning on extra information beyond the selection event, giving rise to a tractable univariate truncated normal likelihood. Set a linear contrast $\eta \in \mathbb{R}^p$ and define:

$$\tau := \frac{\boldsymbol{\Sigma}\,\eta}{\eta^T\,\boldsymbol{\Sigma}\,\eta}, \qquad Z := \tau\eta^T Y, \qquad W := (I - \tau\eta^T)Y. \tag{1.3}$$

This construction provides us with a representation $y = z + w$, where $\eta^T Z \perp W$.

Proposition 1.1. *Let $T \sim N(\mu, \boldsymbol{\Sigma})$ and let Z and W be as defined in* (1.3)*. Then $Cov(\eta^T Z, W) = 0$.*

The next step in the derivation of the polyhedral lemma is to notice that we can now write the selection event of M as:

$$\boldsymbol{V}\,y = \boldsymbol{V}\,z + \boldsymbol{V}\,w < b. \tag{1.4}$$

Examining a single arbitrary row j of the selection event in (1.4),

$$\boldsymbol{V}_{j,\cdot}\, z + \boldsymbol{V}_{j,\cdot}\, w < b_j \Leftrightarrow \begin{cases} \eta^T y < (b_j - \boldsymbol{V}_{j,}\, w)/(\boldsymbol{V}_{j,\cdot}\, \tau) & \text{if } \boldsymbol{V}_{j,\cdot}\, \tau > 0, \\ \eta^T y > (b_j - \boldsymbol{V}_{j,\cdot}\, w)/(\boldsymbol{V}_{j,\cdot}\, \tau) & \text{if } \boldsymbol{V}_{j,\cdot}\, \tau < 0, \cdot \\ \boldsymbol{V}_{j,\cdot}\, w < b_j & \text{if } \boldsymbol{V}_{j,\cdot}\, \tau = 0 \end{cases} \tag{1.5}$$

Defining,

$$\nu^- := \max_{j:\boldsymbol{V}_{j,\cdot}\, \tau<0} \frac{b_j - \boldsymbol{V}_{j,\cdot}\, w}{\boldsymbol{V}_{j,\cdot}\, \tau}, \qquad \nu^+ := \min_{j:\boldsymbol{V}_{j,\cdot}\, \tau>0} \frac{b_j - \boldsymbol{V}_{j,\cdot}\, w}{\boldsymbol{V}_{j,\cdot}\, \tau},$$

we are now ready to state and prove the Polyhedral Lemma.

Proposition 1.2. *Suppose that $Y \sim N(\mu, \boldsymbol{\Sigma})$, and that M is selected iff*

$$\boldsymbol{V}\, Y < b$$

for some fixed $\boldsymbol{V}$ and b. Then,

$$\eta^T Z | M, W = w \sim TN\left(\eta^T \mu, \eta^T\, \boldsymbol{\Sigma}\, \eta, \nu^-(w), \nu^+(w)\right),$$

where $TN(\mu, \sigma^2, a, b)$ denotes a univariate normal distribution truncated to the interval (a, b).

Proof. The pdf of $\eta^T y$ given M and w is

$$\begin{aligned} f(\eta^T y | S(Y) = M, W = w) &= \frac{P(M|\eta^T y, W = w)\varphi(\eta^T y; \eta^T \mu, \eta^T\, \boldsymbol{\Sigma}\, \eta)}{P(M|W = w)} \\ &= \frac{\varphi(\eta^T y; \eta^T \mu, \eta^T\, \boldsymbol{\Sigma}\, \eta)}{\Phi(\nu^+; \eta^T \mu, \eta^T\, \boldsymbol{\Sigma}\, \eta) - \Phi(\nu^-; \eta^T \mu, \eta^T\, \boldsymbol{\Sigma}\, \eta)} I\{\nu^- \le \eta^T y \le \nu^+\}. \end{aligned} \tag{1.6}$$

In (1.6), φ and Φ denote the pdf and CDF of the univariate normal distribution. The second equality in (1.6) is by Proposition 1.1 and equation (1.5). □

Proposition 1.2 provides us with an easy to follow recipe for constructing conditional confidence intervals.

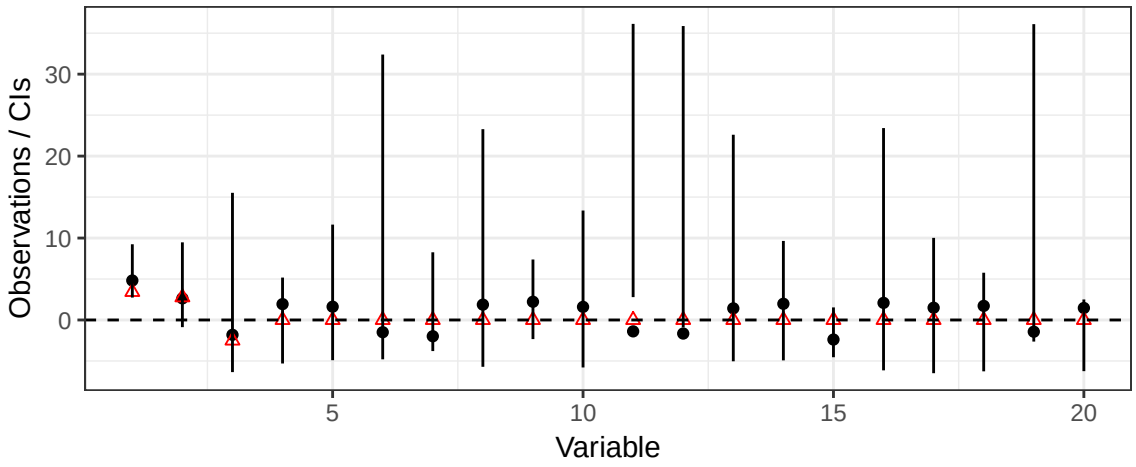

Figure 1.3: Polyhedral Confidence Intervals for the top 20 normal coordinates selected out of 100. Black dots mark the normal means, red triangles the coordinates of the mean vector that we are trying to estimate, and the vertical black lines mark the polyhedral confidence intervals. The polyhedral confidence intervals cover 19 out of the 20 selected means, and reject the null hypothesis of two coordinates at a 5% significance level.

Proposition 1.3. *Define the Polyhedral confidence interval for a contrast of the mean vector* $\eta^T\mu$ *as*

$$\mathrm{CI}_\eta(M, w) = \left\{\eta^T\mu : 1 - \alpha/2 < F_{\eta^T\mu}(\eta^T y | M, W = w) < \alpha/2\right\}. \tag{1.7}$$

Then

$$P(\eta^T\mu^* \in \mathrm{CI}_\eta(M, w)|M) = 1 - \alpha.$$

Proof. Conditioning on w in addition to M, it easy to see that the confidence interval has the correct coverage rate because at the true parameter value $\eta^T\mu^*$, $F_{\eta^T\mu^*}(\eta^T Y|M, W = w) \sim U(0, 1)$. Marginalizing over the distribution of w, we have:

$$\begin{aligned} P(\eta^T\mu \in \mathrm{CI}_\eta(M, w)|M) &= \int P(\eta^T\mu \in \mathrm{CI}_\eta(M, w)|M, W = w) f(w|M) dw \\ &= \int (1 - \alpha) f(w|M) dw = 1 - \alpha. \end{aligned}$$

□

To demonstrate the polyhedral confidence intervals, we apply them to a normal means problem. We generate data vectors from $N_{100}(\mu, \mathbf{\Sigma})$ with $\mathbf{\Sigma}_{i,j} = 0.3^{|i-j|}$, and μ a vector of zeros except for five coordinates selected at random that were sampled from a $N(0, 2)$ distribution. We select the 20 largest coordinates of y in absolute value for estimation. We set our threshold c in an ad-hoc manner to the middle point between the $|y_{(20)}|$ and $|y_{(21)}|$; this is a refinement of the selection event which maintains the validity of the post-selection confidence intervals (Fithian et al., 2014). The data and the confidence intervals are plotted in Figure 1.3. The polyhedral confidence intervals cover 19 out of the 20 selected means.

Even though they achieve the desired coverage rate, the polyhedral confidence intervals leave something to be desired. The polyhedral confidence intervals tend to be oversized. In Figure 1.3 the confidence intervals are orders of magnitude wider than 95% non-selective confidence intervals that have a width of $\approx 4\sigma$. This however is not a glitch, but an artifact of conditioning on W and s as demonstrated by Kivaranovic and Leeb (2018) who show that for model selection by the Lasso, the expected width of the confidence intervals is infinite. As partial solution to this issue Liu et al. (2018) proposed a method for constructing polyhedral confidence intervals for the Lasso without conditioning on the signs of the selected regression coefficients. This however, does not resolve the issue as polyhedral confidence intervals tend to be oversized even if one does not condition on signs.

1.2.3 Extensions and alternatives

After its formulation and application to the Lasso by Lee et al. (2016), the polyhedral lemma was applied in a large number of additional settings where the selection event can be represented as an affine selection event. Lee and Taylor (2014) conduct inference after model selection with marginal screening; Tibshirani et al. (2016b) apply the polyhedral lemma to sequential model selection procedures; Hyun et al. (2016) conduct inference following change point detection; Chen and Bien (2017) conduct valid inference following outlier removal; Reid

et al. (2016a) leverage post-selection techniques to conduct data adaptive aggregate testing; Heller et al. (2017a) conduct post-selection inference for groups of independent p-values.

Some notable generalizations of the polyhedral lemma are those by Taylor and Tibshirani (2016) who propose a heuristic for applying the polyhedral lemma to generalized linear models. Loftus and Taylor (2015) and Yang et al. (2016) generalize the polyhedral lemma to quadratic selection events, and Tian and Taylor (2018) and Tian et al. (2016) develop methods and theory for conducting post-selection inference following model selection with randomization. Some notable theoretical works focusing on conditional inference are those by Fithian et al. (2014) who show that the inference methods based on orthogonal projection methods such as the polyhedral lemma are the only admissible unbiased testing methods, and formulate optimal unbiased post-selection tests; Tibshirani et al. (2015) and Tian and Taylor (2015) develop asymptotic theory for conducting post-selection inference with the polyhedral lemma when the data is not gaussian.

Polyhedral type approaches which project the post-selection distribution onto a line are by far the most ubiquitous in conditional post-selection inference, and relatively few alternatives have been proposed to date. Some notable examples of works that do take an alternative approach to conditional post-selection inference are those by Charkhi and Claeskens (2018) who construct confidence regions for parameters of regression models selected using AIC based on the full conditional distribution, Benjamini et al. (2016) compute confidence intervals for selected regions in DNA Methylation data based on a profile likelihood type heuristic andPanigrahi et al. (2016) and Panigrahi et al. (2017) conduct inference based on an approximation of the conditional likelihood, though the quality of this approximation is unknown.

1.2.4 Related problems

When conducting post-selection inference we assume the existence of a known and fixed model selection procedure S for which we are able to tailor a post-selection inference solution. However, more often than not the model selection function S is not well defined or

is dependent on the whims of a researcher who examines the data and identifies a model that seems subjectively appropriate. To protect against selection in such situations Berk et al. (2013) and Bachoc et al. (2016), among others, propose to widen the naive confidence intervals in such a way as to ensure that the post-selection confidence intervals will have the desired coverage rate regardless of the model selection procedure used. These methods are similar in spirit to the well known Scheffe confidence intervals (Scheffé, 1999), except that they are less conservative as they protect against selection of any possible regression models rather than any possible linear contrast.

Post-selection inference is closely related to the problem of selective inference where the goal is to identify sets of interesting hypotheses, usually with statistical guarantees. One of the first works to consider the problem of inferring on parameters selected based on data is that by Benjamini and Yekutieli (2005) who define the concept False Coverage Rate Control (FCR). The technique they propose shares a duality with the concept of False Discovery Rate (FDR) Control (Benjamini and Hochberg, 1995), in the sense that confidence intervals corresponding to hypotheses that were selected by the Benjamini-Hochberg procedure do not cover zero.

A relatively new set of techniques for conducting selective inference, and that are closely related to conditional post-selection techniques, are those based on Knockoff variables. The Knockoff method first proposed by Barber et al. (2015) uses a set of 'fake' covariates that are exchangeable with the original study variables under the null in order to differentiate between covariates that are truly associated with a response of interest, and those who are not. See also Candes et al. (2016) and Barber et al. (2018). In the multiple testing literature, Lei and Fithian (2016) and Lei et al. (2017) use similar masking techniques in order to allow for interactive hypothesis testing. The key ingredient of the knockoff techniques is the act of leaving out an important piece of information that is orthogonal to the information used by the researcher for decision making. In that sense, Knockoff methods can be viewed as data splitting methods that condition on all of the information used in determining which hypotheses are to be tested.

1.3 Outline of the dissertation

In the rest of the dissertation we seek to address some of the issues we highlighted in the introduction. In Chapter 2 we address the problem of post-selection estimation, a subject that has received comparatively little attention in the post-selection inference literature. There, we propose a stochastic optimization algorithm for computing the conditional maximum likelihood estimate in multivariate problems, and apply it to the normal means problem and the problem of estimating the regression coefficients in a linear model selected by the Lasso. We also propose a method for computing post-selection confidence intervals based on a quadratic approximation heuristic.

In Chapter 3 we consider the problem of conducting inference following aggregate tests. There, we introduce two methods for conducting post-selection inference with less conditioning. One based on conducting inference based on *conservative parametrization*, and the other a *regime switching* procedure which is guaranteed to consistently estimate the post-selection distribution of the data asymptotically in a point-wise manner.

In Chapter 4 we generalize the regime switching method proposed in Chapter 3 to a more general regression setting. Specifically, we will propose a *conditional bootstrap* method for approximating the post-selection distribution of selected regression coefficients. To do so, we build on the work of Chatterjee and Lahiri (2011) on bootstrapping the Lasso estimators to construct consistent post-selection confidence intervals in selected regression models.

In Chapter 5 we conclude, and discuss some interesting open problems.

1.A Proof of Theorem 1.1

Suppose that $Y \sim N(\mu, \sigma^2)$ and that we estimate μ iff $Y < c_1 \leq 0$ or $Y > c_2 \geq 0$. Assume w.l.o.g that we observed $y > c_2$. Furthermore, assume that $c_1 > -\infty$. The log-likelihood of this problem is given by:

$$\log \varphi(y; \mu, \sigma^2) - \log P_\mu(Y \in (-\infty, c_1) \cup (c_2, \infty))$$

The first term of the log-likelihood is the normal log-likelihood and it is maximized at $\hat{\mu} = y$. The second term of the conditional log-likelihood is minus the log probability of observing y. The probability is minimized at the center of the interval $[c_1, c_2]$:

$$\hat{\mu} = \frac{c_1 + c_2}{2}.$$

It also holds that:

$$\log \varphi(y; \mu', \sigma^2) < \log \varphi(y; (c_1 + c_2)/2, \sigma^2), \qquad \forall \mu' < \frac{c_1 + c_2}{2},$$

$$-\log P_{\mu'}(M) < \log P_y(M), \qquad \forall \mu' > y.$$

Thus, it must be that

$$\hat{\mu} \in \left[\frac{c_1 + c_2}{2}, y \right].$$

We now explain why the conditional estimator for a normal mean that has been selected via one-sided testing tends to minus infinity as the observed value approaches the threshold. Let $Y \sim N(\mu, \sigma^2)$, and assume that we estimate μ if and only if $Y > c$. According to Lemma 2.1, the conditional estimator $\hat{\mu}$ given that $Y > c$ solves the equation:

$$y - \mathbb{E}_{\hat{\mu}}(Y|Y > c) = 0.$$

The expectation $\mathbb{E}_{\mu}(Y|Y > c)$ is strictly increasing in μ because

$$\frac{\partial}{\partial \mu} \mathbb{E}_{\mu}(Y|Y > c) = \mathcal{I}_{\mu}(Y|Y > c) > 0,$$

where $\mathcal{I}_{\mu}$ is the information. For any finite $\mu \in \mathbb{R}$, we have $\mathbb{E}_{\mu}(Y|Y > c) > c$ as $P(Y > c) > 0$ for any $\mu \in \mathbb{R}$. Thus, as $y \to c$, it must be the case that $\hat{\mu} \to -\infty$. $\square$

Chapter 2

TRACTABLE POST-SELECTION MAXIMUM LIKELIHOOD INFERENCE

This Chapter was adapted from an arXiv draft paper co-written with Mathias Drton (Meir and Drton, 2017).

2.1 Introduction

I was first introduced to the problem of post-selection inference by Yoav Benjamini during an internship at the end of the first year my PhD studies. In our first joint project we tackled the problem of post-selection estimation following voxel-wise selection in fMRI studies (Benjamini and Meir, 2014). That project was motivated by the work of Vul et al. (2009) who point to the fact that correlations reported in behavioral neuroscience studies between measured brain activation and experimental conditions are implausible, given the expected magnitude of the measurement error and the underlying biological processes.

As we already pointed out, univariate conditional estimation is a relatively simple problem. However, in multivariate problems of non-trivial size the conditional likelihood appeared to be hopelessly intractable. Perhaps this is the reason that the post-selection estimation problem has received relatively little attention in the post-selection inference literature, with the work of Panigrahi et al. (2016) being a rare exception.

A solution to this seemingly intractable problem revealed itself during JSM 2016, where Gerda Claeskens presented an inference method based on sampling from the multivariate truncated normal distribution. We paired these Monte-Carlo sampling techniques with stochastic optimization methods which do not require the evaluation of the likelihood to obtain a tractable optimization routine that made it possible to compute the conditional

maximum likelihood estimators even in high dimensional problems. We present this methodology here. For completeness, we also propose a method for computing post-selection confidence intervals based on a quadratic expansion of the conditional log-likelihood, though this approach lacks a rigorous theoretical justification.

The structure of the rest of the introduction is as follows. In Section 2.1.1 we discuss the distinction between the true parameter and the selected parameter in regression analysis. In Section 2.1.2 we give a short description of the conditional estimation problem in the context of regression and point out the need to avoid conditioning on extra information beyond the selection of a model. In Section 2.1.3 we give an overview of the rest of the Chapter.

2.1.1 Targets of inference

In the context of variable selection in regression, let $\mathcal{M} := \mathcal{P}(\{1, \ldots, p\})$ be the set of models under consideration, defined as the power set of the indices of the columns of the design matrix $\boldsymbol{X}$. Further, let $S : \mathbb{R}^n \to \mathcal{M}$ be a model selection procedure that selects a model $M \in \mathcal{M}$ based on the observed data $y \in \mathbb{R}^n$.

When discussing estimation after model selection in linear regression, one may consider two different targets for inference. The first are the 'true' parameter values in correct models where all variables with non-zero coefficient are present. An alternative target for estimation is the vector of regression coefficients in the selected model

$$\beta_0(y) = (\boldsymbol{X}_M^T \boldsymbol{X}_M)^{-1} \boldsymbol{X}_M^T \mathbb{E}(Y). \tag{2.1}$$

In (2.1), $M = S(y)$ is the selected model, and $\boldsymbol{X}_M$ is the sub-matrix of $\boldsymbol{X}$ made up of the columns indexed by M. These two targets of estimation coincide when the selected model is true, meaning that it contains all variables that have a non-zero regression coefficient. Indeed, if the observed value y is such that $S(y) = M$ for a model M that contains all covariates with non-zero coefficients, then $\mathbb{E}(y) = \boldsymbol{X}_M \beta_0^M$ and $\beta_0^M = \beta_0(y)$. Here β_0^M is the vector of non-zero true coefficients padded with zeros to make it a vector of length $|M|$.

Pötscher (1991) and Leeb and Pötscher (2003) study the behavior of least squares coefficients as estimators of the true regression coefficients in a sequential testing setting. In contrast, works such as Berk et al. (2013) and Leeb et al. (2015) consider inference with respect to the regression coefficients in the selected model. In this work, we follow the latter point of view, taking the stance that a true model does not necessarily exist or, even if one exists, may be difficult to identify. Thus, the interest is in the parameters of the model the researchers have decided to investigate.

2.1.2 Conditioning on selection

A data-driven model selection procedure tends to choose models that are especially suited for the observed data rather than the data-generating distribution. In linear regression this would often be in the form of inclusion of variables that are correlated with the dependent variable only due to random variation. A promising approach for correcting for this bias towards the observed data is to condition on the selection of a model.

Let $y \sim f_\theta$ follow a distribution from an exponential family with sufficient statistic $T(y) \in \mathbb{R}^p$. The likelihood of $T(y)$ given that model M has been selected is

$$\mathcal{L}_M(\theta) = \frac{P(M|T(y))f(T(y))}{P(M)} I_M,$$

where we use the shorthand $P\left(M|T(y)\right) := P\left(S(y) = M|T(y)\right)$ for the conditional probability of selecting model M given $T(y)$. Similarly, $P(M) := P(S(y) = M)$ is the unconditional probability of selecting M, $f(T(y))$ is the unconditional density function of $T(y)$, and $I_M = I_{\{S(y)=M\}}$ is the indicator function for the selection event.

The main obstacle in performing post-selection maximum likelihood inference is the computation of the probability of model selection $P(M)$, which is typically a p dimensional integral. Such integrals are difficult to compute when p is large, and much of the work in the field of post-selection inference has been concerned with getting around the computation of these integrals. For example, Lee et al. (2016) propose to condition on the signs of the selected

variables as well as some additional information contained in the sub-space orthogonal to the quantity of interest in order to obtain a tractable post-selection likelihood. Panigrahi et al. (2016) approximate $P(M)$ with a barrier function.

Conditioning on information beyond the selection of the model of interest, while having the benefit of providing tractable solutions to the post-selection inference problem, may drastically change the form of the likelihood. Consider once again the post-selection estimators for the univariate normal problem (Figure 1.2). Suppose that we observe $y > 0$. Then the right-hand panel plots the conditional estimator for the scenario where two-sided testing is performed. On the left-hand side we plot the conditional estimator for μ as a function of y when we condition on the two-sided selection event as well as the sign of y. Indeed, since our observed value is positive we condition on $\{|Y| > c, Y > 0\} = \{Y > c\}$. This second estimator is close to the observed value y when y is far from the threshold but approaches negative infinity as $y \to c$. Thus, even in the univariate normal case, conditioning on the sign of y in two-sided testing, may drastically alter the resulting conditional estimator.

2.1.3 Outline

In this work, instead of working with the intractable post-selection likelihood, we base our inference on the post-selection score function which can be approximated efficiently even in multivariate problems. The following lemma describes the post-selection score function for exponential family distributions.

Lemma 2.1. Suppose the observation y is drawn from a distribution f_θ that belongs to an exponential family with natural parameter θ and sufficient statistic $T(y)$. If the model selection procedure $S(y)$ satisfies $P(S(y) = M \,|\, T(y)) \in \{0, 1\}$ for a given model M, then the conditional (post-selection) score function is given by:

$$\frac{\partial}{\partial \theta} \log \mathcal{L}(\theta) = T(y) - \mathbb{E}_\theta \left(T(y) | M \right). \tag{2.2}$$

Proof. This result follows directly from the fact that the conditional distribution of an expo-

nential family distribution is also an exponential family distribution as long as $P(M|T(y)) \in \{0, 1\}$. See Fithian et al. (2014) for details. □

In the specific setup we consider subsequently, the conditional distribution of $T(y)$ given M is a multivariate truncated normal distribution. While it is then difficult to compute $\mathbb{E}(T(y)|M)$, we are able to sample efficiently from the multivariate truncated normal distribution using a Gibbs sampler (Geweke, 1991). The main idea behind the method we propose is to use the samples from the truncated multivariate normal distribution as noisy estimates of $\mathbb{E}(T(y)|M)$ and take small incremental steps in the direction of the estimated score function, resulting in a fast stochastic gradient ascent algorithm. Our framework has similarities with the contrastive divergence method of Hinton (2002).

The rest of the chapter is structured as follows. In Section 2.2 we present the proposed inference method in detail and apply it to selective inference on the mean vector of a multivariate normal distribution. In Section 2.3 we describe how the proposed framework can be adapted for post-selection inference in a linear regression model that was chosen by the Lasso. In Section 2.4 we formulate conditions under which the conditional MLE is consistent. A simulation study in Section 2.5 demonstrates that the proposed approach yields improved point estimates for the regression coefficients, and that our confidence intervals, despite lacking a rigorous theoretical justification, achieve close to nominal coverage rates. Finally, in Section 2.6 we conclude with a discussion.

2.2 Inference for selected normal means

Before considering the Lasso, we first discuss the simpler problem of selectively estimating the means of a multivariate normal distribution. Let $y \sim N(\mu, \mathbf{\Sigma})$ with mean vector $\mu \in \mathbb{R}^p$ and a *known* covariance matrix $\mathbf{\Sigma}$. Observing y, we select the model

$$M = \{j \in \{1, \dots, p\} : \ y_j \leq l_j \text{ or } y_j \geq u_j\}, \tag{2.3}$$

where $l_1, \ldots, l_p, u_1, \ldots, u_p \in [-\infty, \infty]$ are predetermined constants with $l_1 < u_1, \ldots, l_p < u_p$. We then perform inference for the coordinates μ_j with $j \in M$ (or possibly inference for a function of these coordinates).

This seemingly simple problem has garnered much attention. For the univariate case of $p = 1$, Weinstein et al. (2013) propose a method for constructing valid confidence intervals, and Benjamini and Meir (2014) compute the post-selection MLE for μ. For $p \gg 1$, Lee et al. (2016) develop a recipe for constructing valid confidence intervals for the selected means or linear functions thereof. Reid et al. (2014) discuss ML estimation when $\mathbf{\Sigma} = \sigma^2 \boldsymbol{I}$. To the best of our knowledge, the method we propose below is the first to address the computation of the conditional MLE when $p \gg 1$ and the covariance matrix $\mathbf{\Sigma}$ is of general structure.

Conditionally on selection, the distribution of y is truncated multivariate normal, as the jth coordinate of y is constrained to lie in the interval (l_j, u_j) if $j \notin M$ or in its complement if $j \in M$. In Section 2.2.1 we describe the Gibbs sampler we use to sample from a truncated multivariate normal distribution, in Section 2.2.2 we describe how such samples can be used to compute the post-selection estimator and in Section 2.2.3 we propose a method for constructing confidence intervals based on the conditional MLE and samples obtained from the truncated normal distribution.

2.2.1 Sampling from a truncated normal distribution

Sampling from the truncated multivariate normal distribution is a well studied problem (Griffiths, 2004; Pakman and Paninski, 2014). We choose to use the Gibbs sampler of Kotecha and Djuric (1999), as it is especially suited to our needs and simple to implement. We describe the sampling algorithm next.

Assume we wish to generate a draw from the univariate truncated normal distribution constrained to lie in the interval $[l, u] \subseteq [-\infty, \infty]$. This distribution has CDF

$$\Phi(y; \mu, \sigma^2, l, u) := \frac{\Phi(y; \mu, \sigma^2) - \Phi(l; \mu, \sigma^2)}{\Phi(u; \mu, \sigma^2) - \Phi(l; \mu, \sigma^2)},$$

where $\Phi(y; \mu, \sigma^2)$ denotes the CDF of the (untruncated) univariate normal distribution with mean μ and variance σ^2. A simple method for sampling from the truncated normal distribution samples a uniform random variable $U \sim U(0, 1)$ and sets

$$y = \Phi^{-1}(U; \mu, \sigma^2, l, u) = \Phi^{-1}\left(U\left(\Phi(u) - \Phi(l)\right) - \Phi(l); \mu, \sigma^2\right). \tag{2.4}$$

Next, consider sampling from the truncated normal constrained to the set $(-\infty, l] \cup [u, \infty)$. In this case, we may first sample a region within which to include y and then sample from a truncated univariate normal distribution constrained to the selected region using the formula given in (2.4).

Given this preparation, we may implement a Gibbs sampler for a truncated multivariate normal distribution as follows. Let $y \sim N(\mu, \mathbf{\Sigma})$, and let $f(y|M)$ be the conditional distribution of y given the selection event. While the marginal distributions of $f(y|M)$ are not truncated normal, the full conditional distribution $f(y_j|M, y_{-j})$ for a single coordinate y_j is truncated normal with parameters

$$\mu_{j,-j} = \mu_j + \mathbf{\Sigma}_{j,-j}\, \mathbf{\Sigma}_{-j,-j}(y_{-j} - \mu_{-j}), \qquad \sigma^2_{j,-j} = \mathbf{\Sigma}_{j,j} - \mathbf{\Sigma}_{j,-j}\, \mathbf{\Sigma}^{-1}_{-j,-j}\, \mathbf{\Sigma}_{-j,j}\,.$$

The Gibbs sampler repeatedly iterates over all coordinates of y and draws a value for y_j conditional on M and y_{-j}. So at the tth iteration we sample

$$y^t_j \sim f(y_j|M, y^t_1, \ldots, y^t_{j-1}, y^{t-1}_{j+1}, \ldots, y^{t-1}_p), \qquad j = 1, \ldots, p.$$

The support of the truncated normal distribution is determined by whether or not $j \in M$.

2.2.2 A stochastic gradient ascent algorithm

The Gibbs sampler described above can be used to closely approximate $\mathbb{E}(y|M)$ but computation of the likelihood $\mathcal{L}_M(\mu)$ remains intractable. However, for optimization of the likelihood,

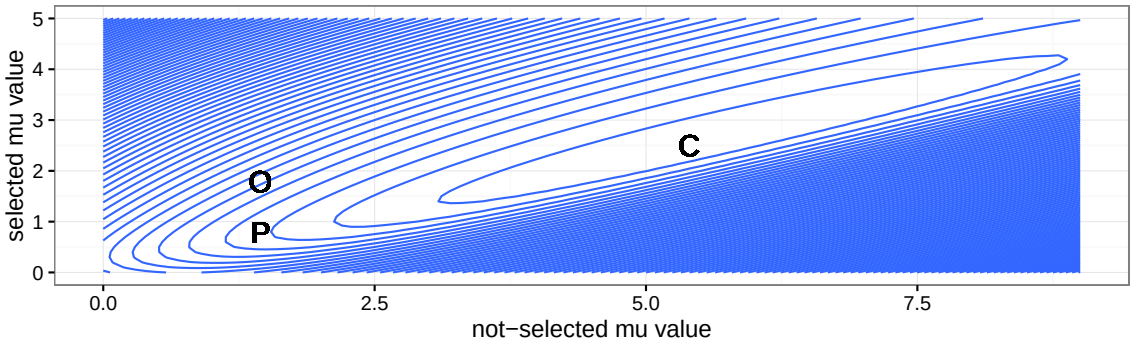

Figure 2.1: The contours of the conditional log-likelihood of a two-dimensional normal distribution. A selection rule $|y_j| > 1.65$ was applied to the observed value $y = (1.45, 1.8)$ marked with 'O'. The conditional MLE where all coordinates are estimated is marked with 'C' at $(5.4, 2.5)$, and the plug-in conditional MLE which does not estimate the coordinates that were not selected is marked as 'P' at $(1.45, 0.8)$. The plug-in estimator, unlike the full conditional MLE, is an adpative shrinkage estimator as in the univariate case.

we can simply take steps of decreasing size in the direction of the evaluated gradient

$$\mu^i = \mu^{i-1} + \gamma_i \, \boldsymbol{\Sigma}^{-1} \left(y - y^i(\mu^{i-1}) \right), \tag{2.5}$$

where y is the observed data, $y^i(\mu^{i-1})$ is a sample from the truncated multivariate normal distribution taken at μ^{i-1} and the step size γ_i satisfies:

$$\sum_{i=1}^{\infty} \gamma_i = \infty, \qquad\qquad \sum_{i=1}^{\infty} \gamma_i^2 < \infty. \tag{2.6}$$

We emphasize that while it is technically possible to compute an MLE for the entire mean vector of the observed random variable, it is not necessarily desirable. To see why, consider once again the left-hand panel of Figure 1.2 where the estimator tends to $-\infty$ as the observed value approaches the threshold. Such erratic behavior may arise when we estimate the coordinates of μ which were not selected, based on observations that are constrained to lie in a convex set, resulting in poor estimates also for the selected coordinates.

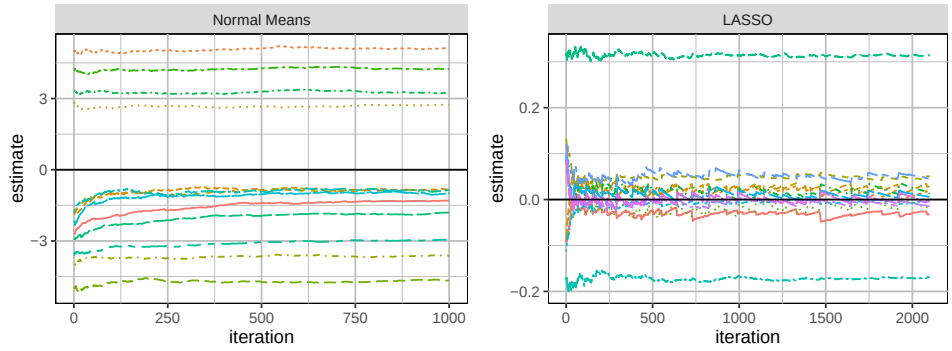

Figure 2.2: Convergence of the stochastic optimization algorithms. We plot the parameter estimates as a function of the number of gradient steps taken for the post-selection normal means estimation problem (left panel) and the post-selection regression estimation problem (right panel). The algorithms tend to converge to the neighborhood of the MLE in a few hundred iterations.

Example 2.1. We plot the conditional log-likelihood for a two-dimensional normal model in Figure 2.1. In such a low-dimensional case, the likelihood function can be computed using routines from the 'mvtnorm' R package (Genz et al., 2016). Our plot is for a setting where we observe $y = (1.45, 1.8)$ with $\mathbf{\Sigma}_{ij} = 0.5^{I\{i \neq j\}}$, and only the first coordinate of μ was selected based on the thresholds $l_1 = l_2 = -1.65$, $u_1 = u_2 = 1.65$. The point y is marked in the figure as an 'O', and the log-likelihood is maximized at the point marked with 'C', which is $\hat{\mu} = (5.4, 2.5)$. We see that instead of performing shrinkage on the observed selected coordinate, the selected coordinate was estimated to be far larger than the observed value.

In order to mitigate this behavior, we propose using a plug-in estimator for the coordinates outside of M. Particularly, we limit ourselves to performing steps of the form

$$\mu_j^i = \begin{cases} \mu^{i-1} + \gamma_i \, \mathbf{\Sigma}_{j,\cdot}^{-1} \left(y - y^i(\mu^{i-1})\right) & \text{if } j \in M, \\ y_j & \text{if } j \notin M, \end{cases} \tag{2.7}$$

where $\mathbf{\Sigma}_{j,\cdot}^{-1}$ is the jth row of $\mathbf{\Sigma}^{-1}$. In other words, we impute the unselected coordinates of μ

with the corresponding observed values of y, and maximize the likelihood only with respect to the selected coordinates of μ. These plug-in estimates for the unselected coordinates of μ are consistent, as we show in Section 2.4. The plug-in conditional MLE for Example 2.1 is shown as a 'P' in Figure 2.1. It is approximately $\hat{\mu} = (1.45, 0.8)$.

Next, we give a convergence statement for the proposed algorithm. Since our gradient steps are based on $y^i(\mu^{i-1})$, a noisy estimate of $\mathbb{E}_{\mu^{i-1}}(y|M)$, the resulting algorithm fits into the stochastic optimization framework of Bertsekas and Tsitsiklis (2000). In short, the theory for stochastic optimization guarantees that taking steps in the form of (2.5) leads to convergence to the MLE as long as the variance of the gradient steps can be bounded.

Theorem 2.1. Let $y \sim N(\mu, \mathbf{\Sigma})$, and let M be defined as in (2.3). Then for all $j \in M$:

$$E_\mu \left(y_j^i(\mu) - \mathbb{E}_\mu(y_j|M)\right)^2 \leq \frac{\mathrm{tr}(\mathbf{\Sigma})}{P\left(\bigcap_{j\notin M}\{l_j < y_j < u_j\}\right)\prod_{j\in M}\Phi(l_j; u_j, \sigma^2_{j,-j})}.$$

The algorithm described in (2.7) converges to the Z-estimator given by the root of the function

$$\psi(\mu)_j = \begin{cases} \mathbf{\Sigma}^{-1}_{j,\cdot}\left(y_j - \mathbb{E}_\mu(y_j|M)\right) & \text{if } j \in M, \\ y_j - \mu_j & \text{if } j \notin M. \end{cases} \tag{2.8}$$

A precise description of the optimization algorithm is given in Algorithm 1 in the appendix. Figure 2.2 shows typical optimization paths for Algorithm 1 as well as the stochastic gradient method for the Lasso described in Section 2.3.

2.2.3 Conditional confidence intervals

In the absence of model selection, the MLE is typically asymptotically normal, and it is common practice to construct Wald confidence intervals based on this limiting distribution:

$$\hat{\mu}^{\text{naive}} = y, \qquad \text{CI}_j^{\text{naive}} = (\hat{\mu}_j^{\text{naive}} - z_{j,1-\alpha/2}, \hat{\mu}_j^{\text{naive}} - z_{j,\alpha/2}), \tag{2.9}$$

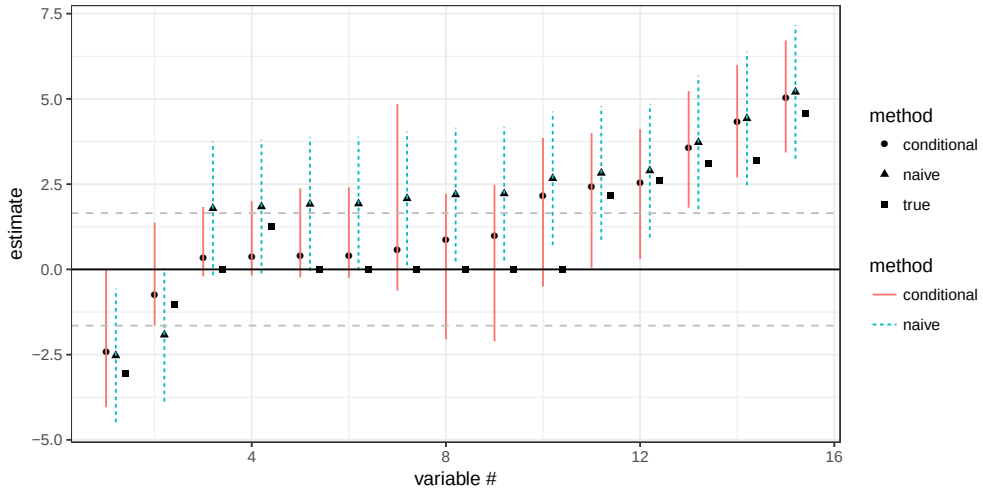

Figure 2.3: Post-selection estimates and confidence intervals for the normal means problem. In this example, 15 means were selected for a threshold of 1.65. The observed values are marked by triangles and the conditional estimators are marked by circles. Naive confidence intervals are marked by a dashed blue line and Conditional-Wald confidence intervals are marked by solid red lines. The true values of the parameter are shown as squares.

where $z_{j,\alpha}$ denotes the $(1-\alpha)$ quantile of the asymptotic normal distribution for the jth coordinate. The post-selection setting is more complicated, however, because we can no longer rely the asymptotic normality of the estimators. Instead, we propose to construct confidence intervals based on the second order Taylor expansion of the conditional likelihood.

In order to describe our proposed approximation to the distribution of the conditional MLE, we extend the normal means problem to the setting of an n-sample. So assume that instead of observing a single vector $y \in \mathbb{R}^p$, we have a set of observations $y_1, \ldots, y_n \in \mathbb{R}^p$ and perform model selection and inference based on $\bar{y}_n = n^{-1} \sum_{i=1}^n y_i$. Our confidence intervals are based on the approximation

$$\sqrt{n}(\hat{\mu}_n^M - \mu_0^M) \approx \sqrt{n}\mathrm{Var}_{\mu_0^M}\left(\sqrt{n}\,\boldsymbol{\Sigma}^{-1}\,\bar{y}_n \middle|\, M\right)^{-1} \boldsymbol{\Sigma}^{-1} \left(\bar{y}_n - \mathbb{E}_{\mu_0^M}(\bar{y}_n|M)\right). \tag{2.10}$$

Based on this approximation, we construct confidence intervals

$$\hat{\mathrm{CI}}_j = \left(\hat{\mu}^M_{j,n} - \hat{\mathrm{TN}}_{j,1-\alpha/2}/\sqrt{n},\ \hat{\mu}^M_{j,n} - \hat{\mathrm{TN}}_{j,\alpha/2}/\sqrt{n} \right). \tag{2.11}$$

Here, $\hat{\mathrm{TN}}$ stands for the conditional distribution given selection of

$$\mathrm{Var}_{\hat{\mu}^M} \left(\sqrt{n}\, \boldsymbol{\Sigma}^{-1}\, \bar{y} \middle|\, M \right)^{-1} \boldsymbol{\Sigma}^{-1}\, \sqrt{n} \left(\bar{y}_n - \mathbb{E}_{\hat{\mu}^M}(\bar{y}_n | M) \right). \tag{2.12}$$

We estimate the quantiles $\hat{\mathrm{TN}}_{j,1-\alpha/2}$ and $\hat{\mathrm{TN}}_{j,\alpha/2}$ using empirical quantiles of samples from the truncated normal distribution. While we are unable to provide theoretical justification for these confidence intervals, a comprehensive simulation study will reveal that they obtain coverage rates that are significantly better than those of the naive confidence intervals, and are surprisingly close to the desired level (Section 2.5).

Example 2.2. Figure 2.3 shows point estimates and confidence intervals for selected means in a selected normal means problems. The figure was generated by sampling $y \sim N(\mu, \boldsymbol{\Sigma})$ with $\mu_1, \ldots, \mu_{20} \sim N(0, 4)$ i.i.d., $\mu_{21} = \cdots = \mu_{100} = 0$ and $\boldsymbol{\Sigma}_{i,j} = 0.3 I_{i \neq j} + 1 I_{i=j}$. The applied selection rule was $S(y) = \{j : \ |y_j| > 1.65\}$. The plotted estimates are the conditional estimates computed using the algorithm defined by (2.7) along with the 95% confidence intervals described in (2.11). In addition, we plot the estimates and confidence intervals described in (2.9) which we term *naive*. These were not adjusted for selection.

As we had seen in the univariate case, the conditional estimator acts as an adaptive shrinkage estimator. When the observed value is far away from the threshold, then no shrinkage is performed and when it is relatively close to the threshold then it is shrunk towards zero.

2.3 Maximum likelihood estimation for the Lasso

In this section we demonstrate how the ideas from the previous section can be adapted for computing the post-selection MLE in linear regression models selected by the Lasso. The

Lasso estimator minimizes the squared error loss augmented by an ℓ_1 penalty,

$$\hat{\beta}_{\text{Lasso}} = \arg\min_{\beta} \frac{1}{2}\|y - \boldsymbol{X}\,\beta\|_2^2 + \lambda\|\beta\|_1$$

with $\lambda \geq 0$ being a tuning parameter. Model selection results from the fact that the ℓ_1 penalty may shrink a subset of the regression coefficients to zero. As in Lee et al. (2016), we perform inference on the non-zero regression coefficients in the Lasso solution, that is, the selection procedure is $S(y) = \{j : \hat{\beta}_{\text{Lasso},j} \neq 0\}$.

Given selection of a model M, we are interested in estimating the unconditional mean of the regression coefficients $\beta = (\boldsymbol{X}_M^T\,\boldsymbol{X}_M)^{-1}\,\boldsymbol{X}_M^T\,\mathbb{E}(Y)$. We begin by describing the Lasso selection event (Section 2.3.1) and then give a Metropolis-Hastings sampler for the post-selection distribution of the least-squares estimates (Section 2.3.2). In Section 2.3.3, we describe a practical stochastic ascent algorithm for estimation after model selection with the Lasso.

2.3.1 The Lasso selection event

Let $M \subseteq \{1, \ldots, p\}$ be a given model. In order to develop a sampling algorithm for a normal distribution truncated to the event that $S(\boldsymbol{X}, y) := \{j : \hat{\beta}_{\text{Lasso},j} \neq 0\} = M$, we invoke the work of Lee et al. (2016) who provide a useful characterization of this Lasso selection event. Let $s \in \{-1, 1\}^{|M|}$ be the vector of signs of $\hat{\beta}_{\text{Lasso}}$ over the active set. We will consider two sets

$$A_1(M, s) := \{\boldsymbol{A}_1(M, s)y < u_1(M, s)\}\,, \tag{2.13}$$

$$A_0(M, s) := \{l_0(M, s) < \boldsymbol{A}_0(M)y < u_0(M, s)\}\,, \tag{2.14}$$

where in the first event

$$\boldsymbol{A}_1(M, s) = -\operatorname{diag}(s)(\boldsymbol{X}_M^T\,\boldsymbol{X}_M)^{-1}\,\boldsymbol{X}_M^T, \quad u_1(M, s) = -\lambda \operatorname{diag}(s)(\boldsymbol{X}_M^T\,\boldsymbol{X}_M)^{-1}s, \tag{2.15}$$

and in the second event

$$\boldsymbol{A}_0(M) = \frac{1}{\lambda}\,\boldsymbol{X}_{-M}^T(I - \boldsymbol{X}_M(\boldsymbol{X}_M^T\,\boldsymbol{X}_M)^{-1}\,\boldsymbol{X}_M^T), \tag{2.16}$$

$$l_0(M,s) = -\mathbf{1} - \boldsymbol{X}_{-M}^T\,\boldsymbol{X}_M(\boldsymbol{X}_M^T\,\boldsymbol{X}_M)^{-1}s, \qquad u_0(M,s) = \mathbf{1} - \boldsymbol{X}_{-M}^T\,\boldsymbol{X}_M(\boldsymbol{X}_M^T\,\boldsymbol{X}_M)^{-1}s.$$

Here, $\boldsymbol{X}_M$ is the submatrix of the design matrix $\boldsymbol{X}$ made up of the columns indexed by the selected model M and the columns in the submatrix $\boldsymbol{X}_{-M}$ correspond to variables which were not selected. It can be shown that

$$\{S(\boldsymbol{X},Y) = M \text{ and sign vector equal to } s\} \;=\; A_0(M,s) \cap A_1(M,s). \tag{2.17}$$

Suppose that $Y \sim (\boldsymbol{X}\,\beta, \sigma^2\,\boldsymbol{I})$, then conditional score function for a model selected by the Lasso is given by

$$\sigma^2\frac{\partial}{\partial\beta}\log\mathcal{L}(\beta|M) = \boldsymbol{X}_M^T\,y - \mathbb{E}(\boldsymbol{X}_M^T\,Y|M) = \boldsymbol{X}_M^T\,y - \frac{\sum_s P(M,s)\,\mathbb{E}\left(\boldsymbol{X}_M^T\,Y|A_1(M,s)\right)}{\sum_s P(M,s)},$$

where for a given set of signs $P(M,s) = P\left(A_0(M,s)\right) \times P\left(A_1(M,s)\right)$.

As in the normal means problem, parameters related to the set of variables excluded from the model play a role in the conditional likelihood. In the normal means problem we advocated excluding those from the optimization of the conditional likelihood. For the Lasso, we similarly must compute a conditional expectation which is a function of $\boldsymbol{A}_0(M)\,\mathbb{E}(Y)$. We again advocate for avoiding conditional likelihood-based estimation of this quantity. In computational experiments we observed that the value of $\boldsymbol{A}_0(M)\,\mathbb{E}(Y)$ tends to be very small and rather well approximated by a vector of zeros. For more on this and some numerical examples see Appendix B.

In the next subsection, we devise an algorithm for sampling from the post-selection distribution of the regression coefficients selected by the Lasso without conditioning on the

sign vector s. The sampler will operate by updating the two quantities

$$\eta := (\boldsymbol{X}_M^T \boldsymbol{X}_M)^{-1} \boldsymbol{X}_M^T y, \qquad \xi := \frac{1}{\lambda} \boldsymbol{X}_{-M} \left(I - \boldsymbol{X}_M (\boldsymbol{X}_M^T \boldsymbol{X}_M)^{-1} \boldsymbol{X}_M^T \right) y.$$

2.3.2 Sampling from the Lasso post-selection distribution

With a view towards Gibbs sampling, we examine the region where a single regression coefficient may lie given the signs of all other coefficients. Let $j \in M$ be an arbitrary index. Denote by s^{+j} and s^{-j} vectors of signs where the signs for all coordinates but j are held constant and the jth coordinates are set to either 1 or -1, respectively. A necessary condition for the selection of M is that $\eta_j \leq \lambda (\boldsymbol{X}_M^T \boldsymbol{X}_M)^{-1}_{j,\cdot} s^{-j}$ or $\eta_j \geq \lambda (\boldsymbol{X}_M^T \boldsymbol{X}_M)^{-1}_{j,\cdot} s^{+j}$. Ideally, we would be able to implement a Gibbs sampler that allows for the change of signs as we have done in Section 2.2.1 by setting

$$l_j = \lambda (\boldsymbol{X}_M^T \boldsymbol{X}_M)^{-1}_{j,\cdot} s^{-j}, \qquad u_j = \lambda (\boldsymbol{X}_M^T \boldsymbol{X}_M)^{-1}_{j,\cdot} s^{+j}. \tag{2.18}$$

However, an important way in which the Lasso selection event differs from the one described in Section 2.2 is that when a single coordinate of s is changed, the thresholds for all other variables change. Thus, in order for a single coordinate of η to change its sign, all other variables must be in positions that allow for that.

In order to explore the entire sample space (and sign combinations) we propose a delayed rejection Metropolis-Hastings algorithm (Mira, 2001; Tierney and Mira, 1999). The algorithm works by attempting to take a Gibbs step for each selected variable in turn. If the proposed Gibbs step for the jth variable satisfies the constraints induced by the selection event then the proposal is accepted. Otherwise, we keep the proposal for the jth variable and make a global proposal for all selected variables keeping their signs fixed. We use the notation:

$$\eta \sim N_p(\beta, \boldsymbol{\Sigma}_1), \qquad \beta = (\boldsymbol{X}_M^T \boldsymbol{X}_M)^{-1} \boldsymbol{X}_M^T \mathbb{E}(Y), \qquad \boldsymbol{\Sigma}_1 = \sigma^2 (\boldsymbol{X}_M^T \boldsymbol{X}_M)^{-1},$$

$$\xi \sim N(0, \boldsymbol{\Sigma}_0), \qquad \boldsymbol{\Sigma}_0 = \sigma^2 \, \boldsymbol{A}_0(M) \, \boldsymbol{A}_0(M)^T.$$

At some arbitrary iteration t, our sampler first makes the draw

$$\xi^t \sim f(\xi | M, \eta^{t-1}). \tag{2.19}$$

This sampling task is quite simple in the sense that $\xi | M, \eta$ has a multivariate normal distribution constrained to a convex set. Next, we make a proposal for each selected variable. For the jth selected variable we sample:

$$r_j \sim f\left(\eta_j \middle| \{\eta_j < l_j\} \cup \{u_j < \eta_j\}, \eta_1^t, \ldots, \eta_{j-1}^t, \eta_{j+1}^{t-1}, \ldots, \eta_p^{t-1}\right),$$

where l_j and u_j are as defined in (2.18). If the sign of r_j differs from the sign of η_j^{t-1}, then we must verify that ξ^t from (2.19) satisfies the constraints imposed by the new set of signs. If the constraints described in (2.14) are not satisfied, then the proposal is rejected. If the proposal yields a point that satisfies both (2.14) and (2.13) then no further adjustment is necessary and the acceptance probability is 1 because the proposal is full conditional (Chib and Greenberg, 1995). On the other hand, if the proposed point is not in the set from (2.13), then a sign change has been performed and we must update the values for other coordinates.

Denote by $\text{TN}(a, b, \mu, \sigma^2)$ a univariate normal distribution with mean μ and variance σ^2 constrained to the interval (a, b). For all variables $k \neq j$ we sample a proposal from the following distribution:

$$r_k \sim \text{TN}(a_k, b_k, \eta_k^t, \sigma_{k,-k}^2), \tag{2.20}$$

where $a_k = u_k$ and $b_k = \infty$ if $s_k^t = 1$, and $a_k = -\infty$ and $b_k = l_k$ if $s_k^t = -1$. Note that in (2.20) l_k and u_k must be recomputed according to the proposed sign change.

The Metropolis-Hastings algorithm in its entirety is described in Algorithm 2 in the Appendix. The following Lemma describes the transitions of the proposed sampler.

Lemma 2.2. For the jth variable at the tth iteration define:

$$r_1^{\rightarrow} = \left(\eta_1^t, \ldots, \eta_{j-1}^t, r_j, \eta_{j+1}^{t-1}, \ldots, \eta_p^{t-1}, \xi^t\right), \quad r_2^{\rightarrow} = \left(r_1, \ldots, r_{j-1}, r_j, r_{j+1}, \ldots, r_p, \xi^t\right),$$
$$r_1^{\leftarrow} = \left(r_1, \ldots, r_{j-1}, \eta_j^{t-1}, r_{j+1}, \ldots, r_p, \xi^t\right), \quad r_2^{\leftarrow} = \left(\eta_1^t, \ldots, \eta_{j-1}^t, \eta_j^{t-1}, \eta_{j+1}^{t-1}, \ldots, \eta_p^{t-1}, \xi^t\right).$$

Here, $r_2^{\leftarrow}$ represents the current state of the sampler after the Gibbs step from (2.19). If ξ^t from (2.19) is not in the set from (2.14), then the proposal for r_j is rejected and the sampler stays in state $r_2^{\leftarrow}$. If ξ^t is in (2.14) and $r_1^{\rightarrow}$ is in the set from (2.13) then the sampler moves to $r_1^{\rightarrow}$. Otherwise, if $r_1^{\leftarrow}$ is in the set from (2.13) then the sampler stays in state $r_2^{\leftarrow}$. Finally if neither $r_1^{\rightarrow}$ nor $r_1^{\leftarrow}$ are in the set from (2.13) then the sampler either moves to $r_2^{\rightarrow}$ or stays put at $r_2^{\leftarrow}$. In this case, the move to $r_2^{\rightarrow}$ occurs with probability

$$p_j^t = \min\left(\frac{\varphi(r_2^{\rightarrow}; \beta, \mathbf{\Sigma})}{\varphi(r_2^{\leftarrow}; \beta, \mathbf{\Sigma})} \frac{q(r_2^{\rightarrow}, r_2^{\leftarrow})}{q(r_2^{\leftarrow}, r_2^{\rightarrow})}, 1\right), \tag{2.21}$$

where

$$q(x, y) = f\left(y_j | \{Y_j < l_j\} \cup \{u_j < Y_j\}, x_{-j}\right) \prod_{k \neq j} \frac{\varphi(y_k; x_k, \sigma^2_{k,-k})}{P(Y_k \in (a_k, b_k); x_k, \sigma^2_{k,-k})}.$$

2.3.3 A stochastic ascent algorithm for the lasso

We now propose an algorithm for computing the post-selection MLE when the model is selected via Lasso. We begin by defining the gradient ascent step, which uses samples from the post-selection distribution of the refitted regression coefficients. We give a convergence statement for the resulting algorithm, and we discuss practical implementation for which we address variance estimation and imposing sign constraints.

Let $M = S(y)$ be the Lasso-selected model. Given a sample $\eta^i \sim f_{\hat{\beta}^{i-1}}(\eta|M)$ from the post-selection distribution of the least squares estimator, we take steps of the form:

$$\hat{\beta}^i = \hat{\beta}^{i-1} + \gamma_i \left(\boldsymbol{X}_M^T y - (\boldsymbol{X}_M^T \boldsymbol{X}_M)\eta^i\right), \tag{2.22}$$

where the γ_i satisfy the conditions from (2.6). In Theorem 2.2 we give a convergence statement for the algorithm defined by (2.22). As in Theorem 2.1, the main challenge is bounding the variance of the stochastic gradient steps.

Theorem 2.2. Let η follow the conditional distribution of $\eta \sim N(\beta, \mathbf{\Sigma})$ given the Lasso selection $S(y) = M$. Then there exists a constant A such that for all $\beta \in \mathbb{R}^p$:

$$\mathbb{E}_\beta \left\| (\boldsymbol{X}_M^T \boldsymbol{X}_M)\eta - \boldsymbol{X}_M^T \mathbb{E}_\beta(Y|M) \right\|_2^2 \leq A.$$

Furthermore, the sequence $(\hat{\beta}^i)$ from (2.22) converges, and its limit $\hat{\beta}^\infty := \lim_{i\to\infty} \hat{\beta}^i$ satisfies $\psi(\hat{\beta}^\infty) := \boldsymbol{X}_M^T y - \mathbb{E}_{\hat{\beta}^\infty}(\boldsymbol{X}_M^T Y|M) = 0$.

Before we exemplify the behavior of the proposed algorithm we first discuss some technicalities. The sampling algorithm proposed in the previous section assumes knowledge of the residual standard error σ, a quantity that in practice must be estimated from the data. We we find that the cross-validated Lasso variance estimate recommended by Reid et al. (2016b), works well for our purposes.

As in the univariate normal case, the post-selection estimator for the Lasso performs adaptive shrinkage on the refitted regression coefficients. However, the asymmetry between the thresholds dictated by different sign sets may cause the sign of the conditional coefficient estimate to be different than the one inferred by the Lasso. Empirically we have found some benefit for constraining the signs of the estimated coefficients to those of the refitted least-squares coefficient estimates.

Example 2.3. We illustrate the proposed method via simulated data that are generated as follows. We form a matrix of covariates by sampling n rows independently from $N_p(0, \mathbf{\Sigma})$ with $\mathbf{\Sigma}_{i,j} = \rho^{|i-j|}$. We then generate a coefficient vector β by sampling k coordinates from the *Laplace*(1) distribution and setting the rest to zero. Next, we sample a response vector $y \sim N(\mu, \sigma^2 \boldsymbol{I})$, where $\mu = \boldsymbol{X}\beta$ and σ^2 is chosen to obtain a certain signal-to-noise ratio defined as $\text{snr} := \text{Var}(\mu)/\sigma^2$. We set $n = 400$, $p = 1000$, $k = 5$, $\rho = 0.3$ and $\text{snr} = 0.2$.

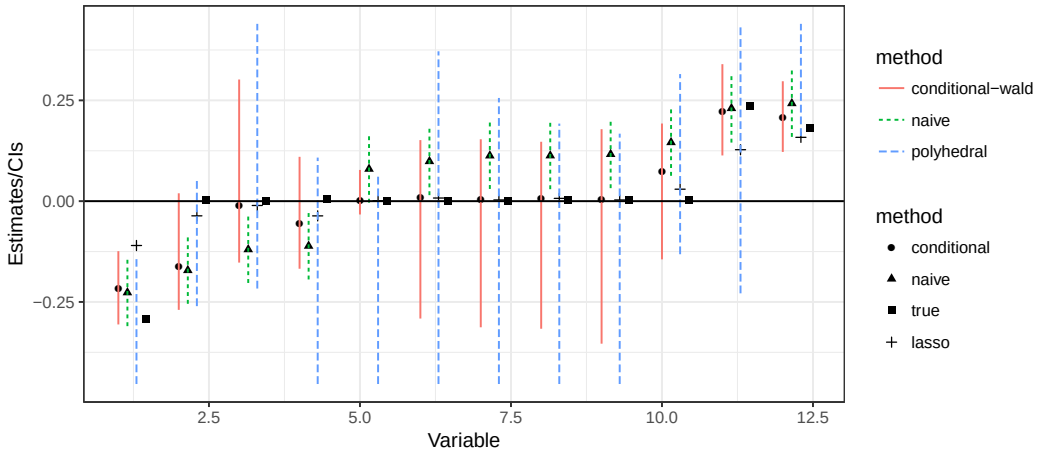

Figure 2.4: Post-selection estimates and confidence intervals for the Lasso. For simulated data, we plot the conditional MLE (circles), refitted least-squares estimates (triangles) and Lasso estimates (squares). The true coefficient values are marked by plus signs. We also plot three types of confidence intervals, Conditional-Wald (solid red line), Refitted-Wald (dashed green line) and Polyhedral confidence intervals (dashed blue lines).

Given a simulated dataset we select a model using the Lasso as implemented in the R package 'glmnet' (Friedman et al., 2010). Following common practice and the default of the package, the tuning parameter λ is selected via cross-validation. Strictly speaking, this yields another post-selection problem.

In Figure 2.4 we plot three types of estimates for the regression coefficients selected by the Lasso. The conditional estimator proposed here, the refitted least-squares estimates and the Lasso estimates. In addition to the point estimates, we also plot three types of confidence intervals. The first are the Conditional-Wald confidence intervals analogous to the ones described in Section 2.2.3. They are given by:

$$\hat{\mathrm{CI}}_j = \left(\hat{\beta}^M_{j,n} - \hat{\mathrm{TN}}_{j,1-\alpha/2}/\sqrt{n},\ \hat{\beta}^M_{j,n} - \hat{\mathrm{TN}}_{j,\alpha/2}/\sqrt{n}\right),$$
$$\hat{\mathrm{TN}} =^D \sigma^2 \operatorname{Var}_{\hat{\beta}^M_n}\left(n^{-0.5}\, \boldsymbol{X}_M^T\, Y \middle| M\right)^{-1} n^{-0.5} \left(\boldsymbol{X}_M^T\, y - \mathbb{E}_{\hat{\beta}^M_n}(\boldsymbol{X}_M^T\, Y | M)\right).$$

The second intervals are the Refitted-Wald confidence intervals obtained from fitting a linear

regression model to the selected covariates without accounting for selection. Finally, we also include the intervals of Lee et al. (2016) as implemented in the R package 'selectiveInference' (Tibshirani et al., 2016a). We term these *Polyhedral* confidence intervals.

In Figure 2.4, black circles mark the conditional estimates, triangles the refitted least squares estimates, squares the Lasso estimates and plus signs the true coefficient values. The conditional estimator tends to lie between the refitted and the Lasso estimates. When the refitted estimate is far from zero the conditional estimator applies very little shrinkage, and when the refitted estimator is closer to zero the conditional estimator is shrunk towards the Lasso estimate. The conditional confidence intervals also exhibit a behavior that depends on the estimated magnitude of the regression coefficients. When the conditional estimator is far from zero the size of the confidence intervals is similar to the size of the refitted confidence interval. When the conditional estimator is shrunk towards zero, its variance tends to be the smallest. The confidence intervals are the widest when the conditional estimator is just in-between the Lasso and refitted estimates. The Polyhedral confidence intervals tend to be the largest in most cases. Section 2.5 gives a more thorough examination of these estimates and confidence intervals.

2.4 Asymptotics for conditional estimators

We now present asymptotic distribution theory that supports the estimation method proposed in the previous sections. Such theory is complicated by the fact that model selection induces dependence between the previously i.i.d. observations. In Section 2.4.1 we first give a consistency result for naive unconditional estimates, which in particular justifies our plug-in likelihood method for the normal means problem. We then outline conditions under which the conditional MLE is consistent for the parameters of interest in a general exponential family setting. In Section 2.4.2 we adapt the theory to the Lasso post-selection estimator. We remark that theory on the efficiency of conditional estimators can be found in Routtenberg and Tong (2015). Proofs for this section are deferred to the appendix.

2.4.1 Theory for exponential families

Suppose we have an i.i.d. sequence of observations $(Y_i)_{i=1}^{\infty}$ drawn from a distribution f^*. As a base model for the distribution of each observation y_i, consider a regular exponential family $\{p_\theta : \theta \in \Theta\}$ with sufficient statistic $T \in \mathbb{R}^p$ and θ being the natural parameter. So, $\Theta \subset \mathbb{R}^p$. For the sample $y_1, \ldots, y_n$, define $\bar{T}_n := n^{-1} \sum_{i=1}^n T(y_i)$. Now, let $\mathcal{M}$ be a countable set of submodels, which we denote by $M = \{p_{\theta^M} : \theta^M \in \Theta^M\}$ with parameter space $\Theta_M \subset \Theta$. We consider a model selection procedure $S_n : \mathbb{R}^p \to \mathcal{M}$ that selects a model M as a function of $\bar{T}_n$. Based on the true distribution f^* the sample is taken from, the selection procedure S_n induces a distribution $P_n(M) := P(S_n(\bar{T}_n) = M)$ over $\mathcal{M}$. We emphasize that f^* need not belong to any model in $\mathcal{M}$ nor the base family $\{p_\theta : \theta \in \Theta\}$.

Example 2.4. In the normal means problem, p_θ is a normal distribution with θ being the mean vector. The sufficient statistic is $T(y) = \mathbf{\Sigma}^{-1} y$, where $\mathbf{\Sigma}$ is the known covariance matrix. Each model $M \in \mathcal{M}$ corresponds to a set of mean vectors with a subset of coordinates equal to zero. The selection procedure S_n is based on comparing the coordinates of $\bar{T}_n$ to predetermined thresholds l_j and u_j, recall (2.3). In an asymptotic setting l_j and u_j will often scale with the sample size to obtain desired type-I error rates.

We consider estimation of a parameter θ_0^M of a fixed model M, which represents the model selected in the data analysis. If the data-generating distribution f^* belongs to M, then $f^* = p_{\theta_0^M}$ for a parameter value $\theta_0^M \in \Theta^M$ and consistency can be understood as referring to the true data-generating distribution. If $f^* \notin M$, then the parameter in question corresponds to the distribution in M that minimizes the KL-divergence from f^*, so

$$\theta_0^M := \arg \inf_{\theta^M \in \Theta^M} - \mathbb{E}_{f^*} \left[\log p_{\theta^M}(y) - \log f^*(y) \right] = \arg \sup_{\theta^M \in \Theta^M} \mathbb{E}_{f^*} \left[\ell_{\theta^M}(y) \right].$$

Note that even under model misspecification we have $\mathbb{E}_{f^*}(\bar{T}_n) = \mathbb{E}_{\theta_0^M}(\bar{T}_n)$ because θ_0^M is the solution to the expectation of the score equation.

The post-selection setting is unusual in the sense that we are only interested in a specific

model M if $S_n(\bar{T}_n) = M$. Hence, it only makes sense to analyze the asymptotic properties of an estimator of θ_0^M if model M is selected infinitely often as $n \to \infty$. This justifies our subsequent focus on conditions that involve the probability of selecting M.

Our first result applies in particular to the normal means problem and is concerned with the post-selection consistency of the unconditional/naive MLE for θ_0^M.

Theorem 2.3. Let M be a fixed model with $P_n(M)^{-1}e^{-\delta n} = o(1)$ for all $\delta > 0$. Let $\tilde{\theta}_n^M = (\tilde{\theta}_{n,j}^M)_{j=1}^p$ be an estimator that unconditionally is unbiased for θ_0^M. Suppose there is a constant $C \in (0, \infty)$ such that for all $1 \le j \le p$ and $n \ge 1$ the distribution of $\sqrt{n}(\tilde{\theta}_{n,j}^M - \theta_{0,j}^M)$ is sub-Gaussian for parameter C. Then $\tilde{\theta}_n^M$ is post-selection consistent, that is,

$$\lim_{n\to\infty} P(\|\tilde{\theta}_n^M - \theta_0^M\|_\infty > \varepsilon \,|\, S_n(\bar{T}_n) = M) = 0 \qquad \forall \varepsilon > 0.$$

Next, we turn to the conditional MLE. Let $\ell_{\theta^M}(y_i)$ be the log-likelihood of y_i as a function of θ^M, and let $P_{n,\theta^M}(M)$ be the probability of $\{S_n(\bar{T}_n) = M\}$ where $y_1, \ldots, y_n$ is an i.i.d. sample from p_{θ^M}. Then the conditional MLE is

$$\hat{\theta}_n^M = \arg\max_{\theta^M \in \Theta^M} \left(\frac{1}{n} \sum_{i=1}^n \ell_{\theta^M}(y_i) \right) - \frac{1}{n} \log P_{n,\theta^M}(M).$$

We first give conditions for its post-selection consistency.

Theorem 2.4. Suppose the fixed model M satisfies

$$P_n(M)^{-1} = o(n), \tag{2.23}$$

$$\lim_{n\to\infty} \inf_{\theta^M} P_{n,\theta^M}(M) e^n = \infty. \tag{2.24}$$

Furthermore, suppose that for a sufficiently small ball $U \subset \Theta$ centered at θ_0^M

$$\sup_{\theta^M \in U(\theta_0^M)} P_{n,\theta^M}(M)^{-1} = o(n). \tag{2.25}$$

Then the conditional MLE is post-selection consistent for θ_0^M, that is,

$$\lim_{n\to\infty} P(\|\hat{\theta}_n^M - \theta_0^M\|_\infty > \varepsilon \,|\, S_n(\bar{T}_n) = M) = 0 \qquad \forall \varepsilon > 0.$$

Condition (2.24) concerns the model-based selection probability and ensures that the conditional MLE exists with probability 1 as $n \to \infty$. Both the plug-in likelihood for the selected means problem and the Lasso likelihood satisfy this condition. We note that this condition excludes examples such as the singly truncated univariate normal distribution, where the probability that an MLE does not exist is positive (del Castillo, 1994). Condition (2.23) concerns the true probability of selecting the considered model M, which is required to not decrease too fast. Condition (2.25) serves to ensure that the conditional score function is well behaved in the neighborhood of the estimand.

2.4.2 Theory for the Lasso

In this section we describe how the theory from the previous section applies to inference in linear regression after model selection with the Lasso. Suppose that we observe an independent sequence of observations

$$(Y_i)_{i=1}^\infty \sim N(\mu_i, \sigma^2). \tag{2.26}$$

Each observation Y_i is accompanied by a vector of covariates $X_i \in \mathbb{R}^p$ which we consider fixed, or equivalently, conditioned upon. The sufficient statistic for the linear regression model is given by $T_n(\boldsymbol{X}, y) = \boldsymbol{X}^T y$ and the model selection function $S_n(\boldsymbol{X}, y)$ is the Lasso, which selects a model:

$$S_n(\boldsymbol{X}, y) = \{j : \hat{\beta}_{\text{Lasso},j} \neq 0\}.$$

For a selected model M, the conditional MLE for the regression coefficients is given by:

$$\hat{\beta}_n^M = \arg\max_\beta \frac{f(\boldsymbol{A}_1 y)}{P_\beta(M)}, \tag{2.27}$$

where $P_\beta(M) = \sum_s P_\beta(\boldsymbol{A}_1(M,s)) \times P_n(\boldsymbol{A}_0(M,s))$. Notice that in our objective function the probabilities for not selecting the null-set are not a function of the parameters over which the likelihood is maximized. Instead, they are defined as a function of the sample size n and are determined by the imputed value for $\boldsymbol{A}_0(M)\mu$. In practice we set $\boldsymbol{A}_0(M)\mu = 0$. This imputation method can be justified by the fact that a model is unlikely to be selected infinitely often if $\lim_{n\to\infty} \boldsymbol{A}_0(M)\mu \neq 0$.

For good behavior of the conditional MLE we made assumptions regarding the probabilities of selecting models of interest. Many previous works have investigated the properties that a data generating distribution must fulfill in order for the Lasso to identify a correct model with high probability. See for example Zhao and Yu (2006), and Meinshausen and Yu (2009). While we do not limit our attention to the selection of the correct model, this line of study sheds light on the conditions a model $M \in \mathcal{M}$ must satisfy in order for it to be selected with sufficiently high probability. In the following we assume that the number of covariates $p_n = p$ is kept fixed while the sample size n grows to infinity. We touch on high-dimensional settings briefly at the end of the section.

The set of models for which we are able to guarantee convergence depends on the scaling of the ℓ_1 penalization parameter. We consider two types of scalings:

$$\lambda_n \propto \sqrt{n}, \tag{2.28}$$

$$\lim_{n\to\infty} \frac{\lambda_n}{\sqrt{n}} = \infty, \qquad \lim_{n\to\infty} \frac{\lambda_n}{n} = 0. \tag{2.29}$$

We begin by discussing the case where the ℓ_1 penalization parameter scales as in (2.28). In this setting, the model selection probabilities can be bounded in a satisfactory manner as long as the expected projection of the model residuals on the linear subspace spanned by the inactive variables is not too large.

Lemma 2.3. Suppose that λ_n scales as in (2.28) and that y follows a normal distribution as defined in (2.26). Suppose further that for an arbitrary model of interest $M \in \mathcal{M}$ there

is a matrix $\mathbf{\Sigma}$ and a vector β_0^M such that following holds:

$$\frac{1}{n}\boldsymbol{X}^T\boldsymbol{X} \to \mathbf{\Sigma}, \tag{2.30}$$

$$(\boldsymbol{X}_M^T\boldsymbol{X}_M)^{-1}\boldsymbol{X}_M^T\mu \to \beta_0^M, \qquad \boldsymbol{A}_0(M)\mu \to 0, \;\; a.s. \tag{2.31}$$

Then there exists an asymptotic lower bound for the probability of selecting M:

$$\lim_{n\to\infty} P_n(M) \geq \lim_{n\to\infty} \inf_{\beta^M} P_{n,\beta^M}(M) = c > 0.$$

Next, we discuss the setting where λ_n grows faster than $\sqrt{n}$. Here we must impose stronger conditions on the selected model because the probability of selecting a model which contains covariates with zero coefficient values may decrease to zero at an exponential rate. Furthermore, we make assumptions similar to the Irrepresentable Conditions of Zhao and Yu (2006) on the selected model in order to make sure that the model selection conditions corresponding to the variables not included in the model are satisfied with high probability.

Lemma 2.4. Suppose that λ_n scales as in (2.29) and that conditions (2.26) and (2.31) hold. Furthermore, assume that:

$$\frac{1}{n}\boldsymbol{X}_M^T\boldsymbol{X}_M \to \mathbf{\Sigma}_M, \;\; a.s., \qquad |\beta_{0j}^M| > 0, \; \forall j \in M,$$

and that

$$\lim_{n\to\infty} \sup |\boldsymbol{X}_{-M}^T\boldsymbol{X}_M(\boldsymbol{X}_M^T X_M)^{-1}s| \leq \nu < \mathbf{1}, \;\; \forall s \in \{0,1\}^{|M|}, \tag{2.32}$$

for some constant ν, where $\mathbf{1}$ is a vector of ones and the inequality holds element wise. Under these conditions the following limits hold:

$$\lim_{n\to\infty} \inf_{\beta^M} P_{n,\beta^M}(M)e^n = \infty, \qquad \lim_{n\to\infty} \inf_{\beta^M \in U(\beta_0^M)} P_{n,\beta^M}(M) = 1.$$

The linear regression model trivially satisfies the modeling assumptions we made in the

previous section. Thus, under the conditions given in the lemmas stated in this section, the conditional MLE for a model selected by the Lasso can be guaranteed to be well behaved.

Corollary 2.1. *Fix a model $M \in \mathcal{M}$ and suppose that the conditions of either Lemma 2.3 or Lemma 2.4 are satisfied. Then the conditional MLE* (2.27) *is consistent for β_0^M.*

Remark 2.1 (High-Dimensional Problems)**.** The Lasso is often used in cases where the number of covariates p is much larger than n. In order to make asymptotic analysis relevant to such cases it is common to assume that p grows with the sample size. While the theory developed here does not explicitly treat such a high-dimensional setting, none of our assumptions prevent us from allowing the model selection function S_n to consider a growing number of covariates as n grows. Specifically, if we assume that the ℓ_1 penalty scales at the rate of $\lambda_n = O\left(\sqrt{n \log p_n}\right)$ as prescribed e.g. by Hastie et al. (2015), then our theory applies as long as the assumptions of Lemma 2.4 are satisfied and $\log p_n = o(n)$.

Remark 2.2 (Normality)**.** While we made a simplifying normality assumption, we expect that for fixed dimension p, non-normal errors can be addressed using conditions similar to those outlined by Tibshirani et al. (2015). For theory for selective inference with non-normal errors in the high-dimensional case, see the work of Tian and Taylor (2015).

2.5 Simulation study

In order to more thoroughly assess the performance of the proposed post-selection estimator for the Lasso, we perform a simulation study, which we pattern after that in Meinshausen (2007). We consider prediction and coefficient estimation using Lasso, our conditional estimator and refitted Lasso. We note already that while some existing theoretical works outline conditions under which the refitted Lasso should outperform the Lasso in prediction and estimation (Lederer, 2013), this does not occur in any of our simulation settings. For confidence intervals we compare our Wald confidence intervals to the confidence intervals of Lee et al. (2016) which we term *Polyhedral*. We find that both selection adjusted methods achieve close to nominal coverage rates.

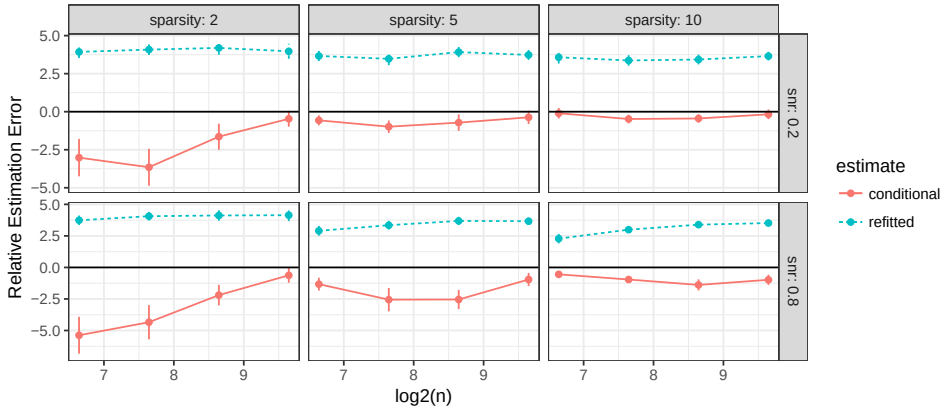

Figure 2.5: The relative estimation error of the regression coefficients compared to the Lasso as defined in (2.33). The error of the conditional estimates (solid red line) is lower than that of the Lasso in all simulation settings and the error of the refitted least-squares estimates (dashed blue line) was worse than that of the Lasso in all simulations.

We generate artificial data for our simulations in a similar manner as we have done for Example 2.3 in Section 2.3.3. We vary the sample size $n = 100, 200, 400, 800$, signal-to-noise ratio snr $= 0.2, 0.8$, and the sparsity level $k = 2, 5, 10$. For each combination of parameter values we generate data and fit models 400 times. We keep the amount of dependence fixed at $\rho = 0.5$ and the number of candidate covariates fixed at $p = 400$.

In Figure 2.5 we plot the log relative estimation error of the refitted-Lasso estimates and the conditional estimates compared to the Lasso as defined by:

$$\frac{1}{|M|}\left(\sum_{j\in M}\log_2(\hat{\beta}_j - \beta_j) - \log_2(\hat{\beta}_{Lasso_j} - \beta_j)\right). \tag{2.33}$$

This measure of error gives equal weights to all regression coefficients regardless of their absolute magnitude. In all simulation settings the refitted least-squares estimates are significantly less accurate than the Lasso or the conditional estimates. The conditional estimates tend to be more accurate than the Lasso estimates in all simulation settings. The conditional estimate tends to do better when there are at least some large regression coefficients in the

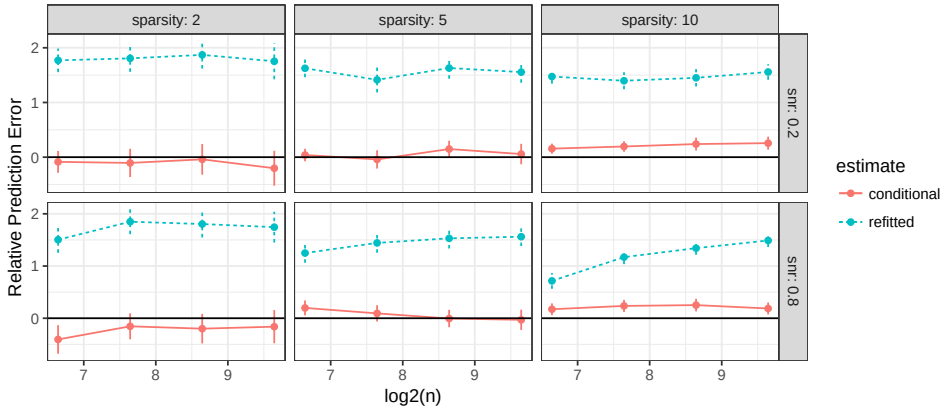

Figure 2.6: The log of the ratio between the prediction errors for the conditional (solid red line) and refitted least-squares regression estimates (dashed blue line) relative to the prediction error of the Lasso as defined in (2.34). The conditional MLE produces better prediction than the Lasso when the signal is spread over a smaller number of variables.

true model.

In Figure 2.6 we present the relative prediction error of the refitted least-squares Lasso estimates and the conditional estimates, as defined by:

$$\log_2 \| \boldsymbol{X}\, \hat{\beta} - \mu \|_2^2 - \log_2 \| \boldsymbol{X}\, \hat{\beta}_{\text{Lasso}} - \mu \|_2^2. \tag{2.34}$$

Here, the Lasso provides more accurate predictions when the true model has more non-zero coefficients and the conditional estimator tends to be more accurate when the true model is sparse.

In Figure 2.7 we plot the coverage rates obtained by the Conditional-Wald confidence intervals proposed here, the Polyhedral confidence intervals and the refitted 'naive' confidence intervals. Both of the selective methods obtain close to nominal coverage rates. The coverage rates of the refitted confidence intervals which were not adjusted for selection were far below the nominal levels in all simulation settings.

While the two types of selection adjusted confidence intervals seem to be roughly on

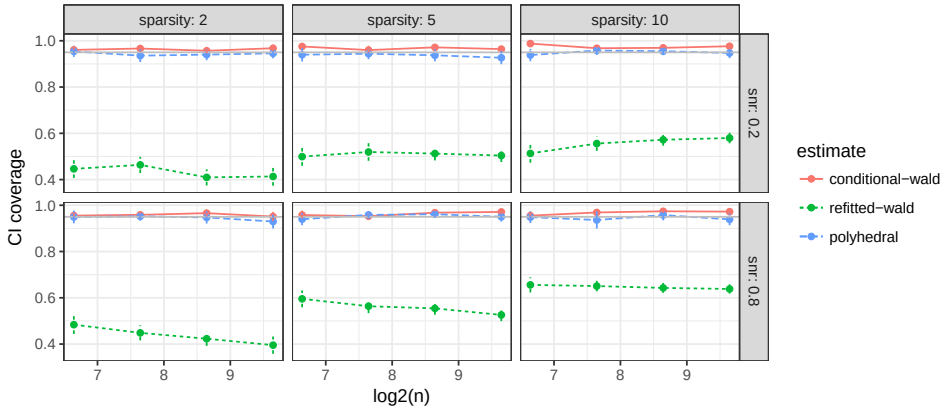

Figure 2.7: Confidence interval coverage rate after model selection. Both the Conditional Wald CIs (solid red line) and the Polyhedral CIs (dashed blue line) achieve the target coverage rate of 95% (horizontal grey line). The coverage rate of the unadjusted Wald confidence intervals (dotted green line) is far below nominal.

par with respect to their coverage rate, they tend to differ in their size. For Figure 2.8 we generate the additional datasets with a smaller number of candidate covariates $p = 200$, a larger range of sample sizes- $n = 40, 75, 150, 300, 600, 1250, 2500, 5000, 10000$, a signal-to-noise ratio of snr $= 0.2$ and $k = 10$ non-zero regression coefficients.

We face some difficulty in assessing the average size of the Polyhedral confidence intervals, as these sometimes have an infinite length. a measure for the length of a typical confidence interval, we take the median confidence interval length in each simulation instance. In Figure 2.8 we plot boxplots describing the distribution of the log relative size of the selection adjusted confidence intervals to that of the unadjusted refitted confidence intervals which tend to be the shortest. We find that as the sample size increases, the sizes of the Conditional-Wald confidence intervals are roughly twice the size the unadjusted confidence intervals, while the typical size of a Polyhedral interval is about twice the size of the Conditional-Wald confidence interval.

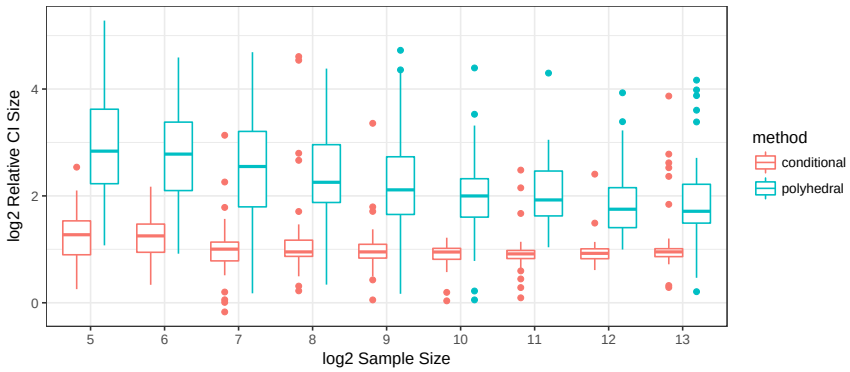

Figure 2.8: Boxplots of the relative median sizes of the selection adjusted confidence intervals to the in relations to the unadjusted ones. The Conditional-Wald confidence are much shorter than the Polyhedral ones under all simulation settings and their size are far less variable.

2.6 Conclusion

In this work we presented a computational framework which enables, for the first time, the computation of correct maximum likelihood estimates after model selection in selection with a possibly large number of covariates. We applied the proposed framework to the computation of maximum likelihood estimates of selected multivariate normal means and regression models selected via the Lasso.

Our methods take the arguably most ubiquitous approach to data analysis, that of computing maximum likelihood estimates and constructing Wald-like confidence intervals. Furthermore, we do not involve conditioning on information additional to the identity of the selected model. A practice which, as shown by Fithian et al. (2014), may lead to a loss in efficiency.

We experimented with the proposed estimators and confidence intervals in a comprehensive simulation study. The proposed conditional confidence intervals were shown to achieve conservative coverage rates and the point estimates were shown to be preferable to the refitted-least squares coefficients estimates in all simulation settings, and preferable to the Lasso coefficient estimates when there are large signals in the data.

While in this work we focused on inference in the linear regression method, our framework and theory are directly applicable to any exponential family distribution. Specifically, it is immediately applicable to estimation of parameters of selected generalized linear models using the normal approximations proposed by Taylor and Tibshirani (2016).

2.A Proof of theorems

2.A.1 Proof of Theorem 2.1

In their work on the convergence of stochastic gradient methods, Bertsekas and Tsitsiklis (2000) formulate a general stochastic gradient method as an iterative optimization method consisting of steps of the form:

$$x_{t+1} = x_t + \gamma_t(s_t + w_t),$$

where γ_t satisfies the condition from (2.6), s_t is a deterministic quantity related to the true gradient and w_t is a noise component. They outline conditions regarding s_t and w_t that ensure the convergence of the ascent algorithm to an optimum of a function $f(x)$ which possesses a gradient $\nabla f(x)$. The conditions require that there exist positive scalars c_1 and c_2 such that for all t:

$$c_1\|\nabla f(x_t)\|^2 \leq \nabla f(x_t)^T s_t, \qquad \|s_t\| \leq c_2(1 + \|\nabla f(x_t)\|), \tag{2.35}$$

and that

$$\mathbb{E}\left[w_t \,|\, \mathcal{F}_t\right] = 0, \tag{2.36}$$

$$\mathbb{E}\left[\|w_t\|^2 \,\middle|\, \mathcal{F}_t\right] \leq A\left(1 + \|\nabla f(x_t)\|\right), \tag{2.37}$$

where $\mathcal{F}_t$ is the filtration at time t, representing all historical information available at time t regarding the sequence $(w_t, s_t)_{i=1}^{\infty}$.

In our case, the function of interest is the conditional log-likelihood $f(x) = l(\mu) := \log \mathcal{L}(\mu)$, where the coordinates of μ which were not selected are imputed with the corresponding observed coordinates of y. The conditions regarding the deterministic component in (2.35) hold as $s_t = \nabla l(\mu|M)$, is the gradient itself. In Theorem 2.1 we assumed that we are able to take independent draws from the truncated multivariate normal distribution, meaning that

$$\mathbb{E}\left[w_t \,|\, \mathcal{F}_t\right] = \mathbb{E}\left[y^t - \nabla l(\mu|M)\right] = 0.$$

In practice, we should make sure that we run the Markov chain for a sufficiently large number of iterations between gradient updates in order for (2.36) to hold in good approximation.

The remaining issue is to bound the variance of w_t. The first step is finding an upper bound for the variance of w_t as a function of μ. In the following, we denote by $f(y)$ the unconditional density of y, by $f(y_j)$ the marginal (unconditional) density of y_j and by $f(y_{-j}|y_j)$ the conditional distribution of y_{-j} given y_j. Since the mean minimizes an expected squared deviation we have

$$\begin{aligned}
\mathbb{E}\left[(y_j - \mathbb{E}(y_j|M))^2 \,\middle|\, M\right] &\leq \mathbb{E}\left[(y_j - \mu_j)^2 \,\middle|\, M\right] \\
&= \int (y_j - \mu_j)^2 \, f(y|M)\, dy \\
&= \int_M (y_j - \mu_j)^2 \, \frac{f(y)}{P(M)}\, dy.
\end{aligned}$$

Let $C(y_j) = \int_M f(y_{-j}|y_j)\, dy_{-j}$, which satisfies $0 \leq C(y_j) \leq 1$. Then

$$\begin{aligned}
\int_M (y_j - \mu_j)^2 \, \frac{f(y)}{P(M)}\, dy &= \int_M (y_j - \mu_j)^2 \, \frac{C(y_j)}{P(M)} f(y_j)\, dy_j \\
&\leq \int_M (y_j - \mu_j)^2 \, \frac{1}{P(M)} f(y_j)\, dy_j \\
&\leq \int_{\mathbb{R}} (y_j - \mu_j)^2 \, \frac{1}{P(M)} f(y_j)\, dy_j = \frac{\sigma_j^2}{P(M)}. \qquad (2.38)
\end{aligned}$$

The next step in bounding the variance of w_t is bounding $P(M)$ from below. The difficulty

with finding a lower bound $P(M)$ is that one may make it arbitrarily small by varying the coordinates of μ for the non-selected coordinates. This is the motivation behind setting them to the observed values and only estimating the selected coordinates, resulting in the Z-estimator described in (2.8).

Assume without loss of generality that the first k coordinates of μ were not selected and that the last $p-k+1$ were selected. We write

$$P(M) = \int_M f(y)dy = \int_M f(y_1|y_2,\ldots,y_p) \times \cdots \times f(y_p)\, dy.$$

We begin with the integration with respect to y_1:

$$\int_M f(y_1|y_2,\ldots,y_p)\, dy_1 = 1 - \Phi(u_1; \mu_{1,-1}, \sigma^2_{1,-1}) + \Phi(l_1; \mu_{1,-1}, \sigma^2_{1,-1}).$$

Now, denote by $m_j = (l_j + u_j)/2$ the mid-point between l_j and u_j. We have

$$\begin{aligned} 1 - \Phi(u_1; \mu_{1,-1}, \sigma^2_{1,-1}) + \Phi(l_1; \mu_{1,-1}, \sigma^2_{1,-1}) \;&\geq\; 1 - \Phi(u_1; m_1, \sigma^2_{1,-1}) + \Phi(l_1; m_1, \sigma^2_{1,-1}) \\ &\geq\; \Phi(l_1; m_1, \sigma^2_{1,-1}) \;\geq\; \Phi(l_1; u_1, \sigma^2_{1,-1}). \end{aligned}$$

We can apply a similar lower bound to all selected coordinates to obtain:

$$\begin{aligned} P(M) &\geq \prod_{j\in M} \Phi(l_j; u_j, \sigma^2_{j,-j}) \int_M f(y_{p-k+1}|y_{p-k+2},\ldots,y_p) \times \cdots \times f(y_p)\, dy_{p-k+1} \ldots dy_p \\ &= P\left(j \notin S(y)\, \forall j \notin M\right) \prod_{j\in M} \Phi(l_j; u_j, \sigma^2_{j,-j}). \end{aligned} \tag{2.39}$$

Taking (2.38) and (2.39) together, we obtain the desired bound:

$$\mathrm{Var}(y_j) \leq \frac{\mathrm{tr}(\mathbf{\Sigma})}{P(\bigcap_{j\notin M}\{j \notin M\}) \prod_{j\in M} \Phi(l_j; u_j, \sigma^2_{j,-j})}.$$

□

The proof of Theorem 2.2 follows in a similar fashion.

2.A.2 Proof of Lemma 2.2

The proposal vectors defined in the lemma are given by:

$$r_1^{\rightarrow} = \left(\eta_1^t, \ldots, \eta_{j-1}^t, r_j, \eta_{j+1}^{t-1}, \ldots, \eta_p^{t-1}, \xi^t\right), \quad r_2^{\rightarrow} = \left(r_1, \ldots, r_{j-1}, r_j, r_{j+1}, \ldots, r_p, \xi^t\right),$$
$$r_1^{\leftarrow} = \left(r_1, \ldots, r_{j-1}, \eta_j^{t-1}, r_{j+1}, \ldots, r_p, \xi^t\right), \quad r_2^{\leftarrow} = \left(\eta_1^t, \ldots, \eta_{j-1}^t, \eta_j^{t-1}, \eta_{j+1}^{t-1}, \ldots, \eta_p^{t-1}, \xi^t\right).$$

The proposed algorithm for sampling $\eta|M, \xi$ is a two-step Delayed Rejection Metropolis-Hastings sampler. In our case the first step is to propose a sample from the full conditional distribution of η_j given η_{-j}. We denote the first proposal by $r_1^{\rightarrow}$. Note that at this stage only the jth coordinate has been changed. The acceptance probability for this step is given by:

$$\alpha(r_2^{\leftarrow}, r_1^{\rightarrow}) = \frac{f(r_{1,j}^{\rightarrow}|r_{1,-j}^{\rightarrow})}{f(r_{2,j}^{\leftarrow}|r_{2,-j}^{\leftarrow})} \frac{f(r_{2,j}^{\leftarrow}|r_{2,-j}^{\leftarrow})}{f(r_{1,j}^{\rightarrow}|r_{1,-j}^{\rightarrow})} I\{S_n(X, r_1^{\rightarrow}) = M\} = I\{S_n(\boldsymbol{X}, r_1^{\rightarrow}) = M\}.$$

That is, the acceptance probability of the first proposal is either 1 or 0 depending on whether the proposal satisfies conditions (2.13) and (2.14).

If the first proposal is not accepted and (2.14) is satsifeid, then we make a second proposal $r_2^{\rightarrow}$. The acceptance probability for the second proposal as defined by Mira (2001) is given by:

$$\alpha(r_2^{\leftarrow}, r_1^{\leftarrow}, r_2^{\leftarrow}) = \frac{f(r_2^{\rightarrow}) q_1(r_2^{\rightarrow}, r_1^{\leftarrow}) q_2(r_2^{\rightarrow}, r_1^{\leftarrow}, r_2^{\leftarrow}) \left(1 - \alpha(r_2^{\rightarrow}, r_1^{\leftarrow})\right)}{f(r_2^{\leftarrow}) q_1(r_2^{\leftarrow}, r_1^{\rightarrow}) q_2(r_2^{\leftarrow}, r_1^{\rightarrow}, r_2^{\rightarrow}) \left(1 - \alpha(r_2^{\leftarrow}, r_1^{\rightarrow})\right)},$$

where $q_1(x, y)$ is the density of the first proposal and $q_2(x, z, y)$ is the density of the second proposal. We only make a second proposal if $\alpha(r_2^{\leftarrow}, r_1^{\rightarrow}) = 0$ and therefore the ratio is always zero if $r_1^{\leftarrow}$ is a legal value. If both $r_1^{\leftarrow}$ and $r_1^{\rightarrow}$ are illegal then $\alpha(r_2^{\leftarrow}, r_1^{\leftarrow}, r_2^{\leftarrow})$ is non-zero and

the proposal densities are given by:

$$q_1(x, y) = f\left(y_j | \{y_j < l_j\} \cup \{u_j < y_j\}, x_{-j}\right),$$
$$q_2(x, z, y) = \prod_{k \neq j} \frac{\varphi(y_k; x_k, \sigma^2_{k,-k})}{P(y_k \in (a_k, b_k); x_k, \sigma^2_{k,-k})}.$$

Put together, we get:

$$q(x, y) := q_1(x, y) q_2(x, z, y),$$

which yields the desired result. □

2.A.3 Proof of Theorem 2.3

Under the assumptions of Theorem 2.3 we show that the unadjusted MLE is consistent even in the presence of model selection, in the sense that:

$$\lim_{n \to \infty} P(\|\hat{\theta}^M_n - \theta^M_0\|_\infty \geq \varepsilon | M) = 0.$$

We prove this result by showing that it holds for a model $M \in \mathcal{M}$ that satisfies the conditions of the theorem. Assume without loss of generality that $\theta^M \in \Theta^M \subseteq \mathbb{R}^p$. In the following we will use the shorthand $I_n(M) = I_{\{S_n(y)=M\}}$. The results follows from the fact that as long as the probability of model selection can be bounded from below, then the selection thresholds cannot be too far a way from the true parameters.

$$
\begin{aligned}
&\lim_{n\to\infty} P(\|\hat{\theta}_n^M - \theta_0^M\|_\infty \geq \varepsilon | M) \\
&= \lim_{n\to\infty} \frac{P_n(M|\{\|\hat{\theta}_n^M - \theta_0^M\|_\infty \geq \varepsilon\})P(\|\hat{\theta}_n^M - \theta_0^M\|_1 \geq \varepsilon)}{P_n(M)} \\
&\leq \lim_{n\to\infty} \frac{P(\|\hat{\theta}_n^M - \theta_0^M\|_\infty \geq \varepsilon)}{P_n(M)} \\
&= \lim_{n\to\infty} \frac{P\left(\bigcup_{j=1}^p \{|\hat{\theta}_{nj}^M - \theta_{0j}^M| \geq \varepsilon\}\right)}{P_n(M)} \\
&\leq \lim_{n\to\infty} \sum_{j=1}^p \frac{P(|\hat{\theta}_{nj}^M - \theta_{0j}^M| \geq \varepsilon)}{P_n(M)} \\
&= \lim_{n\to\infty} \sum_{j=1}^p \frac{P(|\sqrt{n}(\hat{\theta}_{nj}^M - \theta_{0j}^M)| \geq \sqrt{n}\varepsilon)}{P_n(M)} \\
&\leq^{(*)} \lim_{n\to\infty} \sum_{j=1}^p \frac{2e^{\frac{-n\varepsilon^2}{2\sigma_{Mj}}}}{P_n(M)} =^{(**)} 0,
\end{aligned}
$$

where σ_{Mj}^2 is the jth diagonal element of $\mathbf{\Sigma}^M$ and $(*)$ holds by subgaussian concentration. The equality $(**)$ holds by our assumption regarding the rate at which $P_n(M)$ is allowed to tend to zero. □

2.A.4 Proof of Theorem 2.4

Before we prove the theorem, we first state and and prove a couple of Lemmas that will come in handy in the proof of Theorem 2.4. Lemma 2.5 to follow states that the conditional MLE is consistent for θ_0^M even when used in the non-conditional setting (when the model to be estimated is pre-determined).

Lemma 2.5. Set a family of distributions M and assume that no data-driven model selection has been performed. Then under the conditions of Theorem 2.4 the conditional MLE is consistent for θ_0^M, that is,

$$P(\|\hat{\theta}_n^M - \theta_0^M\|_\infty > \varepsilon) \to 0.$$

Proof. Consider once again the conditional MLE

$$\begin{aligned}\hat{\theta}_n^M &= \arg\max_{\theta^M} G_n^M = \arg\max_{\theta^M} \frac{1}{n}\sum_{i=1}^{n}\left[\ell_{\theta^M}(y_i) - \frac{1}{n}\log P_{n,\theta^M}(M)\right] \\ &:= \bar{\ell}_n(\theta^M) - \frac{1}{n}\log P_{n,\theta^M}(M).\end{aligned}$$

where $\ell_{\theta^M}(y_i)$ is the unconditional log-likelihood of y_i. We are evaluating the properties of the conditional estimator in the unconditional setting where M is designated for inference before the data are observed. In this setting, the conditional MLE can be considered an M-estimator obtained from performing inference under a misspecified likelihood.

We now show that $\hat{\theta}_n^M$ is consistent for the θ_0^M. We have

$$\sup_{\theta^M} G_n^M(\theta^M) \geq G_n^M(\theta_0^M),$$

which implies that

$$\bar{\ell}_n(\hat{\theta}_n^M) \geq \bar{\ell}_n(\theta_0^M) - \frac{1}{n}\log P_{n,\theta_0^M}(M) + \frac{1}{n}\log P_{n,\hat{\theta}_n^M}(M). \tag{2.40}$$

Equation (2.40) together with assumption (2.24) gives

$$\bar{\ell}_n(\tilde{\theta}_n^M) \geq \bar{\ell}_n(\theta_0^M) - o(1). \tag{2.41}$$

Thus, the conditions for consistency as given by van der Vaart (1998) (Theorem 5.14 p. 48) are satisfied. The implication of (2.41) is that in the unconditional setting the conditional M-estimator is a consistent estimator. □

Next, we show that the difference between the conditional expectation of the sufficient statistic $\bar{T}_n$ converges to the unconditional expectation. This result will assist us later in proving a law-of-large number type statement for $\bar{T}_n$ under the conditional distribution.

Lemma 2.6. Under the assumptions of Theorem 2.4, for all $\delta < 1/2$,

$$n^{\delta} \| \mathbb{E}_{\theta_0^M}(\bar{T}_n) - \mathbb{E}_{\theta_0^M}(\bar{T}_n|M) \| \to 0.$$

Proof. According to Lemma 2.1, if $y_i \sim f_{\theta^M}$ with f_{θ^M} an exponential family distribution and $P_{n,\theta^M}(M|\bar{T}_n) \in \{0,1\}$ then the first derivative of the conditional log-likelihood is

$$\frac{\partial}{\partial \theta^M} G_n(\theta^M) = \frac{1}{n}\sum_{i=1}^{n} T(y_i) - \mathbb{E}_{\theta^M}(T(y_i)|M) := \bar{T}_n - \mathbb{E}_{\theta^M}(\bar{T}_n|M).$$

At the maximizer of $G_n(\theta^M)$, for any $\delta < 1/2$, we have:

$$n^{\delta}\left[\bar{T}_n - \mathbb{E}_{\hat{\theta}_n^M}(\bar{T}_n|M)\right] = 0,$$

which implies that

$$n^{\delta}(\bar{T}_n - \mathbb{E}_{\theta_0^M}(\bar{T}_n)) + n^{\delta}(\mathbb{E}_{\theta_0^M}(\bar{T}_n) - \mathbb{E}_{\hat{\theta}_n^M}(\bar{T}_n|M)) = 0.$$

Since $n^{\delta}(\bar{T}_n - \mathbb{E}_{\theta_0^M}(\bar{T}_n)) = o_p(1)$ by law of large numbers, we obtain that

$$n^{\delta}(\mathbb{E}_{\theta_0^M}(\bar{T}) - \mathbb{E}_{\hat{\theta}_n^M}(\bar{T}|M)) = o_p(1).$$

Finally in order to prove the desired results we must show that

$$E_{\theta_0^M}(\bar{T}|M) - E_{\hat{\theta}_n^M}(\bar{T}|M) \to 0.$$

It is clear that since $\theta_n^M \to \theta_0^M$, a fixed continuous function of $\hat{\theta}_n^M$ will converge as the sample size grows. However, $E_{\hat{\theta}_n^M}(\bar{T}|M)$ is a function of both $\hat{\theta}_n^M$ and n, and we must make sure that it does not vary too much with n in order for the desired convergence to hold. Define

$t = a^T \bar{T}_n$. By assumption (2.25) we have that for some sufficiently large n:

$$\sup_{\theta^M : \|\theta^M - \theta_0^M\| < \frac{1}{\sqrt{n}}} |\, \mathbb{E}_{\theta_0^M}(t|M) - \mathbb{E}_{\theta^M}(t|M)| \leq \sup_{\theta^M : \|\theta^M - \theta_0^M\| < \frac{1}{\sqrt{n}}} \frac{\mathrm{Var}_{\theta^M}(t)}{P_{n,\theta^M}(M)} \frac{1}{\sqrt{n}}.$$

Because y is of an exponential distribution and t is an average we can bound the unconditional variance in the neighborhood of θ_0^M. For a sufficiently small $\varepsilon > 0$ there exists a constant $C > 0$ such that,

$$\sup_{\theta^M : \|\theta^M - \theta_0^M\| \leq \varepsilon} \mathrm{Var}_{\theta^M}(t) < \frac{C}{n}$$

because $\mathrm{Var}_{\theta^M}(t)$ is a continuous function and the supremum is taken over a compact set. Thus, by the $\sqrt{n}$ consistency of $\hat{\theta}_n^M$ for θ_0^M, the difference satisfies $n^\delta |\, \mathbb{E}_{\theta_0^M}(t|M) - \mathbb{E}_{\hat{\theta}_n^M}(t|M)| = o(1)$ for any vector a as well as for $\bar{T}_n$ itself and the claim follows. □

We are now ready to prove Theorem 2.4. The first step in the proof is showing that $\bar{T}_n$ converges in probability conditionally on M. This result is a simple consequence of Markov's inequality and our assumption that $P_n(M)^{-1} = o(n)$. Set an arbitrary vector $a \in \mathbb{R}^p$ and define $t = a^T \bar{T}_n$. By Markov's inequality,

$$P_n(|t - \mathbb{E}_n(t|M)| > \varepsilon | M) \leq \frac{\mathrm{Var}_n(t|M)}{\varepsilon^2} \leq \frac{O(n^{-1})}{\varepsilon^2 P_n(M)} = o(1). \tag{2.42}$$

To see why (2.42) holds, write:

$$\mathrm{Var}_n(t|M) = \int \frac{(t - \mathbb{E}(t))^2}{P_n(M)} I\{S_n(\bar{T}_n) = M\} f(t) d(t) \; - \; [\mathbb{E}(t) - \mathbb{E}_n(t|M)]^2$$

$$\leq \frac{a^T \mathrm{Var}(T(y_i)) a}{n P_n(M)}.$$

By the fact that (2.42) holds for any arbitrary vector a, together with Lemma 2.6, we can determine that conditionally on M, $\bar{T}_n \to_p \mathbb{E}(\bar{T}_n)$.

By our assumption that the log-likelihood $l_{\theta^M}(y)$ is a continuous mapping of $T(y)$, as-

sumption (2.24) and Lemma 2.6, conditionally on the selection of M we have:

$$\frac{1}{n}\sum_{i=1}^{n}\ell_{\theta^M}(y_i) - \frac{1}{n}\log P_{n,\theta^M}(M) \to_p \mathbb{E}[\ell_{\theta^M}(y_i)].$$

The rest of the proof follows in a similar manner to the proof of Lemma 2.5 where the law of large numbers in the proof of Theorem 5.14 in van der Vaart (1998) is replaced by (2.42) and our assumption that $\bar{\ell}_n(\theta^M)$ is a continuous function of $\bar{T}_n$. □

2.A.5 Proof of Lemma 2.3

In the context of this proof we use the following notation:

$$A_0(M,s) := \{l_o(M,s) \leq \boldsymbol{A}_0(M,s)y < u_0(M,s)\},$$

$$A_1(M,s) := \{\boldsymbol{A}_1(M,s)y < u_1(M,s)\}.$$

For ease of exposition, we make a simplifying assumption that

$$\lim_{n\to\infty}\frac{\lambda_n}{n^{\frac{1}{2}}} = \lambda^*.$$

We begin by bounding the probability of not selecting the null-set. By our assumption that $n^{-1}\,\boldsymbol{X}^T\,\boldsymbol{X}$ converges, we have that the thresholds $l_0(M,s)$ and $u_0(M,s)$ also convergence for all candidate models and sign permutations. Furthermore, by our assumption regrading the rate in which λ_n grows and the expectation of $\boldsymbol{A}_0(M)y$,

$$\boldsymbol{A}_0(M)y \to^D N(0, \boldsymbol{\Sigma}(\boldsymbol{A}_0)),$$

where,

$$\boldsymbol{\Sigma}(\boldsymbol{A}_0) = \lim_{n\to\infty}\frac{\sigma^2}{\lambda_n^2}\,\boldsymbol{X}_{-M}^T(I - \boldsymbol{X}_M(\boldsymbol{X}_M^T\,\boldsymbol{X}_M)^{-1}\,\boldsymbol{X}_M^T)\,\boldsymbol{X}_{-M}\,.$$

Thus,

$$\lim_{n\to\infty} P_n(\boldsymbol{A}_0(M,s)) = c_0(M,s) > 0, \;\; \forall M, s.$$

Since the probability $\boldsymbol{A}_0(M,s)$ can be bounded in a uniform manner, we can set

$$c_0(M) := \min_s c_0(M,s),$$

and obtain a lower bound for the probability of selecting M by bounding

$$P_n(M) \geq c_0(M) P_n\left(\cup_s \boldsymbol{A}_1(M,s)\right) := c_0(M) P_n\left(\boldsymbol{A}_1(M)\right).$$

We bound $P_n(\boldsymbol{A}_1(M))$ next. Recall that the threshold a regression coefficient must cross is given by

$$u_1(M,s) = -\lambda_n \, \mathrm{diag}(s)(\boldsymbol{X}_M^T \boldsymbol{X}_M)^{-1} s.$$

This threshold is a bit unwieldy, as it depends on the signs of the active set and an exact realization of $\boldsymbol{X}_M$. Since we are interested in asymptotic behavior of random quantities, it will be sufficient to work with the limiting value of the threshold:

$$u_1^*(M,s) = \lim_{n\to\infty} \sqrt{n} u_1(M,s) = -\lambda^* \, \mathrm{diag}(s) \, \boldsymbol{\Sigma}_M^{-1} \, s,$$

Now, in order to eliminate the dependence on the signs of the active set define:

$$u_1^*(M) := \sup_s \sup_j \left| \left(\lambda^* \, \mathrm{diag}(s) \, \boldsymbol{\Sigma}_M^{-1} \, s\right)_j \right|,$$

and define an event:

$$\tilde{\boldsymbol{A}}_1 := \{\sqrt{n}|\eta_j| > u_1^*(M), \;\; \forall j \in M\}.$$

In $\tilde{A}_1$ we replaced all coordinate thresholds with the largest threshold, and so it is clear that:

$$\limsup_{n\to\infty} \frac{P_{n,\beta^M}(\tilde{A}_1)}{P_{n,\beta^M}(A_1)} \leq 1.$$

Furthermore, we have the lower bound

$$P_{n,\beta^M}(\tilde{A}_1) \geq \prod_{j \in M} \left(\Phi(-u^*(M); 0, \sigma^2_{j,-j}) + 1 - \Phi(u^*(M); 0, \sigma^2_{j,-j}) \right), \quad \forall \beta^M \in \mathbb{R}^{|M|}, \quad (2.43)$$

where $\sigma^2_{j,-j} := Var(\sqrt{n}\eta_j | \eta_{-j})$. See the proof of Theorem 2.1 for details on how this bound is derived. The rest follows by our normality assumption and the fact that (2.43) holds for all β^M including β_0^M. □

2.A.6 Proof of Lemma 2.4

We begin by treating the probability of satisfying the conditions for not selecting the variables not in the model. Using the same notations as in the proof of Lemma 2.3, the following limit holds:

$$\boldsymbol{\Sigma}(A_0) \to 0,$$

and consequently, by assumption (2.32):

$$\lim_{n \to \infty} P_n(\boldsymbol{A}_0(M, s)) = 1, \quad \forall s.$$

Next, we treat the probabilities of satisfying the conditions for selecting the variables included in the model. As before, we make a simplifying assumption that there exists a constant $0 < \delta < 0.5$ such that:

$$\frac{\lambda_n}{n^{0.5+\delta}} = \lambda^*,$$

In the fast scaling case, a lower bound on $P_{n,\beta^M}(\tilde{A}_1)$ no longer exists because the threshold $u^*(M)$ grows with the sample size. However, we can show that a satisfactory bound exists at β_0^M. Since in this setting λ_n grows faster than $\sqrt{n}$, we redefine the limit of the selection threshold:

$$u_1^*(M, s) = \lim_{n \to \infty} \frac{\sqrt{n}}{n^\delta} u_1(M, s) = -\lambda^* \operatorname{diag}(s) \boldsymbol{\Sigma}_M^{-1} s.$$

We can redefine $u_1^*(M)$ in an analogous manner. Now, we rewrite the bound (2.43) at the

point $\beta^M = \beta_0^M$ and with $u_1^*(M)$ properly scaled as

$$P_{n,\beta_0^M}(\tilde{A}_1) \geq \prod_{j \in M} \left(\Phi(-u^*(M)n^\delta; \sqrt{n}\beta_0^M, \sigma^2_{j,-j}) + 1 - \Phi(u^*(M)n^\delta; \sqrt{n}\beta_0^M, \sigma^2_{j,-j}) \right)$$

With no loss of generality assume that $\beta_0^M < 0$ to obtain the desired bound:

$$\lim_{n\to\infty} P_{n,\beta_0^M}(\tilde{A}_1) \geq \lim_{n\to\infty} \prod_{j \in M} \Phi\left(-u^*(M)n^\delta; \sqrt{n}\beta_0^M, \sigma^2_{j,-j}\right) = 1,$$

where the limit holds because $\delta < 0.5$. A similar result holds in a small neighborhood U of β_0^M because the probability of selection is continuous in β^M.

In order to bound the infimum of $P_{n,\beta^M}(A_1)$, we again start from (2.43) to get:

$$\begin{aligned}
P_{n,\beta^M}(\tilde{A}_1) &\geq \prod_{j \in M} \left(\Phi\left(-u^*(M)n^\delta; 0, \sigma^2_{j,-j}\right) + 1 - \Phi\left(u^*(M)n^\delta; 0, \sigma^2_{j,-j}\right) \right) \\
&\geq \prod_{j \in M} \Phi\left(-u^*(M)n^\delta; 0, \sigma^2_{j,-j}\right) \\
&\geq C \left(\frac{n^\delta u_1^*(M)/\sigma_{j,-j}}{1 + u_1^*(M)^2 n^{2\delta}/\sigma^2_{j,-j}} \right)^{|M|} \prod_{j \in M} e^{-\frac{u_1^*(M)^2 n^{2\delta}}{2\sigma^2_{j,-j}}} \qquad (2.44) \\
&= O\left(\frac{e^{-n^{2\delta}}}{n^{\delta|M|/2}} \right).
\end{aligned}$$

The lemma follows by our assumption that $\delta < 0.5$. In (2.44) we used the inequality:

$$\Phi(t; 0, \sigma^2) \geq C \frac{t/\sigma}{1 + t^2/\sigma^2} e^{-\frac{t^2}{2\sigma^2}}.$$

□

2.B Numerical examples for the Lasso MLE

In Section 2.3.1 we discuss the conditions that must hold in order for a specific model to be selected by the Lasso and propose to estimate the mean vector $\boldsymbol{A}_0(M)E(y)$ by 0. Here, we

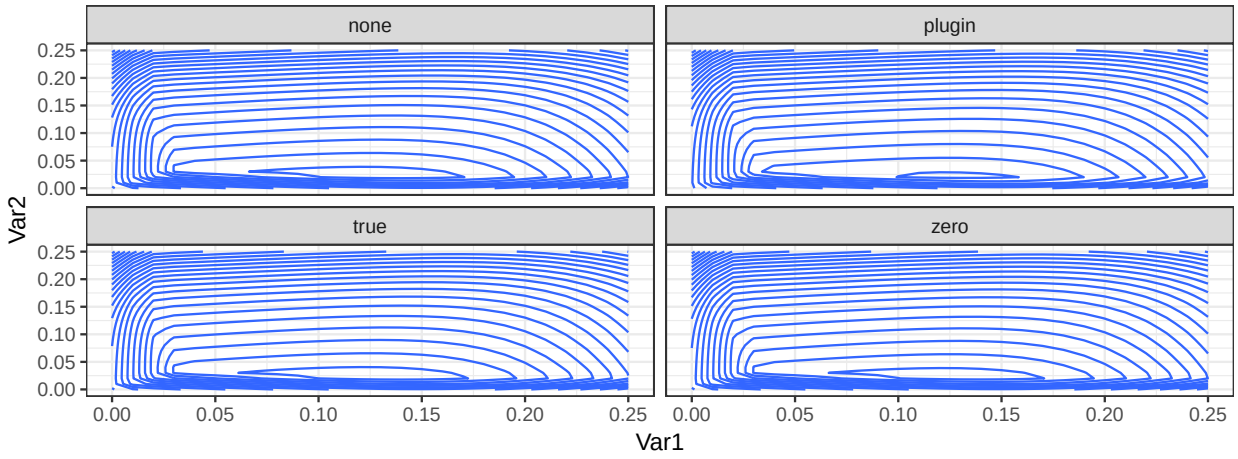

Figure 2.9: Contour plots for the first numerical experiment described in Appendix 2.B. The contour plots describe the log-likelihood of a model selected by the Lasso as a function of the values of the regression coefficients where the probability of not selecting the inactive set was computed in four different ways as described in the text.

propose some alternatives and seek to demonstrate that the proposed method is a reasonable one.

We generate data using the same process as described in Example 2.3 with parameter values $\rho = 0.5$, $n = p = 100$, $k = 3$ and snr $= 0.5$. We selected a model with two active parameters of positive sign with observed values of 0.17 and 0.13. In order to compute the conditional log-likelihood for this example we must decide on appropriate estimates for $E(\boldsymbol{A}_0(M)y)$. We present results for three options. The first is to use the observed value, $\boldsymbol{A}_0\, y$ as an estimate for its expectation, we term this method 'plug-in'. The second is to work under the assumption that $E(\boldsymbol{A}_0\, y) \approx 0$, estimating the expectation with a vector of zeros, we term this method 'zero'. A third option is to simply assume that $P(l < A_0 y < u) \approx 1$ for all signs sets, we term this method 'none'. Finally, we also compute the likelihood under the truth, setting $E(\boldsymbol{A}_0\, y) = \boldsymbol{A}_0\, E(y)$.

We draw the contour plots for the two-dimensional log-likelihoods as a function of the selected regression coefficients in Figure 2.9. While the contour plots are visually similar,

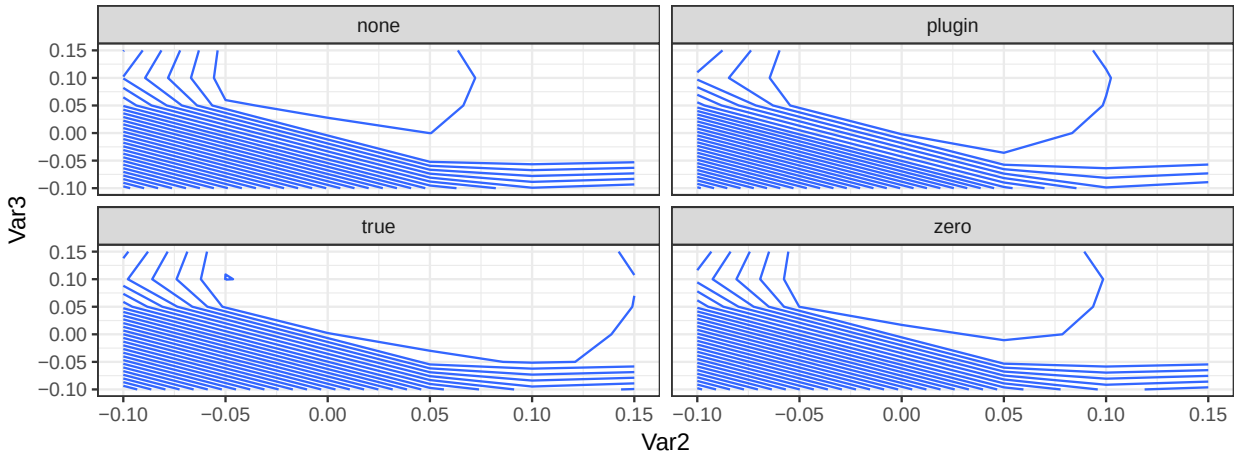

Figure 2.10: Contour plots for the second numerical experiment described in Appendix 2.B. The contour plots describe the log-likelihood of a model selected by the Lasso as a function of the values of the regression coefficients where the probability of not selecting the inactive set was computed in four different ways as described in the text.

the values of the log-likelihoods differ slightly. For the 'none' and 'zero' methods the log-likelihood was maximized at $0.14, 0.02$ at a log-likelihood value of 14.2. This is similar to the log-likelihood computed under the true expectation, where the maximum was also obtained at $0.14, 0.02$ and at a slightly different value of 14.3. Finally, for the plug-in method the maximum was obtained at $0.13, 0.02$ with a value of 16.9. Thus, for this example, the maximum likelihood estimates computed using the different imputation methods yielded results that are essentially equivalent. In this example the true probability of $P(l_0 < \boldsymbol{A}_0\, y < u_0)$ was close to 1 for all sign permutations.

In a second example we generate data using parameter values $\rho = 0.8$, $n = 100$, $p = 500$, $k = 5$ and $\text{snr} = 0.2$. Here we selected a model with four variables where the observed refitted regression coefficients estimates were $0.13, 0.17, 0.21$ and 0.15. For all estimation methods the maximum of the log-likelihood was obtained at approximately $0, -0.05, 0.1, 0$. The values of the log-likelihood function at its maximum was 15.9 when no imputation was used, 19.9 for plugin imputation, 16.1 for the zero imputation and 16.7 when the true parameter value was

used to compute the log-likelihood. The contour of the log-likelihood function are plotted in Figure 2.10 for the second and third variables, keeping the values of the first and last coefficients fixed at zero.

2.C Description of algorithms

Algorithm 1: Stochastic ascent algorithm for the normal means problem.

input : $y, l, u \in \mathbb{R}^n$, $\Sigma^{-1} \in \mathbb{R}^{p \times p}$.
output : $\hat{\mu} \in \mathbb{R}^p$.
initialization: $y^0 \leftarrow y$, $\mu^0 \leftarrow y$.
for $i \in 1 : I$ **do**
  Set $z^0 \leftarrow y^{i-1}$;
  for $t \in 1 : T$ **do**
    for $j \in 1 : p$ **do**
      Sample $z_j^t \sim f_{\mu^i}(z_j | M, z_1^t, \ldots, z_{j-1}^t, z_{j+1}^{t-1}, \ldots, z_p^{t-1})$;
  Set $y^i \leftarrow z^T$;
  for $j \in M$ **do**
    $\mu_j^i \leftarrow \mu_j^{i-1} + \gamma^i \Sigma_{j,\cdot}^{-1}(y - y^i)$;
return μ^I;

Algorithm 2: Sampler for the post-selection distribution under selection by Lasso.

input : $\eta \in \mathbb{R}^{|M|}$, $\lambda \in \mathbb{R}^+$, $X \in \mathbb{R}^{n \times p}$, $\sigma^2 \in \mathbb{R}^+$.
output : A sample point η.

for $t \in 1 : T$ **do**
  Sample $\xi^t \sim f(\xi | M, \eta)$;
  for $j \in 1 : p$ **do**
    Set $r^{\rightarrow} \leftarrow \eta$;
    Sample $r_j^{\rightarrow} \sim f\left(\eta_j | \{\eta_j < l_j\} \cup \{\eta_j > u_j\}, \eta_{-j}\right)$;
    if $l_0(M, \text{sign}(r^{\rightarrow})) < \xi^t < u_0(M, \text{sign}(r^{\rightarrow}))$ **then**
      if $r^{\rightarrow}$ *is in the set from* (2.13) **then**
        Set $\eta \leftarrow r^{\rightarrow}$;
      else
        for $k \neq j$ **do**
          Sample $r_k^{\rightarrow} \sim \text{TN}(a_k, b_k, \eta_k, \sigma^2_{k,-k})$;
        Set $r^{\leftarrow} \leftarrow r^{\rightarrow}$;
        Set $r_j^{\leftarrow} \leftarrow \eta_j$;
        if $r^{\leftarrow}$ *is not in the set from* (2.13) **then**
          Compute p_j^t as in (2.21) ;
          Sample $U \sim \text{Unif}(0, 1)$;
          if $U < p_j^t$ **then**
            Set $\eta \leftarrow r^{\rightarrow}$;

return η ;

Algorithm 3: Stochastic ascent algorithm for the Lasso.

input : $I \in \mathbb{N}$, $\lambda, \sigma^2 \in \mathbb{R}^+$, $X \in \mathbb{R}^{n \times p}$, $y \in \mathbb{R}^n$.
output : $\hat{\beta} \in \mathbb{R}$.

initialization: Set $\hat{\beta}^0, \eta^0 \leftarrow (X_M^T X_M)^{-1} X_M^T y$.

for $i \in 1 : I$ **do**
  Sample η^i using Algorithm 2 ;
  Set $\hat{\beta}^i \leftarrow \hat{\beta}^{i-1} + \gamma_i (X_m^T y - (X_M^T X_M)\eta^i)$;
  for $j \in 1 : p$ **do**
    Set $\hat{\beta}_j^t \leftarrow \text{sign}(\hat{\beta}_j^0) \max(0, \text{sign}(\hat{\beta}_j^0)\hat{\beta}_j^i)$;

return $\hat{\beta}^I$;

Chapter 3

POST-SELECTION TESTING AND ESTIMATION FOLLOWING AGGREGATED ASSOCIATION TESTING

This chapter was adapted from an arXiv pre-print co-written with Ruth Heller and Nilanjan Chatterjee (Heller et al., 2017b).

3.1 Introduction

After working on the estimation problem, I gave a talk at Tel-Aviv University where I met Ruth Heller who, at the time, was already working on the problem of inference following aggregate testing. This problem is an excellent case-study in post-selection inference, as it is a natural generalization of the univariate truncation problem that is more tractable than the general regression problem.

In our work we formulated the polyhedral lemma for inference following aggregate testing problem, but also proposed two additional approaches that guarantee asymptotic error control while improving upon the polyhedral lemma in terms of power to detect sparse signals:

1. Inference under a *most conservative* parametrization.
2. Regime switching methods that approximate the conditional distribution to obtain asymptotically consistent error rates.

We also showed that the problem of computing the conditional maximum likelihood estimator can be decomposed into the problem of finding the model selection probability that maximizes the conditional likelihood, and the problem of computing the parameters values that maximize the unconditional likelihood under the constraint that they obtain a specific model selection problem. This result comes in handy, for example, if the selection

event is affine and imposes a smaller number of constraints than the number of parameters in the unconditional likelihood. In particular, for aggregate testing with a Wald test we show that the problem of computing the conditional maximum likelihood estimate can be cast as a line search problem.

3.1.1 Inference following aggregate testing

Many modern scientific investigations involve simultaneous testing of many thousands of hypotheses. Valid testing of large number of hypotheses requires strict multiple-testing adjustments, making it difficult to identify signals in the data if the signal is weak or sparse. One possible remedy is to pool groups of related test statistics into aggregate tests. This practice reduces the amount of multiplicity correction that needs to be applied and may assist in identifying weak signals that are spread over a number of test statistics. However, once an 'interesting' group of hypotheses has been identified, it may also be of interest to perform inference within the group in order to identify the individual test statistics that drive the signal.

In many scientific fields, there exist a natural predefined grouping of features of interest. In neuroscience, functional magnetic resonance imaging (fMRI) studies aim to identify the locations of activation while a subject is involved in a cognitive task. The individual null hypotheses of no activation are at the voxel level, and regions of interest can be tested for activation by aggregating the measured signals at the voxel level (Benjamini and Heller, 2007; Penny and Friston, 2003). Following identification of the regions of interest, it is meaningful to localize the signal within the region. In microbiome research, the operational taxonomic units (OTUs) are grouped into taxonomic classifications such as species level and genus level. The data for the individual null hypotheses of no association between OTU and phenotype can be aggregated in order to test the null hypotheses of no association at the species or genus level (Bogomolov et al., 2017). Here as well, following identification of the family associated with the phenotype, it is of interest to identify the OTUs within the family that drive the association. In genome-wide association studies (GWAS), the

disease being analyzed may have multiple subtypes of interest. The standard analysis aim is to identify the SNPs associated with the overall disease, but another important aim is to identify associations with specific sub-types of the disease (Bhattacharjee et al., 2012). In genetic association studies, there is also a natural grouping of the genome, since genes are comprised of single variants. The test statistics of single variants within a gene can be aggregated into a test statistic for powerful identification of associations at the gene level (Bhattacharjee et al., 2012; Derkach et al., 2014; Wu et al., 2011; Yoo et al., 2016). Following identification at the gene level, it may be of interest to identify the single variants within the gene that drive the association.

For a single group of features, let $\hat{\beta} = (\hat{\beta}_1, \ldots, \hat{\beta}_m)$ be the estimator of the vector of parameters of interest in the group, β. Much research has focused on developing powerful aggregate tests for selecting the groups of interest, i.e., for testing at the group level the null hypothesis that $\beta = 0$. When $\hat{\beta}$ has (approximately) a known covariance and a normal distribution, classical test-statistics are the score, Wald, and likelihood ratio statistics, all of which have an asymptotic χ^2_m distribution. A recent example is the work by Reid et al. (2016a), which suggested novel tests that improve on classical tests. Other examples come from the field of statistical genetics, where many gene level tests have been recently proposed based on weighted linear or quadratic combinations of score statistics for analyzing genomic studies of rare variants, see Derkach et al. (2014) for a review. In this work we seek to develop methods for conducting inference on the coordinates of β following selection by an aggregate test.

Recently, Heller et al. (2017a) addressed the problem of identifying the individual studies with association with a feature, following selection of potential features by a meta-analysis of multiple independent studies. We generalize the work of Heller et al. (2017a) to allow for (approximately) known dependence across the individual test statistics. In particular, this allows for valid testing of predictors in a generalized linear model that was selected via an aggregated test. We further develop methods for obtaining post-selection point estimates and confidence intervals. We also discuss computation of post-selection confidence intervals

which are based on inversion of the post-selection tests. Finally, we develop regime switching post-selection tests and confidence intervals that adapt to the unknown underlying sparsity of the signal, and thus have good power when the signal is sparse as well as when it is non-sparse.

The paper is organized as follows. In Section 3.2 we formally introduce our inference framework and goals. We develop theory for post-selection testing and estimation in Section 3.3 and Section 3.4, respectively. We conduct empirical evaluation of our test-statistics and post-selection estimates in Section 3.5. In Section 3.6, we apply our methods to a genomic application. Finally, Section 3.7 concludes.

3.2 The set-up and the inferential goals

Let $\hat{\beta} \sim N(\beta, \mathbf{\Sigma})$ with $\mathbf{\Sigma}$ known, and suppose that we are interesting in performing inference on $\beta \in \mathbb{R}^m$ if and only if we can reject an aggregate test of the global-null hypothesis that $\beta = 0$. For testing the global-null hypothesis, we use a quadratic test of the form $S = \hat{\beta}^T \boldsymbol{K} \hat{\beta} > S_{1-t_1}$ where $\boldsymbol{K}$ is a positive semi-definite matrix and S_{1-t_1} is the $1 - t_1$ quantile of S under the null-hypothesis. Setting $\mathbf{K} = \mathbf{\Sigma}^{(-1)}$ results in the well known Wald test statistic. The developments when group selection is by a linear aggregate test $S = a^T \hat{\beta}$ are similar.

The value of t_1 comes from the analysis at the group level. For example, in genomics, when the group is the gene, then typically $t_1 \approx \alpha/20000$. This is because the Bonferroni procedure is commonly used for identifying genes associated with phenotypes using aggregate tests, so the FWER on the family of $\sim 20,000$ genes is controlled at level α.

Given that an aggregate test has been rejected at a level t_1, our aim is to infer on the parameters $\beta_1, \ldots, \beta_m$. For $j \in \{1, \ldots, m\}$, let

$$H_j : \beta_j = 0.$$

Our first aim is to test the family of hypotheses $\{H_j : j = 1, \ldots, m\}$ if $p^G \leq t_1$, with FWER

or FDR control. The conditional FWER and FDR (introduced in Heller et al., 2017a) for the selected group are, respectively, $E(I[V > 0]|S > S_{1-t_1})$ and $E(V/\max\{R, 1\}|S > S_{1-t_1})$, where V and R are the number of false and total rejections in the group. We provide procedures for conditional FWER/FDR control in Section 3.3.

Our second aim is to estimate the magnitude of the regression coefficients $\beta_1, \ldots, \beta_m$ given selection. Denoting the likelihood for β by $\mathcal{L}(\beta)$, the conditional likelihood can be written as

$$\mathcal{L}(\beta|S > S_{1-t_1}) = \frac{\mathcal{L}(\beta)}{P_\beta(S > S_{1-t_1})} I\{S > S_{1-t_1}\}.$$

We propose to use the maximizer of the conditional likelihood as a point estimate, and we show how to obtain it, as well as confidence intervals, in Section 3.4. The following example demonstrates how our framework can be applied to inference following aggregate-testing in genetics.

Example 3.1. Generalized Linear Models. *Suppose we observe a response vector* $y = (y_1, \ldots, y_n) \in \mathbb{R}^n$*, and* m *predictors of interest in a group (e.g., the single variants in a gene),* $X_j, j = 1, \ldots, m$*. Let* $V_j, j = 1, \ldots, k$ *be a set of additional covariates to be accounted for in the model (e.g., environmental factors or ancestry variables in GWAS). Suppose that we are interested in modeling the relationship of the predictors in a group with the response vector using a generalized linear model. So, we assume that* $y_i \sim f_{\theta_i}$*, an exponential family distribution with canonical parameter* $\theta = g^{-1}(\eta_i) \in \Theta$ *for some continuous link function* $g : \Theta \to \mathcal{R}$ *and*

$$\eta = \alpha_0 + \sum_{l=1}^{k} V_l \alpha_l + \sum_{j=1}^{m} X_j \beta_j. \tag{3.1}$$

In the case of linear regression, $g(\eta_i) = \eta_i$ *is the identity function and* $y_i \sim N(\eta_i, \sigma^2)$*. If* $X_1, \ldots, X_m$ *explain little of the variance in* y *(e.g., in genomic applications) it is reasonable to estimate* σ^2 *by the empirical variance of the residuals from the linear model with* $\beta = 0$*.*

When y *is not assumed to be normal, the maximum likelihood estimator for the regression coefficients has an asymptotic normal distribution* $\sqrt{n}(\hat{\beta} - \beta) \to^D N(0, \mathbf{I}^{-1}(\alpha, \beta))$ *and*

an asymptotic truncated-normal distribution post-selection. While $\mathbf{I}^{-1}(\alpha, \beta)$ *depends on* β *and therefore cannot be assumed to be known in general, if* $X_1, \ldots, X_m$ *explain little of the variance in* y *it is reasonable to estimate the variance of* $\hat{\beta}$ *under the assumption that* $\beta = 0$.

3.3 Testing following selection

In the absence of selection, we can test for $H_j : \beta_j = 0$ using the p-value of the test statistic β_j/SE_j: $p_j = 2(1 - \Phi(|\hat{\beta}_j/SE_j|))$ where $SE_j = \sqrt{e_j^T \mathbf{\Sigma} e_j}$ and e_j is the $m \times 1$ unit vector with a single entry of one in position $j \in \{1, \ldots, m\}$. However, conditionally on selection, P_j will often have a distribution that is stochastically smaller than uniform, meaning that its realization p_j will no longer be a valid p-value for testing H_j.

To correct for selection, it appears necessary to evaluate the probability that $S \geq S_{1-t_1}$. However, this probability depends on the unknown β, and hence it cannot be evaluated when H_j is true unless we assume that all other entries in β are zero. In the special case of $\beta = 0$ the distribution of $\hat{\beta}_j/SE_j$, conditional on $S \geq S_{1-t_1}$ is known. Of course, in practice we do not know whether any of the entries of β are non-zero.

In Section 3.3.1 we suggest a way around this problem, by computing a valid conditional p-values using the polyhedral lemma first introduced by Lee et al. (2016). In practice, we find that statistical tests based on the polyhedral lemma tend to have relatively low power if β is sparse, and thus, in Section 3.3.2 we suggest an inference method that automatically adapts to the sparsity level of β. In Section 3.3.3 we discuss applying multiple testing procedures to the valid conditional p-values.

3.3.1 Conditional p-values based on the polyhedral lemma

Denote by $TN(\mu, \sigma^2, \mathcal{A})$ the truncated normal distribution constrained to $\mathcal{A} \subseteq \mathbb{R}$, i.e. the conditional distribution of a $N(\mu, \sigma^2)$ random variable conditional on it being in $\mathcal{A}$. Let $F^{\mathcal{A}}_{\mu,\sigma^2}$ be the CDF of $TN(\mu, \sigma^2, \mathcal{A})$. The following theorem, which is a direct result of the polyhedral lemma of Lee et al. (2016), provides us with a conditional distribution of any linear contrast of $\hat{\beta}$ that we can use for post-selection inference.

Theorem 3.1. Let $\eta^T\hat{\beta}$ be a linear combination of $\hat{\beta}$, and $t_1 \in (0,1]$ a fixed selection threshold. Let $W = (\mathbf{I}_m - c\eta^T)\hat{\beta}$, where $\tau = (\eta^T\mathbf{\Sigma}\eta)^{-1}\mathbf{\Sigma}\eta$ and $\mathbf{I}_m$ is the $m \times m$ identity matrix. Then

$$\eta^T\hat{\beta} \mid S \geq S_{1-t_1}, W \sim TN(\eta^T\beta, \eta^T\mathbf{\Sigma}\eta, \mathcal{A}(W)), \tag{3.2}$$

where $\mathcal{A}(W)$ is defined in Lemma 3.2.

See Appendix 3.A.1 for the proof.

Since the only unknown parameter in the truncated distribution of (3.2) is $\eta^T\beta$, it is straightforward to compute a p-value under the null hypothesis and construct confidence intervals via test inversion.

Corollary 3.1. *For the estimation of β_j, let $W = W_j = (\mathbf{I}_m - \tau {e_j}^T)\hat{\beta}$, $\tau = ({\eta_j}^T\mathbf{\Sigma}\eta_j)^{-1}\mathbf{\Sigma}\eta_j$, and $\mathcal{A}(W)$ as defined in Lemma 3.2. Then,*

$$P\left(\beta_j \in \{b:\ \alpha/2 \leq F^{\mathcal{A}}_{b,{\eta_j}^T\mathbf{\Sigma}\eta_j}(\hat{\beta}_j) \leq 1-\alpha/2\}\right) = 1-\alpha.$$

For testing H_j, let

$$P'_j = 1 - F^{\mathcal{A}(W)}_{0,{\eta_j}^T\mathbf{\Sigma}\eta_j}(\hat{\beta}_j). \tag{3.3}$$

Then, if H_j is true the distribution of P'_j given the selection event $S > S_{1-t_1}$ and $W = w$ is $U(0,1)$.

The following example serves to give some intuition as to how the polyhedral lemma works and the possible adverse effects of the extra conditioning on W.

Example 3.2. Independence Model. *Let $\hat{\beta} \sim N(\beta, \mathbf{I}_m)$ and suppose that we are interested in testing $H_1 : \beta_1 = 0$ after rejecting the global-null hypothesis that $\beta = 0$. In this case, the relevant contrast is $\eta = e_1$ and the orthogonal projection is $W = (0, \hat{\beta}_2, ..., \hat{\beta}_m)$. It is clear that $\hat{\beta}_1$ is independent of W and therefore, conditionally on selection the only relevant information contained in W is that $\hat{\beta}_1 > S_{1-t_1} - \sum_{j=2}^m \hat{\beta}_j^2$ so $\mathcal{A} = \{b:\ b^2 - S_{1-t_1} + \sum_{j=2}^m \hat{\beta}_j^2 > 0\}$.*

Note that if we do not condition on W then the support of $\hat{\beta}_1|S > S_{1-t_1}$ is $\mathbb{R}$ and this is why conditioning on W often results in a loss of efficiency (Fithian et al., 2014).

3.3.2 A hybrid conditional p-value

Our empirical investigation in Section 3.5 suggests that p-values that are computed based on the polyhedral lemma tend to have good power when β is not sparse or has a large magnitude. However, when only a single entry in β is nonzero, p-values based on the polyhedral lemma (which are valid for any configuration of the unknown β) tend to be considerably less powerful than p-values computed based on the distribution of $\hat{\beta}$ under the global-null distribution (where it is assumed that $\beta = 0$). Therefore, we would like to consider a test that adapts to the unknown sparsity of the signal, by combining the two approaches for computing p-values into a single test of H_j, allowing for powerful identification of the non-null coefficients. The combined test will be useful in applications where both groups with sparse signals and with non-sparse signals are likely.

Sampling from the truncated multivariate normal distribution is a well studied problem, see for example Pakman and Paninski (2014). Specifically, under the global null, i.e., $\beta = 0$, one can use samples from the truncated distribution to asses the likelihood of the observed regression coefficients, defining

$$p'_{j,GN} = P_{\beta=0}\left(P'_{j,GN} \leq p'_{j,GN} \mid S > S_{1-t_1}\right) = \frac{1}{t_1} P_{\beta=0}\left(P'_{j,GN} \leq p'_{j,GN}, S > S_{1-t_1}\right), \qquad (3.4)$$

$j = 1, \ldots, m$. Under the global-null distribution, both P'_j and the p-value computed under the global-null distribution $P'_{j,GN}$ have a uniform distribution. However, p'_j may be larger than $p'_{j,GN}$ because it requires extra conditioning on W. Thus, if the only non-zero predictor in the model is the jth predictor, the test based on $P'_{j,GN}$ can be expected to be more powerful than the test based on P'_j.

On the other hand, when more than one of the coordinates of β are non-zero, $p'_{j,GN}$ will often be substantially larger than the original p-value p_j, e.g., if $\hat{\beta}_j^2/SE_j^2 \geq S_{1-t_1}$, then

$p'_{j,GN} = p_j/t_1$. This, while P'_j may not suffer any additional loss of power due to the extra conditioning e.g., if the aggregate test passes the selection threshold t_1 regardless of the value of p_j, then $p'_j = p_j$ and it will clearly be smaller than $p'_{j,GN}$.

Since the preference for using $p'_{j,GN}$ instead of p'_j depends on the (unknown) β, we suggest the following test that combines the two valid post-selection p-values,

$$p'_{j,hybrid} = 2\min(p'_j, p'_{j,GN}). \tag{3.5}$$

Clearly, $p'_{j,hybrid}$ would be a valid p-value, i.e., with a null distribution that is either uniform or stochastically larger than uniform, if both p'_j and $p'_{j,GN}$ are valid p-values. In the previous section we indeed showed that p'_j is a valid p-value. But by the definition in equation (3.4) it is only clear that $p'_{j,GN}$ is valid when $\beta = 0$. Intuitively, for $\beta \neq 0$, we may assume that $p'_{j,GN}$ is conservative (i.e., has a null distribution that is stochastically larger than uniform). We shall now provide a rigorous justification.

We start with the special case that the quadratic aggregate test for selection is Wald's test. Following selection by Wald's test, $P'_{j,hybrid}$ is a valid p-value for testing $H_j : \beta_j = 0$. This follows by showing that the marginal null distribution of $P'_{j,GN}$ is at least stochastically as large as the uniform, so the test based on the global null distribution where $\beta = 0$ is conservative.

Theorem 3.2. If $S = \hat{\beta}^T \mathbf{\Sigma}^{(-1)} \hat{\beta}$, and $\hat{\beta}$ has a normal distribution with mean β and variance $\mathbf{\Sigma}$, then

$$P_{(\beta_1,\ldots,\beta_{j-1},0,\beta_{j+1},\ldots,\beta_m)} \left(P'_{j,GN} \leq x\right) \leq x \quad \forall x \in [0,1].$$

See Appendix 3.A.2 for the proof.

More generally, when selection is based on $S = \hat{\beta}^T \mathbf{K} \hat{\beta} > S_{1-t_1}$, where $\mathbf{K}$ is any positive definite symmetric matrix, we can still justify the use of $p'_{j,hybrid}$ for testing $H_j : \beta_j = 0$ for a large enough sample size. This follows since the conditional p-values under the global null are necessarily larger than the original p-values, as formally stated in the following lemma.

Lemma 3.1. If $\mathbf{K}$ is a positive definite matrix, $S = \hat{\beta}^T \mathbf{K} \hat{\beta}$, and $\hat{\beta} \sim N(\beta, \mathbf{\Sigma})$, then

$$Pr_{\beta=0}(\hat{\beta}_j^2 > b | S > s) \geq Pr_{\beta_j=0}(\hat{\beta}_j^2 > b) \tag{3.6}$$

for arbitrary fixed $b, s > 0$.

See Appendix 3.A.3 for the proof. Setting b to be the realized test statistic and $s = S_{1-t_1}$, $p'_{j,GN} = P_{\beta=0}(P_j \leq p_j \mid S > S_{1-t_1})$ is the lefthand side of (3.6) and $p_j = P(\chi_1^2 \geq \hat{\beta}_j^2 / SE_j^2)$ is the righthand side. It thus follows that

$$p'_{j,GN} \geq p_j.$$

Since $\lim_{n\to\infty} P_{(\beta_1,\ldots,\beta_{j-1},0,\beta_{j+1},\ldots,\beta_m)}(S > S_{1-t_1}) = 1$ regardless of the true value of β_j if $\beta_k \neq 0$ for at least one $k \neq j$, the probability of getting a smaller value than p_j given selection, if H_j is true, coincides with p_j asymptotically. So p_j is an asymptotically valid p-value if $\beta_k \neq 0$ for at least one $k \neq j$. Since $p'_{j,GN} \geq p_j$, it follows that $p'_{j,GN}$ and $p'_{j,hybrid}$ are asymptotically valid p-values for any β.

3.3.3 Controlling the conditional error rate

In order to identify the non-null entries in β, we can apply a valid multiple testing procedure on the conditional p-values computed as in 3.3.1 or 3.3.2. We can then achieve conditional error control.

The Bonferroni-Holm procedure will control the conditional FWER, since the conditional p-values are valid p-values and the procedure is valid under any dependency structure among the test statistics.

For conditional FDR control, we recommend using the Benjamini-Hochberg (BH) procedure. Although the BH procedure does not have proven FDR control for general dependence among the p-values, it usually controls the FDR for dependencies encountered in practice. We believe that the robustness property of the BH procedure carries over to our setting, and

that the conditional FDR will be controlled in practice. The robustness guarantee follows from empirical and theoretical results (Reiner-Benaim, 2007), which suggest that the FDR of the BH procedure does not exceed its nominal level for test statistics with a joint normal distribution, and our simulations in 3.5, which suggest that this holds also following selection.

A conservative procedure that will control the conditional FDR is the Benjamini-Yekutieli procedure for general dependence, introduced in Benjamini and Yekutieli (2001). The theoretical guarantee follows since the conditional p-values are valid p-values and the procedure is valid under any dependency structure among the test statistics.

If the individual test statistics are independent, as occurs when the design matrix $\mathbf{X}$ is orthogonal in the linear model, and the aggregate test statistic is monotone increasing in the absolute value of each test statistic (keeping all others fixed), then we have a theoretical guarantee that the BH procedure on $p'_1, \ldots, p'_m$ controls the conditional FDR, even though these conditional p-values are dependent. This is a direct result of Theorem 3.1 in Heller et al. (2017a), and it is formally stated in the following theorem.

Theorem 3.3. If $S = \hat{\beta}^T \mathbf{\Sigma}^{(-1)} \hat{\beta}$, $\hat{\beta} \sim N(\beta, \mathbf{\Sigma})$, and $\mathbf{\Sigma}$ is a diagonal matrix, then the BH procedure at level α on $p'_1, \ldots, p'_m$ controls the conditional FDR at level $m_0/m \times \alpha$, where m_0 is the number null coefficients in β.

3.4 Estimation following selection

So far we focused on valid testing after selection by an aggregate test. But it is often also desirable to asses the absolute magnitude of parameters of interest. Just as model selection causes an inflation of test statistics, it also has an adverse effect on the accuracy of point estimates. In fact, inflation of estimated effect sizes is the main cause for the increased type-I error rates that are encountered in naive inference following selection. In Section 3.4.1 we discuss the computation of post-selection of maximum likelihood estimators which are defined as the maximizers of the likelihood of the data conditional on selection and serve to correct for some of the selection bias.

Beyond point estimates, valid post-selection confidence intervals can be constructed by inverting the post-selection tests described in Section 3.3. These however, may be either underpowered in the case of confidence intervals based on the polyhedral lemma or too conservative in the case of the hybrid confidence intervals. Thus, in Section 3.4.2 we propose novel regime switching confidence intervals that maintain the validity and power of the hybrid method intervals while ensuring the desired level of confidence asymptotically.

3.4.1 Conditional maximum likelihood estimation

Let $\ell(\beta)$ be the log-likelihood for β, and $\ell(\beta|S > S_{1-t_1})$ the corresponding conditional log-likelihood. Define the conditional MLE as the maximizer of the conditional likelihood:

$$\tilde{\beta} = \arg\max_{\beta} \ell(\beta) - \log P_{\beta}(S > S_{1-t_1}). \tag{3.7}$$

For notational convenience, we suppress the dependence of $\tilde{\beta}$ on the selection threshold t_1. While difficult to compute in many practical cases, computing the conditional MLE following selection by aggregate testing is a relatively simple task. For the special case where $\mathbf{K} = \boldsymbol{\Sigma}^{(-\mathbf{1})}$, we are able to show that the maximum likelihood estimator is given by the solution to a simple line search problem.

Theorem 3.4. Under the conditions of Theorem 3.2, the conditional maximum likelihood estimator is given by:

$$\begin{aligned}\tilde{\beta} &= \arg\max_{\beta} \ell(\beta) - \log P_{\beta}(S > S_{1-t_1}) \\ &= \arg\max_{\lambda \in [0,1]} \ell(\lambda\hat{\beta}) - \log P_{\lambda\hat{\beta}}(S > S_{1-t_1})\end{aligned}$$

where $\hat{\beta}$ is the observed value.

See appendix 3.A.4 for the proof.

The Theorem shows that the maximum likelihood estimation is reduced to maximizing

the likelihood only with respect to a scalar factor. This follows when $\mathbf{K} = \mathbf{\Sigma}^{(-\mathbf{1})}$ because the distribution of the test-statistic is governed by one unknown parameter. In the general case, the distribution of S is a sum of chi-square random variables which depends on rank($\mathbf{K}$) parameters, making the optimization problem slightly more involved. So, for $\mathbf{K} \neq \mathbf{\Sigma}^{-\mathbf{1}}$ we use the stochastic optimization approach proposed in Chapter 2.

$$z(\beta) \sim f_\beta(\hat{\beta}|S > S_{1-t_1})$$

be a sample from the post selection distribution of $\hat{\beta}$ for a mean parameter value β. Then, taking gradient steps of the form

$$\tilde{\beta}^{t+1} = \tilde{\beta}^t + \gamma_t \mathbf{\Sigma}^{(-1)} \left(\hat{\beta} - z(\tilde{\beta}^t) \right) \tag{3.8}$$

will lead to convergence to the conditional MLE as long as

$$\sum_{t=1}^{\infty} \gamma_t = \infty \qquad \text{and} \qquad \sum_{t=1}^{\infty} \gamma_t^2 < \infty.$$

Theorem 3.5. Suppose that $\hat{\beta} \sim N(\beta, \mathbf{\Sigma})$ and that inference is conducted only if $S > S_{1-t_1}$ with $S = \hat{\beta}^T \mathbf{K} \hat{\beta}$. Then, the algorithm defined by (3.8) converges to the conditional MLE for the post-aggregate testing problem which satisfies

$$\lim_{t \to \infty} \hat{\beta} - E_{\tilde{\beta}^t}(\hat{\beta}|S > S_{1-t_1}) = 0.$$

Proof. The result follows from the fact that the variance of the post-selection distribution of $\hat{\beta}$ can be uniformly bounded from above by $\mathbf{\Sigma}/t_1$. □

The conditional MLE is consistent assuming the following. Suppose that we observe a sequence of regression coefficient estimates $\hat{\beta}_1, \ldots, \hat{\beta}_n, \ldots$ such that

$$\hat{\beta}_n \sim N(\beta, \mathbf{\Sigma}_n), \qquad n\mathbf{\Sigma}_n \text{ converges in probability.} \tag{3.9}$$

Furthermore, suppose that we perform inference on the individual coordinates of $\hat{\beta}_n$ if and only if

$$S_n > S_{1-t_1}, \qquad S_n = \hat{\beta}_n^T \mathbf{K_n} \hat{\beta}_n. \tag{3.10}$$

The good behaviour of the conditional MLE hinges on the probability of passing the selection by the aggregate test. The lower bound on this probability is given trivially by t_1 and therefore the conditional MLE is consistent.

Corollary 3.2. *Assume that* (3.9) *and* (3.10) *hold. Then, the conditional MLE is consistent for β, satisfying:*

$$\lim_{n\to\infty} P\left(\|\tilde{\beta}_n - \beta\|_\infty > \varepsilon | S_n > S_{1-t_1}\right) = 0, \qquad \forall \varepsilon > 0.$$

Proof. The result follows from the theory developed in the previous chapter for selective inference in exponential families and the fact that

$$\inf_\beta P_\beta(S_n > S_{1-t_1}) = t_1, \qquad \forall n.$$

□

3.4.2 Confidence intervals following selection by an aggregate test

From Theorem 3.1 it is clear that the truncated normal distribution can be used to construct confidence intervals post-selection in a straightforward manner. However, the extra conditioning (on W) may lead to wide confidence intervals relative to confidence intervals based on the sampled distributions. As an alternative, it is possible to invert a global type test (specifically, the test with null hypothesis $\beta = e_j b$ for coefficient β_j) and construct a hybrid type confidence interval in order to obtain a confidence interval with more power to determine the sign of the regression coefficients (Weinstein et al., 2013).

For constructing a confidence interval at a $1-\alpha$ level, let $L'_j(\alpha)$ and $U'_j(\alpha)$ be the lower

and upper bounds of the polyhedral confidence interval for the jth variable, so:

$$F^{\mathcal{A}}_{L',e_j{}^T\mathbf{\Sigma}e_j}(\hat{\beta}_j) = 1 - \frac{\alpha}{2}, \qquad F^{\mathcal{A}}_{U',e_j{}^T\mathbf{\Sigma}e_j}(\hat{\beta}_j) = \frac{\alpha}{2},$$

where $F^{\mathcal{A}}_{b,\sigma^2}$ is as defined in 3.3.1. Similarly, let $L^T_{GN,j}(\alpha)$ and $U^T_{GN,j}(\alpha)$ be the lower and upper limit of the global-null confidence interval for the jth variable:

$$\left\{b: \ \alpha/2 \leq F_{\beta=e_jb}(\hat{\beta}_j|S > S_{1-t_1}) \leq 1-\alpha/2\right\},$$

where e_j is the unit vector and $F_{\beta=e_jb}(\hat{\beta}_j|S > S_{1-t_1})$ is the CDF of $e_j^T\hat{\beta}$ given selection, for the parameter vector $\beta = e_jb$. We use the Robbins-Monroe process to find $L^T_{GN,j}$ and $U^T_{GN,j}$ (Garthwaite and Buckland, 1992). As in testing, the polyhedral confidence interval tends to be shorter and more efficient if there are several variables in the model that are highly correlated with the response variable and the global-null confidence intervals tend to be more powerful when the model is sparse or if the global-null hypothesis holds (approximately). As we have done in Section 3.3.2, we propose a hybrid method for constructing a confidence interval, as defined by the lower and upper bounds:

$$L_{hybrid,j}(\alpha) = \max\{L'_j(\alpha/2), L'_{GN,j}(\alpha/2)\}, \qquad U_{hybrid,j}(\alpha) = \min\{U'_j(\alpha/2), U'_{GN_j}(\alpha/2)\}.$$

The hybrid confidence intervals, while possessing a good degree of power to determine the sign regardless of the true underlying model, tend to be inefficient when there is strong signal in the data. To see why, consider the case of a regression model where $\beta_1, \beta_2 > 0$. Then, for a sufficiently large sample size the polyhedral confidence interval will apply no correction and hybrid confidence interval will be conservative, with an asymptotic level of $1 - \alpha/2$: $\lim_{n\to\infty} P\{(L', U')_{j,hybrid}(\alpha) = (L', U')_j(\alpha/2)\} = 1$. As a remedy, we propose a regime switching scheme for constructing confidence intervals in which we first determine whether $\|\beta\| \approx 0$ or $\|\beta\| \gg 0$ and then construct confidence intervals accordingly.

Procedure 3.1. The post-selection level $1-\alpha$ confidence interval for β_j, with switching regime at level $t_2 < \alpha \times t_1$ (with default value $t_2 = \alpha^2 \times t_1$):

1. Compute $S_{1-t_2} > S_{1-t_1}$.

2. If $S < S_{1-t_2}$, i.e., the aggregate test does not pass the more stringent threshold t_2, then compute the hybrid conditional confidence interval at level $1-\alpha^* = 1-(\alpha - t_2/t_1)$.

3. If $S \geq S_{1-t_2}$, compute the unconditional confidence interval, at level $1-\alpha^* = 1-\alpha$.

Theorem 3.6. Post-selection confidence intervals constructed with Procedure 3.1 have a confidence level at least $1-\alpha$ if $\beta = 0$, and an asymptotic level $1-\alpha$ if $\beta \neq 0$.

See Appendix 3.A.5 for the proof.

Remark 3.1. *Ours is not the first regime switching procedures proposed for inference in the presence of data-driven variable selection, see for example the works of Chatterjee and Lahiri (2011) and McKeague and Qian (2015). In both these cases, one has to determine whether some (or all) of the parameters are zero and construct a test in an appropriate manner. The usual prescription for selecting tuning parameters in such procedures is to scale the tuning parameter of the test (in our case, t_2) in such a way so the correct regime is selected with probability approaching one as the sample size grows. In our case, this would amount to setting $t_{2,n}$ in such a way so that $t_{2,n} \to 0$ and $S_{1-t_{2,n}} = o(n)$. However, in practice it is necessary to select a single a value for t_2 and so we chose to fix t_2 to a small value as to maintain a good degree of power when there is only limited amount of signal in the data and to modify our procedure in such a way as to ensure some finite sample coverage guarantees.*

Example 3.3. *Figure 3.4.2 shows point estimates and confidence intervals for normal means vector which was selected via a quadratic aggregate test. The figure was generated by sampling $\hat{\beta} \sim N_8(\beta, \mathbf{\Sigma})$ with $\mathbf{\Sigma}_{i,j} = 0.3 I_{i \neq j} + 1 I_{i=j}$, $\beta_1 = -2.5$, $\beta_2 = 0.5$ and $\beta_3 = \cdots = \beta_8 = 0$. The aggregate test applied was a Wald test at an $\alpha = 0.001$ level. The naive*

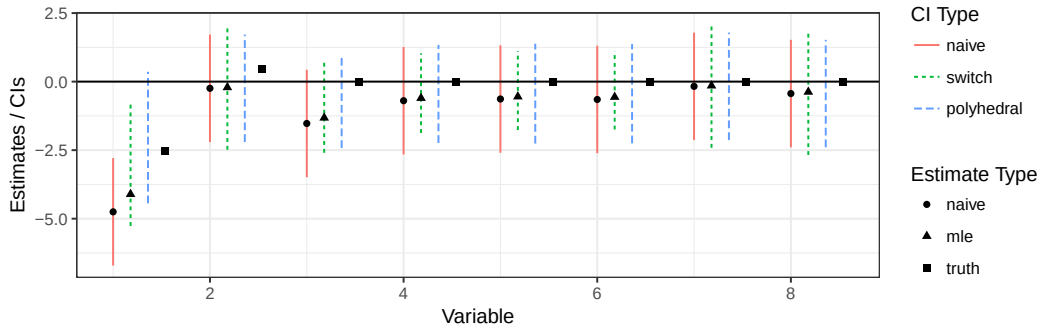

Figure 3.1: Point estimates and confidence intervals for the artificial data example described in Example 3.3. The naive point estimate is marked as a circle, the conditional MLE as a triangle and the true value of the parameter is marked as a square. The confidence intervals are the naive (solid red line), hybrid (dotted green line) and polyhedral (dashed blue line).

and conditional estimates are plotted along with naive, polyhedral and hybrid 95% *confidence intervals. The conditional MLE applies the same multiplicative shrinkage of* 0.86 *to all of the coordinates of* $\hat{\beta}$ *and so the shrinkage is more visible for the larger observed values. Because the selection is driven by* $\hat{\beta}_1$ *corresponding to the large negative coordinate* β_1*, the polyhedral confidence intervals for the other coordinates of* β *are similar in size to the naive confidence intervals. The naive confidence intervals overestimates the magnitude of* β_1*, the polyhedral confidence intervals cover the true parameter value but fails to determine its sign and the regime switching confidence intervals both cover the true parameter value and succeed in determining the sign.*

3.5 Simulations

In this section we conduct a simulation study where we assess the methods proposed in this work and verify our theoretical findings. In Section 3.5.1 we asses the post-selection tests proposed in Section 3.3 with respect to their ability to control the FDR. In Section 3.5.2 we compare the different testing method with respect to their power to detect true signal in the data. In Section 3.5.3 we compare the conditional MLE and the unadjusted MLE with respect to their estimation error. Finally, in Section 3.5.4 we asses the coverage rates of the

polyhedral and regime switching confidence intervals.

In all of our simulations we generate data in a similar manner. We first generate a design matrix in a manner meant to approximate a rare-variant design. We sample marginal expression proportions for our variants from $g_1, ..., g_m \sim Gamma(1, 300)$ constrained to $[2 \times 10^{-4}, 0.1]$ and for each subject we sample two multivariate normal vectors $r_{i,\cdot} \sim N(0, \mathbf{U})$ with $\mathbf{U_{i,j}} = \mathbf{0.8^{|i-j|}}$. We then set $X_{i,} = \sum_{k=1}^{2} I\{\Phi(r_{i,j}) \leq g_j\}$ to obtain a design matrix with dependent columns and a marginal distribution $X_{ij} \sim Bin(2, g_j)$. We generate a sparse regression coefficients vector with $m-s$ zero coordinates and s coordinates which are sampled form the Laplace(1) distribution. We normalize the values of the regression coefficients such that the signal to noise ratio

$$\text{snr} = \sqrt{\beta'(X'X)\beta} \tag{3.11}$$

equals some pre-specified value. Finally, we generate a response variable $y = \boldsymbol{X}\,\beta + \varepsilon$ with $\varepsilon \sim N(0, \boldsymbol{I})$. In all of our simulations we use a Wald aggregate test with a significance level of $t_1 = 0.001$.

3.5.1 Assessment of false discovery rate control

We assess how well the proposed testing procedures control the FDR under the assumed model as well as under model misspecification. We generate datasets with $m = 50$, $n = 10^4$, $s = 3$, snr $\in \{0, 0.032\}$ and three types of distributions for the model residuals, all of which have a variance of 1:

$$\varepsilon_i^{(1)} \sim N(0, 1), \qquad \varepsilon_i^{(2)} \sim Laplace(\sqrt{2}), \qquad \varepsilon_i^{(3)} \sim Unif\left(-\sqrt{12}/2, \sqrt{12}/2\right).$$

We compare four testing procedures. BH on naive p-values which are not adjusted for selection, BH on the polyhedral p-values as computed in equation (3.3), BH on the p-values based on the global null distribution as computed in equation (3.4), and BH on the hybrid p-values as computed in equation (3.5).

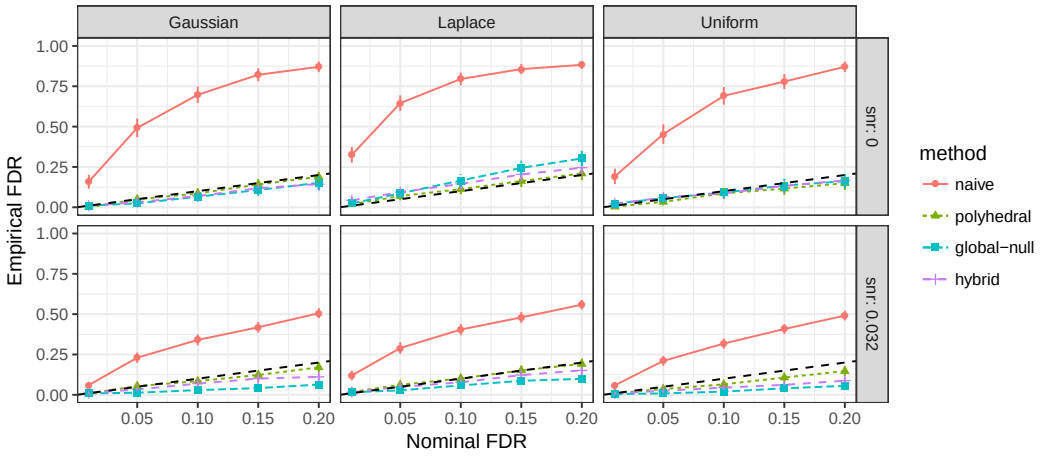

Figure 3.2: False Discovery Rates after aggregate testing. We plot the nominal FDR vs. the empirical FDR for the unadjusted naive p-values (red solid line), the polyhedral p-values (dotted green line), p-values based on the exact post-selection null distribution (dashed-blue) and the hybrid method (dashed purple line). The diagonal line is in dashed black. The figure is faceted according to the distribution of the noise and the signal to noise ratio in the data as defined in equation (3.11). Details about the data generation are in 3.5.1.

We plot the target FDR versus the empirical FDR in Figure 3.2. When there is no signal in the data and the noise is not heavy tailed, all selection adjusted methods obtain close to nominal FDR levels (top left and right panels). When the noise is heavy tailed (Laplace), the methods based on the null Gaussian distribution have higher than nominal FDR rates, while the p-values computed with the polyhedral method exhibit a more robust behavior (top center panel). When there is some signal in the data, all selection adjusted p-values control the FDR at nominal or conservative rate (bottom row). The naive p-values do not control the FDR in any of the simulation settings. Thus, we conclude that the polyhedral p-values may be preferable to the hybrid and global null p-values if the distribution of the data is heavy-tailed. However, as we show in the next section, the hybrid method tends to have more power compared to the polyhedral method and is preferable when the residual distribution is well behaved.

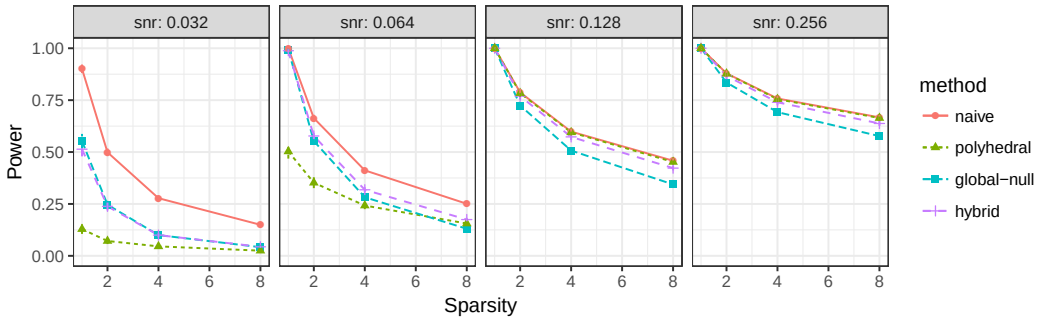

Figure 3.3: Power to detect true signals after aggregate testing. We plot the power of the different inference method as a function as the number of non-zero regression coefficients for the unadjusted naive p-values (red solid line), the polyhedral p-values (dotted green line), p-values based on the exact post-selection null distribution (dashed-blue) and the hybrid method (dashed purple line). The figure is faceted according to the strength of the signal as defined in equation (3.11). Details about the data generation are in 3.5.2.

3.5.2 Assessment of power to detect true signal

We compare the power to detect signal of the proposed testing procedures. We generate datasets with $m = 50$, $n = 10^4$, $s \in \{1, 2, 4, 8\}$, snr $\in \{0.032, 0.064, 0.128, 0.256\}$ and $\varepsilon_i \sim N(0, 1)$. We compare the same testing procedures as in 3.5.1. We measure the power to identify true signals at a nominal FDR level of 0.1.

We plot the results of the simulation in Figure 3.3. In all of the simulations the naive unadjusted p-values have the most power, at the cost of an inflated FDR. When the number of non-zero regression coefficients is small, the global null and hybrid methods tend to have the most power, while all methods have a similar power when the signal is spread over a large number of regression coefficients. The method based on the global null distribution is the most powerful when the signal is sparse and low. The polyhedral method has more power when the signal is not too sparse or low. The hybrid method seems to adapt to the sparsity and signal strength well, exhibiting comparatively good power in all settings.

3.5.3 Assessment of estimation error

We compare the conditional MLE to the naive, unadjusted point estimate, $\hat{\beta}$ itself. We set $n \in \{5000, 10000, 15000, 20000\}$, $m \in \{5, 10, 20\}$, $s = 2$, snr $\in \{0, 0.025\}$ and sample model residuals from the normal distribution with a standard deviation of 1

We plot the results of the simulation in Figure 3.4. When the dimension of β is small, the conditional MLE estimates the vector of regression coefficients better than the unadjusted MLE. The gap between the conditional and naive estimator is roughly constant across the different sample sizes when $\beta = 0$ because the probability of selection remains constant for all sample sizes. However, when there is some signal in the data the probability of passing the aggregate increases in the sample size and the gap between the estimators shrinks. The difference between the conditional MLE and the naive MLE decreases in the size of β, to the extent that for $m = 20$ the two estimators are indistinguishable from one another.

To see why this occurs, consider the following example. Let $y \sim N_m(0, \boldsymbol{I})$, suppose that we perform selection using a Wald test at a fixed level t_1 and consider the conditional likelihood function:

$$\mathcal{L}(y) \propto -\frac{1}{2}\sum_{i=1}^{m}(y_i - \mu_i)^2 - \log P_\mu(S > S_{1-t_1}).$$

As we let the dimension m grow, the decrease in the value of the (unconditional) gaussian log-likelihood due to a possible shrinkage of μ grows linearly in m. At the same time, the additional penalty term $-\log P_\mu(S > S_{1-t_1})$ remains bounded below by $-\log t_1$ regardless of the dimension of the problem.

3.5.4 Assessment of confidence interval coverage rates

In the last set of simulations, we evaluate the regime switching and polyhedral confidence intervals with respect to their coverage rates and power to determine the sign of the non-zero coefficients. We set the parameters of the simulation to $m = 20$, $n = 10^4$, $s \in \{1, 2, 4, 8\}$,

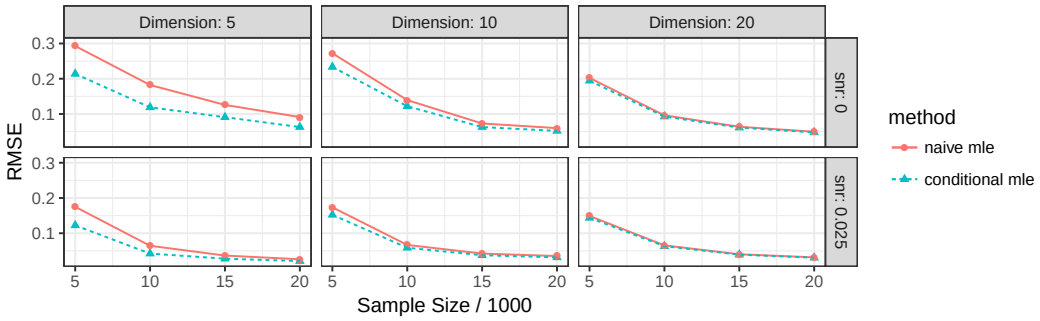

Figure 3.4: Root mean squared error for estimation after aggregate testing. We plot the RMSE for estimating the vector of regression coefficients β with the naive unadjusted estimator $\hat{\beta}$ (solid red line) and the conditional MLE $\tilde{\beta}$ (dashed blue line). The figure is faceted according to the signal-to-noise ratio as defined in (3.11) and the size of β, m.

snr $\in \{0.001, 0.002, 004, 008, 0.016, 0.032, 0.064, 0.128, 0.256\}$ and sample the residuals from a normal distribution with a standard deviation of 1.

We plot the results of the simulation in Figure 3.5. The naive confidence intervals have a coverage rate far below nominal for signal-to-noise ratios less than 1. As could be expected, the polyhedral method achieves the correct coverage rates up to Monte-Carlo error in all simulation settings. When there is no signal in the data the regime switching confidence intervals have close to nominal coverage. When the signal to noise ratio is moderate, the regime-switching confidence intervals are conservative because the polyhedral confidence intervals are superior to the ones based on the global-null assumption with high probability while the probability of S exceeding S_{1-t_2} is still not overwhelmingly large. When the signal to noise ratio is high the regime switching confidence intervals are mostly identical to the naive ones, because the selection occurs with probability of close to 1, and so they have the correct coverage rate. Despite being more conservative than the polyhedral confidence intervals, the regime switching confidence intervals can have better power to determine the sign. Specifically, the regime switching confidence intervals tend to have more power when the true model is sparse and the signal to noise ratio is low or moderate.

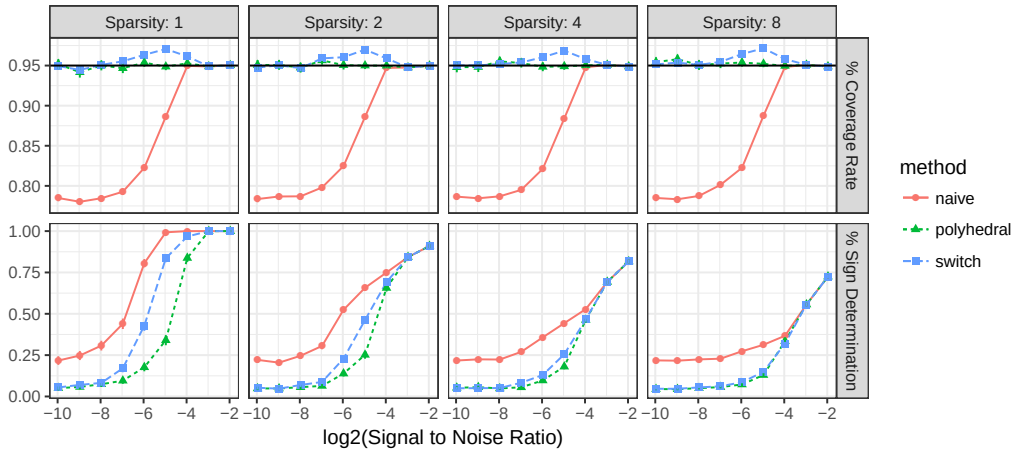

Figure 3.5: Coverage rates and power to determine the sign of confidence intervals constructed after aggregate testing. We plot results rates for the naive unadjusted confidence intervals (solid red line), polyhedral confidence intervals (dotted green line) and regime switching intervals with $t_2 = t_1\alpha^2$ (dashed blue line).

3.6 Application to variant selection following gene-level testing

Genome-wide association studies (GWAS) involves large scale association testing of genetic markers with underlying traits. Large GWAS of uncommon and rare variants are now becoming increasingly feasible with the advent of newer genotyping chips, cheaper sequencing technologies and sophisticated algorithms that allow imputation of low-frequency variants based on combinations of common variants that are already genotyped in large GWAS. Thus, association studies of rare variants is a very active area of research and some of the early studies have already begun to report their findings, e.g., Consortium et al. (2015) and Fuchsberger et al. (2016).

As the statistical power for testing association of traits with individual rare variants may be low, it has been suggested that tests for genetic associations be performed at an aggregated level by combining signals across multiple variants within genomic regions such as those defined by functional units of genes (Lee et al., 2012; Madsen and Browning, 2009; Morris and Zeggini, 2010; Neale et al., 2011; Sun et al., 2013; Wu et al., 2011). These

tests can be divided into sum-based tests (which aggregate the variant statistics by a linear combination), variance component tests (which aggregate the squared variant statistics by a linear combination), or combined (sum-based and variance component) tests. See Derkach et al. (2014) for a review. There is, however, currently a lack of rigorous methods for variant selection following gene-level association testing.

The Dallas Heart Study (DHS) (Romeo et al., 2007) considered four genes of potential interest, genotyped in 3549 individuals (601 hispanic, 1830 non-hispanic black, 1043 non-hispanic white, 75 other ethnicities). We focus on the 32 variants in ANGPTL4, which includes both rare and common variants. Table 3.1, column 2, shows the number of subjects with rare variants.

To detect associations with triglyceride (TG), a metabolism trait, we applied the variance component test SKAT of Wu et al. (2011) , with outcome TG on a logarithmic scale, while adjusting for the covariates race, sex, and age on a logarithmic scale. ANGPTL4 is one of the four genes in the ANGTPTL family (Romeo et al., 2009). Using a Bonferroni correction for testing the genes in the family, ANGTPL4 is selected for post-selection inference if the SKAT test p-value is at most 0.05/4. To identify the potentially susceptible variants, we proceeded as suggested in section 3.3.

The SKAT p-value for ANGTPL4 was 7.5×10^{-5} and therefore the gene was selected. Table 3.1, column 6 lists the weights assigned to each variant in the SKAT test. These weights were obtained using the default settings of the publicly available R library SKAT. Figure 3.6 and Table 3.1 provide, respectively, a graphical display and the actual numbers for the naive (i.e., unconditional, not corrected for selection) and conditional p-values. When using the polyhedral method described in Section 3.3.1, one variant, $E40K$, passes the Bonferroni threshold for FWER control at the 0.05 level. When using the hybrid method described in Section 3.3.2, two variants, $E40K$ and $R278Q$ are identified at an FDR level of 0.1. This example demonstrates that it is possible to make further discoveries in a follow-up analysis after aggregate testing, to identify which underlying variants drive the signal. The variant $E40K$ is indeed associated with TG, as validated by external studies (Dewey et al., 2016).

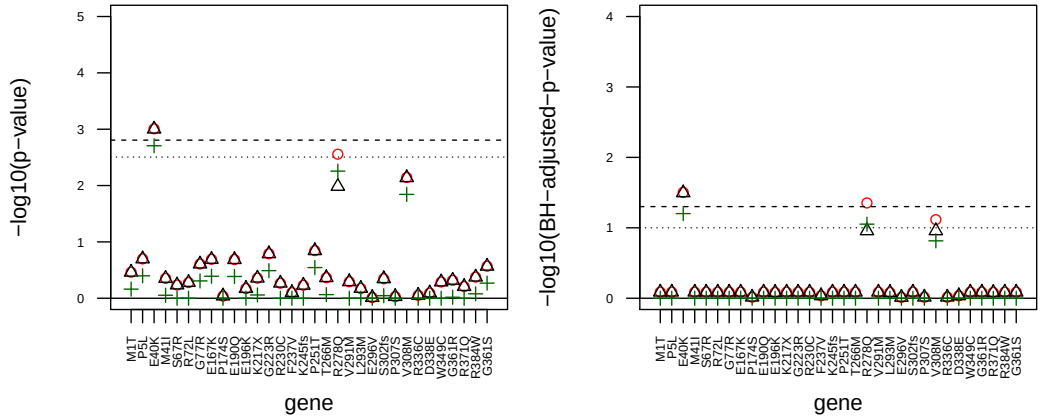

Figure 3.6: The naive and two types selection adjusted p-values on a $-\log 10$ scale (left panel) and FDR adjusted p-values on a $-\log 10$ scale (right panel) for the 32 variants. The p-values plotted are Naive unadjusted p-values (red circles), conditional p-values based on the polyhedral lemma (black triangles) and conditional p-values based on the hybrid method (green plus). The dotted line marks a multiplicity adjusted threshold of 0.1 (FWER in the left panel, FDR in the right) and the dashed line marks a multiplicity adjusted threshold of 0.05.

3.7 Discussion

In this work, we provided valid inference for linear contrasts of estimated parameters, after an aggregate test has passed a pre-defined threshold. For the post-selection inference we suggest in this paper, we only need the summary statistics for the selected group of interest, and knowledge of the selection threshold t_1. The selection threshold does not have to be fixed. For example, a data dependent threshold will be valid if the groups are independent and via the BH procedure, or any other simple selection rule (as defined in Benjamini and Bogomolov, 2014). If the data of all the groups is available, then there remains an open question of how to choose t_1 in order to maximize the chance of discovery for individual hypotheses (assuming that an error control guarantee at the group level is not necessary). Data adaptive methods for choosing t_1 may invalidate the post-selection inference. We are currently investigating potential approaches, but they are outside the scope of this manuscript.

Our methods can be extended to tree structured hypothesis tests in a straightforward manner. See Bogomolov et al. (2017) and the references within for state-of-the-art work on hierarchical testing when there are more than two layers. An interesting genomic application is the following. Within a selected gene, the tests may be further divided naturally into subgroups. For example, clusters of SNPs within a gene (Yoo et al., 2016). It may be of interest to develop a multi-level analysis, where following selection we first examine the subgroups, and only then the individual effects.

In this work we suggested switching regimes to adapt to the different unknown sparsity of the estimated effects. We observed that by combining a powerful method for the sparse setting with a powerful method for the non-sparse setting, we get a method that has overall good performance. Such an approach can be very useful in genomic applications, where the signal is expected to be sparse in some groups but non-sparse in others. The switching regime approach may benefit other post-selection settings as well, e.g., confidence intervals for the selected parameters in a regression model.

Supplementary material

An R implementation of the methods in this paper is available in `https://github.com/ammeir2/PSAT` and will be available (soon) in the Bioconductor package PSAT.

Table 3.1: For the 32 varaints in ANGPTL4, the number of subjects with rare variants (column 2), the estimated effect size (column 3), the conditional p-value (column 4), the original p-value (column 5), the default Beta-density weight in SKAT (column 6), and the contribution of the variant to the SKAT statistic $\sum_{j=1}^{32} w_{m,j}U_j^2$ (column 7). Variant $E40K$ has conditional p-value below the $0.05/32 = 0.0016$, and is therefore discovered by Bonferroni, with a guarantee of conditional FWER control at the 0.05 level. The contribution of variant R278Q is by far the largest towards the SKAT statistic, and therefore the conditional p-value is larger than the naive p-value for R278Q. For all other variants in this gene, the conditional p values coincide with the naive p-values. This is expected when by conditioning on the test statistics of all the other variants, the SKAT test singificance is guaranteed regardless of the single variant test statistic value.

Variant	# rare variants	$\hat{\beta}$	hybrid PV	conditional PV	naive PV	SKAT weight	$w_{m,j}U_j^2$
M1T	1.0000	0.8967	0.6872	0.3434	0.3434	4.9831	19.9657
P5L	2.0000	0.8588	0.3999	0.1995	0.1995	4.9663	74.7223
E40K	50.0000	-0.4490	0.0020	0.0010	0.0010	4.2172	7687.4667
M41I	28.0000	0.1403	0.8860	0.4333	0.4333	4.5466	288.0531
S67R	2.0000	0.3681	1.000	0.5827	0.5827	4.9663	15.9910
R72L	3.0000	-0.3468	1.000	0.5262	0.5262	4.9495	22.0274
G77R	1.0000	-1.0913	0.4928	0.2489	0.2489	4.9831	29.5739
E167K	1.0000	-1.2002	0.4072	0.2049	0.2049	4.9831	34.4125
P174S	1.0000	0.1040	1.000	0.9125	0.9125	4.9831	0.4010
E190Q	32.0000	0.2140	0.4108	0.2054	0.2054	4.4850	1199.4289
E196K	1.0000	-0.3991	1.000	0.6733	0.6733	4.9831	3.5122
K217X	1.0000	-0.7300	0.8704	0.4406	0.4406	4.9831	12.4116
G223R	1.0000	-1.3242	0.3239	0.1621	0.1621	4.9831	40.5690
R230C	1.0000	-0.5784	1.000	0.5412	0.5412	4.9831	7.6586
F237V	1.0000	-0.2453	1.000	0.7956	0.7956	4.9831	1.2266
K245fs	1.0000	0.5157	1.000	0.5858	0.5858	4.9831	6.6047
P251T	1.0000	1.3860	0.2854	0.1432	0.1432	4.9831	49.3013
T266M	1887.0000	0.0230	0.8619	0.2454	0.2454	0.0006	0.0002
R278Q	207.0000	-0.1945	0.0055	0.0103	0.0023	2.4309	10667.1021
V291M	1.0000	-0.6203	1.000	0.5123	0.5123	4.9831	9.5533
L293M	1.0000	0.4021	1.000	0.6717	0.6717	4.9831	1.3213
E296V	1.0000	-0.0308	1.000	0.9741	0.9741	4.9831	0.0235
S302fs	1.0000	-0.7089	0.9012	0.4539	0.4539	4.9831	12.4794
P307S	1.0000	-0.0793	1.000	0.9333	0.9333	4.9831	0.0274
V308M	3.0000	1.4719	0.0143	0.0071	0.0071	4.9495	477.6972
R336C	7.0000	-0.0469	1.000	0.8957	0.8957	4.8829	2.5696
D338E	1.0000	0.2314	1.000	0.8073	0.8073	4.9831	0.0339
W349C	1.0000	-0.6120	1.000	0.5179	0.5179	4.9831	9.3008
G361R	2.0000	0.4744	0.9598	0.4784	0.4784	4.9663	23.2973
R371Q	1.0000	0.4724	1.000	0.6177	0.6177	4.9831	5.5410
R384W	1.0000	-0.7610	0.8420	0.4225	0.4225	4.9831	22.6686
G361S	1.0000	-1.0389	0.5388	0.2724	0.2724	4.9831	26.7995

3.A Proof of theorems

3.A.1 Proof of Theorem 3.1

In order to compute the distribution of a linear combination of $\hat{\beta}$ within selected regions, it is useful to first represent the selection event in simple form (the representation is based on the one used for post-model selection in Lee et al., 2016).

Lemma 3.2. For an arbitrary linear combination $\eta^T\hat{\beta}$, let $\tau = (\eta^T\mathbf{\Sigma}\eta)^{-1}\mathbf{\Sigma}\eta$ and $W = (\boldsymbol{I}_m - \tau\eta^T)\hat{\beta}$, where $\boldsymbol{I}_m$ is the $m \times m$ identity matrix. The selection event $S \geq S_{1-t_1}$ can be rewritten in terms of $\eta^T\hat{\beta}$ as follows:

$$\mathcal{A}(W) = \left\{\eta^T\hat{\beta} \geq A(W), \eta^T\hat{\beta} \leq B(W)\right\},$$

where

$$A(W) = \begin{cases} \frac{-2W^T\mathbf{K}\tau + \sqrt{\Delta}}{2\tau^T\mathbf{K}\tau} & \text{if } \Delta \geq 0, \\ -\infty & \text{if } \Delta < 0, \end{cases} \quad B(W) = \begin{cases} \frac{-2W^T\mathbf{K}\tau - \sqrt{\Delta}}{2\tau^T\mathbf{K}\tau} & \text{if } \Delta \geq 0, \\ \infty & \text{if } \Delta < 0, \end{cases}$$

for $\Delta = 4(W^T\mathbf{K}\tau)^2 - 4(\tau^T\mathbf{K}\tau)(W^T\mathbf{K}W - S_{1-t_1})$.

Proof. Decomposing $\hat{\beta} = W + \tau\eta^T\hat{\beta}$, the result is immediate from rewriting $S = (W + \tau\eta^T\hat{\beta})^T\mathbf{K}(W + \tau\eta^T\hat{\beta})$ as a quadratic polynomial with argument $\eta^T\hat{\beta}$. The selection event is therefore

$$\{S \geq S_{1-t_1}\} = \left\{[\eta^T\hat{\beta}]^2\tau^T\mathbf{K}\tau + [\eta^T\hat{\beta}]2W^T\mathbf{K}\tau + W^T\mathbf{K}W - S_{1-t_1} > 0\right\}.$$

□

Since the covariance between W and $\eta^T\hat{\beta}$ is zero, it follows from Lemma 3.2 that if $\eta^T\hat{\beta}$ is (approximately) normal, then the boundaries of the selection event are independent of $\eta^T\hat{\beta}$. Therefore, conditional on W and on the selection event $\{S \geq S_{1-t_1}\}$, $\eta^T\hat{\beta}$ has a truncated normal distribution, truncated at the values $(-\infty, B(W)] \cup [A(W), \infty)$.

3.A.2 Proof of Theorem 3.2

Proof. We shall show this without loss of generality for $j = 1$, i.e., for testing $H_1 : \beta_1 = 0$. Since $\mathbf{K} = \mathbf{\Sigma}^{(-1)}$, it follows that $W^T\mathbf{K}\tau = \mathbf{0}$. Therefore, the selection event is

$$\{S \geq S_{1-t_1}\} = \left\{[\eta^T\hat{\beta}]^2\tau^T\mathbf{K}\tau + W^T\mathbf{K}W - S_{1-t_1} > 0\right\}.$$

Clearly, the truncation will be smaller (i.e., $A(W)$ smaller and $B(W)$ larger), the larger $W^T\mathbf{K}W$ is. It is clear from the dependence of the distribution of $W^T\mathbf{K}W$ on $(0, \beta_2, \ldots, \beta_m)$ that it will be stochastically smallest for $(0, \beta_2, \ldots, \beta_m) = 0$. Therefore,

$$
\begin{aligned}
&P_{(0,\beta_2,\ldots,\beta_m)}(P_1 \leq x \mid S > S_{1-t_1}) = E\left[P_{(0,\beta_2,\ldots,\beta_m)}(P_1 \leq x \mid S > S_{1-t_1}, W) \mid S > S_{1-t_1}\right] \\
&= E_{(0,\beta_2,\ldots,\beta_m)}\left[1 - F_{0,SE_1^2}^{\{\hat{\beta}_1 \geq A(W), \hat{\beta}_1 \leq B(W)\}}(x) \mid S > S_{1-t_1}\right] \\
&\leq E_0\left[1 - F_{0,SE_1^2}^{\{\hat{\beta}_1 \geq A(W), \hat{\beta}_1 \leq B(W)\}}(x) \mid S > S_{1-t_1}\right] = P_0(P_1 \leq x \mid S > S_{1-t_1}) = x,
\end{aligned}
$$

where the inequality follows since $1 - F_{0,SE_1^2}^{\{\hat{\beta}_1 \geq A(W), \hat{\beta}_1 \leq B(W)\}}(x)$ is an increasing function of $A(W)$ and a decreasing function of $B(W)$, i.e., a decreasing function of $W^T\mathbf{K}W$, so the expecation would be largest when $W^T\mathbf{K}W$ is stochastically smallest, i.e., at $\beta = 0$. □

3.A.3 Proof of Lemma 3.1

Proof. For arbitrary fixed $b, s > 0$, define the following sets for some fixed index $j \in \{1, \ldots, m\}$

$$A := \{\hat{\beta} : \ S < s\}, \qquad B := \{\hat{\beta} : \hat{\beta}_j^2 < b\}$$

The sets A and B are both convex and symmetric about the origin if $\beta = 0$. By the Gaussian correlation inequality (Latala and Matlak, 2017; Royen, 2014) we have:

$$P_{\beta=0}(A, B) \geq P_{\beta=0}(A)P_{\beta=0}(B) \tag{3.12}$$

The left-hand side of equation (3.12) can be written as

$$P_{\beta=0}(A, B) = 1 - P_{\beta=0}(\hat{\beta}_j^2 > b) - P_{\beta=0}(S > s) + P_{\beta=0}(\hat{\beta}_j^2 > b, S > s),$$

and similarly the right-hand side can be written as

$$P_{\beta=0}(A)Pr_{\beta=0}(B) = 1 - P_{\beta=0}(\hat{\beta}_j^2 > b) - P_{\beta=0}(S > s) + P_{\beta=0}(\hat{\beta}_j^2 > b)P_{\beta=0}(S > s).$$

Subtracting $1 - P_{\beta=0}(\hat{\beta}_j^2 > b) - P_{\beta=0}(S > s)$ from both sides of (3.12) yields:

$$P_{\beta=0}(\hat{\beta}_j^2 > b, S > s) \geq P_{\beta=0}(\hat{\beta}_j^2 > b) P_{\beta=0}(S > s).$$

Finally,

$$P_{\beta=0}(\hat{\beta}_j^2 > b | S > s) = \frac{P_{\beta=0}(\hat{\beta}_j^2 > b, S > s)}{P_{\beta=0}(S > s)} \geq \frac{P_{\beta=0}(\hat{\beta}_j^2 > b) Pr_{\beta=0}(S > s)}{P_{\beta=0}(S > s)} = P_{\beta=0}(\hat{\beta}_j^2 > b)$$

□

3.A.4 Proof of Theorem 3.4

Proof. Let $\hat{\beta} \sim N(\beta, \mathbf{\Sigma})$. Now, suppose that we are interested in solving the optimization problem:

$$\max_{\beta} \ell(\beta) - \log P_\beta(S > S_{1-t_1}).$$

The above optimization problem can be rewritten as:

$$\max_{\gamma} \max_{\beta \in B(\gamma)} \ell(\beta) - \log \gamma, \tag{3.13}$$

where

$$B(\gamma) := \{\beta : P_\beta(S > S_{1-t_1}) = \gamma\}.$$

In (3.13) we divided our optimization problem into two parts. First, we must compute the maximizer of the likelihood for each power level γ and then, we must maximize over γ to find the global maximizer of the likelihood. The theorem hinges on the fact that this inner optimization problem has a closed form solution which we derive next.

If $\mathbf{K} = \mathbf{\Sigma^{-1}}$ then the distribution of the test statistic is a non-central chi-square distribution, the parameters of which are the degrees of freedom (a known quantity) and the non-centrality parameter:

$$\beta^T \mathbf{\Sigma^{(-1)}} \beta.$$

Thus, for each value of $\gamma \geq t_1$, there exists a $\delta \geq 0$ such that:

$$\begin{aligned} &\max_{\beta \in B(\gamma)} \ell(\beta) \\ =&\max_{\beta} \ell(\beta), \qquad \text{s.t. } \beta^T \mathbf{\Sigma^{(-1)}} \beta = \delta. \end{aligned} \tag{3.14}$$

Now, for any $\delta \leq \hat{\beta}^T \mathbf{\Sigma^{-1}} \hat{\beta}$ there exists a $c \geq 0$ such that the solution to (3.14) is given by:

$$\max_{\beta} \ell(\beta) - c\beta^T \mathbf{\Sigma^{-1}} \beta.$$

This last problem, is a simple Tikhonov regularization problem, the solution which is given by $(1+c)^{-1}\hat{\beta}$ with:

$$c = \sqrt{\frac{\hat{\beta}^T \mathbf{\Sigma^{-1}} \hat{\beta}}{\delta}} - 1.$$

Thus, for $\delta = 0$, $c = \infty$ and for $\delta = \hat{\beta}^T \mathbf{\Sigma^{(-1)}} \hat{\beta}$, $c = 0$ and we recover the least squares solution. From this, we can infer that $(1+c)^{-1} \in [0, 1]$

Because $(1+c)^{-1} \in [0,1]$ and all of the solution to the inner problem in (3.13) are of the form $(1+c)^{-1}\hat{\beta}$, we can conclude that the maximum likelihood estimator is given by:

$$\tilde{\beta} = \arg \max_{\lambda \in [0,1]} \ell(\lambda \hat{\beta}) - \log P_{\lambda \hat{\beta}}(S > S_{1-t_1}).$$

□

3.A.5 Proof of Theorem 3.6

Denote by NC_j the event in which a non-covering confidence interval was constructed for β_j, by NNC_j the event in which a naive confidence interval does not cover β_j and by CNC_j the event in which a conditional confidence did not cover β_j.

In our procedure, if $S \geq S_{1-t_2}$, the confidence interval is based on the unconditional likelihood, and it is at level $1-\alpha$; if $S_{1-t_1} \leq S < S_{1-t_2}$, the confidence interval is based on

the exact conditional likelihood at $\beta = 0$, and it is at level $1 - (\alpha - t_2/t_1)$; otherwise, no confidence interval is constructed. Therefore,

$$\begin{aligned}
&P_\beta(NC_j \mid S > S_{1-t_1}) = \\
&= P_\beta(NNC_j, S > S_{1-t_2} \mid S > S_{1-t_1}) + P_\beta(CNC_j, S < S_{1-t_2} \mid S > S_{1-t_1}) \\
&\leq P_\beta(S > S_{1-t_2} \mid S > S_{1-t_1}) + P_\beta(CNC_j \mid S > S_{1-t_1}).
\end{aligned}$$

If $\beta = 0$, then $P_0(S \geq S_{1-t_2} \mid S > S_{1-t_1}) = t_2/t_1$ and $P_0(CNC_j \mid S < S_{1-t_2}) = \alpha - t_2/t_1$. Therefore, $P_0(NC_j \mid S > S_{1-t_1}) \leq \alpha$.

If $\beta \neq 0$, then $\lim_{n\to\infty} Pr_\beta(S > S_{1-t_2}) = 1$. This follows for the linear model, since $E(\hat{\beta} - \beta) = 0$ and $var\left(\hat{\beta}\right) = (\mathbf{X}^T\mathbf{X})^{(-1)}var(\epsilon_1)$. This also follows for the logistic model, since large n, $E(\hat{\beta} - \beta) = O(\frac{1}{n})$ and $var\left(\hat{\beta}\right) = (\mathbf{X}^T\mathbf{W}\mathbf{X})^{(-1)}(1 + O(\frac{1}{n}))$. Therefore,

$$\lim_{n\to\infty} P_\beta(NC_j \mid S > S_{1-t_1}) = \lim_{n\to\infty} P_\beta(NNC_j \mid S \geq S_{1-t_2}) = \alpha.$$

□

3.B Most conservative tests following linear aggregate testing

In this section we consider the problem of constructing a conservative test following model selection with a linear aggregate test. Suppose that we observe $y \sim N(\mu, \mathbf{\Sigma})$ and that we estimate μ if and only if we can reject a global null test that:

$$H_0 : a^T\mu \neq 0$$

for some fixed vector a. As thresholds for selection we fix constants $-\infty < l \leq u < \infty$ and define a model selection event:

$$\mathcal{S} = (a^TY < l \cup a^TY > u).$$

The analysis of this problem yields an interesting and somewhat surprising result. Specifically, we find that the most conservative parametrization of β for testing $H_{0\eta} : \eta^T\mu$ is not necessarily the global-null, but any parameterization such that $\eta^T\mu = 0$ and

$$E(a^T W) = \frac{l+u}{2}. \tag{3.15}$$

This parametrization can be used to compute hybrid p-values after testing with an asymmetric two-sided aggregate test. However, we note that if the test is highly imbalanced (e.g. $l = -10, u = 1$) then the most conservative parametrization will yield a very conservative test and it might be preferable to use the asymptotically valid global-null p-value.

Theorem 3.7. Let $y \sim N(\mu, \mathbf{\Sigma})$ and suppose that $a \neq \eta$, $\eta^T\mu = 0$ and that $-\infty < l \leq u < \infty$ and let:

$$p_\mu(b) = 2\min\left(P_\mu(\eta^T Y \geq b|\mathcal{S}), P_\mu(\eta^T Y \leq b|\mathcal{S})\right).$$

Then, for $\tilde{\mu}$ that also satisfies (3.15) we have

$$P_\mu\left(p_{\tilde{\mu}}(\eta^T Y) < t\middle|\mathcal{S}\right) \leq P_{\tilde{\mu}}\left(p_{\tilde{\mu}}(\eta^T Y) < t\middle|\mathcal{S}\right) = t, \qquad \forall t \in (0,1). \tag{3.16}$$

If $a = \eta$ then any parametrization of μ that satisfies $\eta^T\mu = 0$ yields a valid p-value.

Before we get into the technicalities of the proof of Theorem 3.7, let us break down the component of equation (3.16). We have two quantities that depend on the parameter values under which we evaluate the p-value. The first is the p-value itself $p_\beta(b)$ which effectively determines the threshold for declaring that a test is rejected at a level t and it is evaluated under the same parameter value on both sides of the inequality. The second component is the parameter value under which we evaluate the probability of crossing a threshold $P_\mu(p_\mu(\eta^T Y) < t)$ which determines under what set of parameters we evaluate the probability of crossing the thresholds determined by p_μ. If we can show that equation (3.16) holds then this implies that evaluating the probability of crossing the threshold at $\tilde{\mu}$ is the

most conservative, making our procedure conservative.

Assume w.l.o.g that $a^T\tau = 1$. We begin by noting that the joint distribution of $Z := a^TW$ and η^TY is that of independent normal vector, and so, examining the marginal density of Z under the null and the assumption that $E(Z) = (l+u)/2$, we can see that Z has a symmetric distribution about $(l+u)/2$:

$$\begin{aligned} f(z|\mathcal{S}) &= \frac{P(\mathcal{S}|z)}{P(\mathcal{S})}\varphi(z) \\ &= \frac{P(\{\eta^TY < l-z\} \cup \{\eta^TY > u-z\})}{P(\mathcal{S})}\varphi(z) \\ &= \frac{P(\{\eta^TY < \frac{l-u}{2} - (z - \frac{l+u}{2})\} \cup \{\eta^TY > \frac{u-l}{2} - (z - \frac{l+u}{2})\})}{P(\mathcal{S})}\varphi(w) \end{aligned}$$

and so, the truncation has a symmetric distribution about 0 in the sense that

$$A(W) = \frac{l-u}{2} - \left(z - \frac{l+u}{2}\right) =^D - \left(\frac{u-l}{2} - \left(z - \frac{l+u}{2}\right)\right) = -B(W).$$

Thus, there exists a constant $b(t)$ such that:

$$P_{\tilde{\mu}}(p_{\tilde{\mu}} < t|\mathcal{S}) = P_{\tilde{\mu}}(\eta^TY > b(t)|\mathcal{S}) + P_{\tilde{\mu}}(\eta^TY < -b(t)|\mathcal{S}).$$

Fixing a value for Z, we have

$$P_{H_0}(p_{\tilde{\mu}}(\eta^TY) < t|\mathcal{S}, a^TW = z) = \frac{Pr(\eta^TY < -b(t), \mathcal{S}(z)) + Pr(\eta^TY > b(t), \mathcal{S}(z))}{P(\mathcal{S}(z))} \quad (3.17)$$

$$\mathcal{S}(z) := \{\eta^TY < l-z\} \cup \{\eta^TY > u-z\}$$

and

$$P_{\mu}(p_{\tilde{\mu}}(\eta^TY) < t|\mathcal{S}) = \int_{\mathbb{R}} P_{H_0}(p_{\tilde{\mu}}(\eta^TY) < t|\mathcal{S}, a^TW = z) f_{\mu}(z|\mathcal{S})dz.$$

Notice that (3.17) is symmetric about $z = (l+u)/2$.

Taking a derivative,

$$\frac{\partial}{\partial E(Z)} P_\beta(p_{\tilde{\mu}}(\eta^T Y) < t|\mathcal{S}) = \int_{\mathbb{R}} P_{H_0}(p_{\tilde{\mu}}(\eta^T Y) < t|\mathcal{S}, a^T W = z) \frac{\partial}{\partial E(Z)} f(z|\mathcal{S}) dz. \tag{3.18}$$

The inner derivative equals:

$$\frac{\partial}{\partial E(Z)} f(z|\mathcal{S}) = P(\mathcal{S}|z)\varphi(z)\frac{z - E(Z|\mathcal{S})}{\sigma_z^2 P(\mathcal{S})}.$$

The derivative in (3.18) equals zero at $E(Z) = (l+u)/2$ because for such a parameter value $E(Z|\mathcal{S}) = (l+u)/2$, $f'(z - (l+u)/2|S) = -f'((l+u)/2 - z|S)$ and

$$P_{H_0}(p_{\tilde{\mu}}(\eta^T Y) < t|\mathcal{S}, a^T W = z - (l+u)/2) = P_{H_0}(p_{\tilde{\mu}}(\eta^T Y) < t|\mathcal{S}, a^T W = (l+u)/2 - z),$$

this is also the only maximum because the inner derivative is symmetric only at $E(Z) = (l+u)/2$.

Finally, if $a = \eta$ then $a^T W = 0$ by definition and therefore the distribution of $\eta^T \hat{\mu}$ is always a truncated normal constrained to $(-\infty, l] \cup [u, \infty)$. □

Chapter 4

THE CONDITIONAL BOOTSTRAP FOR POST-SELECTION INFERENCE

4.1 Introduction

In the previous chapter we proposed two heuristics for conducting more powerful post-selection inference. The first is based on inference under a conservative parametrization and the second is based on a regime switching approach which is guaranteed to consistently estimate the post-selection distribution of the data in a point-wise manner. In this chapter we will generalize the second idea to a more general post-selection setting.

We propose a generalized regime switching procedure for approximating the post-selection distribution that takes inspiration from the work of Chatterjee and Lahiri (2011) who bootstrap the unconditional distribution of the Lasso regression coefficient estimates. The residual bootstrap is intractable post-selection and so, in its place we use an MCMC procedure to sample from the estimated conditional distribution of the data. This approximate distribution can be used to construct confidence intervals that are asymptotically (point-wise) consistent.

The rest of the chapter proceeds as follows. In Section 4.2 we describe the challenges inherent in trying to bootstrap conditional post-selection distributions in the context of inferring on a univariate selected normal mean. There, we also propose a simple modified bootstrap procedure that will serve as a basis for our multivariate method. In Section 4.3 we describe the types of post-selection problems we consider, describe the post-selection distribution of the data under our model, and propose a conditional bootstrap procedure for approximating it. In Section 4.4 we show that our proposed method produces point-wise consistent confidence intervals, and theory for conducting post-selection inference with data

carving. In Section 4.5 we apply our method to the problem of inferring on regression models that were selected via marginal screening, and in Section 4.6 we conclude with a discussion.

4.2 Bootstrapping the univariate post-selection estimator

Suppose that we are interesting in conditionally inferring on the mean of a single observed normal observation $y \sim N(\mu, \sigma^2)$, σ^2 known, if and only if $|y| > c > 0$ for some fixed constant c. Knowing that selection changes the distribution of the data, we may be tempted to utilize the bootstrap in order to overcome the model misspecification problem. Consider the following naive procedure:

Procedure 4.1. A Naive Post-Selection Bootstrap

1. Estimate $\hat{\mu} = y$.
2. Generate a parametric bootstrap sample $y^1, \ldots, y^B \sim TN(\hat{\mu}, \sigma^2, |y| > c)$.
3. For $b \in \{1, \ldots, B\}$ compute the residuals $r^b = y^b - \hat{\mu}$
4. Compute a level $1 - \alpha$ confidence interval:

$$CI_B(|Y| > c) = (y - r_{1-\alpha/2}, y - r_{\alpha/2})$$

where r_q is the qth quantile of the bootstrapped residuals.

We plot the results of the applying Procedure 4.1 to simulated data in the lefthand side panel of Figure 4.1. To generate the figure, we sampled $10,000$ observations from the truncated normal distribution $y_1, \ldots, y_{10,000} \sim N(0, 1, |Y| > 1.96)$ and constructed $10,000$ naive bootstrap 95% confidence intervals. The confidence intervals that cover zero (the true parameter value) are colored in blue. About 60% of the confidence intervals cover zero. This is an improvement over the gaussian confidence intervals that ignore selection altogether and would have a coverage rate of 0% by definition, but are still unsatisfactory given that the target coverage rate was 95%. To give some intuition as to why this procedure fails, we plot the conditional bootstrap distributions of the residuals for $\mu \in \{0, 2, 2.5, 3\}$ in the righthand

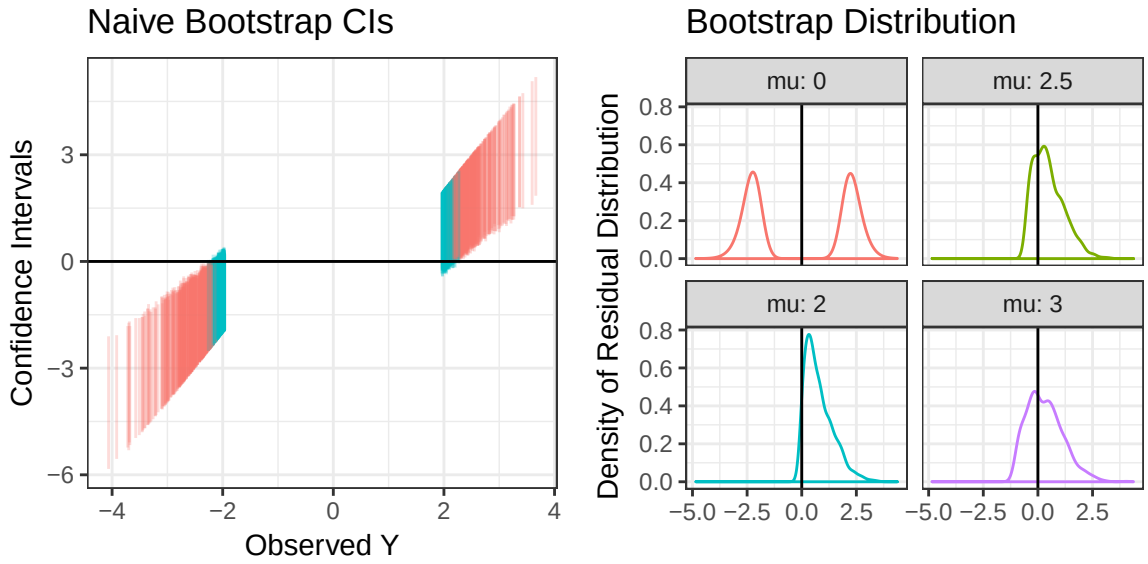

Figure 4.1: Naive Bootstrap distribution for a univariate truncated normal. In the lefthand panel we plot the densities of four truncated normal distributions with different mean parameters. The one for $\mu = 0$ is the true data generating distribution, while the other three are a result of bootstrapping based on inaccurate point estimates.

side panel of Figure 4.1. The bootstrap distribution is different for every parameter value, with the bimodal distribution at $\mu = 0$ being the distribution we are trying to estimate. However, since we never observe $y = 0$, the naive bootstrap procedure will never give a good approximation to the true residual distribution.

The problem of location non-invariance is related to problem of non-convergence of bootstrap distributions: If the post-selection bootstrap distribution converged to the correct limiting distribution of Y conditionally on selection then our selective inference problems would disappear for large enough sample sizes. This however is not always the case. For example, if we observe $y_1, \ldots, y_n, \cdots \sim N(0, \sigma^2)$ and reject the null if $\sqrt{n}|\bar{y}| > c > 0$ then the conditional distribution of $\sqrt{n}\bar{y}$ is truncated for any n. A similar problem was encountered by Chatterjee and Lahiri (2011) who devised a method for bootstrapping the distribution of the lasso estimators. In their work, they found that the bootstrap distribution of the lasso

estimators converges to a random measure, but that the correct distribution can be estimated by thresholding small coefficients to zero. A similar approach also works for post-selection inference. Consider the following modified conditional bootstrap.

Procedure 4.2. A Modified Univariate Conditional Bootstrap

1. Given a significance level α and a threshold c, set $s_1 = 1 - \Phi(\sigma^2 c)$ and choose $0 < s_2 \leq \alpha/2$.
2. Set $\tilde{\mu} = yI\{|y| > z_{1-s_1 s_2}\}$.
3. Generate a parametric bootstrap sample $y^1, \ldots, y^B \sim TN(\tilde{\mu}, \sigma^2, |Y| > c)$.
4. For $b \in \{1, \ldots, B\}$ compute the residuals $r^b = y^b - \tilde{\mu}$.
5. Compute a level $1 - \alpha$ confidence interval:

$$CI_B(|Y| > c) = (y - r_{1-\alpha/2}, y - r_{\alpha/2})$$

   where r_q is the qth quantile of the bootstrapped residuals.

Procedure 4.2 is similar to the procedure proposed by Chatterjee and Lahiri (2011) in the sense that it obtains consistent bootstrap inference by thresholding small observed values to zero. In the case of our procedure, the thresholding is done based on a level s_2 post-selection test of the null hypothesis. The procedure is consistent in the sense that if $\mu \neq 0$ and we collect a large number of observations then we will eventually reject the null-hypothesis with probability 1, and if $\mu = 0$ then the confidence intervals have the correct coverage rate for any sample size. We also note that Procedure 4.2 is a special case of the regime switching procedure proposed in Chapter 3, where the 'aggregate test' is composed of a single test statistic.

To demonstrate Procedure 4.2 we again take $10,000$ samples from $y \sim N(0, 1, |Y| > 1.96)$ and construct post-selection confidence intervals. We set $s_2 = 0.0005$ in order to guarantee that $\sup_\mu \mathbb{P}_\mu(\mu \notin CI_B(|Y| > c) \mid |Y| > c) \leq \alpha = 0.05$. The results of the simulation are presented in Figure 4.2. Given that we sampled from the null and that s_2 is very small, it is

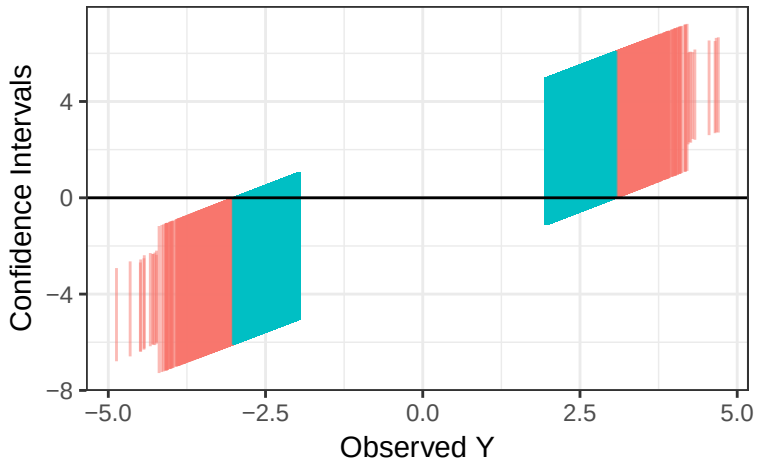

Figure 4.2: Modified bootstrap confidence intervals for selected, univariate normal observations. The majority of the observed values did not pass the post-selection test and the confidence intervals were therefore computed based on bootstrap residuals sampled at $\tilde{\mu} = 0$. For the largest observed values ($|y| > 4.2$) we rejected the null hypothesis and computed confidence intervals based on bootstrap samples taken at $\tilde{\mu} = y$. The confidence intervals obtain a 95% coverage rate of the true parameter value $\mu = 0$.

perhaps not surprising that most of the post-selection tests were not rejected and that most of the confidence intervals were computed under the null. There is a discontinuity in the length of the confidence intervals, as those computed for very large observed values ($|y| > z_{1-s_1 s_2}$) are much smaller than those computed under the null. The confidence intervals have the correct coverage for $\mu = 0$ as would be the case for any modified bootstrap confidence interval computed with $s_2 \leq \alpha/2$. Also, notice that setting $\tilde{\mu}$ to zero does not imply that a confidence interval will cover zero.

The modified bootstrap procedure, while asymptotically consistent, does not offer much over the conditional confidence intervals based on test inversion that we described in Chapter 1 for solving the selected mean problem, as they are only asymptotically consistent and tend to be much wider. However, unlike the test-inversion approach, the modified bootstrap approach can be extended to multivariate problems in a straightforward manner.

4.3 Bootstrapping conditional multi-parameter distributions

In this section we will describe a modified bootstrap procedure for conducting post-selection inference in problems involving multi-parameter distributions. In Section 4.3.1 we describe our model and type of model selection procedures we consider; In Section 4.3.2 we describe the limiting conditional distribution of the data under our assumed model; In Section 4.3.3 we describe the conditional bootstrap procedure, and demonstrate its application to the problem of inference following model selection with marginal screening.

4.3.1 A Simple (and non-comprehensive) setup

Let $y_1, \ldots, y_n, \cdots \in \mathbb{R}^k$ be i.i.d. observations that are distributed according to some distribution with density $f_\theta(y_i)$. Let $T_n := \frac{1}{n}\sum_{i=1}^n T(y_i) \in R^p$ be a statistic that satisfies:

$$\sqrt{n}(T_n - \mu) \to^D N(0, \mathbf{\Sigma}). \tag{4.1}$$

Upon observing $y_1, \ldots, y_n$, we select a model based on a model selection function $S_n : \mathbb{R}^p \to \mathcal{M}$ that defines a set of null hypotheses to be tested $H_{01} : g_1^M(\mu) = 0, \ldots, H_{0J} : g_m^M(\mu) = 0$ based on the magnitude of the observe test statistics T,

$$M = \{j : |T_{nj}| > c_j n^{-\delta} \geq 0\}, \qquad \delta \in (0, 0.5],$$

and the number of hypotheses to be tested J may depend on M. Finally, we assume that $g_i^M(\mu)$ is a differentiable function of μ, and use $g_i^M(t_n)$ as its estimate.

This setup, while not sufficiently comprehensive to capture all post-selection problems of interest, is general enough to describe a large number of interesting problems, examples for some of which we give next. Our setup can be extended in a straight forward manner to encompass other problems. For example, to describe the Lasso selection event we will allow T to depend on the selected model.

Example 4.1. Marginal Screening. Suppose that we observe i.i.d. observations

$$(y_1, X_1), \ldots, (y_n, X_n) \in \mathbb{R} \times \mathbb{R}^p$$

and are interested in fitting a linear regression model to the data:

$$y = \boldsymbol{X}\,\beta + \varepsilon, \qquad \varepsilon \sim N(0, \sigma^2\,\boldsymbol{I}), \qquad \sigma^2 \text{ known.}$$

If p is large, we may want to select only a subset of the columns of $\boldsymbol{X}$ to include in the model. One possible way to select variables is via marginal screening. The test statistic is $T_n(\boldsymbol{X}, y) = n^{-1}X^T y$ and the model selection function $S_n(\boldsymbol{X}, y)$ selects an element of $\mathcal{M} := \{0,1\}^p$. If $M \neq \emptyset$, then we test the hypotheses:

$$H_{0j} : \beta_j^M \neq 0,\ \forall j \in M.$$

In this case the test statistic is also sufficient.

Example 4.2. Marginal Screening with Data Carving. Suppose that we are interested in prioritizing inference over model selection. In the case of regression with marginal screening we can do so by using only a part of our dataset for model selection. One such choice is to set:

$$T_n(\boldsymbol{X}, y)_j = \frac{2}{n}\sum_{i=1}^{n/2} \boldsymbol{X}_{i,j}\, y_i,$$

and proceed as before. In this case the test statistic $T_n(\boldsymbol{X}, y)$ is not sufficient.

4.3.2 The limiting conditional distribution

In the example above, the limiting distribution of the data conditional on selecting a specific model M depends on which selected variables are null and which are not. We give a formal

description of the limiting conditional distribution next. As a preparation, we define the sets

$$M_0 = \{j : j \in M, \mu_j = 0\}, \qquad M_1 = \{j : j \in M, \mu_j \neq 0\},$$

$$M_1^c = \{j : j \notin M, \mu_j \neq 0\}.$$

and the events,

$$A_0 = \bigcap_{j \notin M} \{|T_{nj}| < c_j n^{-\delta}\}, \qquad A_1^0 = \bigcap_{j \in M_0} \{|T_{nj}| \geq c_j n^{-\delta}\}, \qquad A_1^1 = \bigcap_{j \in M_1} \{|T_{nj}| \geq c_j n^{-\delta}\}.$$

These events satisfy

$$\{S(Y) = M\} = \{A_0, A_1^0, A_1^1\}.$$

Theorem 4.1 describes the limiting conditional distribution of a selected model M.

Theorem 4.1. Suppose that the assumptions made in Section 4.3.1 hold with $\delta = 0.5$, and that $M_1^c = \emptyset$. Then,

$$\lim_{n \to \infty} \sup_{\xi \in \mathbb{R}^k} \left| \mathbb{P}_n(\sqrt{n}(T_n - \mu) < \xi \mid S_n(Y) = M) - \frac{\Phi(\xi; 0, \mathbf{\Sigma})}{\mathbb{P}_n(A_0, A_1^0)} \right| = 0,$$

where $\mathbb{P}_n(A)$ denotes the true probability of an event as a function of the sample size. If $\delta \in (0, 0.5)$ then $\lim_{n \to \infty} \mathbb{P}_n(M^*) = 1$, where $M^* = \{j : \mu_j \neq 0\}$.

The proof can be found in the appendix. Notice that in the theorem we do not treat models where some of the non-null test statistics do not pass the selection threshold. The reason for this is that the probability of such an occurrence converges to zero in the sample size, making such settings somewhat irrelevant from an asymptotic point of view. According to Theorem 4.1, if we can distinguish between the null and non-null test statistics, then we will be able to sample from the limiting conditional distribution of the data. We propose a procedure for doing so in the next section.

4.3.3 A conditional bootstrap procedure

We begin this section by formally presenting our conditional bootstrap procedure for multivariate post-selection inference. We will then apply the procedure to the problem of inference following model selection with marginal screening, and unpack the details regarding its implementation.

Procedure 4.3. A Multivariate Conditional Bootstrap

1. Set a thresholding parameter $s_n \in (0, 1)$.
2. Compute (approximate) post-selection p-values $p_1, \ldots, p_p$,

$$p_j = \mathbb{P}_0(|T_{nj}| > |t_{nj}| \mid j \in M),$$

   where t_{nj} are the observed values of T_{nj}.
3. Set $\tilde{\mu}_j = t_{nj} I\{p_j \leq s_n \cap j \in M\}$.
4. For $b \in \{1, \ldots, B\}$ sample,

$$t^b \sim TN(\tilde{\mu}, n^{-1}\,\boldsymbol{\Sigma}, S_n(Y) = M),$$

   and compute the residuals $r^b = g'(\tilde{\mu})^T(t^b - \tilde{\mu})$, where g' is the first derivative of g with respect to μ.
5. Compute a level $1 - \alpha$ confidence interval:

$$CI_B(S(Y) = M) = (g(t_n) - r_{1-\alpha/2}, g(t_n) - r_{\alpha/2})$$

   where r_q is the qth quantile of the bootstrapped residuals.

We demonstrate the conditional bootstrap on a simulated regression dataset. For $i = 1, \ldots, 200$ we sample $X_i \sim N_{50}(0, \mathbf{U})$ with $\mathbf{U}_{k,l} = 0.4^{|k-l|}$, $y_i \sim N(X_i\beta, \sigma^2)$ with all coefficients zero except for $\beta_9 = -0.11$ and $\beta_{22} = -0.72$. We set the variance to $\sigma^2 = 1.75$. To

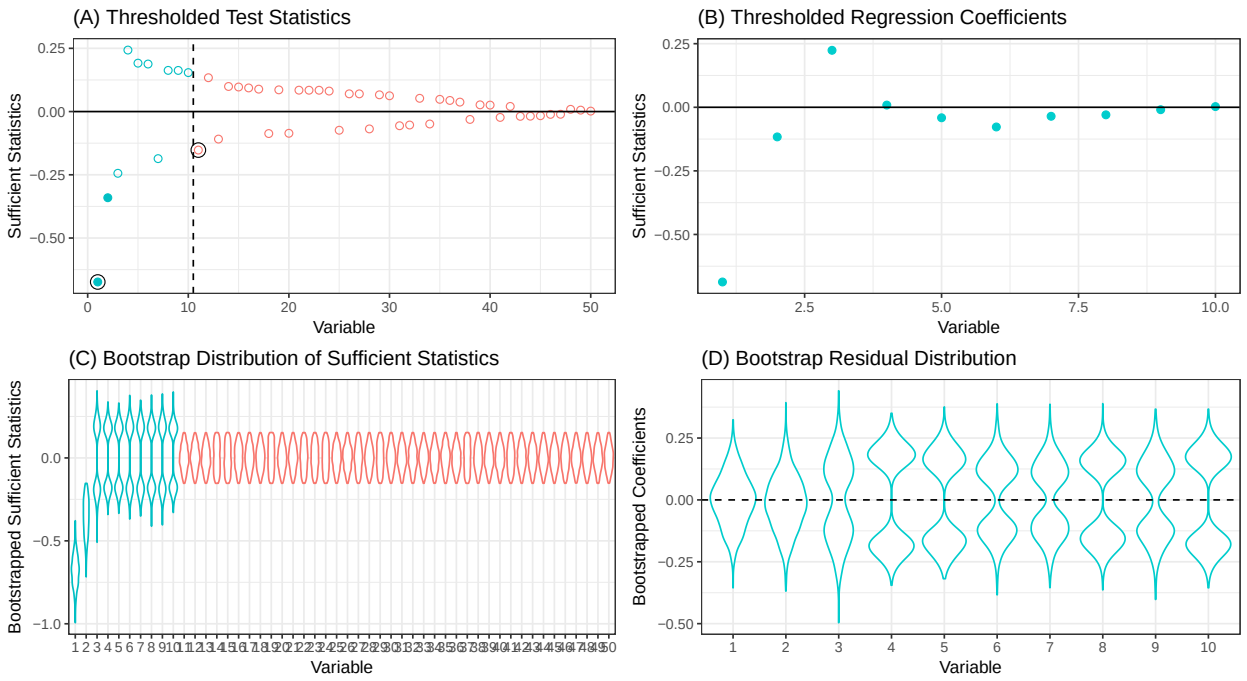

Figure 4.3: Illustration of the conditional bootstrap procedure, applied to marginal screening. **(A)** We observe a sufficient statistic for each column of $\boldsymbol{X}$ and include the 10 covariates corresponding to the largest ones in the model (marked in blue). The non-null variables are marked in black. We then apply hard thresholding to the sufficient statistics based on a post-selection test. In this case only the two top covariates passed the secondary thresholding (marked by full circles). **(B)** We estimate regression coefficients for the selected model based on the thresholded means. **(C)** We sample observations from the conditional distribution of the test statistics under the thresholded parametrization. The selected covariates are truncated to have a large observed value and the sufficient statistics corresponding to covariates that were not selected are truncated to have a value below the selection threshold. **(D)** We transform the sampled sufficient statistic to regression coefficients and compute the bootstrap residuals.

select a model, we normalize the columns of $\boldsymbol{X}$ and select the ten variables with the largest inner product with y in absolute value. We plot the sufficient statistics according to their size in panel A of Figure 4.3, where the non-null variables are marked with black circles. We set the thresholding parameter to $s_n = 0.01/10$, so as to obtain an approximate probability

of rejecting any true nulls of 0.01. We compute the approximate p-values,

$$p_j = \mathbb{P}_0(|\, \boldsymbol{X}_{\cdot,j}^T\, Y| > |\, \boldsymbol{X}_{\cdot,j}^T\, y| \mid j \in M).$$

and set $\tilde{\mu}_j = (\boldsymbol{X}_{\cdot,j}^T\, y) I\{s_j < t_n \cap j \in M\}$. In the context of regression, we infer on regression coefficients which are linear functions η_j of μ, and so in practice,

$$g'(\tilde{\mu})\sqrt{n}(T_n - \mu) = \sqrt{n}(\eta_j^T T_n - \eta_j^T \mu) = \sqrt{n}(\hat{\beta}_j - \beta_j).$$

To compute the bootstrap residuals, we compute a *thresholded coefficient estimate* $\tilde{\beta}^M := n(\boldsymbol{X}_M^T\, \boldsymbol{X}_M)^{-1}\tilde{\mu}_M$ (panel B). We then sample from the truncated distribution conditionally on the selected model (panel C), and compute the bootstrap residuals (panel D),

$$r_j^b = \hat{\beta}_j^M - \tilde{\beta}_j^M.$$

We plot the resulting confidence intervals in Figure 4.4, along with the naive Gaussian confidence intervals and the polyhedral confidence intervals. We had to truncate some of the polyhedral confidence intervals for them to fit in the figure in a reasonable way. In this example, the bootstrap confidence intervals are much shorter than the polyhedral confidence intervals and not much larger than the naive confidence intervals. Setting the thresholding parameter s_n to a smaller value is likely to makes the bootstrap confidence intervals larger, though their expected size is guaranteed to always be finite. For sufficiently large observed regression coefficients we will reject all univariate post-selection tests and the bootstrap confidence intervals will coincide with the Gaussian ones. In a simulation study in Section 4.5, we show that the bootstrap confidence intervals also offer some improved power over the polyhedral confidence intervals.

Remark 4.1. Approximate and exact p-values. In our procedure we compute p-values that are marginally correct for each test-statistic conditionally on that specific test-statistic

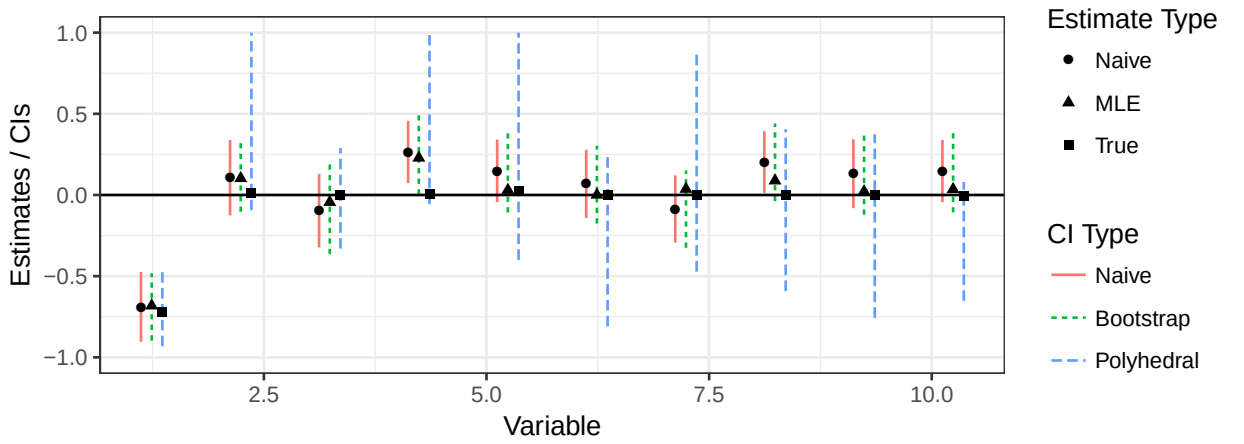

Figure 4.4: Conditional bootstrap confidence intervals following selection with marginal screening. We plot the bootstrap confidence intervals (dashed green line), polyhedral confidence intervals (dashed blue line) and naive confidence intervals (solid red line) for a regression model selected with marginal screening. Both types of selection adjusted confidence intervals cover all true parameter values, while the naive confidence intervals only cover 8 out of 10. The bootstrap confidence intervals tend to be much shorter than the polyhedral confidence intervals. We truncated the second, fourth and sixth polyhedral confidence intervals so they will fit in the figure (they are larger in practice).

crossing the thresholding. However, in our example we applied a Bonferroni correction to the p-values in order to avoid rejecting a false null with high probability. In many post-selection problems the number of variables in the model may not be fixed ahead of time and then the Bonferroni correction may be random with respect to the event of the jth test-statistic crossing the threshold[1]. As we show in Section 4.4, we do not require that our p-values are valid post-selection in order for our procedure to be consistent. However, if one desires a secondary thresholding procedure with finite sample guarantees then it is possible to use polyhedral p-values such as the ones proposed by Lee and Taylor (2014) for inference following marginal screening.

[1]Ruth Heller, personal communication.

4.4 Theory for the conditional bootstrap

In this section we will give some consistency results for the conditional bootstrap procedure proposed in the previous section. In Section 4.4.1 we give a relatively general consistency result, and in Section 4.4.2 we briefly discuss the applicability of our theory to model selection in regression, and data carving. We relegate all of the proofs to the appendix.

4.4.1 A general consistency result

The asymptotic correctness of our bootstrap technique depends on the scaling of the thresholding parameter s_n. A consistent estimation procedures requires two things. First, we need to be able to threshold all null variables and so we ask that,

$$\lim_{n\to\infty} \frac{s_n}{\mathbb{P}_n(M)} = 0. \tag{4.2}$$

Second, we need to make sure that the non-null variables cross the threshold with increasing probability as the sample size grows. Define $a_{nj} = \Phi(-c_j n^{-\delta+0.5})$ and $a_n = \max_j a_{nj}$. We ask that

$$\Phi^{-1}(a_n s_n) = o(\sqrt{n}). \tag{4.3}$$

For some of our results we will also require a bound on the rate at which the distribution of T_n converges to normality compared to the rate at which the probability of selecting M converges to zero, and so we assume that,

$$\mathbb{E}[T(y_i)_j^3] < \infty,\ \forall j, \qquad \lim_{n\to\infty} \sqrt{n}\,\mathbb{P}_n(M) = \infty. \tag{4.4}$$

We begin by showing that our method approximates the distribution of the residuals $\sqrt{n}(T_n - \mu)$ consistently.

Lemma 4.1. Suppose that conditions (4.1), (4.2), (4.3), and (4.4) hold, and also that

$M_1^c = \emptyset$. Then,

$$\lim_{n\to\infty} \sup_{\xi\in\mathbb{R}^p} \left|\mathbb{P}(\sqrt{n}(T_n - \mu) \leq \xi | M) - \mathbb{P}(\sqrt{n}(T_n^b - \tilde{\mu}) \leq \xi | M)\right| = 0.$$

Lemma 4.1 states that our modified bootstrap procedure estimates the distribution of the residuals $\sqrt{n}(T_n - \mu)$ consistently. In order to construct consistent confidence intervals for differentiable functions of μ, we require a delta method type result which we give next.

Lemma 4.2. Under the conditions of Lemma 4.1,

$$\lim_{n\to\infty} \sup_{\xi\in\mathbb{R}} \left|\mathbb{P}\left(g'(\mu)\sqrt{n}(T_n^b - \tilde{\mu}) \leq \xi \mid M\right) - \mathbb{P}\left(\sqrt{n}(g(T_n) - g(\mu)) \leq \xi \mid M\right)\right| = 0.$$

We are now ready to state our main consistency result. In Theorem 4.2 we make two types of statements. The first, is a statement of consistency conditionally on the selection of a specific model where we require assumptions regarding the probability of selecting the model being sufficiently large. The other is a marginal statement regarding the probability of constructing covering confidence intervals on average across experiments. For ease of exposition we formulate our result to address the case where only a single confidence interval is constructed for each selected model, though it is straight forward to generalize our result to the case where several confidence intervals are constructed.

Theorem 4.2. Suppose that conditions (4.1), (4.2), and (4.3) hold, then confidence intervals computed using Procedure 4.3 at a level $1 - \alpha$ satisfy

$$\lim_{n\to\infty} \mathbb{P}\left(g^{S(\mu)}(\theta) \in CI_B(S(Y))\right) = 1 - \alpha.$$

Conditioning on a specific model M, if in addition to the above, assumption (4.4) holds and $M_1^c = \emptyset$ then we also have,

$$\lim_{n\to\infty} \mathbb{P}(g^M(\theta) \in CI_B(M) \mid S(Y) = M) = 1 - \alpha.$$

Remark 4.2. Distributional assumptions. In our proofs, assumption (4.4) can be replaced by the assumption that there exists a constant $a > 0$ such that,

$$\lim_{n\to\infty} \inf \mathbb{P}_n(M) \geq a.$$

This condition holds, for example, if $\delta = 0.5$ and M contains all non-null variables. Conversely, if we are willing to assume that $T_n \sim N(\mu, \mathbf{\Sigma})$, then we only require assumptions (4.2) and (4.3).

4.4.2 Application to regression and data carving

We begin by outlining conditions under which we expect our Conditional Bootstrap to be applicable to inference in regression models. Suppose that we observe $(y_1, X_1), \ldots, (y_n, X_n) \in \mathbb{R} \times \mathbb{R}^p$. We treat X as fixed, or equivalently, conditioned upon. We make the following assumptions,

$$\lim_{n\to\infty} \frac{\sigma^2}{n} \sum_{i=1}^{n} X_i X_i^T = \mathbf{\Sigma}, \qquad \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_i y_i = \mu, \quad a.s. \tag{4.5}$$

with $\sigma^2 = Var(Y_i)$.

Corollary 4.1. *Under the assumptions of Theorem 4.2 as well as condition* (4.5)*,*

$$\lim_{n\to\infty} \mathbb{P}(\beta_j^M \in CI_B(M)|M) = 1 - \alpha, \ \forall\, j \in M$$

and,

$$\lim_{n\to\infty} \mathbb{E}\left[|S(Y)|^{-1} \sum_{j\in S(Y)} I\left\{ \beta_j^{S(Y)} \in CI_B(S(Y)) \right\} \right] = 1 - \alpha.$$

Next we discuss the topic of data carving. Data carving is a term coined by Fithian et al. (2014) referring to the practice of leaving some data out when selecting a model, and then using the left-out data in combination with the information left over from the model selection step to conduct inference. As in Example 4.2 we assume that we assign $m < n$

observation to select a model with,

$$T_n^S = \frac{1}{m}\sum_{i=1}^{m} X_i y_i, \qquad S(Y) = \{|T_{nj}^S| > c_j m^{-\delta}\},$$

and conduct inference based on a weighted average of T_n^s and an estimate based on the left out information,

$$T_n^I = \frac{1}{n-m}\sum_{i=n-m+1}^{n} X_i y_i, \qquad T_n = w_S T_n^S + w_I T_n^I, \qquad w_S + w_I = 1.$$

Our goal is to approximate to distribution of $\sqrt{n}(T_n - \mu)$. We propose a procedure for doing so next. Underlying our procedure is the assumption that

$$\lim_{n\to\infty} \frac{m}{n} = c > 0.$$

Procedure 4.4. Conditional Bootstrap with Data Carving

1. Set a thresholding parameter $s_n \in (0, 1)$.

2. Compute p-values,
$$p_j = \mathbb{P}_0(|T_{nj}^I| \geq |t_{nj}^I|)$$

3. Set $\tilde{\mu}_j^S = t_{nj} I\{p_j \leq s_n \cap j \in M\}$ and $\tilde{\mu}^I = t_n^I$.

4. For $b \in \{1, \ldots, B\}$ sample,

$$t^{Sb} \sim TN(\tilde{\mu}^S, m^{-1}\,\mathbf{\Sigma}, S(Y) = M), \qquad t^{Ib} \sim N(\tilde{\mu}^I, (n-m)^{-1}\,\mathbf{\Sigma}),$$

and compute the residuals,

$$r^b = g'(\tilde{\mu}^I)\left[w_S(t^{Sb} - \tilde{\mu}^S) + w_I(t^{Ib} - \tilde{\mu}^I)\right].$$

5. Compute a level $1-\alpha$ confidence interval:

$$CI_B(S(Y)=M) = (g(t_n) - r_{1-\alpha/2}, g(t_n) - r_{\alpha/2}).$$

Using data carving, we can prove the consistency of our procedure under less stringent conditions because our secondary thresholding procedure no longer relies on conditional p-values.

Theorem 4.3. Assume that condition (4.1) holds, as well as,

$$s_n = o(1), \qquad \Phi^{-1}(s_n) = o(\sqrt{n-m}), \qquad \lim_{n\to\infty} m/n = c > 0.$$

Then, confidence intervals constructed based on Procedure 4.4 satisfy,

$$\lim_{n\to\infty} \mathbb{P}\left(g^{S(\mu)}(\mu) \in CI_B(S(Y))\right) = 1-\alpha.$$

Conditioning on a specific model M, if $M_1^c = \emptyset$ then we also have,

$$\lim_{n\to\infty} \mathbb{P}(g^M(\mu) \in CI_B(M) \mid S(Y) = M) = 1-\alpha.$$

4.5 Simulation study

In this section we conduct a simulation study to verify our theoretical findings, as well as to evaluate the finite sample properties of our methodology. We sample covariates $X_1, \ldots, X_{200} \overset{\text{i.i.d.}}{\sim} N_{100}(0, \mathbf{U})$ where $\mathbf{U}_{i,j} = \rho^{|i-j|}$ with $\rho \in \{0, 0.7\}$. We then generate a vector of regression coefficient β, sampling either 2 or 8 coefficients from the $Laplace(1)$ distribution and setting the rest to zero. We then compute $\mu = \boldsymbol{X}\,\beta$ and set,

$$\sigma^2 = \frac{Var(\mu)}{\text{snr}}$$

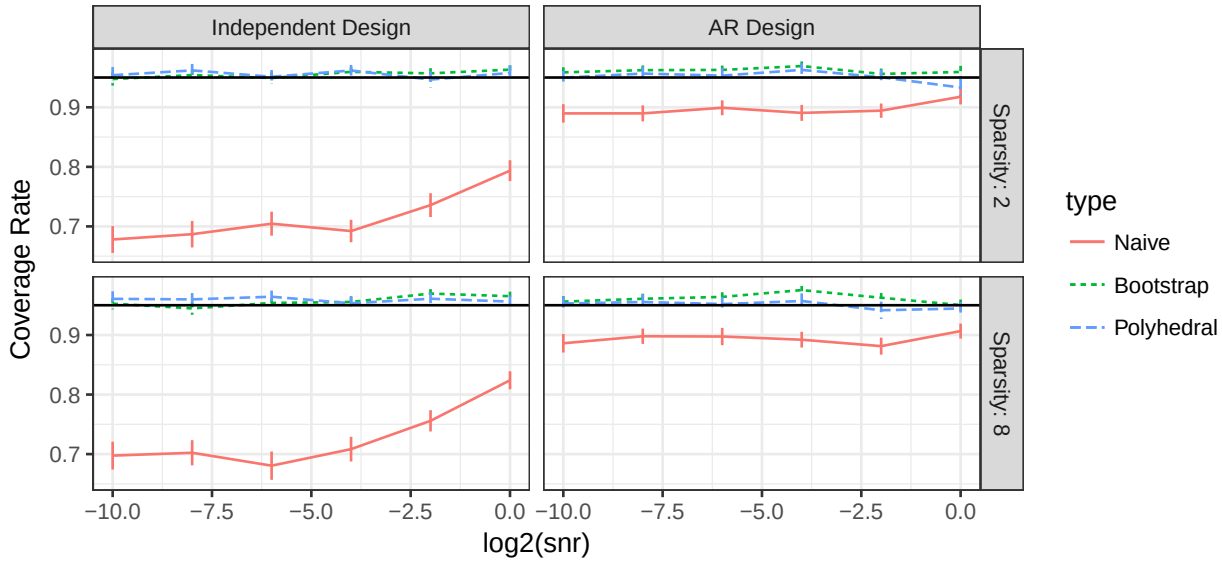

Figure 4.5: Coverage rates of naive, conditional bootstrap, and polyhedral confidence intervals. Both of the selection adjusted confidence intervals obtain the desired coverage rate (or higher) in all simulation settings. The naive confidence intervals have a below nominal coverage rate in all simulation settings, though the coverage rate is worse with independent design.

for a range of signal to noise ratios snr $\in \{0.001, 0.004, 0.016, 0.063, 0.25, 1\}$. Finally, we sample $y_i \sim N(X_i\beta, \sigma^2)$ for $i = 1, \ldots, 200$. For each simulated dataset we select the ten variables that have the highest correlation with a response and estimate a joint model. When constructing bootstrap confidence intervals we set the thresholding parameter to $s_n = 0.01/10$. We emphasize that while we use a very conservative thresholding procedure, as in the univariate case, thresholding a mean parameter to zero does not necessarily imply that the confidence intervals for the corresponding covariate will cover zero.

In Section 4.5.1 we assess the coverage rate of the post-selection inference procedures, in Section 4.5.2 we compare the different methods on their ability to detect true signals, and in Section 4.5.3 we assess the size of the resulting confidence intervals.

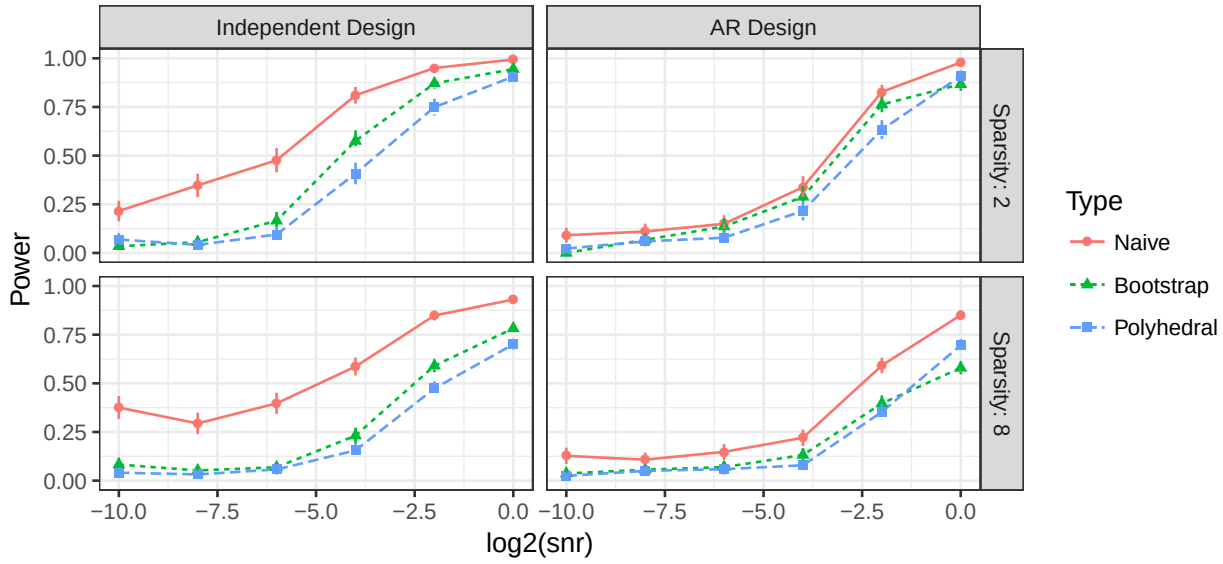

Figure 4.6: Comparison of power to detect true signals. We plot the power of the different inference methods to detect true signal. In the context of our experiments we define power to detect true signals as the power to reject the null for covariates that have non-zero effects in the full model and were included in the selected model. The naive confidence intervals have the highest power, at the cost of inflated type I error rates. The bootstrap confidence intervals have equivalent or higher power than the polyhedral confidence intervals, except for experiments with autoregressive design and the highest signal to noise ratios.

4.5.1 Assessment of coverage rates

We begin by assessing the coverage rates of the naive, bootstrap and polyhedral confidence intervals. The coverage rates are plotted in Figure 4.5 along with 95% confidence intervals for the estimated coverage rates. Both the polyhedral and the bootstrap methods obtain the desired coverage rates or slightly higher in some settings. The naive confidence intervals have a lower than nominal coverage rate in all simulation settings, and have a lower coverage rate when the design is independent. This can be attributed to the fact that when the design is correlated the selected models are less spurious because null covariates that are highly correlated with the true signals have a high probability of being selected, whereas in the independent design only the true signals have a high probability of being selected.

4.5.2 Assessment of power to detect true signals

Having verified that our proposed inference method obtains the nominal coverage rate in the simulation settings considered, it may be of interest to compare the post-selection inference techniques with respect to their ability to detect true signals. There is some difficulty in defining true signals in selected regression models because if a true signal (in the full model) is left out of the selected model, then its explanatory power may be transferred to null variables that are correlated with it. Thus, we work with narrower definition for power given by

$$\text{Power} := \frac{\left|\left\{j : \{\beta_j \neq 0\} \cap \{j \in M\} \cap \{H_{0j}^M \text{ rejected}\}\right\}\right|}{|\{j : \{\beta_j \neq 0\} \cap \{j \in M\}\}|}.$$

That is, we count the number of full model true signals that were selected and for which we rejected the null hypothesis (a null hypothesis is rejected if a confidence interval does not cover zero). If no true signals are selected in a simulation instance, then we do not include that instance in the power calculation.

We plot the result of our experiment in Figure 4.6. The first detail worth noting is that when the covariates are dependent the selection is less severe because it operates on the highly correlated sufficient statistics $\boldsymbol{X}^T y$ rather than the regression coefficients. While β is a sparse vector, μ is not, meaning that the risk of picking up null covariates is much lower (in the marginal sense). Thus, the difference between the selection adjusted methods and the naive confidence intervals is lower when the covariates are dependent, and the difference between the polyhedral and bootstrap confidence intervals is smaller. The bootstrap confidence intervals have better or equivalent power compared to the polyhedral confidence intervals, with the setting with dependent design and very high signal to noise ratio being the sole exception. The difference between the bootstrap and polyhedral methods is at most 20% for the independent design 10% for the autoregressive design.

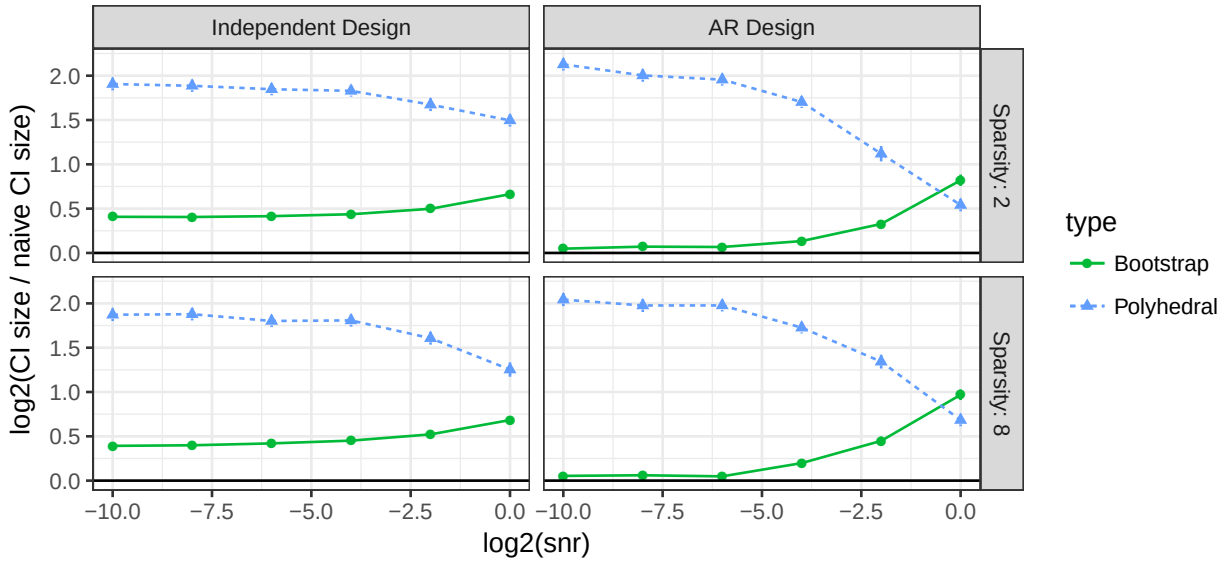

Figure 4.7: Relative size of selection adjusted confidence intervals. We plot the relative size of the polyhedral and bootstrap confidence intervals compared to the naive unadjusted confidence intervals. The bootstrap confidence intervals are between 40% and 100% larger than the naive confidence intervals in most of the simulation settings, while the polyhedral confidence intervals are anywhere between 40% and 900% larger.

4.5.3 Comparing the size of the confidence intervals

Lastly, we compare the selection adjusted confidence intervals with respect to their size. As we discussed in Chapter 2, it is a challenge to quantify the size of the polyhedral confidence intervals, because their expected size is infinity. Thus, we define the 'typical size' of the confidence intervals as the median of the median confidence interval size within each simulations. In Figure 4.7 we plot the typical confidence interval sizes relative to the naive, unadjusted confidence intervals. In almost all of the simulations settings the bootstrap confidence intervals are significantly shorter than the polyhedral confidence intervals, with the exception being the setting with high dependence and high signal-to-noise ratio as before. The bootstrap confidence intervals are typically about 40% larger than the naive confidence intervals, and are guaranteed to have a finite size because they are constructed based on

residuals sampled from the truncated normal distribution.

4.6 Discussion

In this work we presented a method for constructing post-selection confidence intervals based a modified conditional bootstrap approach. The confidence intervals are computed based on the post-selection distribution of the data without any extra conditioning, they are guaranteed to have a finite length and tend to have better power than the polyhedral confidence intervals. In Section 4.4 we outlined conditions under which our method produces consistent confidence intervals. We also proposed a bootstrap approach for inference based on data carving and showed that if an investigator is willing to set aside some data to be used only for inference, then consistency can be guaranteed under milder conditions.

The main drawback of our method is that it is only point-wise consistent even if the data is normally distributed. In contrast, the polyhedral confidence intervals are exact if the data is Gaussian and are otherwise uniformly consistent under certain distributional assumptions (Tibshirani et al., 2015). This issue relates to the result of Leeb and Pötscher (2006) who showed that conditional distributions cannot be approximated in a uniform manner. In order to overcome this drawback we advocated for constructing confidence in a conservative manner, imposing a stringent secondary threshold and setting most of the regression coefficients in the selected model to zero. This proved to be a satisfactory solutions in our simulation study, as the distribution of the bootstrapped residuals tends to be wider for zero valued coefficients. However, in order to give rigorous uniform guarantees we would need to identify specific values of parameters that yield conservative post-selection tests, and that is a difficult problem which we leave for future investigation.

4.A Proof of theorems

4.A.1 Proof of Theorem 4.1

We start by proving the case of $\delta = 0.5$. When $\delta = 0.5$, $\lim_{n\to\infty} \mathbb{P}_n(S_n(Y) = M) = \mathbb{P}_n(A_0, A_1^0) > 0$ because $\lim_{n\to\infty} \mathbb{P}_n(A_1^1) = 1$. To see why this is the case, assume without loss of generality that $\mu \geq 0$,

$$\begin{aligned}
&\lim_{n\to\infty} \mathbb{P}_n(A_1^1) \geq \lim_{n\to\infty} \mathbb{P}_n\left(\bigcap_{j\in A_1^1} T_{nj} > \frac{c_j}{\sqrt{n}}\right) = \lim_{n\to\infty} \mathbb{P}_n\left(\bigcap_{j\in A_1^1} \sqrt{n}(T_{nj} - \mu_j) > c_j - \sqrt{n}\mu_j\right) \\
\geq & 1 - \lim_{n\to\infty} \sum_{j\in A_1^1} \mathbb{P}_n\left(\sqrt{n}(T_{nj} - \mu_j) < c_j - \sqrt{n}\mu_j\right) = 1.
\end{aligned}$$

Thus,

$$\begin{aligned}
&\lim_{n\to\infty} \sup_{\xi\in\mathbb{R}^k} \left| \mathbb{P}_n(\sqrt{n}(T_n - \mu) < \xi \mid S_n(Y) = M) - \frac{\Phi(\xi; 0, \mathbf{\Sigma})}{\mathbb{P}_n(A_0, A_1^0)} \right| \\
\leq & \lim_{n\to\infty} \sup_{\xi\in\mathbb{R}^k} \left| \frac{\mathbb{P}(\sqrt{n}(T_n - \mu) < \xi)}{\mathbb{P}_n(S_n(Y) = M)} - \frac{\Phi(\xi; 0, \mathbf{\Sigma})}{\mathbb{P}_n(S(Y) = M)} \right| + \left| \frac{\Phi(\xi; 0, \mathbf{\Sigma})}{\mathbb{P}_n(S(Y) = M)} - \frac{\Phi(\xi; 0, \mathbf{\Sigma})}{\mathbb{P}_n(A_0, A_1^0)} \right| \quad (4.6) \\
= & 0
\end{aligned}$$

The first term in (4.6) converges to zero because of our assumption that T_n has an asymptotic normal distribution and the second term converges to zero because $\lim_{n\to\infty} \mathbb{P}_n(A_1^1) = 1$.

If $\delta \in (0, 0.5)$ then $\mathbb{P}(A_1^0) \to 0$ whenever $M_0 \neq \emptyset$. To see why, suppose there exists an arbitrary index j such that $j \in M$ and $\mu_j = 0$. Then,

$$\begin{aligned}
\lim_{n\to\infty} \mathbb{P}(S_n(Y) = M) &\leq \lim_{n\to\infty} \mathbb{P}(|T_{nj})| > c_j n^{-\delta}) = \lim_{n\to\infty} \mathbb{P}(\sqrt{n}|T_{nj}| > c_j n^{-\delta+0.5}) \quad (4.7) \\
&\leq \lim_{n\to\infty} \frac{Var(\sqrt{n}T_{nj})}{c_j^2 n^{2(0.5-\delta)}} = 0.
\end{aligned}$$

If T_{nj} is sub-gaussian with sub-gaus then the rate of decay of $\mathbb{P}_n(S(Y) = M)$ is exponential. This result is obtained by replacing the application of Markov's inequality in (4.7) by

Chernouf's bound. Suppose that for all n and j there exists a constant σ_j such that

$$\mathbb{E}\, e^{\lambda(T_{nj}-\mu_j)} \leq e^{\sigma_j^2\lambda^2}, \qquad \forall\lambda \in \mathbb{R}.$$

Then

$$\mathbb{P}(\sqrt{n}|T_{nj}| > c_j n^{-\delta+0.5}) \leq 2e^{-\frac{c_j^2 n^{1-2\delta}}{2\sigma_j^2}}.$$

4.A.2 Proof of Lemma 4.1

By Theorem 4.1,

$$\lim_{n\to\infty}\sup_{\xi\in R^p}\left|\mathbb{P}(\sqrt{n}(T_n-\mu)<\xi|M) - \frac{\Phi(\xi;0,\mathbf{\Sigma})}{\mathbb{P}(A_0,A_1^0)}\right| = 0.$$

Adding and subtracting the limiting truncated distribution we have,

$$\begin{aligned}&\lim_{n\to\infty}\sup_{\xi\in R^p}\left|\mathbb{P}(\sqrt{n}(T_n-\mu)<\xi|M) - \mathbb{P}(\sqrt{n}(T_n^b-\tilde{\mu})<\xi \mid M)\right| \\ \leq& o(1) + \lim_{n\to\infty}\sup_{\xi\in R^p}\left|\mathbb{P}(\sqrt{n}(T_n^b-\tilde{\mu})<\xi|M) - \frac{\Phi(\xi;0,\boldsymbol{V})}{\mathbb{P}(A_0,A_1^0)}\right|.\end{aligned}$$

So, our goal is to show that T_n and T_n^b have the same limiting distribution.

Our first step will be to show that our secondary thresholding is consistent in the sense that it for a large enough sample size it will always threshold null variables to zero and will not threshold non-null with high probability. We start by showing that our method will always threshold null variables to zero. Set $a_{nj} = \Phi(-c_j n^{-\delta+0.5})$. For a false rejection we need,

$$p_j = \mathbb{P}_0(|T_{nj}| \geq |t_{nj}| \mid j \in M) = \frac{\mathbb{P}_0(|T_{nj}| > |t_{nj}|)}{\mathbb{P}_0(|T_{nj}| \geq c_j n^{-\delta})} \leq \frac{\mathbb{P}_0(\sqrt{n}|T_{nj}| \geq \sigma_j\Phi^{-1}(1-a_{nj}s_n/2))}{\mathbb{P}(|T_{nj}| \geq c_j n^{-\delta})}.$$

That is, we need to observe $\sqrt{n}|t_n| \geq \sigma_j \Phi^{-1}(1-\alpha_{nj}s_n/2)$,

$$\mathbb{P}_0(\sqrt{n}|T_{nj}| \geq \sigma_j \Phi^{-1}(1-\alpha_{nj}s_n/2) \mid M) \leq \frac{\mathbb{P}_0(\sqrt{n}|T_{nj}| \geq \sigma_j \Phi^{-1}(1-\alpha_{nj}s_n/2))}{\mathbb{P}_n(M)}$$
$$\leq \frac{|\,\mathbb{P}_0(\sqrt{n}|T_{nj}| \geq \sigma_j \Phi^{-1}(1-\alpha_{nj}s_n/2)) - s_n a_{nj}|}{\mathbb{P}_n(M)} + \frac{s_n a_{nj}}{\mathbb{P}_n(M)} = o(1).$$

Where the last equality is by the Berry-Essen Theorem and assumptions (4.2) and (4.4).

Next, we show that our thresholding method will identify non-nulls with high probability. Suppose that $\mu_j > 0$. Then,

$$\begin{aligned}
&\mathbb{P}(p_j > s_n|M) \leq \mathbb{P}(\sqrt{n}T_n \leq \sigma_j \Phi^{-1}(1-\alpha_{nj}s_n) \mid M) \\
\leq &\frac{\mathbb{P}(\sqrt{n}T_n \leq \sigma_j \Phi^{-1}(1-\alpha_{nj}s_n))}{\mathbb{P}_n(M)} = \frac{\mathbb{P}(\sqrt{n}(T_n-\mu_j) \leq \sigma_j \Phi^{-1}(1-\alpha_{nj}s_n) - \sqrt{n}\mu_j)}{\mathbb{P}_n(M)} \\
\leq &\frac{|\,\mathbb{P}(\sigma_j^{-1}\sqrt{n}(T_n-\mu_j) \leq \Phi^{-1}(1-\alpha_{nj}s_n) - \sigma_j^{-1}\sqrt{n}\mu_j) - \Phi(\Phi^{-1}(1-\alpha_{nj}s_n) - \sigma_j^{-1}\sqrt{n}\mu_j)|}{\mathbb{P}_n(M)} \\
+ &\frac{\Phi(\Phi^{-1}(1-\alpha_{nj}s_n) - \sigma_j^{-1}\sqrt{n}\mu_j)}{\mathbb{P}_n(M)} = o(1). \qquad (4.8)
\end{aligned}$$

In the last part of (4.8) the first term is $o(1)$ by assumption (4.4) and the Berry-Essen Theorem. We use Chernoff's bound to show that the second term also decreases in n,

$$\frac{\Phi(\Phi^{-1}(1-\alpha_{nj}s_n) - \sigma_j^{-1}\sqrt{n}\mu_j)}{\mathbb{P}_n(M)} \leq \frac{1}{\mathbb{P}_n(M)} e^{-\frac{[\sigma_j^2\sqrt{n}\mu + \Phi^{-1}(s_n a_n)]^2}{2}} = o(1).$$

A similar result holds for $\mu_j < 0$.

The result follows from the fact that we sample T_n^b from a normal distribution conditional on selection, and the fact that under our bootstrapped distribution,

$$\lim_{n\to\infty} \widetilde{\mathbb{P}}(A_1^1) := \lim_{n\to\infty} \mathbb{P}\left(\bigcap_{j\in M_1} |T_{nj}^b| > c_j n^{-\delta}\right) = 0.$$

Without loss of generality assume that $\mu_j > 0$ for all $j \in M_1$. Then,

$$\lim_{n\to\infty} \mathbb{P}\left(\bigcap_{j\in M_1} |T_{nj}^b| > c_j n^{-\delta}\right) \geq 1 - \sum_{j\in M_1} \mathbb{P}(|T_{nj}^b| \leq c_j n^{-\delta})$$
$$\geq \lim_{n\to\infty} 1 - \sum_{j\in M_1} \mathbb{P}(\sqrt{n}(T_{nj}^b - T_n) \leq c_j n^{-\delta+0.5} - \sqrt{n}T_n) = 1.$$

4.A.3 Proof of Lemma 4.2

We start by taking a Taylor expansion

$$\sqrt{n}(g(T_n) - g(\mu)) = \sqrt{n}g'(\mu^*)(T_n - \mu),$$

where μ^* is strictly between μ and $\tilde{\mu}$. By Lemma 4.1 $\lim_{n\to\infty} \sqrt{n}(T_n - \mu) =^D \lim_{n\to\infty} \sqrt{n}(T_n^b - \tilde{\mu})$. If $g(\mu)$ is a linear function of μ then we are done. If $g(\mu)$ does not have a constant derivative, then we need to show that $\tilde{\mu}$ is a consistent estimator of μ. We first show that T_n is a consistent estimator of μ.

$$\lim_{n\to\infty} \mathbb{P}\left(\bigcup_j |T_{nj} - \mu_j| > \varepsilon | M\right) \leq \lim_{n\to\infty} \frac{\mathbb{P}\left(\bigcup_j |T_{nj} - \mu_j| > \varepsilon\right)}{\mathbb{P}_n(M)}$$
$$\leq \lim_{n\to\infty} \frac{\sum_{j=1}^p \mathbb{P}(|T_{nj} - \mu_j| > \varepsilon)}{\mathbb{P}_n(M)} \leq \lim_{n\to\infty} \sum_{j=1}^p \frac{\sigma_j^2}{n\,\mathbb{P}_n(M)\varepsilon^2} = 0.$$

The rest follows from the fact our thresholding procedure is guaranteed to identify the zero and non-zero components of $\tilde{\mu}$ correctly for a large enough sample size, so asymptotically $\tilde{\mu}_j = T_n$ for $\mu \neq 0$ and $\tilde{\mu}_j = 0$ for $\mu_j = 0$.

4.A.4 Proof of Theorem 4.2

We begin by showing that under the conditions of the theorem,

$$\lim_{n\to\infty} \mathbb{P}(g^M(\theta) \in CI_B(M) \mid S(Y) = M) = 1 - \alpha. \tag{4.9}$$

We construct confidence intervals of the form,

$$\left[g(T_n) - \frac{1}{\sqrt{n}}\left(\sqrt{n}g'(\tilde{\mu})(T_n^b - \tilde{\mu})\right)_{1-\alpha/2}, g(T_n) - \frac{1}{\sqrt{n}}\left(\sqrt{n}g'(\tilde{\mu})(T_n^b - \tilde{\mu})\right)_{\alpha/2}\right]$$

The conditional consistency (4.9) holds because by Lemma 4.2,

$$\sqrt{n}g'(\tilde{\mu})(T_n^b - \tilde{\mu}) \to^D \sqrt{n}(g(T_n) - g(\mu)),$$

where the convergence is conditional on selection.

For the second part of the theorem, we divide our (finite) set of model $\mathcal{M}$ into two disjoint sets,

$$\mathcal{M}_1 = \{M : \lim_{n\to\infty} \sqrt{n}\,\mathbb{P}_n(M) = \infty\}, \qquad \mathcal{M}_2 = \{M : \lim_{n\to\infty} \sqrt{n}\,\mathbb{P}_n(M) = O(1)\}.$$

So, $\mathcal{M}_1$ is a set of models that are selected with high probability, $\mathcal{M}_2$ is a set of model that have a probability of selection that decays to zero fast, and $\mathcal{M} = \mathcal{M}_1 \cup \mathcal{M}_2$.

$$\begin{aligned}
&\lim_{n\to\infty} \mathbb{P}(g^{S(Y)}(\mu) \in CI_B(S(Y)) \\
=&\lim_{n\to\infty} \sum_{M\in\mathcal{M}_1} \mathbb{P}(g^M(\mu) \in CI_B(S(Y) = M \mid M)\,\mathbb{P}(M) + \sum_{M\in\mathcal{M}_2} \mathbb{P}(g^M(\mu) \in CI_B(S(Y) = M \mid M)\,\mathbb{P}(M) \\
\leq&\lim_{n\to\infty} \sum_{M\in\mathcal{M}_1} \mathbb{P}(g^M(\mu) \in CI_B(S(Y) = M \mid M)\,\mathbb{P}(M) + \sum_{M\in\mathcal{M}_2} \mathbb{P}(M) \\
=&1 - \alpha.
\end{aligned}$$

4.A.5 Proof of Corollary 4.1

The first part of the statement holds by Theorem 4.2, our assumption (4.5), and the fact that β_j^M is a linear function of μ,

$$\beta_j^M = ne_j^T(\boldsymbol{X}_M^T\,\boldsymbol{X}_M)^{-1}\mu.$$

It is left to show that we cover the selected parameters at a rate of $1-\alpha$ on average.

$$\begin{aligned}
&\lim_{n\to\infty} \mathbb{E}\left[|S(Y)|^{-1}\sum_{j\in S(Y)} I\left\{\beta_j^{S(Y)} \in CI_B(S(Y))\right\}\right] \\
&= \lim_{n\to\infty} \mathbb{E}\,\mathbb{E}\left[|M|^{-1}\sum_{j\in M} I\left\{\beta_j^{M} \in CI_B(M)\right\}\middle| M\right] \\
&= \lim_{n\to\infty} \sum_{M\in\mathcal{M}} |M|^{-1}\sum_{j\in M} \mathbb{P}(\beta_j^M \in CI_B(M) \mid M)\,\mathbb{P}_n(M).
\end{aligned}$$

Dividing $\mathcal{M}$ into $\mathcal{M}_1$ and $\mathcal{M}_2$ as we have done in the proof of Theorem 4.2,

$$\begin{aligned}
&\lim_{n\to\infty} \sum_{M\in\mathcal{M}} |M|^{-1}\sum_{j\in M} \mathbb{P}(\beta_j^M \in CI_B(M) \mid M)\,\mathbb{P}_n(M) \\
&= \lim_{n\to\infty} \sum_{M\in\mathcal{M}_1} |M|^{-1}\sum_{j\in M} \mathbb{P}(\beta_j^M \in CI_B(M) \mid M)\,\mathbb{P}_n(M) + \sum_{M\in\mathcal{M}_2} \mathbb{P}_n(M) = 1-\alpha.
\end{aligned}$$

4.A.6 Proof of Theorem 4.3

Our goal is to approximate the distribution of,

$$T_n - \mu = w_S(T_n^S - \mu) + w_I(T_n^S - \mu). \tag{4.10}$$

The two terms on the righthand side of (4.10) are independent and the second of the two converges to a normal in distribution. The first, converges to a truncated normal as described in Theorem 4.1. So in order to show that our bootstrap distribution converges to the correct truncated distribution we need to show that our thresholding procedure identifies the zero and non-zero coordinates of μ as we have done in the proof of Lemma 4.1. Here however, our this task is made easier by the fact that our p-values are not conditional. Starting with the null coordinates, suppose that $\mu_j = 0$. We threshold μ_j to zero if,

$$\mathbb{P}(\sqrt{n}|T_{nj}^I| < \sigma_j\Phi^{-1}(1-s_n)) = |\,\mathbb{P}(\sqrt{n}|T_{nj}^I| < \sigma_j|\Phi^{-1}(s_n)|) - (1-s_n)| + 1 - s_n = 1 + o(1)$$

Similarly for $\mu_j > 0$,

$$\mathbb{P}(\sqrt{n}|T^I_{nj}| < \sigma_j \Phi^{-1}(1-s_n)) \leq \mathbb{P}(\sqrt{n}(T^I_{nj} - \mu_j) < \sigma_j \Phi^{-1}(1-s_n) - \sqrt{n}\mu_j) = o(1).$$

The rest follows as in the proof of Lemma 4.1.

Next result we rely on the delta method described in (4.2). Here our residuals can be written as,

$$r^b = w_S g(\tilde{\mu}^I)(T^{Sb}_n - \tilde{\mu}^S) + w_I g(\tilde{\mu}^I)(T^{Ib}_n - \tilde{\mu}^I). \tag{4.11}$$

In (4.11) the first element converges to the correct distribution by Lemma 4.2, and by the classic delta method,

$$\sqrt{n-m}(g(T^I_n) - g(\mu)) \to^D N(0, g'(\mu)^T \, \mathbf{\Sigma} \, g'(\mu)),$$

and $\tilde{\mu}^I \to^P \mu$ by the law of large numbers. The rest of the proof follows as in the proof of Theorem 4.2.

Chapter 5
DISCUSSION

5.1 Recap

In this work we explored different approaches for conducting conditional post-selection inference while conditioning on the minimal amount of information required to uniquely identify the selected model. First, we developed a computational framework for computing the conditional maximum likelihood estimator in multivariate problems and showed that it offers improved estimation accuracy over naive unadjusted estimates. Then, in the context of inference following aggregate testing we developed post-selection inference methods with improved power over the commonly used polyhedral lemma. Finally, in the last chapter we developed a modified bootstrap approach aimed at consistently estimating the post-selection distribution of unadjusted post-selection estimates.

While in some ways our methods do obtain the goal we set out for ourselves at the outset, they do leave something to be desired in other ways. A common drawback to all of our methodologies is that they are very computationally intensive and thus, are time consuming to use compared to, for example, techniques based on the polyhedral lemma.

Another difficulty is the fact that conditional inference techniques tend to be application specific: A software package aimed at computing post-selection confidence intervals for the lasso cannot be used to compute confidence intervals for inference after model selection with marginal screening. Exacerbating this issue is the fact that our methods (mostly) require a higher degree of computational aptitude than methods based on the polyhedral lemma because their implementation requires at least some degree of expertise in Monte-Carlo sampling and stochastic optimization. In order for conditional inference techniques to be truly widely applicable, we must develop methods that are flexible and simple to generalize

for practitioners who do not have a deep expertise in post-selection inference.

Finally, while our aim was to offer alternatives to the polyhedral lemma, the conditional bootstrap is not quite satisfactory as it is only point-wise consistent under classical large sample assumptions. Model selection is often conducted in high-dimensional settings where the number of features is much larger than the number of observations ($p \gg n$), and in such settings it is especially important to make uniform convergence/consistency statements.

We close by briefly describing a few possible avenues of future work.

5.2 Future work

5.2.1 Most conservative parameterizations

In the previous section we highlighted the importance of making uniform consistency statements. The location non-invariance of the truncated normal distribution makes uniform statements difficult to obtain. Specifically, it is clear by the result of Leeb and Pötscher (2006) that we cannot hope to consistently approximate the post-selection distributions in a uniform manner. As an alternative to uniformly consistent methods, we can attempt to formulate methods that are uniformly conservative. Suppose that we are interested in testing an hypothesis $H_0 : \theta \in \Theta_0^M$ after observing $S(Y) = M$. We say that a parametrization $\tilde{\theta}$ is most conservative parametrization if $\tilde{\theta} \in \Theta_0^M$ and it satisfies,

$$\sup_{\theta \in \Theta_0^M} \mathbb{P}_\theta(\text{Reject } H_0^M | M) \leq \mathbb{P}_{\tilde{\theta}}(\text{Reject } H_0^M | M).$$

In Chapter 3 we showed that for inference after model selection with a Wald test the most conservative parametrization is $\theta = 0$ and for inference following model selection with a linear aggregate test of the form $a^T y < l \cup a^T y > u$ a most conservatives parametrization is any parametrization such that $a^T \theta = (l + u)/2$. However, finding a most conservative parametrization is a difficult problem in general, as the function $\mathbb{P}_\theta(\text{Reject } H_0^M | M)$ is often not convex, and may have multiple local maxima. For example, in the normal means problem if some selected coordinates of y are negatively correlated then the global-null is not the most

conservative parametrization and there exist several local maxima in the neighborhood of the origin.

5.2.2 Profile likelihood based inference

When conducting conditional inference on a distribution parametrized by a univariate parameter, we can construct valid confidence intervals by inverting conditional hypothesis tests

$$CI(M) = \{\theta : \alpha/2 < F_\theta(y|M) < 1 - \alpha/2\}.$$

Suppose that instead we are interested in conditionally inferring on a linear function of a multivariate parameter $\eta^T\theta$, $\theta \in \Theta \subseteq \mathbb{R}^p$. In such cases test inversion is no longer simple to implement because the conditional CDF $F_\theta(y|M)$ will often depend on the entire vector θ and not only on $\eta^T\theta$. One possible approach to inverting tests in multivariate settings is via a profile likelihood approach. Define,

$$\tilde{\theta}(a) = \arg\max_\theta f_\theta(y|M), \qquad \text{s.t. } \eta^T\theta = a.$$

We can construct confidence intervals of the form,

$$CI(M) = \{a : \alpha/2 < F_{\tilde{\theta}(a)}(\eta^T y|M) < 1 - \alpha/2\}.$$

In preliminary work, this approach has already shown some promise when applied to the problem of inferring on selected regions of interest in fMRI studies, though it remains to be seen if it can be justified theoretically.

5.2.3 The m-out-of-n bootstrap

The m-out-of-n is a modified bootstrap procedure (Bickel and Sakov, 2008) where instead of taking n samples with replacement from a dataset of n observations, we take $m < n$ samples

such that,

$$\lim_{n\to\infty} m/n = 0.$$

The m-out-of-n bootstrap is commonly used in situation where the distribution of the non-parametric bootstrap converges to a non-standard or random limit. Post-selection, we expect the m-out-of-n bootstrap to work because it can be shown that sub-sampling the data breaks the dependence induced by the selection.

Theorem 5.1. Suppose that $Y_1, \ldots Y_n \sim N(\mu, \mathbf{\Sigma})$ are i.i.d., $S(Y) = S(\bar{Y})$, $\liminf_{n\to\infty} \mathbb{P}_n(M) \geq a > 0$, and $\lim_{n\to\infty} m/n = 0$. Let $\bar{Y}_m$ be a mean computed based on a m observations sampled out of n. Then,

$$\sqrt{m}(\bar{Y}_m - \mu) \to^D N(0, \mathbf{\Sigma}).$$

The m-out-of-n bootstrap is an attractive potential solution to the post-selection inference problem because we suspect that it can be applicable in a a wide range of problems in a manner that is, at least to some extent, agnostic to the model selection method used and with uniform guarantees.

Proof of Theorem 5.1

Suppose that $y_1, \ldots, y_n \sim N(\mu, \mathbf{\Sigma})$ and that,

$$S(Y) = S(\bar{Y}), \qquad \liminf_{n\to\infty} \mathbb{P}_n(M) \geq a > 0, \qquad \lim_{n\to\infty} \frac{m}{n} = 0$$

The distribution of $Y|M$ is the same for any permutation of indexes and so, we can leverage the theory of Weber (1980) on central limit theorems for triangular arrays of exchangeable random variables to prove our result. We will assume without loss of generality that y_i is univariate because if y_i is multivariate it will be sufficient to prove the result for any fixed contrast $c^T y$. Set,

$$z_i = \frac{y_i - \mu}{\sigma}.$$

We note that under our conditions $E(y_i|M) \to \mu$ and $Var(y_i|M) \to \sigma^2$. According to Corollary 2 of Weber (1980) the m out of n average converges to a normal if,

1. $\lim_{n\to\infty} mE(Z_1 Z_2|M) = 0$.

2. $\lim_{n\to\infty} E(Z_1^2 Z_2^2|M) = 1$.

3. $\lim_{n\to\infty} E(Z_1^2 I\{|Z_1| > \varepsilon\sqrt{n}\}) \mid M) = 0,\ \forall \varepsilon > 0$.

Because we assumed y is gaussian,

$$E(Z_1 Z_2|M) = E\left[E(Z_1 Z_2|\bar{z})|M\right] = O\left(\frac{1}{n}\right),$$

and so $mE(Z_1 Z_2|M) \to 0$.

Next we show that the second condition is satisfied.

$$E(Z_1^2 Z_2^2|M) \leq \frac{1}{a},$$

and so we can apply the dominated convergence theorem to show that,

$$\lim_{n\to\infty} \int z_1^2 z_2^2 f(z_1, z_2) \frac{P(M|z_1, z_2)}{P(M)} dz_1 dz_2 = \int z_1^2 z_2^2 f(z_1, z_2) \lim_{n\to\infty} \frac{P(M|z_1, z_2)}{P(M)} dz_1 dz_2 = 1.$$

Finally,

$$\lim_{n\to\infty} E(Z_1^2 I\{|Z_1| > \varepsilon\sqrt{n}\}) \mid M) \leq \lim_{n\to\infty} a^{-1} E(Z_1^2 I\{|Z_1| > \varepsilon\sqrt{n}\}) = 0,$$

and the result follows.

VITA

TBD

BIBLIOGRAPHY

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second International Symposium on Information Theory (Tsahkadsor, 1971)*, pages 267–281. Akadémiai Kiadó, Budapest.

Bachoc, F., Preinerstorfer, D., and Steinberger, L. (2016). Uniformly valid confidence intervals post-model-selection. *arXiv preprint arXiv:1611.01043*.

Barber, R. F., Candès, E. J., et al. (2015). Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, 43(5):2055–2085.

Barber, R. F., Candès, E. J., and Samworth, R. J. (2018). Robust inference with knockoffs. *arXiv preprint arXiv:1801.03896*.

Benjamini, Y. and Bogomolov, M. (2014). Selective inference on multiple families of hypotheses. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 76(1):297–318.

Benjamini, Y. and Heller, R. (2007). False discovery rate for spatial signals. *Journal of the American Statistical Association*, 102(480):1272–1281.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. Roy. Statist. Soc. Ser. B*, 57(1):289–300.

Benjamini, Y. and Meir, A. (2014). Selective correlations-the conditional estimators. *arXiv preprint arXiv:1412.3242*.

Benjamini, Y., Taylor, J., and Irizarry, R. A. (2016). Selection corrected statistical inference for region detection with high-throughput assays. *bioRxiv*, page 082321.

Benjamini, Y. and Yekutieli, D. (2001). The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.*, 29(4):1165–1188.

Benjamini, Y. and Yekutieli, D. (2005). False discovery rate-adjusted multiple confidence intervals for selected parameters. *J. Amer. Statist. Assoc.*, 100(469):71–81.

Berk, R., Brown, L., Buja, A., Zhang, K., Zhao, L., et al. (2013). Valid post-selection inference. *The Annals of Statistics*, 41(2):802–837.

Bertsekas, D. P. and Tsitsiklis, J. N. (2000). Gradient convergence in gradient methods with errors. *SIAM J. Optim.*, 10(3):627–642 (electronic).

Bhattacharjee, S., Rajaraman, P., Jacobs, K., Wheeler, W., William, A., Melin, B., Hartge, P., Yeager, M., Chung, C., Chanock, S., and Chatterjee, N. a. (2012). A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits. *The American Journal of Human Genetics*, 90(5):821–835.

Bickel, P. J. and Sakov, A. (2008). On the choice of m in the m out of n bootstrap and confidence bounds for extrema. *Statist. Sinica*, 18(3):967–985.

Bogomolov, M., Peterson, C., Benjamini, Y., and Sabatti, C. (2017). Testing hypotheses on a tree: new error rates and controlling strategies. *arXiv preprint arXiv:1705.07529*.

Candes, E., Fan, Y., Janson, L., and Lv, J. (2016). Panning for gold: Model-free knockoffs for high-dimensional controlled variable selection. *arXiv preprint arXiv:1610.02351*.

Charkhi, A. and Claeskens, G. (2018). Asymptotic post-selection inference for akaikes information criterion.

Chatterjee, A. and Lahiri, S. N. (2011). Bootstrapping lasso estimators. *J. Amer. Statist. Assoc.*, 106(494):608–625.

Chen, S. and Bien, J. (2017). Valid inference corrected for outlier removal. *arXiv preprint arXiv:1711.10635*.

Chib, S. and Greenberg, E. (1995). Understanding the metropolis-hastings algorithm. *The American Statistician*, 49(4):327–335.

Consortium, U. et al. (2015). The uk10k project identifies rare variants in health and disease. *Nature*, 526(7571):82.

Cureton, E. E. (1950). Validity, reliability, and baloney. *Educational and Psychological Measurement*, 10(1):94–96.

del Castillo, J. (1994). The singly truncated normal distribution: a nonsteep exponential family. *Ann. Inst. Statist. Math.*, 46(1):57–66.

Derkach, A., Lawless, J. F., and Sun, L. (2014). Pooled association tests for rare genetic variants: a review and some new results. *Statist. Sci.*, 29(2):302–321.

Dewey, F. E., Gusarova, V., ODushlaine, C., Gottesman, O., Trejos, J., Hunt, C., Van Hout, C. V., Habegger, L., Buckler, D., Lai, K.-M. V., et al. (2016). Inactivating variants in angptl4 and risk of coronary artery disease. *New England Journal of Medicine*, 374(12):1123–1133.

Fan, J. and Lv, J. (2008). Sure independence screening for ultrahigh dimensional feature space. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 70(5):849–911.

Fithian, W., Sun, D., and Taylor, J. (2014). Optimal inference after model selection. *arXiv preprint arXiv:1410.2597*.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

Fuchsberger, C., Flannick, J., Teslovich, T. M., Mahajan, A., Agarwala, V., Gaulton, K. J., Ma, C., Fontanillas, P., Moutsianas, L., McCarthy, D. J., et al. (2016). The genetic architecture of type 2 diabetes. *Nature*, 536(7614):41–47.

Garthwaite, P. H. and Buckland, S. T. (1992). Generating Monte Carlo confidence intervals by the Robbins-Monro process. *J. Roy. Statist. Soc. Ser. C*, 41(1):159–171.

Genz, A., Bretz, F., Miwa, T., Mi, X., Leisch, F., Scheipl, F., and Hothorn, T. (2016). *mvtnorm: Multivariate Normal and t Distributions*. R package version 1.0-5.

Geweke, J. (1991). Efficient simulation from the multivariate normal and student-t distributions subject to linear constraints and the evaluation of constraint probabilities. In *Computing science and statistics: Proceedings of the 23rd symposium on the interface*, pages 571–578. Citeseer.

Griffiths, W. (2004). A gibbs' sampler for the parameters of a truncated multivariate normal distribution. *Contemporary issues in economics and econometrics: Theory and application*, pages 75–91.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical learning with sparsity: the lasso and generalizations*. CRC Press.

Hedges, L. V. and Olkin, I. (1985). *Statistical methods for meta-analysis*. Academic Press, Inc., Orlando, FL.

Heller, R., Chatterjee, N., Krieger, A., and Shi, J. (2017a). Post-selection inference following aggregate level hypothesis testing in large scale genomic data. *J. Amer. Statist. Assoc.*

Heller, R., Meir, A., and Chatterjee, N. (2017b). Post-selection estimation and testing following aggregated association tests. *arXiv preprint arXiv:1711.00497*.

Hinton, G. E. (2002). Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800.

Hocking, R. R. (1976). The analysis and selection of variables in linear regression. *Biometrics*, 32(1):1–49.

Hyun, S., G'Sell, M., and Tibshirani, R. J. (2016). Exact post-selection inference for change-point detection and other generalized lasso problems. *arXiv preprint arXiv:1606.03552*.

Iyengar, S. and Greenhouse, J. B. (1988). Selection models and the file drawer problem. *Statistical Science*, pages 109–117.

Kivaranovic, D. and Leeb, H. (2018). Expected length of post-model-selection confidence intervals conditional on polyhedral constraints. *arXiv preprint arXiv:1803.01665*.

Kotecha, J. H. and Djuric, P. M. (1999). Gibbs sampling approach for generation of truncated multivariate gaussian random variables. In *Acoustics, Speech, and Signal Processing, 1999. Proceedings., 1999 IEEE International Conference on*, volume 3, pages 1757–1760. IEEE.

Latala, R. and Matlak, D. (2017). Royen's proof of the Gaussian correlation inequality. In *Geometric aspects of functional analysis*, volume 2169 of *Lecture Notes in Math.*, pages 265–275. Springer, Cham.

Lederer, J. (2013). Trust, but verify: benefits and pitfalls of least-squares refitting in high dimensions. *arXiv preprint arXiv:1306.0113*.

Lee, J. D., Sun, D. L., Sun, Y., and Taylor, J. E. (2016). Exact post-selection inference, with application to the lasso. *Ann. Statist.*, 44(3):907–927.

Lee, J. D. and Taylor, J. E. (2014). Exact post model selection inference for marginal screening. In *Advances in Neural Information Processing Systems*, pages 136–144.

Lee, S., Wu, M. C., and Lin, X. (2012). Optimal tests for rare variant effects in sequencing association studies. *Biostatistics*, 13(4):762–775.

Leeb, H. and Pötscher, B. M. (2003). The finite-sample distribution of post-model-selection estimators and uniform versus nonuniform approximations. *Econometric Theory*, 19(1):100–142.

Leeb, H. and Pötscher, B. M. (2005). Model selection and inference: facts and fiction. *Econometric Theory*, 21(1):21–59.

Leeb, H. and Pötscher, B. M. (2006). Can one estimate the conditional distribution of post-model-selection estimators? *Ann. Statist.*, 34(5):2554–2591.

Leeb, H., Pötscher, B. M., and Ewald, K. (2015). On various confidence intervals post-model-selection. *Statist. Sci.*, 30(2):216–227.

Lei, L. and Fithian, W. (2016). Adapt: An interactive procedure for multiple testing with side information. *arXiv preprint arXiv:1609.06035*.

Lei, L., Ramdas, A., and Fithian, W. (2017). Star: A general interactive framework for fdr control under structural constraints. *arXiv preprint arXiv:1710.02776*.

Liu, K., Markovic, J., and Tibshirani, R. (2018). More powerful post-selection inference, with application to the lasso. *arXiv preprint arXiv:1801.09037*.

Loftus, J. R. and Taylor, J. E. (2015). Selective inference in regression models with groups of variables. *arXiv preprint arXiv:1511.01478*.

Madsen, B. E. and Browning, S. R. (2009). A groupwise association test for rare mutations using a weighted sum statistic. *PLoS genetics*, 5(2):e1000384.

McKeague, I. W. and Qian, M. (2015). An adaptive resampling test for detecting the presence of significant predictors. *J. Amer. Statist. Assoc.*, 110(512):1422–1433.

Meinshausen, N. (2007). Relaxed Lasso. *Comput. Statist. Data Anal.*, 52(1):374–393.

Meinshausen, N. and Yu, B. (2009). Lasso-type recovery of sparse representations for high-dimensional data. *Ann. Statist.*, 37(1):246–270.

Meir, A. and Drton, M. (2017). Tractable post-selection maximum likelihood inference for the lasso. *arXiv preprint arXiv:1705.09417*.

Mira, A. (2001). On Metropolis-Hastings algorithms with delayed rejection. *Metron*, 59(3-4):231–241 (2002).

Morris, A. P. and Zeggini, E. (2010). An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genetic epidemiology*, 34(2):188–193.

Neale, B. M., Rivas, M. A., Voight, B. F., Altshuler, D., Devlin, B., Orho-Melander, M., Kathiresan, S., Purcell, S. M., Roeder, K., and Daly, M. J. (2011). Testing for an unusual distribution of rare variants. *PLoS genetics*, 7(3):e1001322.

Pakman, A. and Paninski, L. (2014). Exact Hamiltonian Monte Carlo for truncated multivariate Gaussians. *J. Comput. Graph. Statist.*, 23(2):518–542.

Panigrahi, S., Markovic, J., and Taylor, J. (2017). An mcmc free approach to post-selective inference. *arXiv preprint arXiv:1703.06154*.

Panigrahi, S., Taylor, J., and Weinstein, A. (2016). Bayesian post-selection inference in the linear model. *arXiv preprint arXiv:1605.08824*.

Penny, W. and Friston, K. (2003). Mixtures of general linear models for functional neuroimaging. *IEEE Transactions on Medical Imaging*, 22:504–514.

Pötscher, B. M. (1991). Effects of model selection on inference. *Econometric Theory*, 7(2):163–185.

Reid, S., Taylor, J., and Tibshirani, R. (2014). Post-selection point and interval estimation of signal sizes in gaussian samples. *arXiv preprint arXiv:1405.3340*.

Reid, S., Taylor, J., and Tibshirani, R. (2016a). A general framework for estimation and inference from clusters of features. *Journal of the American Statistical Association*, (just-accepted).

Reid, S., Tibshirani, R., and Friedman, J. (2016b). A study of error variance estimation in Lasso regression. *Statist. Sinica*, 26(1):35–67.

Reiner-Benaim, A. (2007). FDR control by the BH procedure for two-sided correlated tests with implications to gene expression data analysis. *Biom. J.*, 49(1):107–126.

Romeo, S., Pennacchio, L. A., Fu, Y., Boerwinkle, E., Tybjaerg-Hansen, A., Hobbs, H. H., and Cohen, J. C. (2007). Population-based resequencing of angptl4 uncovers variations that reduce triglycerides and increase hdl. *Nature genetics*, 39(4):513.

Romeo, S., Yin, W., Kozlitina, J., Pennacchio, L. A., Boerwinkle, E., Hobbs, H. H., and Cohen, J. C. (2009). Rare loss-of-function mutations in angptl family members contribute to plasma triglyceride levels in humans. *The Journal of clinical investigation*, 119(1):70.

Routtenberg, T. and Tong, L. (2015). Estimation after parameter selection: Performance analysis and estimation methods. *arXiv preprint arXiv:1503.02045*.

Royen, T. (2014). A simple proof of the Gaussian correlation conjecture extended to some multivariate gamma distributions. *Far East J. Theor. Stat.*, 48(2):139–145.

Scheffé, H. (1999). *The analysis of variance*. Wiley Classics Library. John Wiley & Sons, Inc., New York. Reprint of the 1959 original, A Wiley Publication in Mathematical Statistics.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist.*, 6(2):461–464.

Sun, J., Zheng, Y., and Hsu, L. (2013). A unified mixed-effects model for rare-variant association in sequencing studies. *Genetic epidemiology*, 37(4):334–344.

Taylor, J. and Tibshirani, R. (2016). Post-selection inference for l1-penalized likelihood models. *Canadian Journal of Statistics*.

Tian, X., Bi, N., and Taylor, J. (2016). Magic: a general, powerful and tractable method for selective inference. *arXiv preprint arXiv:1607.02630*.

Tian, X. and Taylor, J. (2018). Selective inference with a randomized response. *Ann. Statist.*, 46(2):679–710.

Tian, X. and Taylor, J. E. (2015). Selective inference with a randomized response. *arXiv preprint arXiv:1507.06739*.

Tibshirani, R., Tibshirani, R., Taylor, J., Loftus, J., and Reid, S. (2016a). *selectiveInference: Tools for Post-Selection Inference*. R package version 1.1.3.

Tibshirani, R. J., Rinaldo, A., Tibshirani, R., and Wasserman, L. (2015). Uniform asymptotic inference and the bootstrap after model selection. *arXiv preprint arXiv:1506.06266*.

Tibshirani, R. J., Taylor, J., Lockhart, R., and Tibshirani, R. (2016b). Rejoinder: Exact post selection inference for sequential regression procedures [MR3538690; MR3538689]. *J. Amer. Statist. Assoc.*, 111(514):618–620.

Tierney, L. and Mira, A. (1999). Some adaptive monte carlo methods for bayesian inference. *Statistics in medicine*, 18(1718):2507–2515.

van der Vaart, A. W. (1998). *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

Vul, E., Harris, C., Winkielman, P., and Pashler, H. (2009). Puzzlingly high correlations in fmri studies of emotion, personality, and social cognition. *Perspectives on psychological science*, 4(3):274–290.

Weber, N. (1980). A martingale approach to central limit theorems for exchangeable random variables. *Journal of Applied Probability*, 17(3):662–673.

Weinstein, A., Fithian, W., and Benjamini, Y. (2013). Selection adjusted confidence intervals with more power to determine the sign. *J. Amer. Statist. Assoc.*, 108(501):165–176.

Wu, M. C., Lee, S., Cai, T., Li, Y., Boehnke, M., and Lin, X. (2011). Rare-variant association testing for sequencing data with the sequence kernel association test. *The American Journal of Human Genetics*, 89(1):82–93.

Yang, F., Barber, R. F., Jain, P., and Lafferty, J. (2016). Selective inference for group-sparse linear models. In *Advances in Neural Information Processing Systems*, pages 2469–2477.

Yoo, Y., Sun, L., Poirier, J., Paterson, A., and Bull, S. (2016). Multiple linear combination (mlc) regression tests for common variants adapted to linkage disequilibrium structure. *Genetic epidemiology*, 41:108–121.

Zhao, P. and Yu, B. (2006). On model selection consistency of Lasso. *J. Mach. Learn. Res.*, 7:2541–2563.