Swiss Federal Institute of Technology Zurich

Seminar for
Statistics

Department of Mathematics

Semester Paper

Spring 2024

Filip Ilic
Paul Schlossmacher

Computational advances in post-selection inference for high-dimensional data

Submission Date: May 31st 2024

Coadvisor: Christoph Schultheiss
Advisor: Prof. Dr. Peter Bühlmann

Abstract

We compare different post-selection inference methods with respect to power and family-wise error rate (FWER). While it does dominate data splitting in terms of power, data carving, as introduced by Fithian, Sun, and Taylor (2017), relies on inefficient MCMC methods for inference. Recently, Drysdale (2023) proposed a new carving estimator that alleviates the efficiency problems due to its exact parametric distribution. In this paper, we discuss potential problems of the new estimator in settings of many active variables, where it may not be well-defined. In our simulation setups, we further observe that the new carving estimator generally achieves lower power while controlling the FWER than the original carving estimator in the selected viewpoint. Notably, even the computationally comparable carving estimator in the saturated viewpoint from Fithian et al. (2017) outperforms the new carving estimator. Finally, we explore recent developments in post-selection inference and provide an overview of problems they aim to solve.

Contents

Chapter 1

Introduction

In the field of high-dimensional statistics, practitioners face the challenge of dealing with a large number of variables relative to the available observations. Although linear regression is a very powerful technique, it is not directly applicable in this setting and researchers need to first reduce the number of variables using appropriate model selection techniques. In this paper, we will focus on the Lasso, popularized by Tibshirani (1996), as one such method. The Lasso imposes an ℓ_1-penalty on the regression coefficients, which enforces sparsity in the resulting estimator and thus provides a natural way of reducing model complexity. After running the Lasso, researchers may be tempted to perform inference directly on the selected variables. However, this approach can compromise the validity of the inference, as it would require conditioning on the fact that we already "peeked" at the data during the selection process.

One straightforward technique for obtaining valid p-values in this setting is data splitting, where the dataset is randomly split into two parts: one for model selection and the other for inference. It dates back at least as far as Cox (1975). While it is very simple, Fithian et al. (2017) demonstrated that this approach yields inadmissible tests. Alternatively, using the entire dataset for both selection and inference requires adjustments for selection bias to ensure valid p-values. Tibshirani (2013) showed that Lasso selection imposes polyhedral constraints on the response variable. Lee, Sun, Sun, and Taylor (2016) proceeded to show that they can be described by a set of linear inequality constraints. Conditioning on these constraints, the signs of the estimator, and some nuisance parameters, enables the derivation of the exact distribution of the estimator of the regression coefficients, a method we refer to as pure post-selection inference (PoSI).

Data splitting and PoSI meet somewhere in the middle at the so-called "data carving" proposed by Fithian et al. (2017), which also splits the data randomly. In contrast to data splitting, this method does not discard all of the data used in the selection procedure, but reuses the remaining information for the inference stage. Hence, this method is said to "carve" the data. This can be done by using the same polyhedral constraints as for PoSI, but imposing them only on the data used for selection. Data carving outperforms both data splitting and PoSI with respect to power. However, the resulting conditional distribution is intractable. Fithian et al. (2017), and other works such as Schultheiss, Renaux, and Bühlmann (2021), resort to Markov Chain Monte Carlo (MCMC) methods to sample from it. These methods have proven to be computationally very demanding, often limiting their practicality.

Drysdale (2023) proposed a solution to improve efficiency of data carving. He defines a new carving estimator based on a combination of data splitting and PoSI, indicating that its power strictly dominates that of data splitting. We will refer to it as the combined carving estimator. Its main advantage lies in the fact that, as both data splitting and PoSI have tractable distributions, their convex combination will also have a tractable distribution, which can be used to perform valid inference after selection. The availability of an exact distribution, which happens to be simple to sample from, gives this approach a meaningful improvement in terms of efficacy. Our key contribution in this paper is that we empirically compare the combined carving estimator to the other estimators described so far in terms of power and FWER.

Another approach has been proposed by Liu (2023), who tackles the problem of slow convergence of MCMC more directly. She finds that the polyhedral constraints resulting from the Lasso provide a special framework suitable for more efficient sampling methods than regular MCMC. To take advantage of this, she resorts to the use of the randomized Lasso and also mentions its potential to boost inferential power compared to the regular Lasso.

In most of the above approaches, we split the data into two subsets before performing model selection on one subset and subsequent inference on either the other subset or both of them. Although not emphasized, this split is generated randomly. This means that if one tries to repeat the same experiment with a different split, the Lasso selection procedure may pick different variables. This instability in the selection process may yield significantly different p-values if the procedure is repeated multiple times. Meinshausen, Meier, and Bühlmann (2009) call this phenomenon the "p-value lottery" and propose splitting the data multiple times and subsequently aggregating the resulting p-values using quantile functions. This idea was adapted to the carving method by Schultheiss et al. (2021), who proposed the so-called "multicarving", which works similarly but uses the data carving method instead of the data splitting method.

This paper is structured as follows: in Chapter 2, we present the theoretical background of linear regression and model selection, data splitting, PoSI and the carving approaches from Drysdale (2023) and Fithian et al. (2017). Following this, we present our results on power simulations in Chapter 3, comparing all of the estimators introduced in the preceding chapter with respect to power and FWER control. Lastly, in Chapter 4, we outline new ideas on improving MCMC sampling proposed by Liu (2023) and discuss approaches to mitigating the p-value lottery that comes along with methods that split the data randomly and only once.

Chapter 2

Preliminaries

We begin by providing an outline of the Ordinary Least Squares (OLS) estimator and of our model selection procedure in Section 2.1. Following this, we give a brief introduction of various methods for post-selection inference. We start with data splitting (Section 2.2) and pure post-selection inference (Section 2.3), both of which are foundational for the combined estimator presented in Section 2.4. We conclude the preliminaries chapter by presenting the carving estimator from Fithian et al. (2017) (Section 2.5). Throughout all of the paper we will use bold letters to denote vectors in $\mathbb{R}^n$, regular large letters for matrices and small letters for scalars. Furthermore, we write $\mathbf{Y}$ for a random vector and $\mathbf{y}$ for its realization.

2.1 Linear regression and model selection with Lasso

We consider a vector of observations $\mathbf{y} = (y_1, \ldots, y_n)^T \in \mathbb{R}^n$ and a set of fixed predictor variables $(\mathbf{x}_i)_{i=1}^n$ where $\mathbf{x}_i \in \mathbb{R}^p$ for all $i \in \{1, \ldots, n\}$. The design matrix containing all predictor variables stacked row-wise will be denoted by $X \in \mathbb{R}^{n \times p}$, where we write for its columns $(X_1, \ldots, X_p)$. For each single observation, we assume that there is an underlying linear model of the form

$$Y_i = \mathbf{x}_i^T \boldsymbol{\beta} + \varepsilon_i,$$

where the $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ are i.i.d random variables explaining the noise with known or unknown variance σ^2 and $\boldsymbol{\beta} \in \mathbb{R}^p$ is the unknown parameter of interest. For convenience, we will sometimes also use the multivariate expression summarizing all of the above linear relationships as

$$\mathbf{Y} = X\boldsymbol{\beta} + \boldsymbol{\varepsilon},$$

where now $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_n)$. Minimizing the sum of squared residuals $||\mathbf{Y} - X\boldsymbol{\beta}||^2$, we arrive at the OLS solution

$$\hat{\boldsymbol{\beta}}^{OLS} = (X^T X)^{-1} X^T \mathbf{Y} = X^{\dagger} \mathbf{Y},$$

where $X^{\dagger}$ is the so-called Moore-Penrose Inverse. As $(X^T X)^{-1} \in \mathbb{R}^{p \times p}$ has a rank of at most n, this inverse will never be defined in the high-dimensional setting $p \gg n$. Hence, we have to reduce the number of covariates through model selection and use the assumption that the number of true active variables $M = \{j : \beta_j \neq 0\}$ is much smaller than n, i.e. $m = |M| \ll n$. In order to do this, we will focus on model selection via the Lasso popularized by Tibshirani (1996). The Lasso minimizes the residual sum of squares subject to an ℓ_1-penalty on the coefficients of $\boldsymbol{\beta}$. This forces some of the entries of $\hat{\boldsymbol{\beta}}^{Lasso}$ to be exactly zero and thus gives a natural framework for reducing the model size. We will consider the following definition of the Lasso estimator:

$$\hat{\boldsymbol{\beta}}^{Lasso} \in \arg\min_{\boldsymbol{\beta}} \frac{1}{2}||\mathbf{Y} - X\boldsymbol{\beta}||_2^2 + \lambda||\boldsymbol{\beta}||_1, \tag{2.1}$$

following the approach of Lee et al. (2016). Here, $\lambda \in \mathbb{R}_{\geq 0}$ is a parameter controlling the strength of regularization, i.e. higher values of λ will enforce more sparse solutions.

In the following, we will adopt some theory from Schultheiss et al. (2021). There, the model selected by the Lasso is then described as $\hat{M} = \{j : \hat{\beta}_j^{Lasso} \neq 0\}$ with cardinality $\hat{m} = |\hat{M}|$. We write $X_{\hat{M}} \in \mathbb{R}^{n \times \hat{m}}$ for the matrix X containing only the selected columns. Similarly, by $X_{-\hat{M}}$ we mean that all columns except for the ones in $\hat{M}$ get selected.

The condition $M \subseteq \hat{M}$ is called screening. It is of great interest to achieve screening with the selection procedure. This is because we will have no error guarantees at the inference stage if screening is not given. Throughout all of the post-selection inference methods presented in the following sections, our main goal will be to test

$$H_{0,j}: \quad \beta_j = 0 \qquad \text{versus} \qquad H_{A,j}: \quad \beta_j \neq 0 \tag{2.2}$$

for all $j \in \{1, \dots, p\}$. Under screening, we obtain unbiased tests for the above, given that we have valid tests for

$$H_{0,j}^{\hat{M}}: \quad \beta_j^{\hat{M}} = 0 \qquad \text{versus} \qquad H_{A,j}^{\hat{M}}: \quad \beta_j^{\hat{M}} \neq 0, \tag{2.3}$$

where now $\boldsymbol{\beta}^{\hat{M}} \in \mathbb{R}^{\hat{m}}$ is defined as in Schultheiss et al. (2021):

$$\boldsymbol{\beta}^{\hat{M}} \equiv \arg\min_{\mathbf{b}^{\hat{M}}} \mathbb{E} \left\| \mathbf{Y} - X_{\hat{M}} \mathbf{b}^{\hat{M}} \right\|^2 = X_{\hat{M}}^{\dagger} X \boldsymbol{\beta}. \tag{2.4}$$

Thus, $\boldsymbol{\beta}^{\hat{M}}$ is the best linear predictor in the given model. If $M \subseteq \hat{M}$, then (2.4) will find the same entries as the true $\boldsymbol{\beta}$, i.e. $\forall j \in \hat{M} : \beta_j^{\hat{M}} = \beta_j$. This is why, given screening, valid tests on the selected submodel as in (2.3) also yield unbiased tests for the full model in (2.2). In the following, we therefore focus on the "simpler" hypothesis from (2.3).

When testing this hypothesis, we seek type I error guarantees. This means that if we choose a significance level $\alpha \in [0, 1]$, we want that

$$\mathbb{P}_{H_0^{\hat{M}}}(\text{reject } H_0) \leq \alpha. \tag{2.5}$$

The index $H_0^{\hat{M}}$ in (2.5) emphasises that the probability is regarded under the assumption that $\mathbf{y}$ was generated from model $\hat{M}$ and that H_0 is indeed true. Controlling for (2.5) would be correct, if the statistician had predefined the model before looking at the data. But in the case of inference after selection, one needs to condition on the selection event and thus control for the so called selective type-I error rate

$$\mathbb{P}_{H_0^{\hat{M}}}(\text{reject } H_0 \mid (\hat{M}, H_0)) \leq \alpha. \tag{2.6}$$

If the model is misspecified, i.e. the screening condition is not fulfilled, we cannot expect to receive any type I error guarantees.

2.2 Data splitting

A simple way to perform valid inference after model selection is data splitting. As in Section 2.1, we assume the number of truly active variables m to be sufficiently small for OLS to be applicable. For this section, we will adhere to some notation from Drysdale (2023). We separate our data points randomly into two groups of size n_A and n_B respectively, i.e., $n = n_A + n_B$. Let us denote the data from group A by $(X_A, \mathbf{y}_A)$ and the one from group B by $(X_B, \mathbf{y}_B)$. We can use set A for model selection and set B for subsequent inference. As we assumed our data to be i.i.d, we know that those observations are independent and thus allow for valid inference after model selection. More precisely, let us define the Lasso estimator obtained on group A by $\hat{\boldsymbol{\beta}}_A^{Lasso}$ and its resulting selected model as $\hat{M}_A = \{j : \hat{\beta}_{A,j}^{Lasso} \neq 0\}$. If we apply this model reduction on the design matrix from group B, we obtain $X_{B,\hat{M}_A} \in \mathbb{R}^{n_B \times \hat{m}}$ which directly enables us to define

$$\hat{\boldsymbol{\beta}}^{Split} = X_{B,\hat{M}_A}^{\dagger} \mathbf{Y}_B \in \mathbb{R}^{\hat{m}}. \tag{2.7}$$

Note that $\hat{\boldsymbol{\beta}}^{Split}$ is now well-defined only when $\hat{m} = |\hat{M}_A| \leq n_B$. As this is an OLS estimator, we may find an explicit expression for the distribution of $\hat{\boldsymbol{\beta}}^{Split}$, since it inherits the Gaussian nature of $\mathbf{y}_B = X_B \boldsymbol{\beta} + \boldsymbol{\varepsilon}_B$. Hence, looking at each entry of the estimator, we arrive for all $j \in \{1, \dots, \hat{m}\}$ at

$$\hat{\beta}_j^{Split} \sim \mathcal{N}((X_{B,\hat{M}_A}^{\dagger} X_B \boldsymbol{\beta})_j,\ \sigma_B^2 (X_{B,\hat{M}_A}^T X_{B,\hat{M}_A})_{jj}^{-1}),$$

where in our homoscedastic setting $\sigma_B^2 = \sigma^2$ and $\boldsymbol{\beta}$ is the true underlying coefficient vector. To make this distribution usable for inference, we just replace $\boldsymbol{\beta}$ with our null hypothesis from (2.3) and assume screening.

Then, the distribution of the data splitting estimator simplifies under the null to

$$\hat{\beta}_j^{Split} \sim \mathcal{N}(0,\ \sigma^2 (X_{B,\hat{M}_A}^T X_{B,\hat{M}_A})_{jj}^{-1}).$$

The main advantage of having an explicit distribution at hand is that it allows for tests using standard z-scores. Because $\mathrm{Var}(\hat{\beta}_j^{Split}) = \sigma^2 (X_{B,\hat{M}_A}^T X_{B,\hat{M}_A})_{jj}^{-1}$, we can thus obtain the pivotal quantity

$$Z_j^{Split} = \frac{\hat{\beta}_j^{Split}}{\sqrt{\mathrm{Var}(\hat{\beta}_j^{Split})}} \sim \mathcal{N}(0,1).$$

From the universality of the uniform it follows that for the standard normal cumulative distribution function Φ, we have

$$\Phi(Z_j^{Split}) \sim \mathrm{Unif}(0,1).$$

Thus, for a given significance level $\alpha \in [0,1]$, we can obtain two sided p-values via

$$p_j^{Split} = 2 \min\{\Phi(Z_j^{Split}), 1 - \Phi(Z_j^{Split})\},$$

which satisfy $\mathbb{P}(p_j^{Split} \leq \alpha) \leq \alpha$ for all $1 \leq j \leq \hat{m}$ as desired. We will think of the final output of p-values from data splitting as a p-dimensional vector with ones at all positions $\{1, \dots, p\} \backslash \hat{M}_A$ and the entries of $\mathbf{p}^{Split} = (p_j^{Split})_{j=1}^{\hat{m}}$ at the corresponding positions that we selected.

While data splitting provides a simple framework for valid inference after model selection, it has certain limitations. First of all, the resulting p-values are only valid given the screening condition and screening becomes more likely the more data we have for the selection process. If we split the data such that A is large, there is a chance that we will select more variables than there are observations left in the set B. This brings us back to the problem of $X^T_{B,\hat{M}_A} X_{B,\hat{M}_A}$ being singular and hence our estimator not being well-defined. If it still works with a large selection set, the distribution of $\hat{\boldsymbol{\beta}}^{Split}$ will have increased variance due to the small size of the inference set. Drysdale (2023) warns that this will reduce the power of the tests obtained from $\hat{\boldsymbol{\beta}}^{Split}$.

Given screening, the power of data splitting would still be inferior to the methods presented in the next sections, as the type-I error controlled by data splitting is

$$\mathbb{P}_{H_0^{\hat{M}}}(\text{reject } H_0 \mid \mathbf{y}_A) \leq \alpha, \tag{2.8}$$

whereas the other methods control for the selective type-I error as defined in (2.6). This means that data splitting conditions on more than necessary, which leads to a decrease in power. For more details, we refer to Section 2.5. Note that in (2.8) we avoid conditioning on X_A, because the design matrix was assumed to be fixed.

2.3 Pure post-selection inference

Even though it may intuitively be the most straightforward approach for valid post-selection inference, splitting the data is not strictly necessary. As opposed to all the other methods we will see in this paper, "pure post-selection inference", which was introduced by Lee et al. (2016), uses all available observations for the selection step.

To further explain the approach by Lee et al. (2016), let us first go through the definitions they used: Consider a model $M \subseteq \{1, \dots, p\}$ and let $m = |M|$ and $\boldsymbol{s} \in \{-1, 1\}^m$. As seen before, models $\hat{M}$ get selected through the Lasso estimator $\hat{\boldsymbol{\beta}}^{Lasso}$ defined in (2.1) by

$$\hat{M} = \{j : \hat{\beta}_j^{Lasso} \neq 0\}.$$

Let $\hat{m} = |\hat{M}|$ and $\hat{\boldsymbol{s}} = \text{sign}(\hat{\boldsymbol{\beta}}^{Lasso}) \in \mathbb{R}^{\hat{m}}$. Lee et al. (2016) show that the event of model M being selected, that is $\{\hat{M} = M\}$, is equivalent to a union of polyhedra $A(M, \boldsymbol{s})\boldsymbol{Y} \leq \mathbf{b}(M, \boldsymbol{s})$. With $P_M = X_M(X_M^T X_M)^{-1} X_M^T$ being the projection matrix onto the column space of X_M, we have A and $\boldsymbol{b}$ depending on the model M and the vector $\boldsymbol{s}$ as follows:

$$\begin{aligned}
A_0(M, \boldsymbol{s}) &= \frac{1}{\lambda} \begin{pmatrix} X_{-M}^T (I - P_M) \\ -X_{-M}^T (I - P_M) \end{pmatrix}, \\
\boldsymbol{b}_0(M, \boldsymbol{s}) &= \begin{pmatrix} \mathbf{1} - X_{-M}^T (X_M^T)^{\dagger} \boldsymbol{s} \\ \mathbf{1} + X_{-M}^T (X_M^T)^{\dagger} \boldsymbol{s} \end{pmatrix}, \\
A_1(M, \boldsymbol{s}) &= -\operatorname{diag}(\boldsymbol{s})(X_M^T X_M)^{-1} X_M^T, \\
\boldsymbol{b}_1(M, \boldsymbol{s}) &= -\lambda \operatorname{diag}(\boldsymbol{s})(X_M^T X_M)^{-1} \boldsymbol{s}, \\
A(M, \boldsymbol{s}) &= \begin{pmatrix} A_0(M, \boldsymbol{s}) \\ A_1(M, \boldsymbol{s}) \end{pmatrix} \in \mathbb{R}^{(2p-m) \times n}, \\
\boldsymbol{b}(M, \boldsymbol{s}) &= \begin{pmatrix} \boldsymbol{b}_{\mathbf{0}}(M, \boldsymbol{s}) \\ \boldsymbol{b}_{\mathbf{1}}(M, \boldsymbol{s}) \end{pmatrix} \in \mathbb{R}^{(2p-m)}.
\end{aligned}$$

Now, Theorem 4.3 of Lee et al. (2016) states that $\{\hat{M} = M, \hat{\boldsymbol{s}} = \boldsymbol{s}\} = \{A(M, \boldsymbol{s})\boldsymbol{Y} \leq \boldsymbol{b}(M, \boldsymbol{s})\}$. As outlined further, one obtains the event $\{\hat{M} = M\}$ by taking the union over the different sign patterns:

$$\{\hat{M} = M\} = \bigcup_{\boldsymbol{s} \in \{-1,1\}^m} \{A(M, \boldsymbol{s})\boldsymbol{Y} \leq \boldsymbol{b}(M, \boldsymbol{s})\}. \tag{2.9}$$

Furthermore, they point out that for a given vector $\boldsymbol{\eta} \in \mathbb{R}^n$, the random variable $\boldsymbol{\eta}^T \boldsymbol{Y}$ conditioned by the selection event $\{A\boldsymbol{Y} \leq \boldsymbol{b}\}$ has a truncated normal distribution. We can calculate the truncation limits as follows: Let $\Sigma = \mathrm{Cov}(\boldsymbol{Y})$. Then, we set

$$\begin{aligned} \boldsymbol{c} &= \Sigma\boldsymbol{\eta}(\boldsymbol{\eta}^{\boldsymbol{T}}\Sigma\boldsymbol{\eta})^{-1} \quad \text{and} \\ \boldsymbol{z} &= (I_n - \boldsymbol{c}\boldsymbol{\eta}^{\boldsymbol{T}})\boldsymbol{y}. \end{aligned}$$

With A and $\boldsymbol{b}$ as above, we can define

$$\mathcal{V}^-(\boldsymbol{z}) = \max_{j:(A\boldsymbol{c})_j < 0} \frac{b_j - (A\boldsymbol{z})_j}{(A\boldsymbol{c})_j} \quad \text{and} \tag{2.10}$$

$$\mathcal{V}^+(\boldsymbol{z}) = \min_{j:(A\boldsymbol{c})_j > 0} \frac{b_j - (A\boldsymbol{z})_j}{(A\boldsymbol{c})_j}. \tag{2.11}$$

However, notice that in the case of Σ having the homoscedastic structure $\Sigma = \sigma^2 I_n$ and $\boldsymbol{\eta} = X_M(X_M^T X_M)^{-1} \in \mathbb{R}^{n \times m}$, it follows that $A_0\boldsymbol{c} = \boldsymbol{0}$, because $(I - P_M)\boldsymbol{\eta} = \boldsymbol{0}$. Thus, the rows of A_0 in $A\boldsymbol{z}$ also do not get considered in the max and min of (2.10) and (2.11) respectively. In our simulations, we therefore set $A = A_1$, to avoid numerical issues with values close to 0.

To calculate the distribution of the estimator $\hat{\boldsymbol{\beta}}^{PoSI}$, we shrink the union of polyhedra from (2.9) to a single polyhedron, by additionally conditioning on the sign pattern $\hat{\boldsymbol{s}}$ obtained through the Lasso. As described in Lee et al. (2016), this conditioning happens simply by using the truncation limits $\mathcal{V}^-_{\hat{\boldsymbol{s}}}(\boldsymbol{z})$ and $\mathcal{V}^+_{\hat{\boldsymbol{s}}}(\boldsymbol{z})$ of this one sign pattern $\hat{\boldsymbol{s}}$. Note that $\mathcal{V}^-_{\hat{\boldsymbol{s}}}(\boldsymbol{z})$ and $\mathcal{V}^+_{\hat{\boldsymbol{s}}}(\boldsymbol{z})$ depend on $\hat{\boldsymbol{s}}$, because A and $\boldsymbol{b}$ do. By conditioning on $\hat{\boldsymbol{s}}$, our procedure is aligned with the one used by Schultheiss et al. (2021). There it is also stated that, while this leads to a small loss in power, it is easier computationally.

Thus, we can say

$$\forall \boldsymbol{\eta} \in \mathbb{R}^n : \boldsymbol{\eta}^{\boldsymbol{T}}\boldsymbol{Y} | \{A\boldsymbol{Y} \leq \boldsymbol{b} \wedge \boldsymbol{Z} = \boldsymbol{z}\} \sim \mathcal{TN}(\boldsymbol{\eta}^{\boldsymbol{T}} X\boldsymbol{\beta}, \sigma^2 \boldsymbol{\eta}^{\boldsymbol{T}}\Sigma\boldsymbol{\eta}, \mathcal{V}^-(\boldsymbol{z}), \mathcal{V}^+(\boldsymbol{z})).$$

If we then define the PoSI estimator as $\hat{\boldsymbol{\beta}}^{PoSI} = X_{\hat{M}}^{\dagger}\boldsymbol{Y}$, and assume $\boldsymbol{Y} \sim \mathcal{N}(X\boldsymbol{\beta}, \sigma^2 I_n)$, we have

$$\hat{\boldsymbol{\beta}}^{PoSI} | \{A\boldsymbol{Y} \leq \boldsymbol{b} \wedge \boldsymbol{Z} = \boldsymbol{z}\} \sim \mathcal{TN}(X_{\hat{M}}^{\dagger} X\boldsymbol{\beta}, \sigma^2 (X_{\hat{M}}^T X_{\hat{M}})^{-1}, \mathcal{V}^-(\boldsymbol{z}), \mathcal{V}^+(\boldsymbol{z})). \tag{2.12}$$

For the sake of brevity, we will leave out the conditioning events when talking about the distribution of $\hat{\boldsymbol{\beta}}^{PoSI}$ in the following. Also, we will write $\mathcal{V}^-(\boldsymbol{y})$ instead of $\mathcal{V}^-(\boldsymbol{z})$, since $\boldsymbol{y}$ is the only randomly generated input of $\boldsymbol{z}$ and we thus do not have to keep the definition of the auxiliary term $\boldsymbol{z}$ in mind at all times. We can therefore simplify (2.12) to

$$\hat{\boldsymbol{\beta}}^{PoSI} \sim \mathcal{TN}(X_{\hat{M}}^{\dagger} X\boldsymbol{\beta}, \sigma^2 (X_{\hat{M}}^T X_{\hat{M}})^{-1}, \mathcal{V}^-(\boldsymbol{y}), \mathcal{V}^+(\boldsymbol{y})).$$

2.4 Combined carving

Now that we have addressed both data splitting and PoSI estimators, we will explore the carving estimator proposed by Drysdale (2023), which is a convex combination between $\hat{\boldsymbol{\beta}}^{Split}$ and $\hat{\boldsymbol{\beta}}^{PoSI}$. We refer to it as the combined carving estimator, to distinguish it from the carving estimator in Fithian et al. (2017), which will be presented in Section 2.5. Returning to the setup described in Section 2.2, we divide our data into two groups A and B. Model selection is performed on $(X_A, \mathbf{y}_A)$ to obtain the indices $\hat{M}_A$. As the splitting estimator loses power if the inference set B is too small, Drysdale (2023) suggests using the PoSI estimator on the set A to reuse left over information from the selection event. He defines his estimator as

$$\hat{\boldsymbol{\beta}}^{Comb} = \frac{n_A}{n}\hat{\boldsymbol{\beta}}^{PoSI} + \frac{n_B}{n}\hat{\boldsymbol{\beta}}^{Split}. \tag{2.13}$$

Note that since $\hat{\boldsymbol{\beta}}^{Comb}$ incorporates $\hat{\boldsymbol{\beta}}^{Split}$, it is still constrained by the size of the selection set $\hat{M}_A$ and n_B, meaning that the composed carving estimator is only well-defined when $X^T_{B,\hat{M}_A}X_{B,\hat{M}_A}$ is invertible. Using a rank argument, it is necessary that $|\hat{M}_A| \leq n_B$, as we already highlighted in Section 2.1. We have seen that pure post-selection inference and data splitting offer explicit distributions of their estimators, which are the truncated normal and regular normal distribution respectively. Therefore, we can deduce that the distribution of the composed carving estimator will be a sum of a normal and truncated normal distribution (SNTN). To characterize the SNTN distribution, we borrow Lemma 3.1 from *Drysdale* (2023):

Lemma 2.4.1 (SNTN distribution)**.** *Define $X_1 \sim \mathcal{N}(\mu_1, \tau_1^2)$ and $X_2 \sim TN(\mu_2, \tau_2^2, a, b)$, then $Z = c_1X_1 + c_2X_2$, $c_i \in \mathbb{R}$, is said to follow an SNTN distribution denoted as either $\mathcal{SNTN}(\mu_1, \tau_1^2, \mu_2, \tau_2^2, a, b, c_1, c_2) \overset{d}{=} \mathcal{SNTN}(\theta_1, \sigma_1^2, \theta_2, \sigma_2^2, \omega, \delta)$ where $\theta_1 = c_1\mu_1 + c_2\mu_2$, $\sigma_1^2 = c_1^2\tau_1^2 + c_2^2\tau_2^2$, $\theta_2 = \mu_2$, $\sigma_2^2 = \tau_2^2$, $\rho = c_2\sigma_2/\sigma_1$, $\lambda = \rho/\sqrt{1-\rho^2}$, $\gamma = \lambda/\rho$, $m_j(x) = (x-\theta_j)/\sigma_j$ for $j \in \{1,2\}$, $\omega = m_2(a)$ and $\delta = m_2(b)$. The cumulative distribution function F of the SNTN distribution can then be characterized as follows:*

$$F^{\omega,\delta}_{\theta,\sigma^2} = \frac{B_\rho(m_1(z),\delta) - B_\rho(m_1(z),\omega)}{\Phi(\delta) - \Phi(\omega)},$$

where

$$B_\rho(x_1, x_2) = \mathcal{N}\left(\begin{pmatrix}0\\0\end{pmatrix}, \begin{pmatrix}1 & \rho\\ \rho & 1\end{pmatrix}\right)$$

is the CDF of a standard bivariate normal with correlation ρ.

Now that we have characterized the SNTN distribution, we can directly obtain the distribution of the composed carving estimator. Assuming $\mathbf{Y} \sim \mathcal{N}(X\boldsymbol{\beta}, \sigma^2 I_n)$, recall that under the null we have for all $1 \leq j \leq \hat{m}$:

$$\begin{aligned}
\hat{\beta}_j^{Split} &\sim \mathcal{N}(\mu_1, \tau_1^2) = \mathcal{N}(0,\ \sigma^2(X^T_{B,\hat{M}_A}X_{B,\hat{M}_A})^{-1}_{jj}),\\
\hat{\beta}_j^{PoSI} &\sim \mathcal{TN}(\mu_2, \tau_2^2, a, b)\\
&= \mathcal{TN}(0,\ \sigma^2(X^T_{A,\hat{M}_A}X_{A,\hat{M}_A})^{-1}_{jj}, \mathcal{V}^-(\mathbf{y}_A)_j, \mathcal{V}^+(\mathbf{y}_A)_j).
\end{aligned}$$

Following the notation from Lemma 2.4.1 we can then write

$$\begin{aligned}
\hat{\beta}_j^{Comb} = c_1\hat{\beta}_j^{Split} + c_2\hat{\beta}_j^{PoSI} &\sim \mathcal{SNTN}(0, \tau_1^2, 0, \tau_2^2, \mathcal{V}^-(\mathbf{y}_A)_j, \mathcal{V}^+(\mathbf{y}_A)_j, c_1, c_2)\\
&\overset{d}{=} \mathcal{SNTN}(0, \sigma_1^2, 0, \sigma_2^2, \omega, \delta),
\end{aligned}$$

where $c_1 = n_B/n$, $c_2 = n_A/n$ and

$$\begin{aligned}
\sigma_1^2 &= c_1^2\tau_1^2 + c_2^2\tau_2^2 = \frac{n_B^2}{n^2}\sigma^2(X^T_{B,\hat{M}_A}X_{B,\hat{M}_A})^{-1}_{jj} + \frac{n_A^2}{n^2}\sigma^2(X^T_{A,\hat{M}_A}X_{A,\hat{M}_A})^{-1}_{jj},\\
\sigma_2^2 &= \tau_2^2 = \sigma^2(X^T_{A,\hat{M}_A}X_{A,\hat{M}_A})^{-1}_{jj}),\\
\omega &= \frac{a-\theta_2}{\sigma_2} = \frac{\mathcal{V}^-(\mathbf{y}_A)_j - 0}{\sigma(X^T_{A,\hat{M}_A}X_{A,\hat{M}_A})^{-1/2}_{jj}},\\
\delta &= \frac{b-\theta_2}{\sigma_2} = \frac{\mathcal{V}^+(\mathbf{y}_A)_j - 0}{\sigma(X^T_{A,\hat{M}_A}X_{A,\hat{M}_A})^{-1/2}_{jj}},\\
\rho &= \frac{c_2\sigma_2}{\sigma_1} = \frac{c_2\sigma(X^T_{A,\hat{M}_A}X_{A,\hat{M}_A})^{-1/2}_{jj}}{\left[c_1^2\sigma^2(X^T_{B,\hat{M}_A}X_{B,\hat{M}_A})^{-1}_{jj} + c_2^2\sigma^2(X^T_{A,\hat{M}_A}X_{A,\hat{M}_A})^{-1}_{jj}\right]^{1/2}}\\
&= \frac{n_A\cdot(X^T_{A,\hat{M}_A}X_{A,\hat{M}_A})^{-1/2}_{jj}}{\left[n_B^2(X^T_{B,\hat{M}_A}X_{B,\hat{M}_A})^{-1}_{jj} + n_A^2(X^T_{A,\hat{M}_A}X_{A,\hat{M}_A})^{-1}_{jj}\right]^{1/2}}.
\end{aligned}$$

Drysdale (2023) demonstrates in his paper that we can obtain exact p-values from the composed carving estimator, which is done in a similar manner as we presented for data splitting in Section 2.2, by just using the fact that under the null hypothesis:

$$F^{\omega,\delta}_{0,\sigma^2}(\hat{\beta}^{Comb}_j) \sim \text{Unif}(0,1), \quad \forall\, 1 \leq j \leq \hat{m}. \tag{2.14}$$

Furthermore, their computation is very efficient, as the most effort in the computation of the SNTN distribution is spent on the calculation of the two bivariate normal distributions, which can be done efficiently. Drysdale (2023) presents several approaches for estimating the bivariate normal distribution, but for our simulations, we will stick to the `mvtnorm` R-package by Hothorn, Genz, Bretz, Miwa, Mi, Leisch, Scheipl, Bornkamp, and Maechler (2023). We expect the composed carving estimator to strictly dominate the regular data splitting estimator in terms of power. This will be discussed further in Chapter 3.

2.5 Data carving

We have just seen the composed carving estimator, for which we can obtain an explicit distribution and even compute it efficiently. In the timeline of events, Fithian et al. (2017) were the ones to coin the term ”carving” for post selection inference. Again, like in Section 2.4, the data is split into groups A and B, where A is used for selection and all of the data is used for inference. This is highlighted by our usage of $\hat{M}_A$ for the set of selected indices.

The whole idea is based on the observation that the selective type-I error in (2.6) allows for more powerful tests, as less conditioning is done than in (2.8). This can be intuitively explained as follows: the selection event $(\hat{M}_A, H_0)$ just defines a region where $\mathbf{y}_A$ has to lie, such that it yields the same selection event. As we saw in Section 2.3, these regions can be characterized as a union of polyhedra. On the other hand, (2.8) sets one single $\mathbf{y}_A$ into stone, which is a much stronger conditioning event. Fithian et al. (2017) mention that by conditioning on a random variable, we disqualify it as evidence against the null hypothesis. Thus, we would like to condition on as little as possible in the inference stage.

In order to create tests that control for the selective type I error in (2.6), we need to understand the distribution of $\mathbf{Y}|M(\mathbf{Y}_A)$, where we define $M(\mathbf{Y}_A) = \{(\hat{M}_A, H_0) \text{ selected}\}$ as it is done in Schultheiss et al. (2021).

Analogously to Section 2.2, we consider the Lasso selection procedure with the estimator $\hat{\boldsymbol{\beta}}_A^{Lasso}$. For our purposes, we will again use the same simplified scenario, where we condition on the sign pattern $\hat{\mathbf{s}}$ of $\hat{\boldsymbol{\beta}}_A^{Lasso}$, such that the selection event may be fully characterized by a single polyhedron, and thus a single set of linear inequality constraints $A\mathbf{Y} \leq \mathbf{b}$. We then may write $\mathbf{Y}|M(\mathbf{Y}_A) = \mathbf{Y}|A\mathbf{Y} \leq \mathbf{b}$. Note that here the A in the subscript of $\mathbf{Y}$ denotes the selection set A, while the A in the linear inequality is the matrix defined in Section 2.3.

2.5.1 Selected model

To continue, we need to set our assumptions straight. We first present the so-called "selected model" framework, which assumes $\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 I_n)$ and $\mathbb{E}[\mathbf{Y}] = \boldsymbol{\mu} = X_{\hat{M}_A}\boldsymbol{\beta}^{\hat{M}_A}$. Given the screening condition, Schultheiss et al. (2021) elaborate that both the selected model and the saturated model, which we will introduce later, are valid to test (2.2). The selected model is however generally more powerful. Furthermore, the existence of $\boldsymbol{\beta}^{\hat{M}_A}$ such that $\mathbb{E}[\mathbf{Y}] = X_{\hat{M}_A}\boldsymbol{\beta}^{\hat{M}_A}$ is exactly the screening condition.

Fithian et al. (2017) show that for models that correspond to an exponential family, i.e.

$$\mathbf{Y} \sim f_{\boldsymbol{\theta}}(\mathbf{y}) = \exp(\boldsymbol{\theta}^T T(\mathbf{y}) - \psi(\boldsymbol{\theta}))f_0(\mathbf{y}),$$

their conditional distribution given $\mathbf{Y} \in Z$ for any measurable Z will still be an exponential family with the same natural parameters $\boldsymbol{\theta}$ and sufficient statistics $T(\mathbf{y})$, but different carrier measure $f_0(\mathbf{y})$ and normalizing constant $\psi(\boldsymbol{\theta})$. This means that we can get rid of nuisance parameters by conditioning on them. Formally, Fithian et al. (2017) present the following model which we introduce as a Lemma:

Lemma 2.5.1 (Exponential family with nuisance parameters)**.** *Assume* $\mathbf{Y}$ *follows a* p*-parameter exponential family with sufficient statistics* $T(\mathbf{y})$ *and* $U(\mathbf{y})$*, of dimension* k *and* $p-k$ *respectively:*

$$\mathbf{Y} \sim\ f_{\boldsymbol{\theta},\boldsymbol{\zeta}}(\mathbf{y}) = \exp(\boldsymbol{\theta}^T T(\mathbf{y}) + \boldsymbol{\zeta}^T U(\mathbf{y}) - \psi(\boldsymbol{\theta},\boldsymbol{\zeta}))f_0(\mathbf{y}),$$

with $(\boldsymbol{\theta},\boldsymbol{\zeta}) \in \Theta \subseteq \mathbb{R}^p$ *open. If* $\boldsymbol{\theta}$ *corresponds to a parameter of interest and* $\boldsymbol{\zeta}$ *to an unknown nuisance parameter, we may eliminate* $\boldsymbol{\zeta}$ *from the problem by conditioning on* U*. For* $k = 1$*, we obtain a single-parameter family for* T*.*

In the setting of linear regression in a selected submodel $\hat{M}$, we have

$$\mathbf{Y} \sim\ \mathcal{N}(X_{\hat{M}}\boldsymbol{\beta}^{\hat{M}}, \sigma^2 I_n)$$

and hence we can derive the following exponential family:

$$\mathbb{P}_{\hat{M}}(\mathbf{Y} = \mathbf{y}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\|\mathbf{y} - X_{\hat{M}}\boldsymbol{\beta}^{\hat{M}}\|^2}{2\sigma^2}\right) \tag{2.15}$$

$$= \exp\left(\frac{(\boldsymbol{\beta}^{\hat{M}})^T X_{\hat{M}}^T \mathbf{y}}{\sigma^2} - \frac{\|\mathbf{y}\|^2}{2\sigma^2} - \psi(X_{\hat{M}}\boldsymbol{\beta}^{\hat{M}}, \sigma^2)\right) \tag{2.16}$$

where $\psi(X_{\hat{M}}\boldsymbol{\beta}^{\hat{M}}, \sigma^2) = \frac{\|X_{\hat{M}}\boldsymbol{\beta}^{\hat{M}}\|^2}{2\sigma^2} + \log(\sqrt{2\pi\sigma^2})$. Taking $\beta_j^{\hat{M}}$ for some $j \in \hat{M}$ as our parameter of interest and assuming that σ^2 is known, we may read off in line (2.16) that

$(X_j)^T\mathbf{Y}$ serves as a sufficient statistic. Hence, we can apply Lemma 2.5.1 on the sufficient statistics $(X_j)^T\mathbf{Y}$ and $(X_{\hat{M}\setminus j})^T\mathbf{Y}$ in place of $T(\mathbf{Y})$ and $U(\mathbf{Y})$ respectively. In conclusion, inference on $\beta_j^{\hat{M}}$ can be performed by conditioning on the realization of $(X_{\hat{M}\setminus j})^T\mathbf{Y}$, the assumed null hypothesis and the inequality constraints given by the selection event.

Schultheiss et al. (2021) summarize all of this with the following definition of p-values for the carving estimator. Note that this structure allows for one-sided tests:

$$p_j(\mathbf{y}) = \begin{cases} \mathbb{P}\left[(X^\dagger_{\hat{M}_A})_j\mathbf{Y} \geq (X^\dagger_{\hat{M}_A})_j\mathbf{y} \;\middle|\; \beta_j^{\hat{M}_A} = 0, (X_{\hat{M}_A\setminus j})^T\mathbf{Y} = (X_{\hat{M}_A\setminus j})^T\mathbf{y}, \right. \\ \qquad\qquad\qquad\qquad \left. A\mathbf{Y} \leq \mathbf{b} \right], \text{ if } \hat{\beta}_j^{Lasso} > 0 \\ \mathbb{P}\left[(X^\dagger_{\hat{M}_A})_j\mathbf{Y} \leq (X^\dagger_{\hat{M}_A})_j\mathbf{y} \;\middle|\; \beta_j^{\hat{M}_A} = 0, (X_{\hat{M}_A\setminus j})^T\mathbf{Y} = (X_{\hat{M}_A\setminus j})^T\mathbf{y}, \right. \\ \qquad\qquad\qquad\qquad \left. A\mathbf{Y} \leq \mathbf{b} \right], \text{ if } \hat{\beta}_j^{Lasso} < 0. \end{cases}$$

Unfortunately, this distribution is a degenerate truncated multivariate Gaussian, which is not easily tractable. Thus, Fithian et al. (2017) resort to MCMC methods for an approximation of this distribution. For more details on sampling from a linearly constrained Gaussian, we refer to the Appendix from Schultheiss et al. (2021), which shines some light on the algorithm used by Fithian et al. (2017).

2.5.2 Saturated model

As for the selected model in Subsection 2.5.1, we again lean on Fithian et al. (2017) for the theory behind the saturated model. They themselves, however, already acknowledge Berk, Brown, Buja, Zhang, and Zhao (2013) for introducing an approach in the same spirit. The significant advantage offered by both the so-called saturated model by Fithian et al. (2017) and also the approach by Berk et al. (2013) is that in both cases, inference is valid regardless of whether the correct model was chosen. Therefore, no screening assumption is necessary. In fact, the same also holds for the PoSI approach by Lee et al. (2016) that we saw in Section 2.3.

For the saturated model our only assumption is that $\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\mu}, \sigma^2 I_n)$. Compared to the selected model from Subsection 2.5.1, we therefore do not assume here that there exists a $\boldsymbol{\beta}^{\hat{M}_A} \in \mathbb{R}^{\hat{m}}$ such that $\mathbb{E}[\mathbf{Y}] = \boldsymbol{\mu} = X_{\hat{M}_A}\boldsymbol{\beta}^{\hat{M}_A}$. However, having selected a model $\hat{M}_A$ and assuming that $X_{\hat{M}_A}$ has full rank, we can of course still construct the least squares estimator

$$\hat{\boldsymbol{\beta}}^{Sat} = X^\dagger_{\hat{M}_A}\mathbf{Y}.$$

When σ^2 is known, we can still reasonably use $\hat{\boldsymbol{\beta}}^{Sat}$ for inference testing. The choice between the saturated and the selected model is then concretely illustrated by Fithian et al. (2017): Let $P_{\hat{M}_A}$ be the projection onto the column space of $X_{\hat{M}_A}$ and $P^\perp_{\hat{M}_A} = (I - P_{\hat{M}_A})$. The existence of a linear relationship such that $\boldsymbol{\mu} = X_{\hat{M}_A}\boldsymbol{\beta}^{\hat{M}_A}$ is then equivalent to $P^\perp_{\hat{M}_A}\boldsymbol{\mu} = 0$, since

$$\begin{aligned} P^\perp_{\hat{M}_A}\boldsymbol{\mu} &= (I - X_{\hat{M}_A}(X^T_{\hat{M}_A}X_{\hat{M}_A})^{-1}X^T_{\hat{M}_A})X_{\hat{M}_A}\boldsymbol{\beta}^{\hat{M}_A} \\ &= X_{\hat{M}_A}\boldsymbol{\beta}^{\hat{M}_A} - X_{\hat{M}_A}\boldsymbol{\beta}^{\hat{M}_A} \\ &= \mathbf{0}. \end{aligned}$$

Hence, for the selected model from Subsection 2.5.1, we only have to condition on

$$(X_{\hat{M}_A \setminus j})^T \mathbf{Y} = (X_{\hat{M}_A \setminus j})^T \mathbf{y}$$

when doing inference, because we assume that $P^{\perp}_{\hat{M}_A} \boldsymbol{\mu} = 0$. For the saturated model, on the other hand, we need to condition on both of the following:

$$(X_{\hat{M}_A \setminus j})^T \mathbf{Y} = (X_{\hat{M}_A \setminus j})^T \mathbf{y} \quad \text{and}$$
$$P^{\perp}_{\hat{M}_A} \mathbf{Y} = P^{\perp}_{\hat{M}_A} \mathbf{y}.$$

As Fithian et al. (2017) point out, this additional conditioning tends to yield lower power compared to the selected model viewpoint, but it is computationally easier because the distribution is known. Moreover, the saturated viewpoint allows for valid tests of the null hypothesis (2.3) as noted by Schultheiss et al. (2021).

Since the reader is now familiar with the carving approach in both the selected and saturated viewpoints, and as we also saw the PoSI approach in Section 2.3, it is worthwhile to acknowledge the connection between these methods: When performing variable selection on all n data points to obtain the model $\hat{M}_A$ and also conditioning on the signs of $\hat{\boldsymbol{\beta}}^{Lasso}$, the three estimators mentioned above are the same (Fithian et al. (2017)). In this context, $\hat{\boldsymbol{\beta}}^{PoSI}$ - which by default uses all the data for selection - can be considered as a limiting case of both $\hat{\boldsymbol{\beta}}^{Carve}$ and $\hat{\boldsymbol{\beta}}^{Sat}$.

Chapter 3

Results

In this chapter, we first validate the correctness of our code for the combined carving estimator by looking at the empirical distribution of its p-values under screening. Then, we present our simulation results regarding the power and FWER of the estimators introduced in Chapter 2.

For all simulations, we use $n = 100$, $p = 200$ and stick to a design matrix X that is sampled from a multivariate Gaussian with mean zero and Toeplitz covariance matrix Σ, i.e. $\Sigma_{i,j} = \rho^{|i-j|}$ with $\rho = 0.6$. This setup is consistent with the simulations from Schultheiss et al. (2021).

For comparability between different simulation settings, we adopt the concept of signal to noise ratio (SNR), defined as

$$\text{SNR} = \frac{\widehat{\text{Var}}(X\boldsymbol{\beta})}{\sigma^2},$$

where the numerator is the sample variance of the true response variable $\boldsymbol{y}_{true} = X\boldsymbol{\beta}$, and σ^2 is the true underlying variance of the noise. As we fix X in our setting, changes in the SNR can only stem from $\boldsymbol{\beta}$ or the variance of the noise. For each simulation run, we sample new noise $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, I_n)$ and compute $\boldsymbol{y} = X\boldsymbol{\beta} + \sigma\boldsymbol{\varepsilon}$.

All the code for our simulations is written in the R language (R Core Team (2021)). Together with the plots and environments, it can be found in our GitHub repository at https://github.com/Villamaravilla07/power-sim-carving.

3.1 Importance of screening for the combined carving estimator

Existing code for the combined carving estimator was entirely written in the programming language python (Van Rossum and Drake Jr (1995)) by Drysdale (2023). To integrate it into our simulation setup, we rewrote it in the R language. To validate the correctness of our rewritten code, we used that according to the universality of the uniform, given successful screening and under the null hypothesis, the obtained p-values should be uniformly distributed. This means that the empirical cumulative distribution function (ECDF) $F_n : [0, 1] \to [0, 1]$ should under this assumption satisfy $F_n(x) = x$, and thus be a perfect diagonal on the square $[0, 1]^2$. In Figure 3.1, the left plot shows that under successful screening, the p-values follow the predicted uniform distribution. This suggests that our code provides valid type-I error guarantees when screening is given. In contrast,

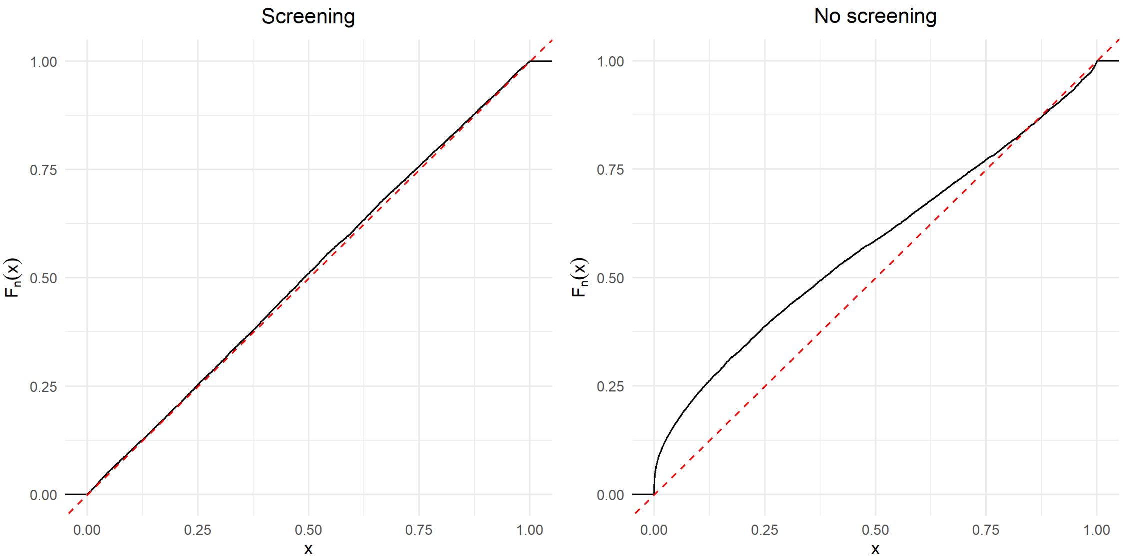

Figure 3.1: *Empirical p-value distribution with* 10000 *p-value samples obtained from the combined carving estimator. On the left, screening is satisfied, whereas on the right it is not. We used the following parameters for both plots:* $n = 100, p = 200, m = 5, SNR = 2$ *and* 70% *of the data used for selection.*

the right plot in Figure 3.1 demonstrates how drastically uniformity of p-values is violated when screening is not fulfilled.

3.2 Power simulations

In this subsection, we compare various estimators for post-selection inference in terms of power. Our simulation setup closely follows the one from Schultheiss et al. (2021), as we partially reuse code from their GitHub repository.[1] Specifically, we use their selection procedure, which performs cross-validated Lasso with the regularization parameter λ_{1se} and simultaneously checks for the validity of the constraints $\boldsymbol{A}\boldsymbol{y} \leq \boldsymbol{b}$. Additionally, the carving estimator in the selected and saturated viewpoint is computed using their code. We refrain from checking successful screening, even though this means that from a theoretical perspective, we wont have any type-I error guarantees.

Given the context of multiple hypothesis testing, we consider the FWER in our simulations. The FWER is defined as the probability of obtaining at least one type-I error among the family of hypotheses being tested. Instead of considering (2.2), we are now interested in

$$\mathbb{P}_{H_0}(\exists j \in \{1, \ldots, p\} \ : \ H_{0,j} \text{ rejected}) \leq \alpha.$$

We will fix the significance level to $\alpha = 0.05$. To control the FWER in our simulations, we apply a Bonferroni correction to the obtained p-values, modifying all p-values to

$$\tilde{p}_j = \min\{1, p_j \hat{m}\},$$

where $\hat{m}$ denotes the size of the selected model $\hat{M}$.

[1] The GitHub repository can be found at https://github.com/cschultheiss/Multicarving

We consider two different configurations of the true $\boldsymbol{\beta}$ for our main power comparison: a 5-sparse and a 15-sparse setup, with true active coefficients set as follows:

$$\beta_j = 1, \text{ for } j \in \{1, 5, 10, 15, 20\} \text{ and}$$
$$\beta_j^{'} = 1, \text{ for } j \in \{1\} \cup \{5i\}_{i=1}^{14}.$$

Furthermore, we vary the fractions f of the data that are used for the selection. For each f in $\{0.5 + 0.05i\}_{i=0}^{9} \cup \{0.99\}$, we run our simulation of power and FWER for different estimators.

3.2.1 Power simulations for varying SNR

In this subsection, we present our results for power simulations that evaluate several estimators: carving in both the selected and saturated viewpoints, combined carving, data splitting, and the PoSI method.

Figure 3.2 displays two power simulations with different signal to noise ratios. It is evident that the power of the estimators is higher in the right plot, because there is less variance in the noise as a consequence of higher SNR. Interestingly, this scenario exhibits higher FWER values for lower fractions than in the left plot, but towards higher fractions it obtains a better FWER control.

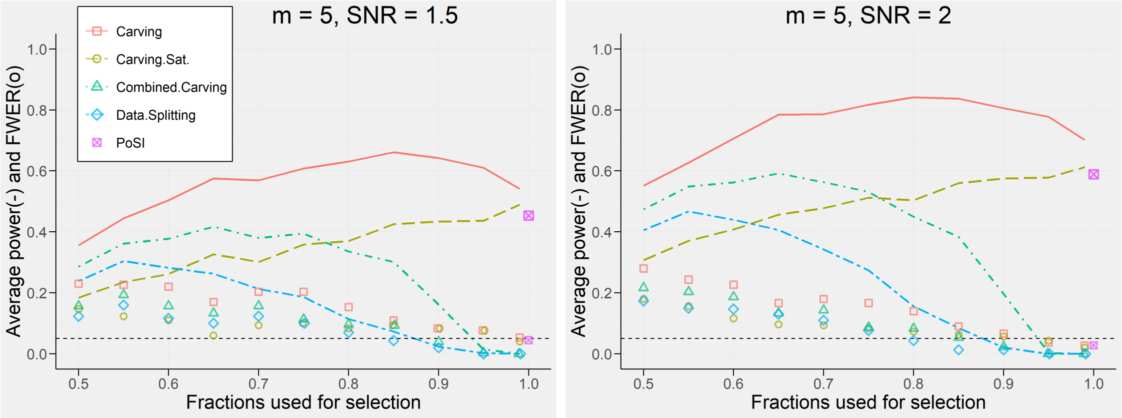

Figure 3.2: *Results from the power simulation. On the left we have SNR* $= 1.5$ *and on the right SNR* $= 2$*. In both plots the number of truly active variables is* $m = 5$*. On the x-axis we have the fractions* f *of the data used for selection. On the y-axis we have the power and FWER displayed as lines and symbols respectively. For* $f = 1$*, we used the "PoSI" estimator which has its power(upper) and FWER(lower) represented by the same symbol. At the significance level* $\alpha = 0.05$ *there is a horizontal dashed line.*

For an SNR of 1.5, the combined carving estimator controls the FWER starting from fraction 0.9. However, at this much data for selection, there are often insufficient observations left for inference. Hence, its p-values often get all set to 1, which is most certainly the cause for the decrease in power and the observed reduction in the FWER. The highest power achieved by this estimator while maintaining FWER control is 0.16 at fraction 0.9. Notably, the pure post-selection inference method computed with $f = 1$ outperforms the combined carving estimator, which itself uses the PoSI estimator as a component, though at lower fractions for selection. The PoSI estimator achieves a power of 0.45 with an

FWER of 0.04. Although the carving estimator in the selected viewpoint achieves the highest power, it never provides valid error control for the tested fractions at SNR $= 1.5$. Hence, in this setting, the estimator with highest power that provides valid FWER control is the carving estimator in the saturated viewpoint, achieving a power of 0.49 at fraction 0.99 with an FWER of 0.04. This suggests that for lower signal to noise ratios, it is sensible to use all or nearly all of the data and thus increase the chances of screening and, consequently, valid FWER control.

Moving to the right plot with SNR $= 2$, the situation is a bit different. All estimators provide FWER control at the given α for fractions higher than 0.95. Now, the maximum power at controlled FWER is achieved by the carving estimator in the selected viewpoint, with 0.78 at fraction 0.95. When considering only the power values with valid FWER, we find that both PoSI and carving in the saturated viewpoint outperform the combined carving estimator. This comparison is particularly interesting, as all of these estimators can be computed in a comparable time. While controlling the FWER, the combined carving estimator achieves a maximum power of 0.20 at fraction 0.9. It is worth mentioning that the data splitting estimator achieves valid FWER control starting at $f = 0.8$, the lowest among all estimators. This is mostly due to the increasing cases were the p-values are all set to 1, as their computation starts to fail because of the rank issues discussed in Section 2.2.

As previously suggested at the end of Subsection 2.5.2, the power of the PoSI estimator should be the exact limiting point for the power of carving in both the saturated and selected viewpoints. At first, it may seem that this is not the case in Figure 3.2. However, Fithian et al. (2017) mention that carving in the selected viewpoint may experience a substantial decrease in power when going from fraction 0.99 to the full dataset for selection. Moreover, our simulations indicate that carving in the saturated viewpoint does not demonstrate strictly monotonically increasing power as fractions increase. This leads us to conclude that the power values we obtained for the PoSI estimator in Figure 3.2 are indeed plausible.

Lastly, we present our results for the 15-sparse setup in Figure 3.3. In contrast to the 5-sparse setup, we had to significantly increase the SNR to enable the estimators to achieve any reasonable power. This increase in SNR results in a low noise variance. Remarkably, although all the estimators initially exhibit very high FWER, they are able to steadily decrease it, ultimately achieving FWER control at $f = 0.9$. Upon comparing all estimators in the range where they control the FWER, we find that carving in the selected viewpoint achieves the highest power, reaching 0.46 at fraction $f = 0.95$. The combined carving estimator and data splitting have negligible power for the fractions where they control the FWER. The maximum power while controlling for FWER achieved by PoSI and carving in the saturated viewpoint is 0.16 and 0.14, respectively. Interestingly, in Figure 3.3 the power of PoSI aligns perfectly with the theoretical expectations, which further supports our argument from the last paragraph.

In conclusion, it appears that in both cases, all estimators struggle to achieve FWER control for fractions below 0.9. The only estimator that obtains valid FWER values for lower fractions is data splitting. The carving estimator in the selected viewpoint dominates all other estimators with respect to power, although we must remain cautious, as it is the least conservative with respect to FWER. Hence, for low SNR values, it may be preferable to use post-selection inference methods that use more data for the selection process. That is, carving in the saturated viewpoint with a high fraction f or even directly PoSI on all

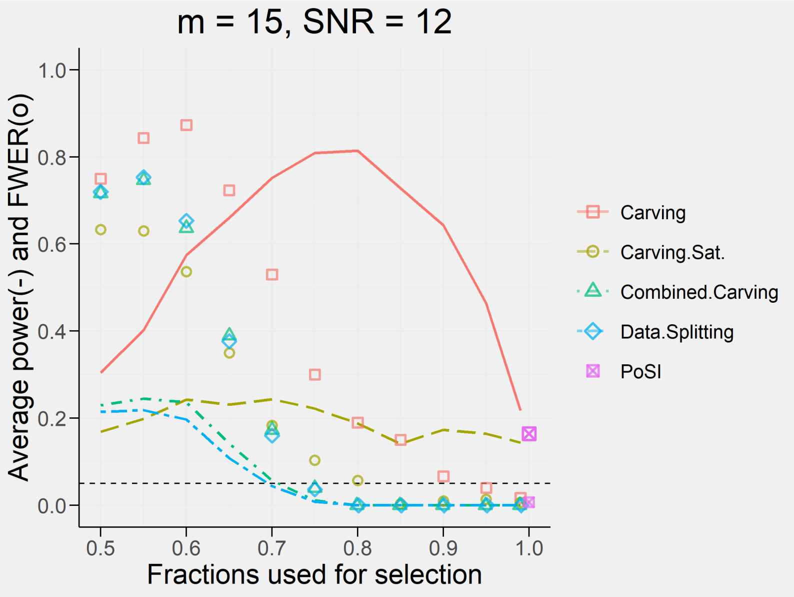

Figure 3.3: *Results from the power simulation on all of the presented estimators. Here we have* 15 *truly active variables and set* $SNR = 12$*. The power is represented as lines and the FWER as symbols. There is a horizontal dashed line at the significance level* $\alpha = 0.05$*.*

of the data. Interestingly, the regular carving estimator initially exhibits an increase in power as f increases, followed by a decrease after peaking at around $f = 0.85$. Schultheiss et al. (2021) explain this by the trade-off between more successful screening and a loss of power at the inference stage due to more imposed constraints. The most conservative choice with respect to FWER is the data splitting estimator, but its power is not on par with the other estimators.

3.2.2 Components of the combined carving estimator

In this subsection, we want to visualize the power of the combined carving estimator together with its components, i.e. the data splitting and the PoSI estimator. For this purpose we additionally computed the power and FWER of PoSI for the same fractions as for $\hat{\boldsymbol{\beta}}^{Split}$ and $\hat{\boldsymbol{\beta}}^{Comb}$. The setup for this subsection is the same as in the right plot of Figure 3.2.

The resulting plot can be seen in Figure 3.4. We would expect that the line representing the power of the PoSI estimator monotonically increases. Indeed, this would be the case if we had not bound the PoSI estimator to the other two estimators in our simulations. To make the relevant powers of the components of the combined carving estimator representative, we set the p-values of all three estimators to 1, if data splitting could not be computed due to too many selected variables. This would not have been necessary for the PoSI estimator, but was done for this illustration.

We see that combined carving mostly dominates in terms of power, but exhibits also the least conservative FWER. Both PoSI and combined carving yield valid FWER for fractions above 0.9. In this range it is interesting that PoSI already obtains larger power than combined carving itself, even while having its power diminished by the decreasing computability of data splitting, which is evident from the sharp decline of PoSI's power at around $f = 0.85$. This phenomenon is surprising, as we expected the combined carving

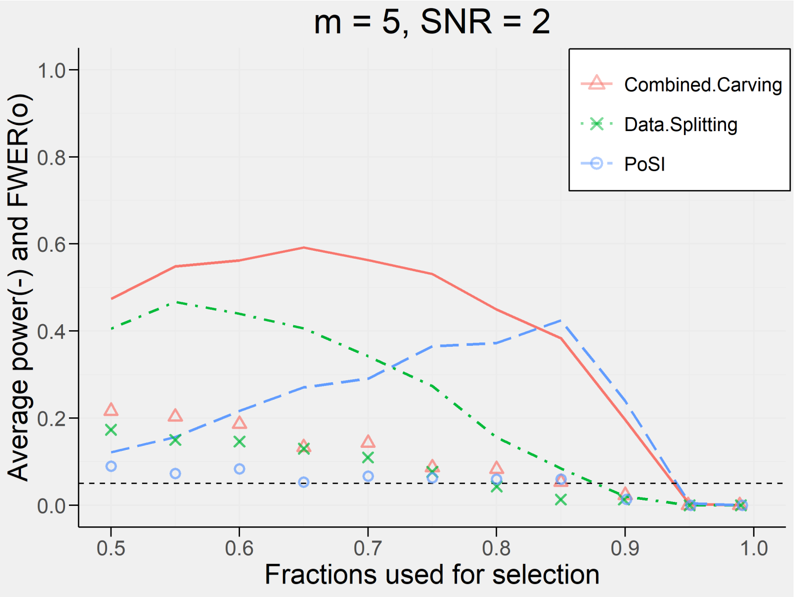

Figure 3.4: *Results from the power simulation on the components of combined carving estimator. Here we have* 5 *truly active variables and set SNR* $= 2$*. The power is represented as lines and the FWER as symbols. There is a horizontal dashed line at the significance level* $\alpha = 0.05$*.*

estimator to dominate both its components with respect to power.

The PoSI estimator surpasses data splitting in terms of power between fractions $f = 0.7$ and $f = 0.75$. Given that they are weighted by f and $1 - f$ respectively in the definition of $\hat{\boldsymbol{\beta}}^{Comb}$, we know that for fractions $f > 0.5$, the combined carving estimator incorporates more information from the PoSI estimator than from the data splitting estimator. These weights also explain why the powers of the carving and PoSI estimators converge at higher fractions.

3.2.3 Multiple attempts for the combined carving estimator

We saw in Subsections 3.2.1 and 3.2.2 that the power of the combined carving estimator tends to drop off quite drastically once we reach a point where it cannot be computed anymore due to the aforementioned rank issues. This occurs because we immediately set the p-values to 1 for all variables in this case. However, the practitioner might be willing to take advantage of the randomness of the model selection process and simply try again instead. That is, if $\hat{\boldsymbol{\beta}}^{Comb}$ cannot be computed, they might hope that fewer variables would be selected by the Lasso upon retrying, allowing $\hat{\boldsymbol{\beta}}^{Comb}$ to be computed once again. It is important to note that from a theoretical standpoint, this approach would of course be invalid, because it would necessitate conditioning on the previously failed attempts. Nonetheless, we thought that the results of the simulation would be of interest, when allowing for five different attempts for the calculation of $\hat{\boldsymbol{\beta}}^{Comb}$. Thus, we implemented the same simulation as described in Subsection 3.2.1, with the difference that whenever $\hat{\boldsymbol{\beta}}^{Comb}$ could not be computed, we ran the Lasso another time on the same $\mathbf{y}$ for $\hat{\boldsymbol{\beta}}^{Comb}$. We kept the results from the initial variable selection for $\hat{\boldsymbol{\beta}}^{Carve}$ and $\hat{\boldsymbol{\beta}}^{Sat}$.

The results can be seen in Figure 3.5. When comparing the results with those from Figure 3.2 on the right, where the same parameters were used, we observe the following: as expected, $\hat{\boldsymbol{\beta}}^{Sat}$, $\hat{\boldsymbol{\beta}}^{Carve}$ and $\hat{\boldsymbol{\beta}}^{PoSI}$ performed almost exactly like before, with slight

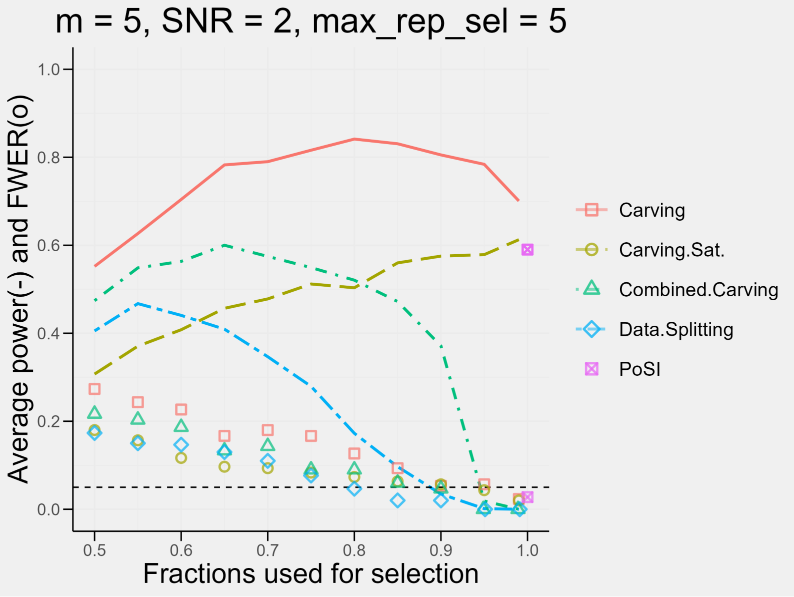

Figure 3.5: *Results from the power simulation, where we allowed the combined carving estimator 5 attempts. Again, we have* 5 *truly active variables and set SNR* $= 2$. *The power is represented as lines and the FWER as symbols. There is a horizontal dashed line at the significance level* $\alpha = 0.05$.

Fractions	Repeated Selections	In %
0.50	0	0.0%
0.55	0	0.0%
0.60	1	0.1%
0.65	7	0.5%
0.70	12	0.8%
0.75	16	1.1%
0.80	79	5.3%
0.85	136	9.1%
0.90	735	49.0%
0.95	1469	97.9%
0.99	1500	100.0%

Table 3.1: *How often did the selection have to get repeated to calculate the combined carving estimator? We allowed a maximum of 5 attempts per simulation run and fraction (nsim* $= 300$*)*

differences only being due to the randomness in the simulation. The combined carving estimator improved substantially and its power fell off only at a much higher fraction compared to before. It reached a peak power of 0.37 while controlling the FWER. Interestingly, the splitting estimator improved only very slightly. We suspect that this may be due to inference becoming more difficult from $f = 0.8$ onwards anyways. Note that there is no contradiction in the combined carving estimator improving strongly and the data splitting estimator not doing so: according to the definition of $\hat{\boldsymbol{\beta}}^{Comb}$ in (2.13), the influence of $\hat{\boldsymbol{\beta}}^{Split}$ becomes smaller, the closer we get to $f = 1$.

In Table 3.1, we see how often the selection had to be repeated. While it only rarely occurs until fraction $f = 0.85$, the need for repeated selections increases sharply immediately afterwards. There is of course a direct connection with the loss in power we observed at the same fractions in Figure 3.5.

3.2.4 Reducing fractions for the combined carving estimator

Another approach that the practitioner might take instead of simply trying to run the selection process again, is reducing the number of observations for selection whenever the combined carving estimator can not be calculated. That way, they would again have more observations left for computing the $\hat{\boldsymbol{\beta}}^{Split}$ component of the estimator. Just as in Subsection 3.2.3, this is of course an invalid approach from a theoretical point of view and we might lose our type-I error guarantees as a consequence. We did however think that it would be an interesting approach to compare the estimators in respectively optimal circumstances as far as the selection-inference split is concerned.

In Figure 3.6 we can see the results of a simulation with the same parameters as in Subsection 3.2.3. The difference is the following: we start off with all the estimators being calculated on the same selection-inference split. However, each time that so many variables are selected that $\hat{\boldsymbol{\beta}}^{Comb}$ cannot be calculated, we reduce the fraction by 0.025 for $\hat{\boldsymbol{\beta}}^{Comb}$, $\hat{\boldsymbol{\beta}}^{Split}$ and $\hat{\boldsymbol{\beta}}^{PoSI}$.

The results for $\hat{\boldsymbol{\beta}}^{Carve}$, $\hat{\boldsymbol{\beta}}^{Sat}$ and $\hat{\boldsymbol{\beta}}^{PoSI}$ are of course again very similar to earlier simulations. As one would expect, the combined carving estimator behaved differently in this new simulation setup. However, its performance does not actually improve and it does not manage to control the FWER at any point. Instead, the estimator has very similar results for all fractions from $f = 0.85$ onwards. We can see in Table 3.2 that this is also the fraction upon which the estimator seems to fall back on. We would like to point out that the combined carving estimator not controlling the FWER is most likely due to this specific simulation approach and the lack of the necessary additional conditioning. We therefore discourage the practice of reducing fractions step-wise in the case of $\hat{\boldsymbol{\beta}}^{Comb}$ not being computable.

Given that carving under the saturated viewpoint is also computationally efficient, requiring no MCMC sampling, it appears to be a more than viable alternative to the combined carving estimator - even in a scenario that is very favourable to the combined carving estimator in terms of power.

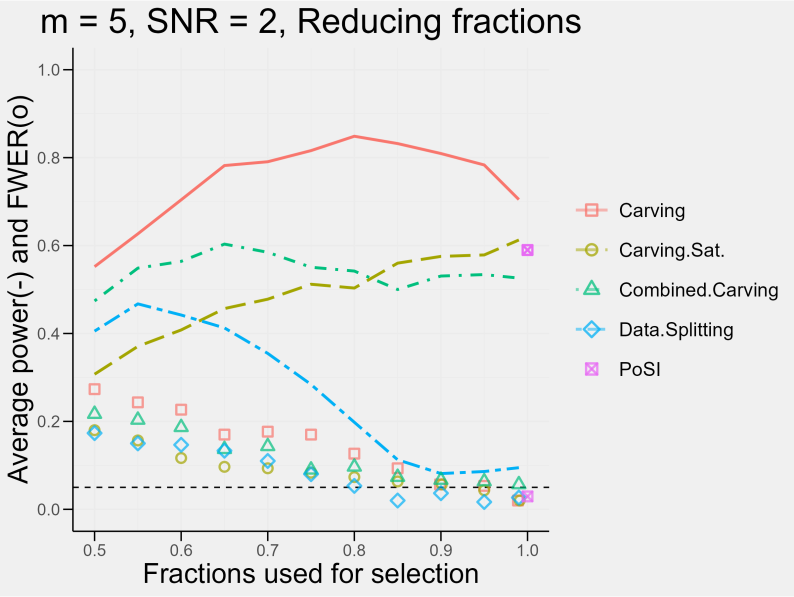

Figure 3.6: *Results from the power simulation, where we reduce the fraction each time that the combined carving estimator can not be computed. Again, we have* 5 *truly active variables and set* $SNR = 2$*. The power is represented as lines and the FWER as symbols. There is a horizontal dashed line at the significance level* $\alpha = 0.05$*.*

Original Fraction	Average fraction $\hat{\beta}^{Comb}$
0.50	0.500
0.55	0.550
0.60	0.600
0.65	0.649
0.70	0.699
0.75	0.749
0.80	0.795
0.85	0.841
0.90	0.867
0.95	0.871
0.99	0.873

Table 3.2: *Original fraction and average fraction reached by the combined carving estimator when lowering the original fraction step-wise by 0.025 at each failed attempt. We restart at the original fraction for each new simulation run.*

Chapter 4

Alternative approaches

4.1 From a polyhedron to a unit cube

As we have seen in the results from Section 3, the carving estimator of Fithian et al. (2017) performs well in terms of power and FWER control when compared to the other estimators we considered in this paper. However, it does have one drawback when trying to do inference on it, which is the big computational effort needed for the MCMC sampling. In her paper, Liu (2023) observes that the MCMC sampling is a very general approach that does not take advantage of the special structure we have as a consequence of the polyhedral constraints. She therefore proposes a procedure, by which the variables get transformed, such that they are then constrained by a unit cube. On this unit cube, existing sampling methods such as randomized quasi-Monte Carlo (RQMC) can be done very efficiently.

We will now give a brief summary of the theoretical methods introduced by her. For more details, we refer the interested reader to her paper. We start off by introducing the randomized Lasso.

4.1.1 The randomized Lasso

One approach to gaining power when doing inference, is to include an additional random variable in the model selection phase. In order to do this, Liu (2023) proposes the following: we operate in the same Lasso setup as described in Section 2.1 with σ^2 being known. Additionally, we assume that $\omega \sim \mathcal{N}(\mathbf{0}, \Omega)$ has a p-dimensional multivariate normal distribution and is independent of the data. With this, let us consider the randomized lasso problem:

$$\hat{\boldsymbol{\beta}}^{Rand} \in \arg\min_{\boldsymbol{\beta}} \frac{1}{2}||\mathbf{Y} - X\boldsymbol{\beta}||_2^2 + \lambda||\boldsymbol{\beta}||_1 \ - \boldsymbol{\omega}^T\boldsymbol{\beta}. \tag{4.1}$$

While the problem described in (4.1) might seem very different to the methods we have seen so far, Liu (2023) shows that the randomized Lasso is in fact asymptotically equivalent to the carving approach of Fithian et al. (2017). She argues this by considering their approach of splitting the data. The Lasso problem is then

$$\hat{\boldsymbol{\beta}} \in \arg\min_{\boldsymbol{\beta}} \frac{1}{2\rho}||\mathbf{Y}_A - X_A\boldsymbol{\beta}||_2^2 + \lambda||\boldsymbol{\beta}||_1,$$

where $\rho = \frac{n_A}{n} \in (0, 1)$. She sets

$$\omega = -X^T(\mathbf{Y} - X\hat{\boldsymbol{\beta}}) + \frac{1}{\rho}X_A^T(\mathbf{Y}_A - X_A\hat{\boldsymbol{\beta}}),$$

such that the randomized Lasso problem in (4.1) has the same form as the Lasso problem in (4.1.1). Then, asymptotically, $\omega \xrightarrow{d} \mathcal{N}(\mathbf{0}, \frac{1-\rho}{\rho}\sigma^2 X^T X)$. Thus, she presents a link between the data carving approach by Fithian et al. (2017) that we saw before and the randomized Lasso that she employs throughout her paper.

4.1.2 Variable transformation

In the setting of the randomized Lasso, Liu (2023) intends to do inference for $\hat{\boldsymbol{\beta}}_{\hat{M}}^{Rand}$, i.e. an estimator obtained through the randomized Lasso (4.1) with the corresponding model being $\hat{M}$. She does this by first calculating the conditional density of $\hat{\boldsymbol{\beta}}_{\hat{M}}^{Rand}|\{\boldsymbol{r}, \boldsymbol{s}\}$, where $\boldsymbol{r} = X^T(X_{\hat{M}}\hat{\boldsymbol{\beta}}_{\hat{M}}^{Rand} - \mathbf{Y})$ and $\boldsymbol{s}$ is the subgradient of $\lambda||\hat{\boldsymbol{\beta}}_{\hat{M}}^{Rand}||$. This density is portrayed as integrals over the orthant $\mathcal{O} = \{\boldsymbol{v} \in \mathbb{R}^{\hat{m}}|\boldsymbol{v} > 0\}$, which are tricky to estimate efficiently. As a next step, she calculates the distribution function $F(x)$ of the estimator, which again contains integrals over $\mathcal{O}$. By using the separation-of-variable method by Genz (1992), she manages to transform $F(x)$ into integrals over the unit cube $[0, 1]^{\hat{m}}$. Then, it is possible to estimate $F(x)$ by generating uniform samples over the unit cube, which is computationally much easier than the MCMC approaches that we needed in Section 2.5.

4.2 “Multi”-approach

A problem that still remains with all the methods discussed in Chapter 2 and also the computationally faster method described in Section 4.1, is the inherent randomness behind the Lasso choosing the variables in the first place. This can easily lead to the Lasso choosing different variables on different occasions and can be a barrier to practical applications. Meinshausen et al. (2009) call this problem the ”p-value lottery”. Their solution is to run the Lasso B times instead. For each individual split, they then use the same sample-splitting setup we also described in Section 2.2 to calculate p-values p_j^b for each split $b \in \{1, \ldots, B\}$ and each variable $j \in \{1, \ldots, p\}$. The $p \cdot B$ many p-values p_j^b then get aggregated into p-values P_j using quantile functions. The spirit of the idea of Meinshausen et al. (2009) was carried over by Schultheiss et al. (2021) into the carving framework by Fithian et al. (2017) in the form of “Multicarving”. With this, Schultheiss et al. (2021) managed to improve upon the FWER of single-carving. However, as Fithian et al. (2017) already point out, data carving requires the use of Monte Carlo methods in most cases, which can become computationally expensive. This problem of course gets exacerbated in the multicarving approach.

Chapter 5

Conclusion

Data carving in the selected viewpoint has some very attractive theoretical properties. However, it can be impractical in very high-dimensional settings due to the need for MCMC sampling and the accompanying computational costs.

Drysdale (2023) proposed a solution by introducing the combined carving estimator, which is computationally inexpensive due to having a parametric distribution. In this paper, we set out to analyze the performance of this estimator. While being computationally faster, we unfortunately found that it is not well-defined in some relatively common scenarios. This problem led to poor performance in terms of power and FWER when compared to the two carving estimators by Fithian et al. (2017) and also the PoSI-estimator. Even when putting systems into place to circumvent this theoretical flaw, its performance made it hard to justify its application in practice.

It is hard to make general statements about which post-selection inference method is best. From our simulations, we conclude that in settings where SNR is reasonably high, carving in the selected viewpoint achieves superior power compared to the other estimators we tested. If in doubt, especially when many truly active variables could potentially be in play, PoSI and carving in the saturated viewpoint with high fractions for the selection may be preferable. They provide a more conservative approach with respect to FWER while still achieving reasonable power.

As for the problem of p-value lottery, since Liu (2023) manages to reduce the computational cost of carving in the selected viewpoint, the implementation of her work in a “Multi”-framework could be very interesting.

Bibliography

Berk, R., L. Brown, A. Buja, K. Zhang, and L. Zhao (2013). Valid post-selection inference. *The Annals of Statistics 41*(2), 802 – 837.

Cox, D. R. (1975). A note on data-splitting for the evaluation of significance levels. *Biometrika 62*(2), 441–444.

Drysdale, E. (2023). A parametric distribution for exact post-selection inference with data carving. *arXiv preprint 2305.12581*.

Fithian, W., D. Sun, and J. Taylor (2017). Optimal inference after model selection. *arXiv preprint 1410.2597*.

Genz, A. (1992). Numerical computation of multivariate normal probabilities. *Journal of Computational and Graphical Statistics 1*(2), 141–149.

Hothorn, T., A. Genz, F. Bretz, T. Miwa, X. Mi, F. Leisch, F. Scheipl, B. Bornkamp, and M. Maechler (2023). *mvtnorm: Multivariate Normal and t Distributions.* R package version 1.2-4.

Lee, J. D., D. L. Sun, Y. Sun, and J. E. Taylor (2016). Exact post-selection inference, with application to the lasso. *Ann. Statist. 44*(3), 907–927.

Liu, S. (2023). An exact sampler for inference after polyhedral model selection. *arXiv preprint 2308.10346*.

Meinshausen, N., L. Meier, and P. Bühlmann (2009). p-values for high-dimensional regression. *Journal of the American Statistical Association 104*(488), 1671–1681.

R Core Team (2021). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing.

Schultheiss, C., C. Renaux, and P. Bühlmann (2021). Multicarving for high-dimensional post-selection inference. *Electron. J. Stat. 15*(1), 1695–1742.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Statist. Soc. Ser. B 58*(1), 267–288.

Tibshirani, R. J. (2013). The lasso problem and uniqueness. *Electron. J. Stat. 7*, 1456–1490.

Van Rossum, G. and F. L. Drake Jr (1995). *Python reference manual.* Centrum voor Wiskunde en Informatica Amsterdam.