Università degli Studi di Padova
Dipartimento di Scienze Statistiche

Corso di Laurea Magistrale in
Scienze Statistiche

Permutation-Based Inference for High-Dimensional Linear Models

Relatore: Professor Livio Finos
Dipartimento di Scienze Statistiche

Correlatrice: Dott.ssa Daniela Corbetta
Dipartimento di Scienze Statistiche

Paolo Della Penna
Matricola N. 2107071

Anno Accademico 2024/2025

Abstract

Hypothesis testing is a fundamental tool of statistical inference, but its validity is compromised when applied after data-driven model selection or in the presence of multiple comparisons. These issues, long recognized in classical statistics, are magnified in modern high-dimensional settings where the number of predictors exceeds the sample size. Standard least-squares procedures become ill-posed, and naïve p-values fail to provide control of type I error. While penalized estimators such as the lasso have become essential for sparse modeling, their integration with valid post-selection inference and multiple-testing adjustment remains an open problem. This thesis addresses these challenges by proposing a permutation-based framework that combines effective scores with sign-flip inference, adapted to high-dimensional linear models. The method relies on a confounder selection step that excludes the variable under test, thereby limiting post-selection bias, and constructs test statistics whose null distribution is approximated through random sign flips of the score contributions. The framework naturally extends to multiple hypothesis testing via maxT correction, ensuring rigorous Familywise Error Rate control. Theoretical results are complemented by extensive simulations, which demonstrate robust type I error control and competitive power across different correlation structures and signal-to-noise regimes. These findings highlight the potential of sign-flip based inference as a flexible and reliable tool for high-dimensional hypothesis testing.

Acknowledgements

I would like to express my deepest gratitude to Professor Livio Finos and Dr. Daniela Corbetta for their guidance, support, and invaluable feedback throughout this work.

I could not have undertaken this journey without the constant encouragement and affection of my family and friends, to whom I am deeply grateful.

Contents

Chapter 1

Introduction

Hypothesis testing has long provided a principled framework for assessing evidence against scientific hypotheses. The formulation of test statistics, the interpretation of p-values, and the quantification of error probabilities are cornerstones of statistical inference, and they have shaped the development of modern science throughout the twentieth century. The early debates among Fisher, Pearson, and Neyman already highlighted both the power of hypothesis testing as a scientific tool and the potential risks of misuse. When properly applied, hypothesis tests enable the rigorous validation of experimental findings; when applied carelessly, however, they may lead to false discoveries.

There are several pitfalls in the hypothesis testing procedure, which can ultimately lead to spurious conclusions and which, although known, are still taken lightly. An example of such an error often occurs when multiple hypotheses must be tested simultaneously: already in Bonferroni (1936), Emilio Bonferroni provided a simple and robust procedure to limit the Familywise Error Rate when testing. While conceptually straightforward, this type of correction illustrates the general challenge of error control. Beyond the problem of multiple testing, an even subtler issue arises when inference is conducted after a model has been selected from the data. In such cases, the assumptions underlying classical theory are no longer valid, and naïve p-values fail to control the nominal error rate. These problems are often underestimated in applied research; yet, they are critical for ensuring the reliability of scientific conclusions.

The rapid expansion of high-dimensional data has made these issues even more pressing. In many modern applications, from genomics to neuroscience, the number of available predictors can far exceed the number of observations. This poses fundamental challenges: classical least-squares estimation is ill-posed,

confidence intervals and standard errors are unreliable, and naïve testing strategies are invalid. To address these difficulties, modern statistics often relies on structural assumptions such as sparsity, under which only a small subset of variables is truly relevant. Penalized estimation methods embody this principle, with the lasso (Tibshirani, 1996) as the most widely known example. The lasso and its extensions have become standard tools for constructing sparse predictive models, focusing on the bias-variance trade-off and on the selection process.

Although significant progress has been made in prediction and model selection, valid statistical inference in high dimensions remains a challenging problem. Two issues are particularly critical. First, post-selection inference requires accounting for the randomness introduced by the selection step; otherwise, standard error rates cannot be guaranteed. Second, multiple testing arises naturally when many hypotheses are simultaneously assessed, and without proper adjustment, the probability of false discoveries escalates rapidly. Addressing both problems in a unified framework is essential for producing reliable scientific results, yet this has proven difficult in practice. This thesis contributes to this ongoing line of research by proposing a testing framework for high-dimensional linear models that simultaneously accounts for post-selection bias and multiple testing. The approach builds on sign-flip tests constructed from score contributions (Hemerik et al., 2020), which provide a non-parametric alternative to classical inference, and on their variance-standardized versions (De Santis et al., 2025), which improve finite-sample calibration. By combining these elements with appropriate selection strategies, we develop procedures that are both empirically justified and practically feasible. The methods are designed to offer robustness under model misspecification and flexibility in the presence of high dimensionality while still providing strong error control when testing many hypotheses at once. The contributions of the thesis are twofold. On the theoretical side, we investigate the properties of sign-flip inference in high-dimensional settings, clarifying the conditions under which type I error control can be maintained. On the empirical side, we conduct extensive simulations to assess the performance of the proposed procedures, with particular attention to the trade-off between power and error control under different correlation structures and signal strengths.

The thesis is organized as follows. Chapter 2 provides an overview of the main challenges in statistical inference when the dimensionality of the data is high. In particular, it reviews classical approaches to model selection, highlights the difficulties that arise when inference is carried out after a selection step, and

discusses the problem of error inflation in multiple testing scenarios. Moreover, it briefly introduces two of the most promising methodologies in this field, the debiased lasso (van de Geer et al., 2014; Zhang & Zhang, 2013) and the ridge projection (Bühlmann, 2013). This chapter sets the stage for the methodological developments that follow by clarifying why standard inferential tools fail in modern applications and why new solutions are needed.

Chapter 3 introduces the concept of the effective score, along with the sign-flip methodology, within the framework of generalized linear models. The discussion covers both the theoretical foundations and the practical implementation of the method, explaining how sign-flipping provides a non-parametric alternative to conventional inference and how variance standardization improves finite-sample performance. These ideas form the theoretical backbone of the procedures later adapted to high-dimensional linear models.

Chapter 4 extends the sign-flip approach to the high-dimensional setting. Different strategies for confounder selection are described, including lasso-based and stepwise approaches, each combined with sign-flip inference. The chapter also explains how the procedure can be generalized to address multiple hypotheses simultaneously through a maxT (Westfall & Young, 1993) correction, thereby achieving rigorous Familywise Error Rate control. This part represents the central methodological contribution of the thesis.

Chapter 5 presents a comprehensive simulation study designed to evaluate the empirical performance of the proposed procedures. The simulations investigate both single and multiple hypothesis testing scenarios, focusing on the control of type I error, statistical power, and computational aspects. Comparisons with alternative approaches are included to highlight strengths and limitations under different correlation structures and signal-to-noise regimes.

Finally, Chapter 6 concludes by summarizing the main contributions and reflecting on the implications of the results. It also discusses open questions that remain unresolved and outlines possible directions for future research, particularly in relation to theoretical guarantees, extensions to other models, and applications in fields where high-dimensional inference is essential.

Chapter 2

Statistical Inference for High-Dimensional Data

Over the past decades, the increasing ability to collect vast amounts of data has made the analysis of high-dimensional datasets a central topic in many applied fields.

High-dimensional statistics deals with inference in settings where the number of parameters p greatly exceeds the sample size n. This includes supervised models like regression or classification with many more covariates than observations, unsupervised scenarios such as clustering or graphical modeling, and multiple testing problems where the number of hypotheses surpasses the available samples. In this thesis, we specifically address the problem of studying the relationship between a large number of predictors and a univariate response variable, without considering the case of multivariate response variables.

Classical inference techniques are not suitable in this context. For example, fitting a linear model via least squares when $p > n$ leads to ill-posed problems and unreliable p-values. Without structural assumptions or model restrictions, statistical inference in high dimensions is fundamentally unfeasible.

To address the challenges of high dimensionality, modern approaches often assume some form of sparsity: only a small subset of variables is truly relevant. Among the methods exploiting this idea, the most well-known is the lasso estimator (Tibshirani, 1996), which imposes an ℓ_1 penalty to shrink some coefficients to zero. This improves prediction accuracy by reducing variance at the cost of some bias, and also performs variable selection. Building on the lasso, several extensions have been proposed, including the elastic net (Zou & Hastie, 2005), the adaptive lasso (Zou, 2006), and non-convex penalties such as SCAD (Fan

& Li, 2001) and MCP (Zhang, 2010), which aim to improve variable selection consistency and predictive performance.

Despite the large literature on prediction and selection, statistical inference (e.g., valid p-values or confidence intervals after model selection) in high-dimensional settings is still developing. In high-dimensional linear models, several strategies have emerged to provide valid error control and significance testing. One key approach involves debiasing the lasso estimator to restore asymptotic normality and enable inference (van de Geer et al., 2014; Zhang & Zhang, 2013). Another method, known as the ridge projection approach, uses ridge regression followed by a correction for bias due to shrinkage (Bühlmann, 2013).

A different line of work uses sample splitting: the data is divided into two parts, one for variable selection and one for inference. Wasserman & Roeder (2009) applied this to the lasso and stepwise procedures, while Meinshausen et al. (2009) enhanced robustness and power by aggregating results across multiple random splits.

More recently, Zhao et al. (2021) showed that a seemingly naïve two-step approach, selecting variables with lasso and then refitting with ordinary least squares, can provide valid inference under certain assumptions, without requiring sample splitting.

In this chapter, we first introduce two model selection methods: stepwise selection procedures and the lasso, along with their main properties. We then discuss hypothesis testing, focusing on the issue of double dipping, reusing data for both selection and inference, which invalidates classical procedures, and showing the problems that occur in multiple testing. Finally, we present two competitive inference methods: ridge projection (Bühlmann, 2013) and the debiased lasso (van de Geer et al., 2014; Zhang & Zhang, 2013).

2.1 Model Selection in High Dimensions

The choice of a statistical model for studying a phenomenon of interest is fundamental. A central problem in statistics is to identify the relationship between a response variable and a set of explanatory variables. The formulation of the model is guided by the research question, and this flexibility means there is not a single correct model but rather one that is appropriate for the goal at hand.

In experimental settings, the researcher often has control over the values of the explanatory variables, which allows careful planning to reduce confounding

and detect interactions. In such cases, a broad variability among experimental conditions typically increases the precision of the analysis. However, in observational studies, variables are not set in advance, and we are often faced with large, complex datasets with many potential predictors. Selecting the appropriate variables becomes a crucial step.

To address this complexity, automatic procedures for data-driven model selection have been developed. These methods aim to balance model fit, interpretability, and generalizability. Including too many variables may lead to excellent in-sample performance but poor out-of-sample predictions and inflated variance of coefficient estimates. This phenomenon, known as overfitting, highlights the trade-off between bias, variance, and model complexity. Model selection procedures attempt to find a good compromise between these competing goals. Two prominent families of methods are stepwise selection algorithms and regularization-based methods such as the lasso.

For clarity, we focus on the linear regression framework with n observations and p covariates:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim \mathrm{N}_n(\mathbf{0}, \sigma^2 \boldsymbol{I}_n), \tag{2.1}$$

where $\boldsymbol{Y} \in \mathbb{R}^n$ is the response vector, $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ is the design matrix, $\boldsymbol{\beta} \in \mathbb{R}^p$ is the vector of unknown coefficients, and $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ is a vector of random errors.

2.1.1 Stepwise Selection

Given p candidate variables, the number of possible models is $2^p - 1$, excluding the model with only the intercept. Even for moderate p, exhaustive evaluation becomes computationally unfeasible. stepwise procedures explore a subset of the model space in a sequential, greedy manner, often leading to suboptimal but computationally efficient solutions. The main methods are:

- Forward selection starts with an empty model and adds variables that most improve fit.
- Backward selection starts with the full model and removes variables contributing least.
- Hybrid methods alternate between adding and removing variables.

A common criterion to guide selection is the Akaike Information Criterion (AIC)(Akaike, 1974):

$$\mathrm{AIC} = -2l(\hat{\boldsymbol{\beta}}; \boldsymbol{y}, \boldsymbol{X}) + 2p,$$

where $l(.)$ denotes the log-likelihood function Similarly, the Bayesian Information Criterion (BIC) (Schwarz, 1978) is defined as:

$$\text{BIC} = -2l(\hat{\boldsymbol{\beta}}; \boldsymbol{y}, \boldsymbol{X}) + p \log(n),$$

These criteria balance goodness-of-fit with model complexity.

Forward and hybrid stepwise selections are attractive in high-dimensional contexts ($p \gg n$) due to computational feasibility. However, they may ignore important variables due to the greedy nature of the search and are sensitive to the order in which variables enter the model.

2.1.2 The Lasso

The lasso (Least Absolute Shrinkage and Selection Operator) (Tibshirani, 1996) performs both shrinkage and variable selection. Given the model in (2.1) the lasso optimization problem is:

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

where $\lambda \geqslant 0$ is a regularisation parameter and $\|.\|_2$, $\|.\|_1$ rappresents respectively the ℓ_2 and the ℓ_1 norm. Larger values of λ enforce greater sparsity by shrinking more coefficients to zero. The power of this penalization lies in its ability to produce sparse estimates while preserving the convexity of the optimization problem, which makes it solvable via efficient algorithms such as Least Angle Regression (LARS) or pathwise coordinate descent.

The lasso produces a solution $\hat{\beta}(\lambda)$ whose support $\hat{S}_\lambda = \{j : \hat{\beta}_j(\lambda) \neq 0, j = 1, \ldots, p\}$ depends on the choice of λ. Under suitable conditions, such as the irrepresentability condition and strong signal strength, it is possible to consistently recover the true active set $S = \{j : \beta_j^0 \neq 0, j = 1, \ldots, p\}$ as $n \to \infty$, where β_j^0, $j = 1, \ldots, p$, rappresent the true value of the j-th coeffient.

The irrepresentable condition requires that

$$\|(X_S^T X_S)^{-1} X_S^T X_{S^C}\|_\infty \leqslant 1 - \varepsilon \quad \text{for some } \varepsilon \in (0, 1).$$

This ensures that irrelevant variables X_{S^c} are not too correlated with relevant ones X_S.

The signal strength condition requires that non-zero coefficients satisfy:

$$\min_{j \in S} |\beta_j^0| \gg \sqrt{\frac{\log(p)}{n}}.$$

When both conditions hold, there exists a $\lambda \gg \sqrt{\log(p)/n}$ such that

$$P(\hat{S}_\lambda = S) \to 1 \quad \text{as} \quad n \to \infty.$$

However, a fundamental trade-off arises in the orthonormal case:

- For consistent variable selection, one needs $\sqrt{n}\lambda \to \infty$.
- For $\sqrt{n}$-consistent estimation, one needs $\lambda = O(n^{-1/2})$.

These goals are incompatible: the lasso cannot achieve both at once. This is one reason why alternatives like SCAD (Fan & Li, 2001) and MCP (Zhang, 2010) have been proposed to bridge the gap between selection and estimation.

Moreover, in practice, λ is typically chosen using cross-validation to minimise prediction error. This often prioritises estimation accuracy over sparsity, resulting in models that include irrelevant variables but still capture the true ones with high probability.

2.2 Hypothesis Testing in High Dimensions

In regression analysis, hypothesis testing is typically used to assess whether specific coefficients are significantly different from zero. Given the linear model (2.1), we are often interested in testing, for each $j = 1, \ldots, p$, the null hypothesis:

$$H_0^j : \beta_j = 0 \quad \text{vs} \quad H_1^j : \beta_j \neq 0. \tag{2.2}$$

Each test yields a test-statistic, a function $t : \mathcal{Y} \to \mathbb{R}$ that identifies two regions in the sample space $\mathcal{Y}$:

- the acceptance region $\mathcal{A}$, where the data do not indicate evidence against H_0;
- the rejection region $\mathcal{R}$, where the data indicate evidence against H_0.

Following the Fisherian view of hypothesis testing, the result of the test quantifies the amount of evidence that the observed data provides against the null hypothesis H_0. This quantity is called the p-value. For the j-th two-sided test in (2.2), with $j = 1, \ldots, p$, the observed significance level (p-value) is defined as:

$$\alpha^{obs} = 2 \cdot \min\{\mathbb{P}_{\beta^j}(t(\boldsymbol{Y}) \leqslant t(\boldsymbol{y}), \mathbb{P}_{\beta^j}(t(\boldsymbol{Y}) \geqslant t(\boldsymbol{y})\},$$

where $t(\boldsymbol{y})$ is the test statistic for the observed data $\boldsymbol{y}$. We reject H_0 if the p-value is less than or equal to α (commonly 0.05).

The outcomes of hypothesis testing are shown in Table 2.1. Two types of errors may occur: Type I error and Type II error.

Definition 2.1 (Error Types)**.** Given a p-value α^{obs} for a hypotesis test correctly specified of level α:

- Type I Error (False Positive): Rejecting H_0 when it is true.

$$\mathbb{P}(\text{Type I Error}) = \mathbb{P}(\alpha^{obs} \leqslant \alpha | H_0).$$

- Type II Error (False Negative): Failing to reject H_0 when it is false.

$$\mathbb{P}(\text{Type II Error}) = \mathbb{P}(\alpha^{obs} \geqslant \alpha | H_1).$$

where the notation $\mathbb{P}(. \mid H_0)$ and $\mathbb{P}(. \mid H_1)$ refers to probabilities conditional on H_0 being true or false, respectively.

Another important quantity is the power of a test, which is the probability of correctly rejectingH_0 under H_1:

$$\text{Power} = \mathbb{P}(\alpha^{obs} \leqslant \alpha | H_1) = 1 - \mathbb{P}(\text{Type II Error}).$$

Table 2.1: Error types in hypothesis testing

		Null Hypothesis	
		False (H_1 true)	True (H_0 true)
Test Decision	Reject H_0	True Positive	Type I Error
	Do not reject H_0	Type II Error	True Negative

2.2.1 The Post-Selection Problem

In classical settings, hypothesis testing assumes that the model under test is fixed a priori. However, in modern applications, models are often selected via data-driven procedures (e.g., lasso, stepwise selection). This results is a selection bias

known as double dipping: the same data is used both to select the model and to test hypotheses within it.

As a result, classical inferential guarantees such as correct type I error control are no longer valid. The selected model is not fixed, but a stochastic function of the data. For instance, in forward stepwise selection, the selected variables depend on the residual sum of squares at each iteration; thus, their inclusion reflects sample-specific noise.

To formalize this, let us denote the active set selected by a generic selection procedure as $\hat{S} = \hat{S}(\boldsymbol{Y}, \boldsymbol{X}) = \{j : \hat{\beta} \neq 0, j = 1, \ldots, p\}$. Classical inference proceeds conditionally on $\hat{S}$, ignoring that is random. This conditioning invalidates the uniformity of the null distribution of test statistics.

A simple illustration of this phenomenon is to perform a stepwise selection and refit the model with least squares on the selected set $\hat{S}$. Despite the apparent significance of estimated coefficients, the nominal p-values do not reflect the true false positive rate, as the model was tuned to the same data. We conducted a simple simulation as follows. We generated an equicorrelated design matrix $\boldsymbol{X}$ with $n = 50$, $p = 10$, and pairwise correlation $\rho = 0.5$. Three predictors were designated as active with coefficients equal to 3, while the remaining coefficients were set to 0; the signal-to-noise ratio was fixed at $\mathrm{SNR} = 5$. From this setup we generated the response $\boldsymbol{y}$ and then randomly chose one predictor that is truly null (i.e., not active) as a reference variable. Starting from the full model, we performed backward stepwise selection. If the reference variable was included in the selected model, we recorded its two-sided Wald p-value; otherwise we set the p-value equal to 1. We repeated this scheme until we collected 10,000 such replicates.

Figure 2.1 reports the empirical rejection curve over the range of α typically used in practice. The curve lies above the diagonal, revealing anti-conservative behavior: the null hypotesis is rejected more often than it should be in theory, indicating that the nominal post-selection p-values fail to control the type I error.

2.2.2 Multiple Testing and Error Inflation

In high-dimensional inference, we often test a large number of hypotheses simultaneously. Even if each individual test is controlled at level α, the probability of obtaining at least one false rejection increases with the number of tests.

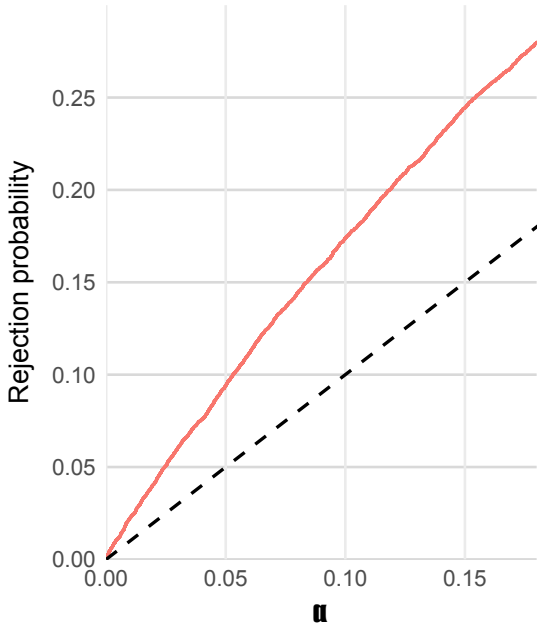

Figure 2.1: Empirical rejection curve (red) of the nominal two-sided Wald test; the dashed line corresponds to the theoretical behaviour valid null p-values.

Let us consider m null hypotheses $H_0^1, \ldots, H_0^m$ (Table 2.2), among which m_0 are true and $m_1 = m - m_0$ are false. Denote:

- V: number of false rejections;
- R: total number of rejections;
- $S = R - V$: true discoveries.

Table 2.2: Multiple testing outcomes

	Null Hypothesis		**Total**
Test Result	False	True	m
Rejected	S	V	R
Not rejected	T	U	$m - R$
Total	m_1	m_0	m

In the multiple testing literature, several measures have been proposed to generalize the concept of type I error. The most common are the FamilyWise Error Rate (FWER) and the False Discovery Rate (FDR).

Definition 2.2. The FamilyWise Error Rate is the probability of making at least one false rejection:

$$\text{FWER} = \mathbb{P}(V \geqslant 1).$$

Definition 2.3. False Discovery Rate The False Discovery Rate is the expected proportion of false rejections among all rejections:

$$\text{FDR} = \mathbb{E}\left[\frac{V}{\max(R, 1)}\right].$$

The FWER is a more stringent measure as it directly controls the probability of any false rejections, which can lead to overly conservative tests. In contrast, the FDR focuses on controlling the expected proportion of errors among rejections, allowing more flexibility at the cost of being less stringent, as it is an expected value rather than a strict probability.

To control these measures, several p-value correction procedures have been developed. The most widely known for FWER control are the Bonferroni correction (Bonferroni, 1936) and the maxT method (Westfall & Young, 1993) based on resampling techniques. For FDR control, the most famous method is the Benjamini-Hochberg procedure (Benjamini & Hochberg, 2018).

We now show a simple example of what happens jointly at the control of type I error. Suppose we test two independent null hypotheses at level $\alpha = 0.05$. The marginal error rate is controlled:

$$\mathbb{P}(p_1 \leqslant 0.05 | H_0^1) \leqslant 0.05, \quad \mathbb{P}(p_2 \leqslant 0.05 | H_0^2) \leqslant 0.05.$$

However, the joint probability of at least one false rejection is:

$$\mathbb{P}(p_1 \leqslant 0.05 \cup p_2 \leqslant 0.05 | H_0^1, H_0^2) = 1 - (1 - 0.05)^2 = 0.0975 > 0.05.$$

Thus, the Familywise Error Rate is not controlled.

2.2.3 maxT correction

A powerful method for controlling the FWER is the maxT procedure, which relies on the distribution of the maximum test statistic under the global null hypothesis. In contrast to Bonferroni-type corrections, which do not exploit dependence and therefore can be conservative in the presence of correlated tests, the maxT method uses the joint distribution of the test statistics and explicitly accounts for their dependence structure. Consider the two-sided testing setup in Equation 2.2.

Concretely, the procedure approximates the null distribution of

$$T_{\max} = \max_{1 \leqslant j \leqslant p} T_j,$$

via permutations. Each resample yields a new set of test statistics, and the largest among them is recorded. The empirical distribution of these maxima is then used to obtain adjusted p-values:

$$p_j^{\text{adj}} = \mathbb{P}\big(|T_{\max}| \geqslant |t_j| \;\big|\; H_0^c\big),$$

where H_0^c denotes the complete null hypothesis under which $\beta_j = 0$ for all $j = 1, \ldots, p$. This construction guarantees strong control of FWER while avoiding the excessive conservativeness of Bonferroni.

Algorithm 2.1. **Single-step maxT procedure**

Input: Matrix of dimension $(B+1) \times p$ of the observed and permuted test statistics, where B is the number of permutation:

$$\begin{bmatrix} t_1^{obs} & \ldots & t_p^{obs} \\ t_1^1 & \ldots & t_p^1 \\ \vdots & \ddots & \vdots \\ t_1^B & \ldots & t_p^B \end{bmatrix}.$$

1. For $b = 1, \ldots, B$ record the maximum absolute statistic

$$t_{\max}^b \leftarrow \max_{1 \leqslant j \leqslant p} |t_j^b|.$$

2. For each hypothesis $j = 1, \ldots, p$, set the single-step adjusted p-value

$$p_j^{adj} = \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}\left\{ t_{\max}^{(b)} \geqslant |t_j| \right\}.$$

Return: $p_1^{adj}, \ldots, p_p^{adj}$.

Following Westfall & Young (1993), we distinguish two variants:

- the single-step procedure, in which all tests are calibrated against the same distribution of T_{max} (Algorithm 2.1);

- the step-down procedure, which orders the tests from most to least significant and proceeds sequentially, updating the reference maximum to the set of unrejected hypotheses. Once an hypothesis fails to be rejected, all the remaining ones are retained (Algorithm 2.2).

Algorithm 2.2. **Step-down maxT procedure based on Algorithm 4.1 in Westfall & Young (1993)**

Input: Matrix of dimension $(B+1) \times p$ of the observed and permuted test statistics, where B is the number of permutation:

$$\begin{bmatrix} t_1^{obs} & \dots & t_p^{obs} \\ t_1^1 & \dots & t_p^1 \\ \vdots & \ddots & \vdots \\ t_1^B & \dots & t_p^B \end{bmatrix}.$$

1. order the absolute value of the observed test statistics $t_1^{obs}, \dots, t_p^{obs}$ decreasingly

$$|t_{r_1}^{obs}| \geqslant |t_{r_2}^{obs}| \geqslant \cdots \geqslant |t_{r_p}^{obs}|$$

2. For $b = 1, \dots, B$ compute the successive maxima (backward recursion)

$$u_m^b \leftarrow |t_{r_m,b}|, \qquad u_j^b \leftarrow \max(u_{j+1}^b, |t_{r_j}^b|) \quad \text{for } j = m-1, \dots, 1.$$

3. For each hypothesis $j = 1, \dots, p$ set the step-down adjusted p-value

$$p_{r_j}^{adj} \leftarrow \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}\left\{u_j^b \geqslant |t_{r_j}^{obs}|\right\}.$$

4. Enforce monotonicity (step-down adjustment):

$$p_{r_1}^{adj} \leftarrow p_{r_1}^{adj}, \qquad p_{r_j}^{adj} \leftarrow \max(p_{r_j}^{adj}, p_{r_{j-1}}^{adj}) \quad \text{for } j = 2, \dots, m.$$

Return: $p_{r_1}^{adj}, \dots, p_{r_m}^{adj}$.

The single-step maxT is simpler and applies a uniform threshold across hypotheses, but it is more conservative. The step-down maxT proceeds sequentially and is more complex, gaining power by exploiting the fact that, as hypotheses are rejected, the set over which the maximum is taken shrinks, leading to less stringent thresholds.

2.3 Inference Methods in High-Dimensional Settings

Several methods have been developed to perform valid statistical inference when working with regularized estimators. These approaches aim to correct the bias introduced by penalization, thereby enabling the construction of reliable confidence intervals and p-values. In what follows, we present two prominent examples: the debiased lasso (van de Geer et al., 2014; Zhang & Zhang, 2013) and the ridge projection method (Bühlmann, 2013).

2.3.1 Debiased Lasso

A very successfull and intuive method to obtain valid p-values and confidence intervals in a high-dimensional setting is to try to come back to classical theory were we have unbiased estimators, taking into account that we have to work with biased estimators, such as the lasso. The method under the name of debiased lasso uses the lasso estimate as a starting point and applies a debiasing operation to yield an estimate that can be used for constructing confidence intervals and p-values. Suppose we start from the model specified in Equation 2.1 and we want to estimate the unknown parameters β_j, $j = 1, \dots, p$ and obtain confidence intervals for each parameter. As mentioned before, in a classical setting $n > p$ the ordinary least square estimation is the way to go, yielding $\hat{\boldsymbol{\beta}}^{\text{OLS}} = (\boldsymbol{X}^\top \boldsymbol{X})^{-1} \boldsymbol{X}^\top \boldsymbol{y}$. The construction of confidence interval for a generic β_j, $j = 1, \dots, p$ is straightforward

$$\hat{\beta}_j^{\text{OLS}} \pm z_\alpha \hat{\sigma} \sqrt{[(\boldsymbol{X}^\top \boldsymbol{X})^{-1}/n]_{jj}} \quad ,$$

where z_α is the percentile of order α of the standard normal distribution, and $\hat{\sigma}^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2/(n-p)$. In the high-dimensional setting where $p > n$, the matrix $\boldsymbol{X}^\top \boldsymbol{X}$ is rank deficient so is not invertible and a possibility is to use biased estimators. One proposal that has been suggested (van de Geer et al., 2014; Zhang & Zhang, 2013; Javanmard & Montanari, 2014), is to use a debiased version of the

lasso estimator. Recalling the definition provided in subsection Subsection 2.1.2, let $\hat{\boldsymbol{\beta}}_\lambda$ the biased lasso estimate for a fixed value of λ. The debiased estimator is given by

$$\hat{\boldsymbol{\beta}}^d = \hat{\boldsymbol{\beta}}_\lambda + \frac{1}{n}\boldsymbol{M}\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_\lambda),$$

where $\boldsymbol{M}$ is an approximate inverse of $(\boldsymbol{X}^\top\boldsymbol{X})/n$. The basic intuition is that $\boldsymbol{X}^\top(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_\lambda)/n\lambda$ is a subgradient of the ℓ_1 norm at the lasso solution $\hat{\boldsymbol{\beta}}_\lambda$, and so, by adding a term proportional to this subgradient, it is possible to compensates the bias introduced.

It can be shown (Javanmard & Montanari, 2014, Section 2.1) that

$$\hat{\boldsymbol{\beta}}^d \dot{\sim} N_p(\boldsymbol{\beta}^d, \sigma^2(\boldsymbol{M}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{M}/n)),$$

where $\boldsymbol{X}^\top\boldsymbol{X}/n$ is the empirical covariance of the feature matrix. This result allows to construct confidence intervals and p-values in complete analogy with classical statistics procedures. There have been different proposals for estimating $\boldsymbol{M}$:

- van de Geer et al. (2014) propose to construct the approximate inverse using the lasso for node-wise regression (Meinshausen & Bühlmann, 2006) on the design matrix $\boldsymbol{X}$;

- Javanmard & Montanari (2014) propose to construct $\boldsymbol{M}$ by solving a convex program that aims at optimizing two objectives: for each j they define $\boldsymbol{m}_j$ to be the solution to the convex program

$$\begin{aligned} \min_{\boldsymbol{m}_j \in \mathbb{R}^p} \quad & \frac{1}{n}\boldsymbol{m}_j^\top\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{m}_j \\ \text{subject to} \quad & \|\frac{1}{n}\boldsymbol{X}^\top\boldsymbol{X}\boldsymbol{m}_j - \boldsymbol{I}_j\|_\infty \leqslant \gamma, \end{aligned}$$

where $\boldsymbol{I}_j$ is the j-th column of the identity matrix. The function to minimize controls the variance of $\hat{\boldsymbol{\beta}}^d$, while the function to minimize controls the bias of $\hat{\boldsymbol{\beta}}^d$. The approximation matrix $\boldsymbol{M}$ it will have as columns the $m_1, \ldots, m_p$ vectors obtained from the minimization problem. Starting from the definition of a confidence interval, one can construct p-values for each component. Based on many single p-values, is it possible to use standard procedures for multiple testing adjustment to control for various type I error measures.

2.3.2 Ridge Projection

An alternative approach to obtain valid p-values and confidence intervals in the high-dimensional setting $p > n$ is based on the ridge projection method. The starting point is the ridge regression estimator, which is particularly attractive for high-dimensional data since it has small variance but is biased. The key idea is to correct ridge regression in order to remove the projection bias, so that valid inference can be performed.

Let $\hat{\boldsymbol{\beta}}^{ridge}$ be the ridge estimator and $\hat{\boldsymbol{\beta}}^{init}$ a sparse initial estimator (e.g. the lasso or scaled lasso). The ridge projection method defines the following debiasing operator:

$$\hat{\beta}_j^d = \hat{\beta}_j^{ridge} - \sum_{k \neq j} [P_{\boldsymbol{X}}]_{jk} \hat{\beta}_k^{init}, \tag{2.3}$$

where $(P_{\boldsymbol{X}})_{jk}$ denotes the (j,k)-entry of the projection matrix $P_{\boldsymbol{X}} = \boldsymbol{X}^T(\boldsymbol{X}\boldsymbol{X}^T)^- \boldsymbol{X}$, with $(\boldsymbol{X}\boldsymbol{X}^T)^-$ denoting the pseudo-inverse of the squared matrix. Intuitively, the correction term removes the leakage induced by the projection of $\boldsymbol{\beta}^{ridge}$ onto the row space of X, which would otherwise create spurious signal under the null $H_0^j : \beta_j = 0$.

Based on this representation, one can construct valid two-sided p-values for testing the null hypothesis $H_0^j : \beta_j = 0$. The p-value is defined as

$$p_j = 2\left(1 - \Phi\left((a_{n,p;j}(\sigma)\,|\hat{\beta}_j^d| - \Delta_j)_+\right)\right),$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal distribution, $a_{n,p;j}(\sigma)$ is the normalization factor, Δ_j is a correction term that accounts for the uncertainty due to the initial sparse estimator $\hat{\boldsymbol{\beta}}^{init}$ (Bühlmann, 2013, Section 2.4), and $(\cdot)_+$ denotes the positive part operator.

To address the problem of multiple testing, an alternative procedure is proposed instead of relying on classical corrections. The key idea is to exploit the Gaussian structure of the debiased estimator: the test statistics have a joint distribution that is approximately known. Each p-value is then "recalibrated" by comparing it to the distribution of the minimum p-value under the Gaussian model.

Formally, one defines the distribution function

$$F_Z(c) = \mathbb{P}\left(\min_{1 \leqslant j \leqslant p} 2\big(1 - \Phi(a_{n,p;j}(\sigma)|Z_j|)\big) \leqslant c\right),$$

where $Z \sim \mathcal{N}_p(0, n^{-1}\sigma^2\Omega)$ and Ω is the covariance matrix of the ridge estimator. This corresponds to the distribution (which can be simulated) of the minimum p-value under the null hypothesis. The corrected p-values are then defined as

$$p_j^{adj} = F_Z(P_j + \zeta), \tag{2.4}$$

with a small constant $\zeta > 0$.

The underlying principle is the same as in the minP procedure of Westfall & Young (1993): each p-value is adjusted by comparing it to the distribution of the minimum across tests. However, while in the minP procedure this distribution is estimated via permutation, the ridge projection approach only requires simulating correlated Gaussian variables, exploiting the known joint distribution of the debiased statistics.

Chapter 3

Effective Score and Sign-Flip Inference in GLMs

Hypothesis testing within regression frameworks is particularly vulnerable to violations of standard distributional assumptions, notably those concerning the variance of the error terms. In the classical linear model, the assumption of homoscedasticity is often unrealistic. When it fails, conventional inference methods may produce misleading conclusions, with test statistics exhibiting either conservative or anti-conservative behavior.

Generalized linear models (GLMs), although more flexible than linear models, are not exempt from these issues. Most commonly, GLMs rely on a single dispersion parameter across all observations, an assumption that may be inappropriate in the presence of overdispersion or other forms of heterogeneity. In such cases, standard parametric inference procedures may suffer from serious validity loss.

To mitigate the consequences of variance misspecification, several robust approaches have been developed. One widely used method is to couple Wald-type statistics with sandwich estimators, which provide asymptotically robust standard errors under weak assumptions on the variance structure (White, 1982). Another strategy involves generalized additive models for location, scale, and shape (GAMLSS), which allow for a more nuanced modeling of the variance function (Rigby & Stasinopoulos, 2005). However, both approaches may perform poorly in small samples, often failing to adequately control the type I error rate.

In addition to heteroscedasticity, regression inference is also complicated by collinearity among covariates. Such dependencies can induce confounding effects, whereby the estimated impact of a variable of interest is biased due to its correlation with unmeasured or uncontrolled predictors, potentially leading to spurious findings.

To address these challenges, Hemerik et al. (2020) and De Santis et al. (2025) introduced a testing procedure based on sign-flipping of the individual contributions to the score function. The method operates on the additive decomposition of the score and introduces random signs to each contribution, yielding a test statistic that enjoys robust theoretical properties even under variance misspecification. This approach is particularly appealing in scenarios involving multiple testing, as it inherently accounts for dependencies among test statistics (Westfall & Young, 1993).

However, when nuisance parameters are estimated from the data, as is typical in practice, the assumption of independence between score contributions is no longer valid. This compromises the asymptotic validity of the sign-flip test. To resolve this, the notion of the effective score has been proposed (Hall & Mathiason, 1990). This corrected score removes the influence of the nuisance parameters, thereby reducing dependency and restoring asymptotic validity.

Building upon this idea, De Santis et al. (2025) developed a standardized version of the sign-flip test, where the variance of the test statistic is estimated conditionally on the applied sign flips. This standardization accelerates convergence to the limiting distribution and substantially improves type I error control in finite samples.

This chapter provides a derivation of the effective scoreand its implementation within the framework of GLMs. We also discuss the properties and advantages of the standardized sign-flip test.

3.1 Effective Scores: Theory and Motivation

3.1.1 Score Functions in Classical Hypothesis Testing

In large-sample settings, hypothesis testing is often based on three classical methods: the likelihood ratio test, the Wald test, and the score test. Foundational contributions by Wilks (1938), Wald (1943), and Rao (1948) established the asymptotic equivalence of these procedures under regularity conditions. Nevertheless, their behavior may differ substantially in finite samples. In this chapter, we focus primarily on the score test.

Let us first define some basic notation. Let $\boldsymbol{y} \in \mathcal{Y}$ denote the observed data, and let $\boldsymbol{\theta} \in \Theta \subseteq \mathbb{R}^p$ be a p-dimensional parameter. The likelihood function is denoted by $L(\boldsymbol{\theta}; \boldsymbol{y})$, and the log-likelihood by $l(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta}; \boldsymbol{y})$.

Definition 3.1. The score function is the gradient of the log-likelihood:

$$\boldsymbol{u}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}) = \left[\frac{\partial l(\boldsymbol{\theta})}{\partial \theta_1}, \ldots, \frac{\partial l(\boldsymbol{\theta})}{\partial \theta_p} \right]^\top.$$

The corresponding test statistic, known as the score test or Rao test, is given by:

$$W_u(\boldsymbol{\theta}) = \boldsymbol{u}(\boldsymbol{\theta})^\top \boldsymbol{I}(\boldsymbol{\theta})^{-1} \boldsymbol{u}(\boldsymbol{\theta}),$$

where $\boldsymbol{I}(\boldsymbol{\theta})$ denotes the Fisher information matrix.

Definition 3.2. The observed information matrix is defined as:

$$\boldsymbol{J}(\boldsymbol{\theta}) = -\nabla^2_{\boldsymbol{\theta}} l(\boldsymbol{\theta}) = \begin{bmatrix} -\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \theta_1^2} & \cdots & -\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \theta_p \, \partial \theta_1} \\ \vdots & \cdots & \vdots \\ -\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \theta_1 \, \partial \theta_p} & \cdots & -\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \theta_p^2} \end{bmatrix},$$

i.e., the negative of the Hessian matrix of the log-likelihood.

Definition 3.3. The Fisher information matrix is the expected value of the observed information:

$$\boldsymbol{I}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}[\boldsymbol{J}(\boldsymbol{\theta})].$$

Theorem 3.1 (Information Identity)**.** *In regular parametric models, the Fisher information equals the variance of the score function:*

$$\boldsymbol{I}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}} \left[\boldsymbol{u}(\boldsymbol{\theta}) \boldsymbol{u}(\boldsymbol{\theta})^\top \right].$$

An important property of the score function under simple random sampling is its additive structure. Let $\boldsymbol{u}(\boldsymbol{\theta}) = \sum_{i=1}^{n} \boldsymbol{\nu}_i(\boldsymbol{\theta})$ be the sum of individual score contributions from n independent observations. Let ν_j denote the j-th component of the score vector. Then:

$$\nu_j = \sum_{i=1}^{n} \nu_{j,i} = \boldsymbol{\nu}_j^\top \mathbf{1}_n \quad ,$$

where $\boldsymbol{\nu}_j = (\nu_{j,1}, \ldots, \nu_{j,n})^\top$ is the vector of individual contributions to the j-th component, and $\mathbf{1}_n$ is an n-dimensional vector of ones. This decomposition is fundamental for defining the sign-flip test.

Following Hemerik et al. (2020), the basic score statistic across all observations can be written as:

$$s(\theta) = s_\theta = \left[n^{-1/2} \sum_{i=1}^{n} v_{\theta_1,i}, \ldots, n^{-1/2} \sum_{i=1}^{n} v_{\theta_p,i} \right]^T = \left[s_{\theta_1}, \ldots, s_{\theta_p} \right]^T, \tag{3.1}$$

which implies that the full score vector is scaled by $n^{1/2}$:

$$n^{1/2} s_\theta.$$

3.1.2 Projection-Based Correction of the Score

A central assumption in the sign-flipping method proposed by Hemerik et al. (2020) is the independence of the individual score contributions v_i.

Assumption 3.1. The random variables v_i, $i = 1, \ldots, n$, are mutually independent, have finite second moments, and satisfy the following Lindeberg-type condition: for every $\epsilon > 0$,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[v_i^2 \cdot \mathbb{1}_{\{|v_i| > \epsilon\sqrt{n}\}} \right] = 0.$$

Under Assumption 3.1, and assuming that H_0 is a point null hypothesis, the score contributions v_i satisfy $\mathbb{E}[v_i] = 0$ under H_0.

However, difficulties arise when nuisance parameters are present. Suppose the parameter vector is partitioned as $\theta = (\psi, \lambda)$, where:

- $\psi \in \mathbb{R}$ is the parameter of interest;
- $\lambda = (\lambda_1, \ldots, \lambda_k)^T \in \mathbb{R}^k$ is the vector of nuisance parameters;
- $Y_i \in \mathbb{R}$ are independent and identically distributed according to $p_{Y_i}(y_i; \psi, \lambda)$.

In practice, nuisance parameters must be estimated, commonly via maximum likelihood under H_0. This estimation induces correlation between the score components for ψ and those for λ, thus violating the independence condition in Assumption 3.1 even asymptotically.

To address this issue, Hall & Mathiason (1990) propose the effective score statistic, which removes the linear dependence of s_ψ on s_λ. It is defined as:

$$e_\psi = s_\psi - I_{\psi\lambda} I_{\lambda\lambda}^{-1} s_\lambda, \tag{3.2}$$

where:

- s_ψ is score for ψ;

- s_λ is the k-dimensional score for λ;

- $I_{\psi\lambda}$ is the $1 \times k$ block of the Fisher information matrix representing the cross-information between ψ and λ;

- $I_{\lambda\lambda}$ is the $k \times k$ Fisher information block for λ.

Let us define the following elements, essential for the derivation:

- $\nu_\psi = (\nu_{\psi,1}, \ldots, \nu_{\psi,n}) \in \mathbb{R}^n$: the vector of individual score contributions for ψ;

- $e^*_\psi = (e^*_{\psi,1}, \ldots, e^*_{\psi,n}) \in \mathbb{R}^n$: the vector of individual effective score contributions;

- $\mathbf{N}_\lambda \in \mathbb{R}^{k\times n}$: the matrix of score contributions for λ, where the j-th row contains the contributions $(\nu_{\lambda_j,1}, \ldots, \nu_{\lambda_j,n})$, with $j = 1, \ldots, k$;

- I: the full Fisher information matrix, partitioned as:

$$I = \begin{bmatrix} I_{\psi\psi} & I_{\psi\lambda} \\ I_{\lambda\psi} & I_{\lambda\lambda} \end{bmatrix}.$$

Conceptually, computing the effective score amounts to removing the projection of $\nu_\psi^\top$ onto the column space of $\mathbf{N}_\lambda^\top$. That is,

$$\begin{aligned} {e^*_\psi}^\top &= \nu_\psi^\top - P_{\mathbf{N}_\lambda^\top}(\nu_\psi^\top) \\ &= \nu_\psi^\top - \mathbf{N}_\lambda^\top(\mathbf{N}_\lambda\mathbf{N}_\lambda^\top)^{-1}\mathbf{N}_\lambda\nu_\psi^\top. \end{aligned}$$

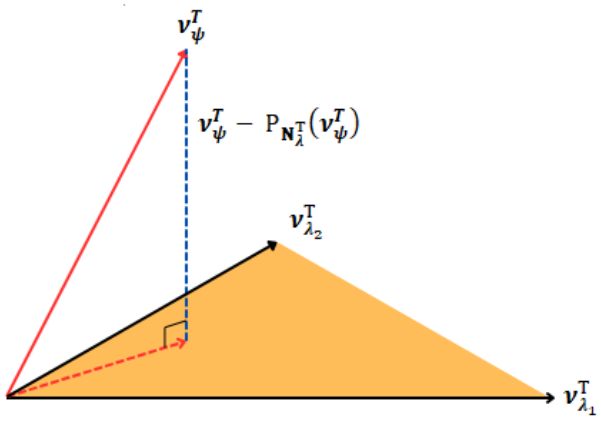

Figure 3.1: Geometrical representation of the projection of $\nu_\psi^\top$ onto the nuisance score space. The residual vector defines the effective score.

Recalling that an empirical covariance matrix estimator for a centered matrix $M \in \mathbb{R}^{n\times p}$ is $\widehat{\mathrm{Cov}}(M) = \frac{1}{n}M^\top M$, we approximate (Scott, 2002):

$$\widehat{I}_{\lambda\lambda} = \frac{1}{n}\mathbf{N}_\lambda\mathbf{N}_\lambda^\top, \quad \widehat{I}_{\lambda\psi} = \frac{1}{n}\mathbf{N}_\lambda\nu_\psi^\top.$$

Hence,

$$\begin{aligned} e_\psi^{*\top} &= \nu_\psi^\top - \mathbf{N}_\lambda^\top(n\widehat{I}_{\lambda\lambda})^{-1}n\widehat{I}_{\lambda\psi} \\ &= \nu_\psi^\top - \mathbf{N}_\lambda^\top\widehat{I}_{\lambda\lambda}^{-1}\widehat{I}_{\lambda\psi}. \end{aligned}$$

Taking the transpose and using symmetry of the Fisher matrix:

$$e_\psi^* = \nu_\psi - \widehat{I}_{\psi\lambda}\widehat{I}_{\lambda\lambda}^{-1}\mathbf{N}_\lambda.$$

Finally, summing over observations and normalizing:

$$\begin{aligned} \sum_{i=1}^n e_{\psi,i}^* &= \sum_{i=1}^n \nu_{\psi,i} - \widehat{I}_{\psi\lambda}\widehat{I}_{\lambda\lambda}^{-1}\sum_{i=1}^n \nu_{\lambda,i}, \\ n^{-1/2}\sum_{i=1}^n e_{\psi,i}^* &= n^{-1/2}\left(\sum_{i=1}^n \nu_{\psi,i} - \widehat{I}_{\psi\lambda}\widehat{I}_{\lambda\lambda}^{-1}\sum_{i=1}^n \nu_{\lambda,i}\right). \end{aligned}$$

Thus, using Equation (3.1), we recover the effective score statistic in Equation (3.2):

$$e_\psi = s_\psi - \widehat{I}_{\psi\lambda}\widehat{I}_{\lambda\lambda}^{-1}s_\lambda.$$

3.2 Effective Score Formulation in GLMs

In this section, we derive the effective score in the context of Generalized Linear Models, following the construction introduced in De Santis et al. (2025).

Let $\{Y_i\}_{i=1}^n$ be independent random variables whose distributions belong to an exponential dispersion family. Each observation admits a density of the form:

$$p(y_i;\theta_i,\phi_i) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{a_i(\phi_i)} + c(y_i,\phi_i)\right\}, \tag{3.3}$$

where:

- $y_i \in \mathcal{Y} \subseteq \mathbb{R}$ is the observed response;
- $\theta_i \in \Theta \subseteq \mathbb{R}$ is the canonical parameter;
- $\phi_i > 0$ is the dispersion parameter;

- $a_i(\phi_i)$, $b(\cdot)$, and $c(\cdot)$ are known functions defining the family.

Under standard regularity conditions (Azzalini, 1996, Chapter 3), and assuming $a_i(\phi_i) = \phi_i$, the first two moments of Y_i are:

$$\mathbb{E}[Y_i] = \mu_i = b'(\theta_i), \qquad Var(Y_i) = b''(\theta_i) \cdot \phi_i.$$

We assume that each μ_i depends on a set of explanatory variables $(\boldsymbol{x}_i, \boldsymbol{z}_i)$ through a link function $g(\cdot)$ such that:

$$g(\boldsymbol{\mu}) = \boldsymbol{\eta} = \boldsymbol{x}\beta + \boldsymbol{Z}\boldsymbol{\gamma} \quad ,$$

where:

- $\boldsymbol{x} = (x_1, \ldots, x_n)^\top \in \mathbb{R}^n$ is the covariate for the parameter of interest $\beta \in \mathbb{R}$;
- $\boldsymbol{Z} \in \mathbb{R}^{n \times k}$ is the design matrix of covariates associated with nuisance parameters $\boldsymbol{\gamma} \in \mathbb{R}^k$;
- $\boldsymbol{\eta} \in \mathbb{R}^n$ is the linear predictor;
- $\boldsymbol{\mu} \in \mathbb{R}^n$ is vector of the means;
- $g(\cdot)$ is the link function.

The dispersion parameters ϕ_i are typically nuisance components. Estimating them is challenging: unless specific structural assumptions are satisfied, consistent estimation is not possible (Neyman & Scott, 1948). Nevertheless, under the assumption that the link function $g(\cdot)$ is correctly specified, the estimators of β and $\boldsymbol{\gamma}$ remain consistent even when ϕ_i is misestimated.

Let us now compute the score for β from the log-likelihood function. For a single observation i, we define:

$$l_i(\boldsymbol{\theta}) = \frac{y_i\theta_i - b(\theta_i)}{\phi_i} + c(y_i, \phi_i).$$

Using the chain rule, we compute:

$$\begin{aligned}
\frac{\partial l_i(\boldsymbol{\theta})}{\partial \beta} &= \frac{\partial l_i}{\partial \theta_i} \cdot \frac{\partial \theta_i}{\partial \mu_i} \cdot \frac{\partial \mu_i}{\partial \eta_i} \cdot \frac{\partial \eta_i}{\partial \beta} \\
&= \frac{1}{\phi_i}(y_i - \mu_i) \cdot \frac{\phi_i}{Var(Y_i)} \cdot \frac{\partial \mu_i}{\partial \eta_i} \cdot x_i \\
&= \frac{(y_i - \mu_i)}{Var(Y_i)} \cdot \frac{\partial \mu_i}{\partial \eta_i} \cdot x_i.
\end{aligned}$$

Thus, summing over all observations, we obtain the full score:

$$\nu_\beta = \sum_{i=1}^{n} \frac{(y_i - \mu_i)}{\mathrm{Var}(Y_i)} \cdot \frac{\partial \mu_i}{\partial \eta_i} \cdot x_i = \boldsymbol{x}^\top \boldsymbol{D} \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{\mu}), \tag{3.4}$$

where:

$$\underset{n\times n}{\boldsymbol{D}} = \mathrm{diag}\left\{\frac{\partial \mu_i}{\partial \eta_i}\right\} \quad \text{and} \quad \underset{n\times n}{\boldsymbol{V}} = \mathrm{diag}\{\mathrm{Var}(y_i)\}, \qquad i = 1, \ldots, n.$$

Analogously, the score for the nuisance parameters is:

$$\boldsymbol{\nu}_\gamma = \boldsymbol{Z}^\top \boldsymbol{D} \boldsymbol{V}^{-1} (\boldsymbol{y} - \boldsymbol{\mu}). \tag{3.5}$$

We now compute the Fisher information blocks required for the effective score in Equation (3.2). Focusing on the mixed term $\boldsymbol{I}_{\beta\gamma_j}$:

$$\begin{aligned}
\boldsymbol{I}_{\beta,\gamma_j} &= \mathbb{E}\left[-\sum_{i=1}^{n} \frac{\partial^2 l_i(\boldsymbol{\theta})}{\partial \gamma_j \partial \beta}\right] \\
&= \sum_{i=1}^{n} \mathbb{E}\left[\frac{\partial l_i(\boldsymbol{\theta})}{\partial \gamma_j} \cdot \frac{\partial l_i(\boldsymbol{\theta})}{\partial \beta}\right] \\
&= \sum_{i=1}^{n} \mathbb{E}\left[\left(\frac{(y_i - \mu_i)}{\mathrm{Var}(y_i)} \cdot \frac{\partial \mu_i}{\partial \eta_i} \cdot z_{ij}\right) \cdot \left(\frac{(y_i - \mu_i)}{\mathrm{Var}(y_i)} \cdot \frac{\partial \mu_i}{\partial \eta_i} \cdot x_i\right)\right] \\
&= \sum_{i=1}^{n} \mathbb{E}\left[\frac{(y_i - \mu_i)^2}{\mathrm{Var}(y_i)^2} \cdot \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 \cdot z_{ij} x_i\right] \\
&= \sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 \frac{z_{ij} x_i}{\mathrm{Var}(y_i)^2} \mathbb{E}\left[(y_i - \mu_i)^2\right] \\
&= \sum_{i=1}^{n} \left(\frac{\partial \mu_i}{\partial \eta_i}\right)^2 \frac{z_{ij} x_i}{\mathrm{Var}(y_i)} \\
&= \boldsymbol{x}^\top \boldsymbol{D} \boldsymbol{V}^{-1} \boldsymbol{D} \boldsymbol{Z} = \boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{Z}.
\end{aligned}$$

The full matrix form is:

$$\boldsymbol{I}(\beta, \boldsymbol{\gamma}) = \begin{bmatrix} \boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{x} & \boldsymbol{x}^\top \boldsymbol{W} \boldsymbol{Z} \\ \boldsymbol{Z}^\top \boldsymbol{W} \boldsymbol{x} & \boldsymbol{Z}^\top \boldsymbol{W} \boldsymbol{Z} \end{bmatrix}, \tag{3.6}$$

where $\boldsymbol{W} = \boldsymbol{D} \boldsymbol{V}^{-1} \boldsymbol{D}$.

Inserting (3.4), (3.5) and (3.6) into the effective score formula, we obtain:

$$
\begin{aligned}
e_\beta &= s_\beta - \boldsymbol{I}_{\beta\gamma}\boldsymbol{I}_{\gamma\gamma}^{-1}\boldsymbol{s}_\gamma \\
&= n^{-1/2}(\nu_\beta - \boldsymbol{I}_{\beta\gamma}\boldsymbol{I}_{\gamma\gamma}^{-1}\boldsymbol{\nu}_\gamma) \\
&= n^{-1/2}[\boldsymbol{x}^\top \boldsymbol{D}\boldsymbol{V}^{-1}(\boldsymbol{y}-\hat{\boldsymbol{\mu}}) - \boldsymbol{x}^\top \boldsymbol{W}\boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{W}\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top \boldsymbol{D}\boldsymbol{V}^{-1}(\boldsymbol{y}-\hat{\boldsymbol{\mu}})] \\
&= n^{-1/2}[\boldsymbol{x}^\top \boldsymbol{D}\boldsymbol{V}^{-1/2}\boldsymbol{V}^{-1/2}(\boldsymbol{y}-\hat{\boldsymbol{\mu}}) - \boldsymbol{x}^\top \boldsymbol{W}^{1/2}\boldsymbol{W}^{1/2}\boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{W}\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top \boldsymbol{D}\boldsymbol{V}^{-1/2}\boldsymbol{V}^{-1/2}(\boldsymbol{y}-\hat{\boldsymbol{\mu}})] \\
&= n^{-1/2}[\boldsymbol{x}^\top \boldsymbol{W}^{1/2}\boldsymbol{V}^{-1/2}(\boldsymbol{y}-\hat{\boldsymbol{\mu}}) - \boldsymbol{x}^\top \boldsymbol{W}^{1/2}\boldsymbol{W}^{1/2}\boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{W}\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top \boldsymbol{W}^{1/2}\boldsymbol{V}^{-1/2}(\boldsymbol{y}-\hat{\boldsymbol{\mu}}) \\
&= n^{-1/2}\boldsymbol{x}^\top \boldsymbol{W}^{1/2}(\boldsymbol{I} - \boldsymbol{W}^{1/2}\boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{W}\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top \boldsymbol{W}^{1/2})\boldsymbol{V}^{-1/2}(\boldsymbol{y}-\hat{\boldsymbol{\mu}}) \\
&= n^{-1/2}\boldsymbol{x}^\top \boldsymbol{W}^{1/2}(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{V}^{-1/2}(\boldsymbol{y}-\hat{\boldsymbol{\mu}}), \qquad (3.7)
\end{aligned}
$$

where:

- $\boldsymbol{H} = \boldsymbol{W}^{1/2}\boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{W}\boldsymbol{Z})^{-1}\boldsymbol{Z}^\top \boldsymbol{W}^{1/2}$ is the hat matrix associated with nuisance covariates;

- $\hat{\boldsymbol{\mu}}$ are the fitted means under $H_0 : \beta = \beta_0$.

Equation (3.7) provides the exact expression of the effective score in GLMs, accounting for nuisance correction via projection on the nuisance score space.

3.3 Inference via Sign Flipping

In recent years, sign-flipping and permutation-based methods have gained attention for hypothesis testing (Pesarin, 2001). These approaches are particularly useful in regression contexts where standard parametric assumptions may not hold.

In classical settings, permutation tests often require exchangeability or residual symmetry (Winkler et al., 2014). For generalized linear models, which lack these properties due to nonlinear link functions or heteroscedasticity, the approach proposed by Hemerik et al. (2020) introduces a powerful alternative: testing via random sign-flips of the individual score contributions. This method operates directly on the additive decomposition of the score statistic, relying on the property that, under the null, the mean of each score contribution is zero. Specifically, it is assumed that the ν_i satisfy Assumption 3.1, which ensures the test to be asymptotically exact.

3.3.1 Test Statistic via Sign-Flipping

The idea is to compute a test statistic based on randomly flipping the signs of the individual score contributions. Let:

$$\boldsymbol{\nu} = (\nu_1, \ldots, \nu_n)^\top,$$

be the vector of score contributions (e.g., basic scores or effective score). For the unflipped data, define the original statistic as:

$$T_1^n = n^{-1/2} \sum_{i=1}^{n} \nu_i.$$

Let $g_1 = (1, \ldots, 1) \in \mathbb{R}^n$ be the unit vector (no flipping), and let $g_2, \ldots, g_w \in \{-1, 1\}^n$ be $w - 1$ independent sign vectors sampled uniformly at random. Then, for $j = 1, \ldots, w$, define:

$$T_j^n = n^{-1/2} \sum_{i=1}^{n} g_{ji} \nu_i.$$

Let $T^n_{[1-\alpha]}$ denote the empirical $(1-\alpha)$-quantile of the collection $\{T_j^n\}_{j=1}^w$. Then, a level-α test rejects the null hypothesis if and only if :

$$T_1^n > T^n_{[1-\alpha]}.$$

The key theoretical justification is given by the following lemma and theorem from Hemerik et al. (2020). For the proofs, please refer to the reference article just cited.

Lemma 3.1. *Suppose that, as $n \to \infty$, the vector $(T_1^n, \ldots, T_W^n)$ converges in distribution to a vector of i.i.d. continuous random variables. Then:*

$$\mathbb{P}\left(T_1^n > T^n_{[1-\alpha]}\right) \to \frac{\lfloor \alpha w \rfloor}{w}.$$

Theorem 3.2. *Under Assumption 3.1 and standard regularity conditions, the test based on T_1^n and the sign-flipped reference distribution satisfies:*

$$\lim_{n\to\infty} \mathbb{P}\left(T_1^n > T^n_{[1-\alpha]}\right) = \frac{\lfloor \alpha w \rfloor}{w}.$$

Moreover, the statistics T_j^n are asymptotically independent and Gaussian with mean zero and variance $\lim_{n\to\infty} s^2$.

This result ensures asymptotic control of the type I error rate at the nominal level, even when the data are not continuous, since the probability of ties vanishes asymptotically:

$$\lim_{n\to\infty} \mathbb{P}(T_j^n = T_l^n) = 0, \quad \forall j \neq l.$$

3.3.2 Implementation of Flip Tests in GLMs

In De Santis et al. (2025), a matrix formulation of the sign-flipping test is presented for GLMs. Let $\mathcal{F}$ be a diagonal matrix of random signs:

$$\mathcal{F} = \text{diag}(f_1, \ldots, f_n), \quad f_i \overset{\text{i.i.d.}}{\sim} \text{Uniform}\{-1, +1\}.$$

Then, the sign-flipped version of the effective score for a given flip matrix $\mathcal{F} = \boldsymbol{F}$ becomes:

$$e_\beta(\boldsymbol{F}) = n^{-1/2} \boldsymbol{x}^\top \boldsymbol{W}^{1/2} (\boldsymbol{I} - \boldsymbol{H}) \boldsymbol{V}^{-1/2} \boldsymbol{F} (\boldsymbol{y} - \hat{\boldsymbol{\mu}}). \tag{3.8}$$

Note that when $\mathcal{F} = \boldsymbol{I}$, we recover the original *effective score* from Equation (3.7).

Under the null hypothesis, the following approximations hold:

$$\mathbb{E}[e_\beta(\boldsymbol{I})] = \mathbb{E}[e_\beta(\mathcal{F})], \qquad Var[e_\beta(\boldsymbol{I})] - Var[e_\beta(\mathcal{F})] \to 0.$$

This formulation provides a robust testing procedure that adapts to the correlation structure induced by the nuisance parameters, while avoiding direct reliance on asymptotic normality of the unstandardized test statistic.

3.4 Standardized Effective Flip Score

As shown in Hemerik et al. (2020), the test statistic based on the sign-flipped score contributions is asymptotically valid under mild assumptions. However, in finite samples, its distribution can significantly differ from the reference distribution obtained by random sign-flipping. In particular, De Santis et al. (2025) observe a lack of second-moment null-invariance, a property that would ensure that both the original and flipped statistics have the same variance under H_0.

Although $\mathbb{E}[e_\beta(\boldsymbol{I})] = \mathbb{E}[e_\beta(\mathcal{F})]$ under H_0, in finite samples it holds that:

$$Var[e_\beta(\boldsymbol{I})] > Var[e_\beta(\mathcal{F})].$$

This variance inflation leads to anti-conservativeness: extreme values of the observed statistic $e_\beta(\boldsymbol{I})$ are more likely than those generated under the null distribution from the sign-flips, thus increasing the probability of false positives.

To correct for this mismatch, De Santis et al. (2025) introduce a *standardized* test statistic, where each flipped score is divided by its estimated standard deviation, conditional on the applied flip. This restores second-moment null-invariance and improves finite-sample calibration.

The key to standardization lies in the computation of the conditional variance of $e_\beta(\mathcal{F})$, given the flip matrix $\mathcal{F}$. The following result is central:

Lemma 3.2 (De Santis et al. (2025), Lemma 5)**.** *The variance of the sign-flipped score is given by:*

$$\mathrm{Var}[e_\beta(\boldsymbol{F})] = \frac{1}{n}\boldsymbol{x}^\top \boldsymbol{W}^{1/2}(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{F}(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{F}(\boldsymbol{I}-\boldsymbol{H})\boldsymbol{W}^{1/2}\boldsymbol{x} + o_p(1).$$

We now define the standardized test statistic as:

$$e^*_\beta(\boldsymbol{F}) = \frac{e_\beta(\boldsymbol{F})}{\mathrm{Var}[e_\beta(\boldsymbol{F})]^{1/2}}.$$

This statistic has mean zero and, by construction, unit variance up to $o_p(1)$ corrections. Importantly, standardization is performed conditionally for each sign-flip configuration. The resulting test uses $e^*_\beta(\boldsymbol{F})$ in the same way as $e_\beta(\boldsymbol{F})$, comparing the unflipped standardized score to its empirical distribution under random flips.

As a result, the standardized sign-flip test retains the advantages of its unstandardized counterpart, robustness to variance misspecification, flexibility in GLMs, and general applicability, while achieving better finite-sample type I error control.

Chapter 4

Sign-Flip approaches for High-Dimensional Linear Models

4.1 Introduction and Motivation

The ultimate goal of this thesis is to provide valid inference for all parameters of the high-dimensional linear model. As a first step, in this chapter we focus on testing a single regression coefficient at a time, denoted by β and associated with the covariate of interest $\boldsymbol{x}$. The remaining covariates are treated as potential confounders $\boldsymbol{Z}$ with coefficients γ. Establishing rigorous inference for β in the presence of high-dimensional $\boldsymbol{Z}$ is a crucial building block towards simultaneous inference on the full parameter vector.

Consider the high-dimensional linear model

$$\boldsymbol{Y} = \boldsymbol{x}\beta + \boldsymbol{Z}\gamma + \varepsilon, \qquad \varepsilon \sim N_n(\mathbf{0}, \sigma^2 \boldsymbol{I}), \tag{4.1}$$

where $\boldsymbol{Y} \in \mathbb{R}^n$, $\boldsymbol{x} \in \mathbb{R}^n$ is the variable of interest with parameter $\beta \in \mathbb{R}$, and $\boldsymbol{Z} \in \mathbb{R}^{n \times d}$ are the confounders with coefficients $\gamma \in \mathbb{R}^d$. We focus on $d > n$, so that the usual projection matrix $\boldsymbol{H} = \boldsymbol{Z}(\boldsymbol{Z}^\top \boldsymbol{Z})^{-1}\boldsymbol{Z}^\top$ is not defined.

Two broad strategies can restore feasibility:

1. Penalized inversion (e.g., ridge): add a positive constant to the diagonal of $\boldsymbol{Z}^\top \boldsymbol{Z}$ so that the inverse exists. This induces bias and typically calls for debiasing steps.

2. Selection of a subset of confounders: pick a subset of columns of $\boldsymbol{Z}$ so that the projection is well-defined and the test on β has enough degrees of freedom.

We follow the second route, coupled with sign-flip inference, which is robust to model misspecification and non-Gaussianity under invariance assumptions (see Chapter 3). However, selection introduces two classical concerns:

- Screening property: the selected set should contain all truly active confounders for $\boldsymbol{Y}$ (and, in a relaxed form, those confounders that drive the association between $\boldsymbol{Y}$ and $\boldsymbol{x}$).
- Double dipping: using the same data to select and then to test may inflate Type I error (cf. Chapter 2, Post-Selection Problem).

A natural remedy to double dipping is the multi-split approach (Meinshausen et al., 2009), which splits the sample multiple times, aggregates p-values, and guarantees family-wise or false discovery control under suitable conditions. We do not adopt multi-split here for two reasons. First, in small samples the repeated splitting substantially reduces effective power; second, the aggregation step can be conservative when the signal is sparse and weak. Instead, we leverage a key observation confirmed by our empirical evidence: if the variable of interest $\boldsymbol{x}$ is not used in the selection of confounders and the screening property holds, standard parametric tests (Wald, LR) and, in particular, our sign-flip test can retain Type I error control. While we do not provide a full proof for all procedures in this chapter, this guiding principle motivates our constructions and is supported by simulations.

A second practical constraint concerns the rank of the design matrix in the final low-dimensional regression used to build the sign-flip statistic. Suppose n observations are available and, after selection, M confounders are included together with an intercept and the variable of interest. In the final model, the number of parameters to be estimated is $M + 2$ (confounders and intercept plus the coefficient β), in addition to the residual variance σ^2. For the projection matrix $(\boldsymbol{I} - \boldsymbol{H})$ to be well-defined and leave a non-degenerate residual space, we require that

$$M \leqslant n - 3. \tag{4.2}$$

This ensures that the residual space has dimension at least one (e.g., with $n = 50$ and $M = 47$, the residual dimension is $50 - (47 + 1 + 1) = 1$). This rank condition is essential for the validity and numerical stability of the sign-flip procedure. We propose three concrete procedures that combine selection and sign-flip inference: lasso flipscores, SVD-lasso flipscores, and stepwise flipscores.

For each method we state assumptions, the testing algorithm, and a conjecture on asymptotic validity. Detailed proofs will be provided in Corbetta (2025).

Throughout, we let $\mathcal{D} = \{1, \dots, d\}$ be the index set of confounders, $\mathcal{D}_1 = \{j \in \mathcal{D} : \gamma_j \neq 0\}$ the truly active confounders for $\boldsymbol{Y}$, and (when useful) $\mathcal{D}_1^x \subseteq \mathcal{D}$ the subset of confounders truly associated with $\boldsymbol{x}$.

4.2 Selection of Confounders via Lasso

Lasso is a natural choice to enforce sparsity and reduce dimensionality before testing. In our setting we exclude the variable of interest $\boldsymbol{x}$ from the selection step, to mitigate post-selection bias on the target parameter β, and then test β in the reduced model via a sign-flip statistic. This design aligns with the "no double use of $\boldsymbol{x}$" insight discussed above. Nevertheless, lasso selection may be unstable with highly correlated confounders, and its active set can be sensitive to the penalty level λ; these limitations motivate the SVD-based variant in Section 4.3.

Assumption 4.1 (Screening property)**.** Asymptotically, all truly active confounders for $\boldsymbol{Y}$ are selected:

$$\lim_{n\to\infty} \mathbb{P}\big(\mathcal{D}_1 \subseteq \mathcal{A}\big) = 1,$$

where $\mathcal{A}$ is the random active set returned by the selection procedure. A relaxed version replaces $\mathcal{D}_1$ with $\mathcal{D}_1 \cap \mathcal{D}_1^x$.

Assumption 4.2 (Lasso regularity)**.** Let $\mathcal{A}_\lambda = \{j : \hat{\gamma}_{\lambda,j} \neq 0\}$ denote the noiseless-lasso (van de Geer & Bühlmann, 2009) active set, where the noiseless-lasso is defined as

$$\hat{\gamma}_\lambda = \arg\min_\gamma \frac{1}{2n}\mathbb{E}\left[\|\boldsymbol{Y} - \boldsymbol{Z}\gamma\|_2^2\right] + \lambda\|\gamma\|_1, \tag{4.3}$$

as in Zhao et al. (2021, Equation 1.7). We require mild regularity to apply a Lindeberg–Feller argument to the sign-flip statistic.

1. (Small omitted signal) $\|\boldsymbol{Z}_{\mathcal{A}_\lambda^c}\gamma_{\mathcal{A}_\lambda^c}\|_2 = o(1)$.
2. (Lindeberg condition) With $r(\boldsymbol{F}) = (\boldsymbol{I} - \boldsymbol{H}_{\mathcal{A}_\lambda})\boldsymbol{F}(\boldsymbol{I} - \boldsymbol{H}_{\mathcal{A}_\lambda})\boldsymbol{x}$,

$$\lim_{n\to\infty} \frac{\|r(\boldsymbol{F})\|_\infty}{\|r(\boldsymbol{F})\|_2} = 0.$$

Algorithm. Let $\boldsymbol{x} = \boldsymbol{Q}_j$ be the variable of interest, $\beta = \delta_j$, and $\boldsymbol{Z} = \boldsymbol{Q}_{-j}$ the confounders ($d = p - 1$). Fix M satisfying constraint (4.2) and fix the number of sign-flips B.

Algorithm 4.1. **Lasso Flipscores for a single variable of interest**

Input: $\boldsymbol{y} \in \mathbb{R}^n$; design $\boldsymbol{Q} = [\boldsymbol{x}\ \boldsymbol{Z}]$.

Initialization: Generate $\boldsymbol{F}_1, \ldots, \boldsymbol{F}_B$ with $\boldsymbol{F}_1 = \boldsymbol{I}_n$.

1. Solve the lasso on $\boldsymbol{Z}$ (exclude $\boldsymbol{x}$):

$$\hat{\gamma}_\lambda = \arg\min_\gamma \frac{1}{n}\|\boldsymbol{y} - \boldsymbol{Z}\gamma\|_2^2 + \lambda\|\gamma\|_1.$$

2. Choose $\hat{\lambda}$ so that exactly M variables are active; let $\boldsymbol{Z}_{\hat{\lambda}}$ be the selected columns and $\boldsymbol{H}_{\boldsymbol{Z}_{\hat{\lambda}}} = \boldsymbol{Z}_{\hat{\lambda}}(\boldsymbol{Z}_{\hat{\lambda}}^T \boldsymbol{Z}_{\hat{\lambda}})^{-1}\boldsymbol{Z}_{\hat{\lambda}}^T$.

3. For each $b = 1, \ldots, B$, compute the standardized statistic

$$e_\beta^{lasso}(\boldsymbol{F}_b) = \frac{\boldsymbol{x}^T(\boldsymbol{I} - \boldsymbol{H}_{\boldsymbol{Z}_{\hat{\lambda}}})\boldsymbol{F}_b(\boldsymbol{I} - \boldsymbol{H}_{\boldsymbol{Z}_{\hat{\lambda}}})\boldsymbol{y}}{\sqrt{\mathrm{Var}\left[\boldsymbol{x}^T(\boldsymbol{I} - \boldsymbol{H}_{\boldsymbol{Z}_{\hat{\lambda}}})\boldsymbol{F}_b(\boldsymbol{I} - \boldsymbol{H}_{\boldsymbol{Z}_{\hat{\lambda}}})\boldsymbol{y}\right]}}.$$

4. Collect the vector of test statistics

$$\left(e_\beta^{lasso}(\boldsymbol{F}_1),\ e_\beta^{lasso}(\boldsymbol{F}_2),\ \ldots,\ e_\beta^{lasso}(\boldsymbol{F}_B)\right),$$

and compute the permutation p-value

$$p_\beta^{lasso} = \frac{1}{B}\sum_{b=1}^{B} \mathbb{1}\left\{|e_\beta^{lasso}(\boldsymbol{F}_1)| \geqslant |e_\beta^{lasso}(\boldsymbol{F}_b)|\right\}.$$

Conjecture 4.1. *If Assumptions 4.1 and 4.2 hold, then*

$$e_\beta^{lasso}(\boldsymbol{F}) \xRightarrow{d} N(0,1),$$

and the test that rejects $H_0 : \beta = 0$ *when* $e_\beta^{lasso}(\boldsymbol{F}_1)$ *exceeds the* $(1-\alpha)$ *empirical quantile of* $\{e_\beta^{lasso}(\boldsymbol{F}_b)\}_{b=1}^B$ *is an asymptotically valid* α*-level test.*

The lasso may fail to screen under strong collinearity, leading to either omitted active confounders (inflating Type I error) or unnecessary inclusions (reducing power). The sign-flip framework is robust to mild model deviations, but the maximum number of variables constraint (4.2) must be enforced. Choice of M trades power (smaller M) against error control (larger M).

4.3 Selection of Confounders via SVD-Lasso

The instability of lasso under collinearity is particularly problematic in our context. Since the lasso objective allows some bias in exchange for variance reduction, it may entirely exclude variables that are strongly correlated with each other—even if both are active. In classical prediction this is acceptable, but for our sign-flip inference procedure we require an unbiased representation of $\boldsymbol{Y}$ under H_0. Any distortion in the residual structure undermines the validity of the test.

To mitigate this, we reparameterize the confounder space through the Singular Value Decomposition $\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$, obtaining orthogonal left singular vectors in $\boldsymbol{U}$. Writing $\boldsymbol{\xi} = \boldsymbol{D}\boldsymbol{V}^\top\gamma$, we consider

$$\boldsymbol{Y} = \boldsymbol{x}\beta + \boldsymbol{U}\boldsymbol{\xi} + \varepsilon. \tag{4.4}$$

Selection is then performed on columns of $\boldsymbol{U}$, which are orthonormal and often yield more stable active sets. A structural caveat is that the signal of any single original variable in $\boldsymbol{Z}$ is spread across multiple components, so an exact screening property on the original scale is not attainable. Nonetheless, if the selected components span the relevant confounding subspace well enough, the resulting sign-flip test behaves favorably in finite samples.

Assumption 4.3 (SVD-Lasso regularity)**.** Let $\mathcal{A}_\lambda^{\mathrm{svd}}$ denote the active set returned by noiseless-lasso on $\boldsymbol{U}$. Assume:

1. (Selection consistency on components) $\lim_{n\to\infty} \mathbb{P}(\hat{\mathcal{A}}_\lambda^{\mathrm{svd}} = \mathcal{A}_\lambda^{\mathrm{svd}}) = 1$.

2. (Small omitted signal in component space) $\|\boldsymbol{U}_{(\mathcal{A}_\lambda^{\mathrm{svd}})^c}\boldsymbol{\xi}_{(\mathcal{A}_\lambda^{\mathrm{svd}})^c}\|_2 = o(1)$.

3. (Lindeberg condition) With $r^{\mathrm{svd}}(\boldsymbol{F}) = (\boldsymbol{I} - \boldsymbol{H}_{\mathcal{A}_\lambda^{\mathrm{svd}}})\boldsymbol{F}(\boldsymbol{I} - \boldsymbol{H}_{\mathcal{A}_\lambda^{\mathrm{svd}}})\boldsymbol{x}$,

$$\lim_{n\to\infty} \frac{\|r^{\mathrm{svd}}(\boldsymbol{F})\|_\infty}{\|r^{\mathrm{svd}}(\boldsymbol{F})\|_2} = 0.$$

Algorithm. Fix M satisfying (4.2) and B permutations.

Algorithm 4.2. **SVD-Lasso Flipscores**

Input: $\boldsymbol{y} \in \mathbb{R}^n$; $\boldsymbol{Q} = [\boldsymbol{x}\ \boldsymbol{Z}]$ with $\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$,

Initialization: Compute SVD $\boldsymbol{Z} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^\top$; generate $\boldsymbol{F}_1, \ldots, \boldsymbol{F}_B$ with $\boldsymbol{F}_1 = \boldsymbol{I}_n$.

1. Solve lasso on $\boldsymbol{U}$ (exclude $\boldsymbol{x}$):

$$\hat{\boldsymbol{\xi}}_\lambda = \arg\min_{\boldsymbol{\xi}} \frac{1}{n}\|\boldsymbol{y} - \boldsymbol{U}\boldsymbol{\xi}\|_2^2 + \lambda\|\boldsymbol{\xi}\|_1.$$

2. Choose $\hat{\lambda}$ so that exactly M components are selected; let $\boldsymbol{U}_{\hat{\lambda}}$ be these columns, with projection $\boldsymbol{H}_{\boldsymbol{U}_{\hat{\lambda}}}$.

3. For each $b = 1, \ldots, B$, compute

$$e_\beta^{svd}(\boldsymbol{F}_b) = \frac{\boldsymbol{x}^\top(\boldsymbol{I} - \boldsymbol{H}_{\boldsymbol{U}_{\hat{\lambda}}})\boldsymbol{F}_b(\boldsymbol{I} - \boldsymbol{H}_{\boldsymbol{U}_{\hat{\lambda}}})\boldsymbol{y}}{\sqrt{\mathrm{Var}\left[\boldsymbol{x}^\top(\boldsymbol{I} - \boldsymbol{H}_{\boldsymbol{U}_{\hat{\lambda}}})\boldsymbol{F}_b(\boldsymbol{I} - \boldsymbol{H}_{\boldsymbol{U}_{\hat{\lambda}}})\boldsymbol{y}\right]}}.$$

4. Collect

$$\left(e_\beta^{svd}(\boldsymbol{F}_1), \ldots, e_\beta^{svd}(\boldsymbol{F}_B)\right),$$

and compute the p-value as in Algorithm 4.1.

Conjecture 4.2. *If Assumption 4.3 holds, then*

$$e_\beta^{svd}(\boldsymbol{F}) \overset{d}{\Rightarrow} N(0,1),$$

and the permutation test is asymptotically valid.

The SVD-lasso gains stability in collinear settings, at the cost of an imperfect screening property in the original variable space. Finite-sample performance is encouraging in simulations, though a full theoretical justification remains open.

4.4 Selection of Confounders via Stepwise Procedure

While stepwise regression is often criticized for its greedy nature and potential overfitting, it remains popular in applied fields. Here we adapt it to the flipscores framework, using AIC to penalize complexity. The procedure iteratively adds or removes covariates to minimize AIC until reaching exactly M selected variables, excluding the variable of interest.

Algorithm. Fix $M \leqslant n - 3$ and B.

***Algorithm 4.3.* Stepwise Flipscores**

Input: $\boldsymbol{y} \in \mathbb{R}^n$; $\boldsymbol{Q} = [\boldsymbol{x}\ \boldsymbol{Z}]$.

Initialization: Generate $\boldsymbol{F}_1, \ldots, \boldsymbol{F}_B$.

1. Run stepwise selection on $\boldsymbol{Z}$, by minimizing

$$\mathrm{AIC} = 2k + n \log\left(\frac{\|(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y}\|_2^2}{n}\right),$$

   until exactly M covariates are selected. Denote them $\boldsymbol{Z}_{\hat{M}}$ with projection $\boldsymbol{H}_{\boldsymbol{Z}_{\hat{M}}}$.

2. For each $b = 1, \ldots, B$, compute

$$e_\beta^{\mathrm{step}}(\boldsymbol{F}_b) = \frac{\boldsymbol{x}^\top(\boldsymbol{I} - \boldsymbol{H}_{\boldsymbol{Z}_{\hat{M}}})\boldsymbol{F}_b(\boldsymbol{I} - \boldsymbol{H}_{\boldsymbol{Z}_{\hat{M}}})\boldsymbol{y}}{\sqrt{\mathrm{Var}\left[\boldsymbol{x}^\top(\boldsymbol{I} - \boldsymbol{H}_{\boldsymbol{Z}_{\hat{M}}})\boldsymbol{F}_b(\boldsymbol{I} - \boldsymbol{H}_{\boldsymbol{Z}_{\hat{M}}})\boldsymbol{y}\right]}}.$$

3. Collect

$$\left(e_\beta^{\mathrm{step}}(\boldsymbol{F}_1), \ldots, e_\beta^{\mathrm{step}}(\boldsymbol{F}_B)\right),$$

   and compute the p-value as before.

Conjecture 4.3. *We conjecture that if (i) the stepwise procedure asymptotically includes all active confounders with probability tending to one, and (ii) the omitted signal is asymptotically negligible, then the stepwise flipscores test has asymptotically correct level.*

A stronger condition, as a substitute for point (i) in Conjecture 4.3 the stepwise procedure asymptotically includes all active confounders as the first component. This implies that the selection based on the AIC is sign-invariant in ε. However, stepwise lacks the convex-analytic guarantees of lasso, and its theoretical analysis remains challenging; we leave this for future work (Corbetta, 2025).

4.5 Multiple Testing Extension

The flipscores framework naturally extends to simultaneous testing of all regression coefficients δ_j in the full high-dimensional model. For each $j = 1, \ldots, p$, we treat $\boldsymbol{Q}_j$ as the variable of interest and $\boldsymbol{Q}_{-j}$ as confounders, apply one of the selection routines above, and compute test statistics. To control the family-wise error rate, we adopt the MaxT correction.

Algorithm 4.4. **Multiple testing extension with MaxT correction**

Input: Response $\boldsymbol{y}$; design $\boldsymbol{Q} = [\boldsymbol{Q}_1, \ldots, \boldsymbol{Q}_p]$; chosen selection routine; model size $M \leqslant n - 3$; sign-flips B.

Initialization: Generate $\boldsymbol{F}_1, \ldots, \boldsymbol{F}_B$.

1. For each $j = 1, \ldots, p$:

   (a) Apply the selected procedure (lasso, SVD-lasso, or stepwise) with $\boldsymbol{x} = \boldsymbol{Q}_j$.

   (b) Obtain the vector
   $$\big(e^j_\beta(\boldsymbol{F}_1),\, e^j_\beta(\boldsymbol{F}_2),\, \ldots,\, e^j_\beta(\boldsymbol{F}_B)\big).$$

2. Stack all results into the $B \times p$ matrix
   $$\begin{bmatrix} e^1_\beta(\boldsymbol{F}_1) & e^2_\beta(\boldsymbol{F}_1) & \ldots & e^p_\beta(\boldsymbol{F}_1) \\ e^1_\beta(\boldsymbol{F}_2) & e^2_\beta(\boldsymbol{F}_2) & \ldots & e^p_\beta(\boldsymbol{F}_2) \\ \vdots & \vdots & \ddots & \vdots \\ e^1_\beta(\boldsymbol{F}_B) & e^2_\beta(\boldsymbol{F}_B) & \ldots & e^p_\beta(\boldsymbol{F}_B) \end{bmatrix}.$$

3. For each j, compute the adjusted p-value
   $$p^{adj}_j = \frac{1}{B} \sum_{b=1}^{B} \mathbb{1}\left\{ |e^j_\beta(\boldsymbol{F}_1)| \leqslant \max_{1 \leqslant \ell \leqslant p} |e^\ell_\beta(\boldsymbol{F}_b)| \right\}.$$

The use of a common set of sign-flip matrices ensures the joint null distribution of all test statistics is preserved. This permits strong control of the family-wise error rate via MaxT, analogously to permutation-based multiple testing in classical designs.

4.6 Discussion

We have proposed three variants of the flipscores procedure for high-dimensional settings: lasso, SVD-lasso, and stepwise. Each balances the screening property and practical feasibility differently. Our conjectures suggest that, under suitable conditions, the resulting sign-flip tests are asymptotically valid. Establishing rigorous proofs and optimal choice of M are left to ongoing work. Importantly, the methods are immediately applicable in practice and their empirical performance is explored in Chapter 5.

Chapter 5

Simulation

In this chapter we present the empirical results of the simulation study. We distinguish between two main settings: the case where a single hypothesis is tested and the case where multiple hypotheses are tested simultaneously.

5.1 Single hypothesis testing

The first aim of our simulation study was to evaluate empirically the control of type I error and the power of the competing procedures. We considered the following settings, with two levels of signal-to-noise ratio (SNR), namely $\text{SNR} = \{1, 5\}$:

A. A simulated dataset with $n = 50$, $p = 100$, and a design matrix drawn from a centered multivariate normal distribution with equicorrelated covariance structure, i.e. $\text{Cov}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \rho$, $\forall i \neq j$, $i, j = 1, ..., p$, with $\rho = \{0.3, 0.7\}$;

B. As in A, but with a Toeplitz covariance structure, $\text{Cov}(\boldsymbol{x}_i, \boldsymbol{x}_j) = \rho^{|i-j|}$, $\forall i \neq j$, $i, j = 1, ..., p$.

For every setting we fixed the number of active variables $|\mathcal{D}_1| = 5$, with coefficient values equal to 5. Each configuration was repeated 1000 times. In each simulation we randomly selected one variable from the active set $\mathcal{D}_1$, tested it, and used the result to compute the power. Likewise, we randomly selected a variable from the inactive set to estimate the empirical type I error.

To implement the various methods, we fixed the following parameters:

- debiased lasso (van de Geer et al., 2014; Zhang & Zhang, 2013): we employed the `lasso.proj()` function from the `hdi` (High-Dimensional Inference) R

package (Ruben Dezeure et al., 2015). For the initial coefficient estimation, we used the scaled lasso (Sun & Zhang, 2012);

- Ridge Projection (Bühlmann, 2013): we employed the `ridge.proj()` function from the `hdi` package. As in the previous case, the scaled lasso was used for the initial estimation. For the ridge penalty parameter we set $\lambda_{ridge} = 1/n$, following the recommendation in Bühlmann (2013, Section 5);
- lasso flipscores: we fixed the number of selected variables at $M_{lasso} = 10$;
- SVD-lasso flipscores: we fixed the number of selected components at $M_{SVD-lasso} = 30$;
- stepwise flipscores: we fixed the number of selected variables at $M_{stepwise} = 10$. The selection proceeded in both directions (forward and backward), starting from the null model containing only the intercept.

In the Toeplitz scenario with $\rho = 0.3$ (top row of Figure 5.1), all methods follow the theoretical rejection curve very closely. Even at higher signal-to-noise ratio (SNR=5), the empirical rejection probabilities remain well aligned with the diagonal, indicating correct calibration. Notably, the three flipscores-based procedures behave as expected, with no sign of error inflation or excessive conservativeness, supporting their finite-sample validity beyond the asymptotic arguments developed earlier.

In the equicorrelated design (Figure 5.2), the situation is less favorable. The lasso flipscores departs substantially from the theoretical benchmark, showing severe anti-conservativeness that worsens as correlation increases. This confirms that the instability of lasso-based selection under strong dependence translates into poor inferential calibration.

Both ridge projection and the debiased lasso lose accuracy, tending to become too liberal. By contrast, the stepwise flipscores displays the most reliable alignment with the theoretical curve, suggesting that its selection mechanism is more robust to the dependence structure of the design matrix.

Figure 5.3 zooms in on the nominal level $\alpha = 0.05$, reporting the empirical type I error rates across scenarios. In the Toeplitz design, all methods control the error well, with values close to the nominal threshold across both SNR levels. This indicates that the Toeplitz correlation structure is less challenging for calibration.

The equicorrelated design, instead, highlights more striking differences. Here the lasso flipscores clearly loses control, with type I error rates well above 5%,

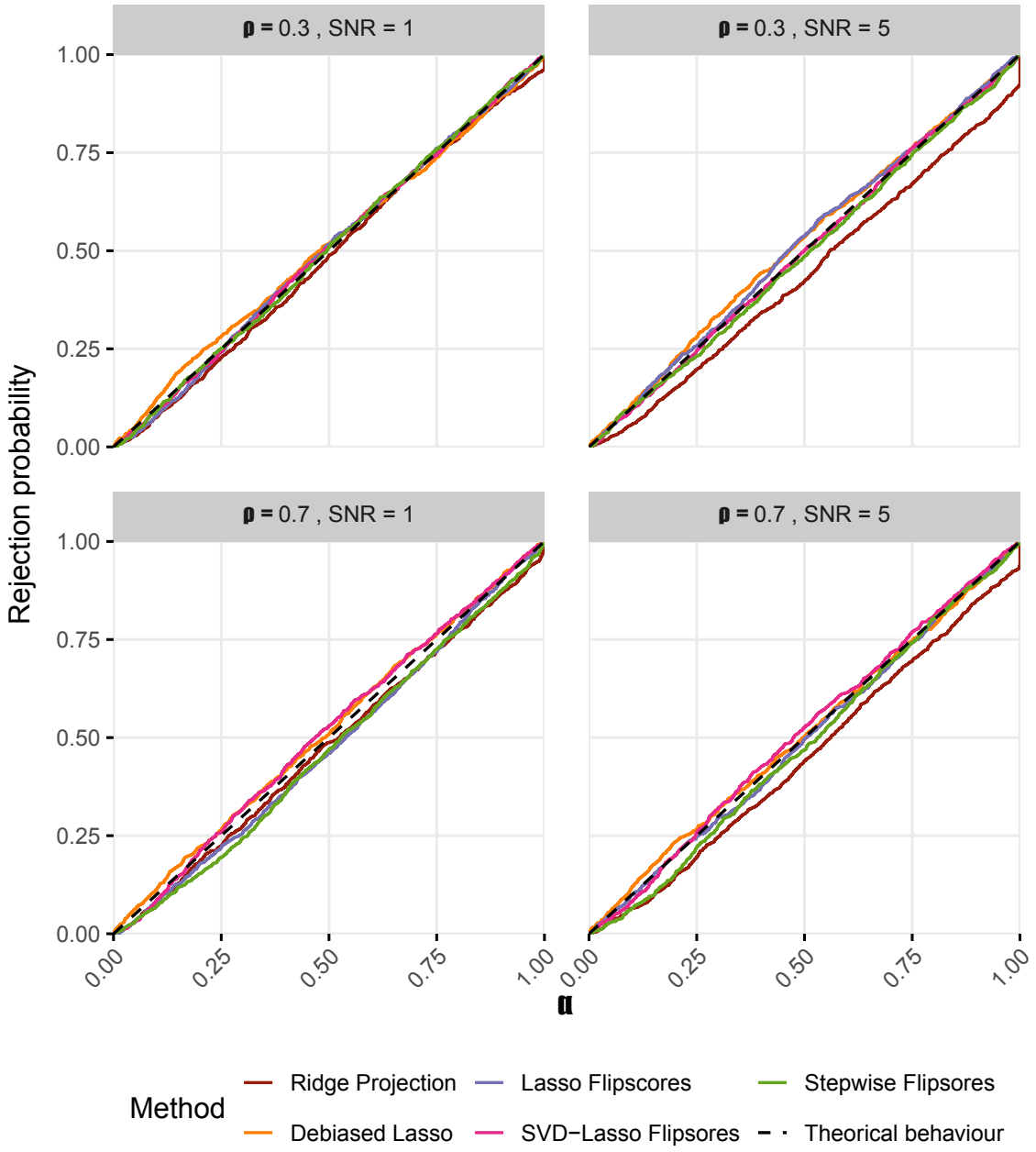

Figure 5.1: Empirical rejection curves under Toeplitz design for different correlation and SNR levels. The dashed line corresponds to the theoretical behaviour.

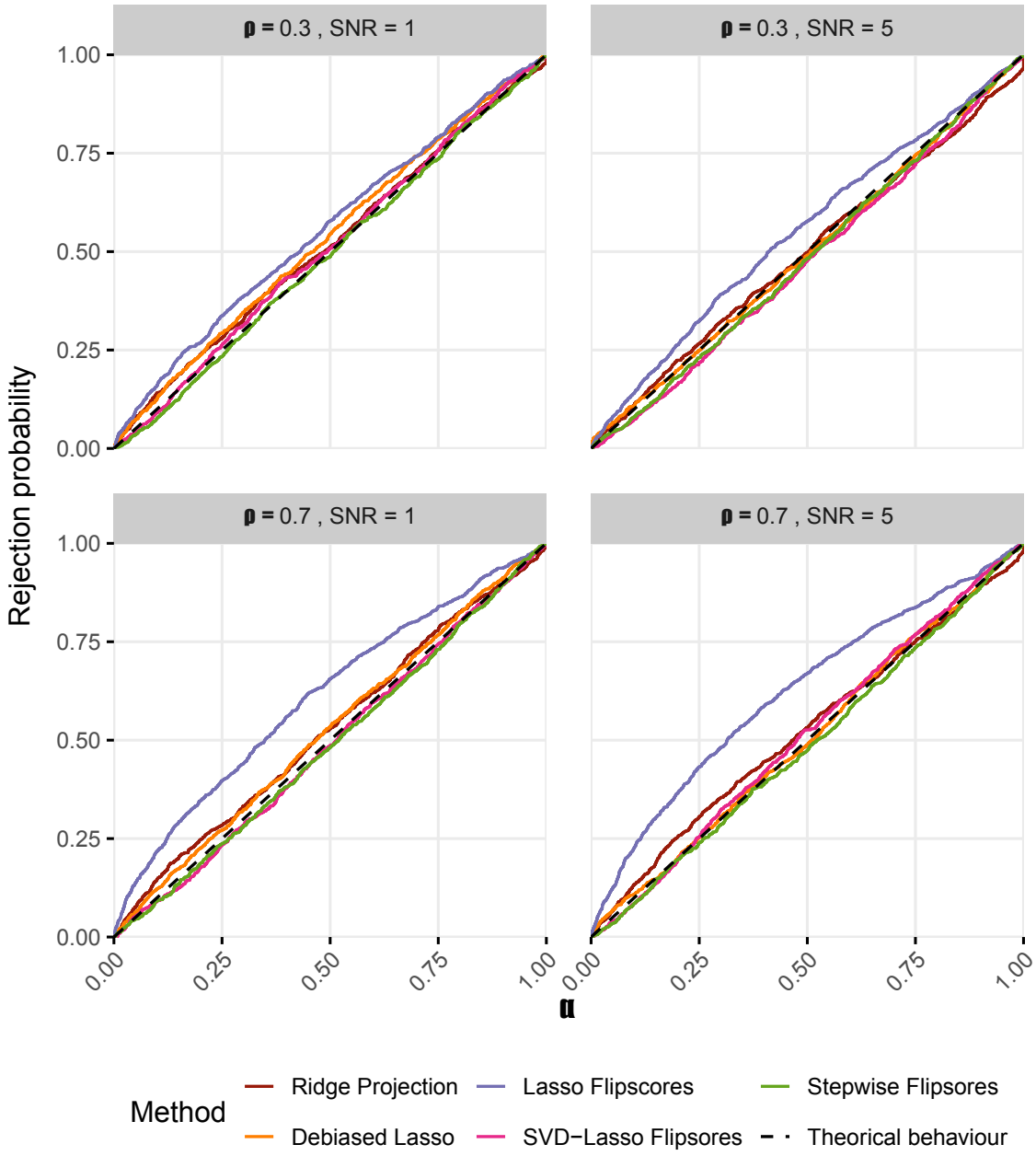

Figure 5.2: Empirical rejection curves under equicorrelated design for different correlation and SNR levels. The dashed line corresponds to the theoretical behaviour.

especially as correlation increases. Ridge projection and the debiased lasso also display mild anti-conservativeness, overshooting the nominal threshold. By contrast, the SVD-lasso and stepwise flipscores remain close to the theoretical level, demonstrating robustness even under strong dependence.

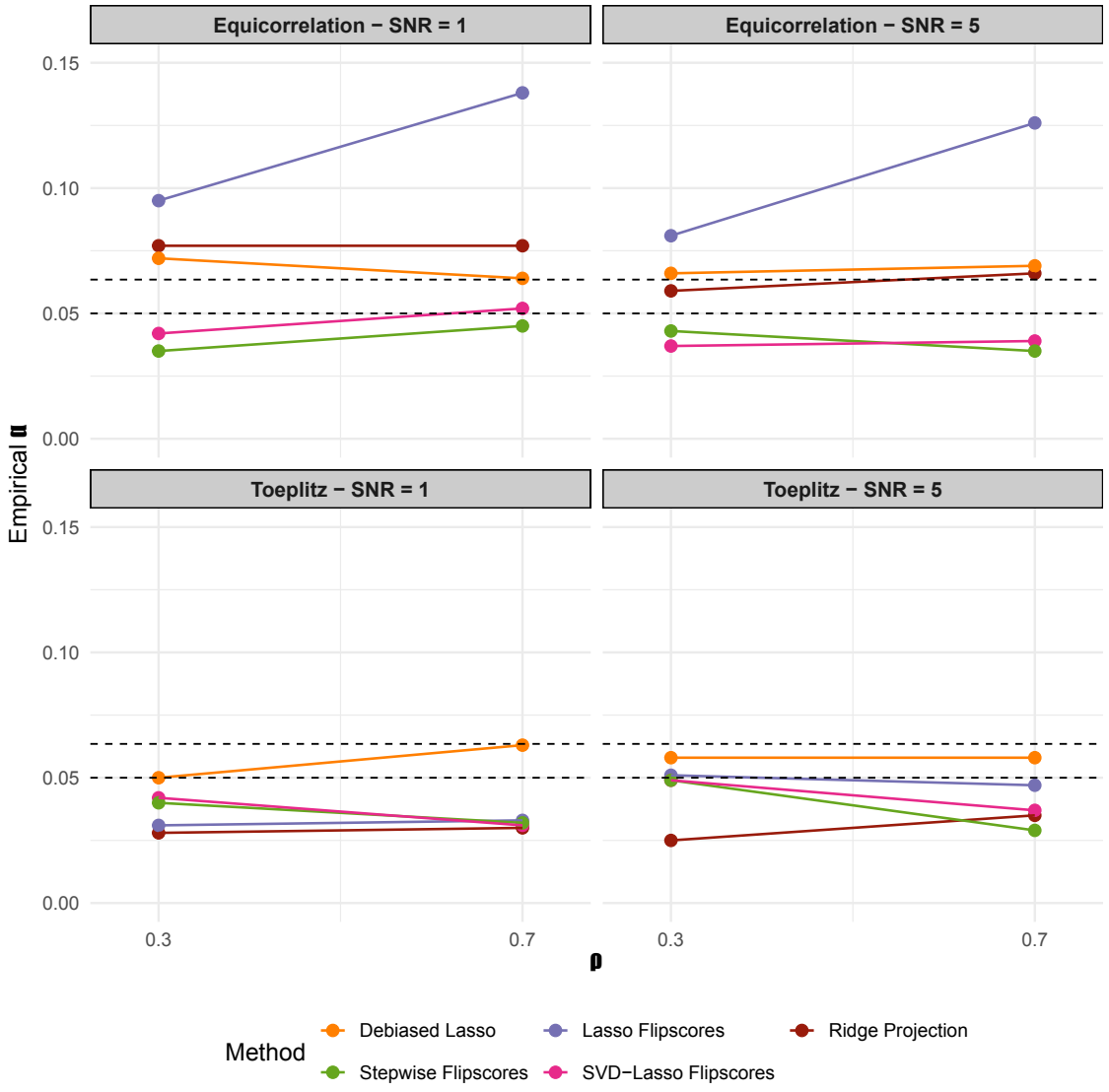

Figure 5.3: Empirical type I error rates. The dashed lines correspond to $\alpha = 0.05$ and to the upper bound at level 95%, $\alpha + 1.96\sqrt{\alpha(1-\alpha)/1000}$.

Figure 5.4 reports the empirical power. The contrast with type I error results is immediate: the debiased lasso and ridge projection achieve the highest power across settings, although this advantage is partly explained by their anti-conservative behaviour in the equicorrelated case. Among the valid methods, the stepwise flipscores is the only competitor with comparable performance, balancing power and validity.

The SVD-lasso flipscores shows very low power in single-test settings. This is plausibly due to the way the signal is distributed across components, together with the fixed choice of 30 components in the selection step. The lasso flipscores, while sometimes competitive in terms of power, cannot be considered reliable because of its inflated type I error under equicorrelation.

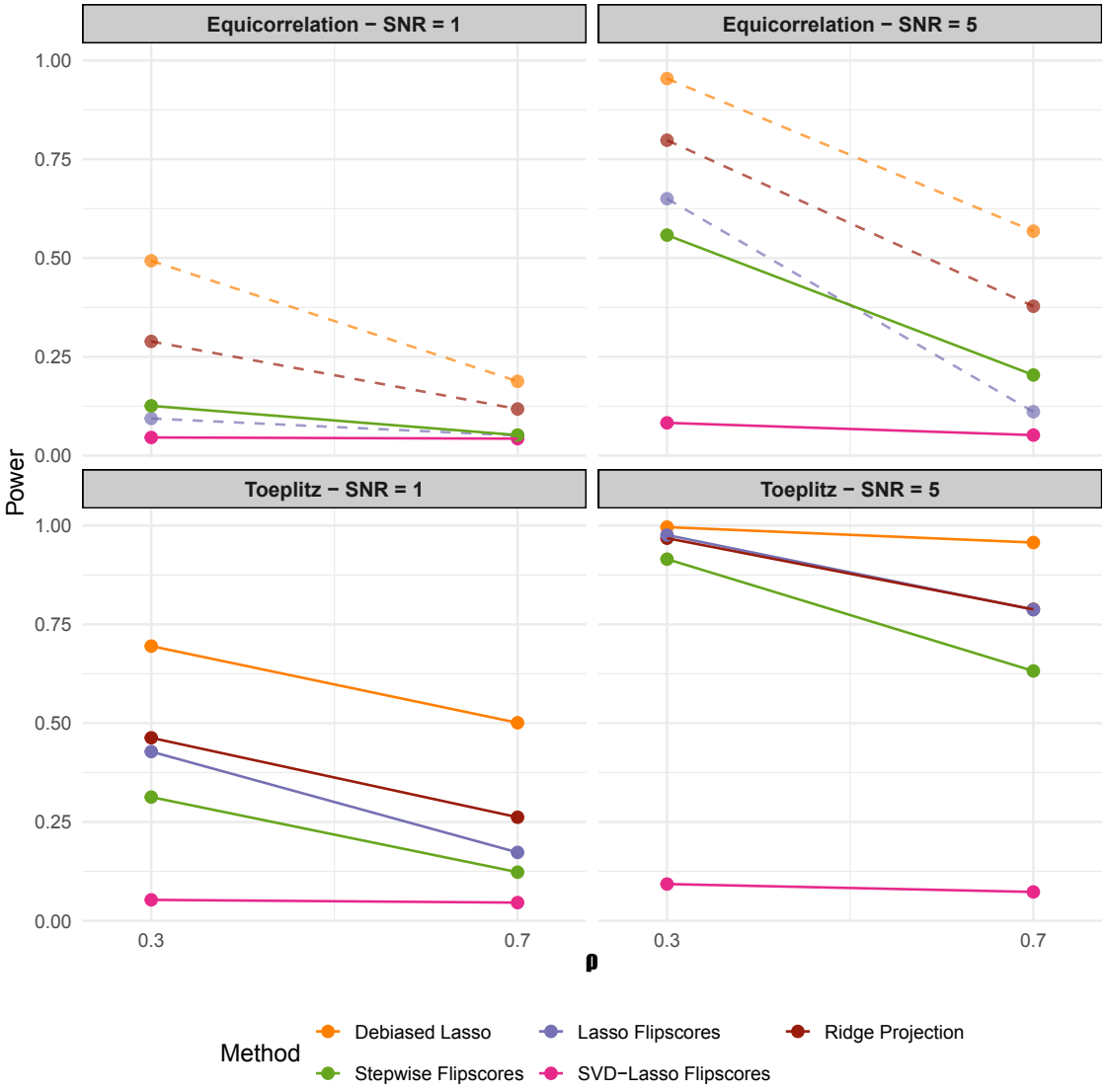

Figure 5.4: Empirical power under equicorrelated and Toeplitz designs for different correlation and SNR levels. Dashed lines correspond to the methods that do not control the Type I error in the respective setting proposed in Figure 5.3.

In summary, the simulations confirm a persistent trade-off between validity and sensitivity. Methods with inflated rejection rates naturally show higher power, while those that control the error correctly, such as stepwise flipscores, sacrifice some detection ability. The SVD-lasso flipscores in particular illustrates

how dimensionality reduction choices can directly affect the balance between robustness and power.

5.2 Multiple hypothesis testing

After studying the single-test behaviour, we turn to the multiple testing problem, which is of primary interest. We excluded the lasso flipscores from this part of the study, due to its instability across scenarios. For ridge projection and debiased lasso, we used the built-in multiple testing corrections presented in Equation (2.4); for our flipscores methods we adopted the maxT correction (Westfall & Young, 1993).

The simulation settings remain the same as before. We fixed the sample size, number of variables, number of active variables and coefficient values as above, focusing on the Toeplitz covariance structure with $\rho = \{0.3, 0.7\}$ and $\text{SNR} = \{3, 7\}$. Each configuration was repeated 1000 times.

We first examined whether the procedures provide strong control of the nominal family-wise error rate (FWER). The FWER was computed as the proportion of simulations in which at least one false rejection occurred. Figure 5.5 reports the results at the 5% level. The differences are stark: the debiased lasso performs very poorly, with error rates exceeding 20% in all cases.

By contrast, the flipscores-based methods succeed in controlling the FWER. stepwise flipscores and ridge projection appear conservative, with empirical values below the nominal threshold. The SVD-lasso flipscores stays closer to the target level, balancing validity and conservativeness. Overall, these results reinforce the robustness of the flipscores framework in multiple testing, in contrast with the instability of the debiased lasso.

We then computed power as the mean proportion of true rejections across simulations. Figure 5.6 shows the results. Excluding the debiased lasso, which is invalid due to its inflation of false positives, the comparison broadly confirms the single-test findings. The SVD-lasso flipscores again exhibits very low power, reflecting its limitations when the signal is spread across components.

The stepwise flipscores, while less powerful than ridge projection, achieves satisfactory results, striking a balance between error control and sensitivity.

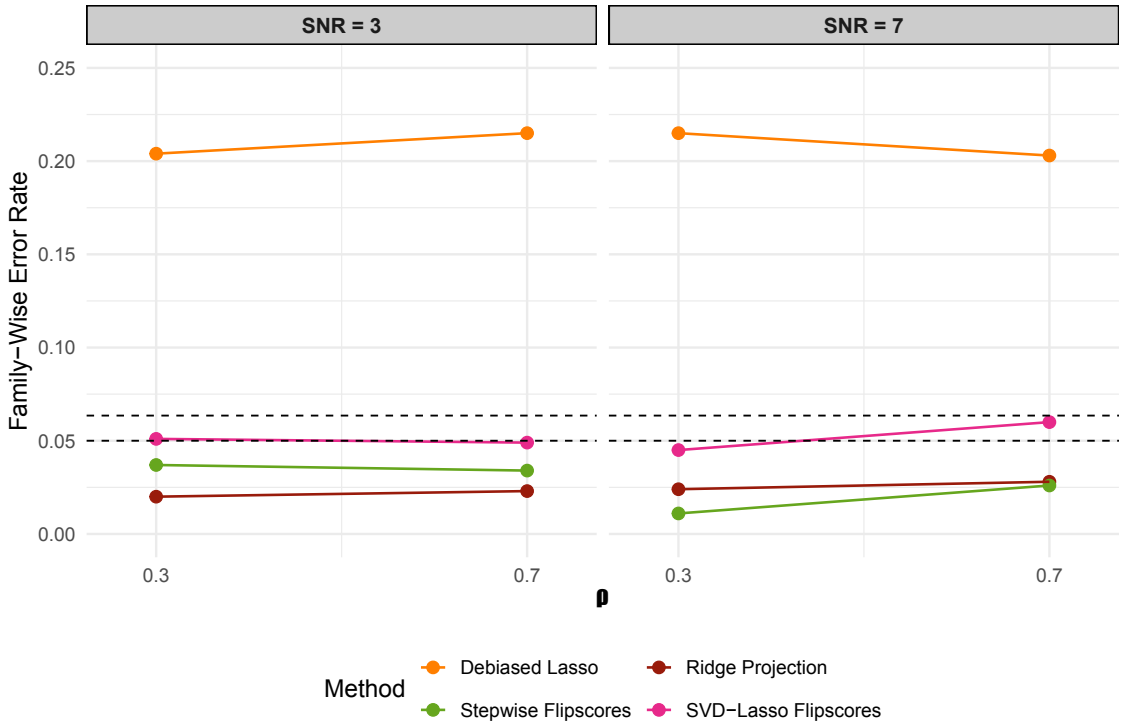

Figure 5.5: Family-wise error rate (FWER) at the 5% level. The dashed lines correspond to $\alpha = 0.05$ and to the upper bound at level 95%,$\alpha + 1.96\sqrt{\alpha(1-\alpha)/1000}$.

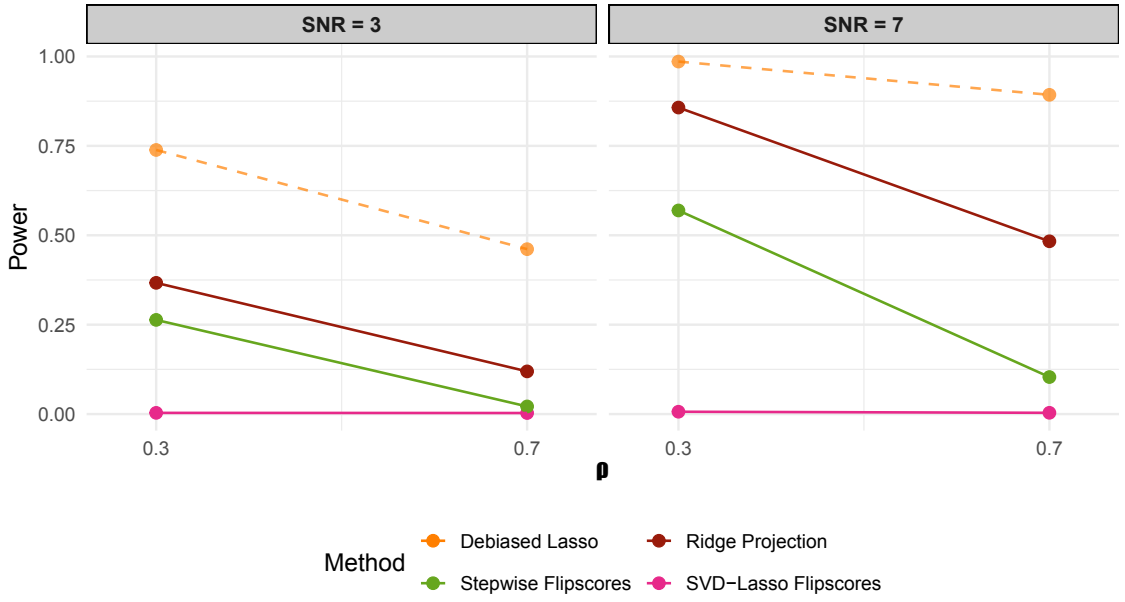

Figure 5.6: Empirical power in the multiple testing setting under Toeplitz design. Dashed lines correspond to the methods that do not control the FWER in the respective setting proposed in Figure 5.5.

5.2.1 Computational time comparison

Tables 5.1–5.2 report the median, mean and standard deviation of computational times across methods. SVD-lasso and stepwise flipscores were both implemented using parallel computation on 6 cores, so as to mitigate the computational burden arising from the sequential structure of their algorithms. Ridge projection is by far the fastest, with negligible execution times in all scenarios. The debiased lasso and SVD-lasso require slightly more computation, but remain within a modest range (around 10 seconds on average). stepwise flipscores, on the other hand, is two orders of magnitude slower, with average times above two minutes and noticeable variability, especially under higher correlation. This result stresses the trade-off between computational efficiency and statistical robustness: while stepwise selection offers competitive inferential properties, its practical usability may be limited by its computational burden.

	$\rho = 0.3$			$\rho = 0.7$		
	median	mean	std. dev.	median	mean	std. dev.
debiased lasso	9.5	11.0	2.7	8.8	9.0	0.9
Ridge Projection	0.3	0.3	0.1	0.3	0.3	0.1
SVD-lasso	11.0	11.1	0.8	10.4	10.7	0.7
stepwise	120.6	121.7	3.0	129.6	129.3	9.7

Table 5.1: Computational times (in seconds) for SNR $= 3$.

	$\rho = 0.3$			$\rho = 0.7$		
	median	mean	std. dev.	median	mean	std. dev.
debiased lasso	9.8	10.8	2.3	8.8	8.8	0.2
Ridge Projection	0.3	0.3	0.1	0.3	0.3	0.1
SVD-lasso	11.4	11.4	1.1	10.4	10.5	0.6
stepwise	120.7	122.3	3.4	121.9	126.6	31.9

Table 5.2: Computational times (in seconds) for SNR $= 7$.

Chapter 6

Conclusion

The research presented in this thesis has the aim to propose a new method to deal with the problem on how to obtain valid inference when the number of predictors is larger than the number of observations. In such situations, the classical tools of regression analysis, such as least squares estimation, no longer provide reliable results. The problems of the bias caused by naïve post-selection inference approaches and multiple testing inflation can lead to false discoveries in research areas where they often occur.

The work has introduced a permutation-based methodology that combines effective scores with sign-flip inference. The key idea from an empirical point of view is the removal of the variable of interest from the selection step: it seems that one can evade the double–dipping problem, and by approximating the null distribution through random sign flips, it is possible to build a test that remains valid under mild assumptions. In addition, the procedure is specifically designed for high-dimensional contexts, utilising selection methods and a correction method for multiple tests based on resampling methods, the maxT procedure.

One of the main contributions of this thesis was to adapt and extend the sign-flip method to high-dimensional linear models. To do this, the work included a selection step for choosing potential confounding factors, exploring three methods: lasso selection, lasso selection after performing an SVD, and stepwise selection. A fundamental feature of the approach is that the variable being tested is left out of the selection step. This might seem like a small detail, but it makes a big difference: By avoiding the use of the same variable for both model selection and inference, the method appears to mitigate post-selection bias, helping to ensure that the final inference remains valid.

From an empirical point of view, the thesis has relied on an extensive set of simulations in order to assess the performance of the proposed procedures. The various scenarios have provided the opportunity to explore different correlation structures among predictors, a variety of signal-to-noise ratios, as well as both single-hypothesis and multiple-hypothesis. The results confirmed that the methods are capable of maintaining control of Type I error in almost every setting tried. One conclusion is that the stepwise based sign flipping methods achieved competitive power compared to established alternatives. At the same time, the comparison across selection strategies revealed that their relative merits depend on the data-generating mechanism, providing useful guidance for practical applications.

Finally, the practical relevance of the framework has also been emphasized, proving that the procedures can be implemented with a computational effort that is reasonable in most high-dimensional contexts of interest. Because they are robust to model misspecification and rely on few distributional assumptions, these procedures appear particularly suitable for applied settings such as genomics, neuroscience, and epidemiology, where reliable inference in the presence of a large number of predictors is one of the most important requests.

The simulation study has provided several insights. First, the importance of avoiding the variable under test in the selection step was confirmed to maintain the control of the type I error in our procedures. Second, the comparison between lasso, SVD-lasso, and stepwise procedures showed that the latter, overall, performed best, achieving both single-test and FWER control, while also obtaining power comparable to competing methods.

However, the proposed framework is not without limitations. Its finite-sample performance is still influenced by the choice of confounder-selection algorithm, showing the importance of this step. Moreover, while the simulations provided strong empirical support, a complete set of theoretical guarantees for all the proposed variants is still lacking. In addition, due to the use of permutations and, in particular, of a procedure that tests one variable at a time, the proposed methods have non-negligible computational costs that must be taken into account in their application.

The main power of this method is its extendability. Extending the proposed framework beyond the linear model, for instance to generalized linear models, is a natural step. Another promising line of work is to strengthen the theoretical results and clarify under which precise conditions the validity of the tests is

ensured in finite samples, and consequently also to explore alternative selection methods that may achieve a better trade-off between computational time and power, while still maintaining error control. Another avenue is that, being a permutation-based method, there is the possibility, in the multiple testing context, to refine this approach to directly control the False Discovery Proportion, rather than the FDR or FWER, through recently developed post-hoc methods that adapt to the correlation structure of the test statistics, such as NoTip (Blain et al., 2022) or pARI (Andreella et al., 2023).

The results presented here suggest that permutation-based approaches, when carefully designed, can offer a viable and reliable alternative to more traditional inferential methods. They provide a way to maintain error control without relying heavily on asymptotic approximations or restrictive assumptions, and therefore have the potential to contribute significantly to the reliability of scientific findings in high-dimensional contexts.

Bibliography

Akaike, H. (1974). A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control* **19**, 716–723.

Andreella, A., Hemerik, J., Finos, L., Weeda, W. & Goeman, J. (2023). Permutation-based true discovery proportions for functional magnetic resonance imaging cluster analysis. *Statistics in Medicine* **42**, 2311–2340.

Azzalini, A. (1996). *Statistical Inference based on the likelihood*. Chapman and Hall.

Benjamini, Y. & Hochberg, Y. (2018). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)* **57**, 289–300.

Blain, A., Thirion, B. & Neuvial, P. (2022). Notip: Non-parametric true discovery proportion control for brain imaging. *NeuroImage* **260**, 119492.

Bonferroni, C. (1936). *Teoria statistica delle classi e calcolo delle probabilità*. Pubblicazioni del R. Istituto superiore di scienze economiche e commerciali di Firenze. Seeber.

Bühlmann, P. (2013). Statistical significance in high-dimensional linear models. *Bernoulli* **19**, 1212–1242.

Corbetta, D. (2025). Ph.D. thesis, Università degli Studi di Padova.

De Santis, R., Goeman, J. J., Hemerik, J., Davenport, S. & Finos, L. (2025). Inference in generalized linear models with robustness to misspecified variances. *Journal of the American Statistical Association* , 1–10.

Fan, J. & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* **96**, 1348–1360.

Hall, W. J. & Mathiason, D. (1990). On large-sample estimation and testing in parametric models. *International Statistical Review* **58**, 77–97.

Hemerik, J., Goeman, J. J. & Finos, L. (2020). Robust testing in generalized linear models by sign flipping score contributions. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* **82**, 841–864.

Javanmard, A. & Montanari, A. (2014). Confidence intervals and hypothesis testing for high-dimensional regression. *J. Mach. Learn. Res.* **15**, 2869–2909.

Meinshausen, N. & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics* **34**, 1436–1462.

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

Neyman, J. & Scott, E. L. (1948). Consistent estimates based on partially consistent observations. *Econometrica* **16**, 1–32.

Pesarin, F. (2001). *Multivariate Permutation Tests : With Applications in Biostatistics*. Wiley.

Rao, C. R. (1948). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Proceedings of the Cambridge Philosophical Society* **44**, 50–57.

Rigby, R. A. & Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society* **54**, 507–554.

Ruben Dezeure, Peter Bühlmann, Lukas Meier & Nicolai Meinshausen (2015). High-dimensional inference: Confidence intervals, p-values and R-software hdi. *Statistical Science* **30**, 533–558.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6**, 461–464.

Scott, W. A. (2002). Maximum likelihood estimation using the empirical fisher information matrix. *Journal of Statistical Computation and Simulation* **72**, 599–611.

Sun, T. & Zhang, C.-H. (2012). Scaled sparse linear regression. *Biometrika* **99**, 879–898.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **58**, 267–288.

van de Geer, S., Bühlmann, P., Ritov, Y. & Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for hih-dimensional models. *The Annals of Statistics* **42**, 1166–1202.

van de Geer, S. A. & Bühlmann, P. (2009). On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics* **3**, 1360 – 1392.

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society* **54**, 426–482.

Wasserman, L. & Roeder, K. (2009). High-dimensional variable selection. *The Annals of Statistics* **37**, 2178–2201.

Westfall, P. H. & Young, S. S. (1993). *Resampling-Based Multiple Testing: Examples and Methods for p-Value Adjustment*. Wiley.

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica* **50**, 1–25.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics* **9**, 60–62.

Winkler, A. M., Ridgway, G. R., Webster, M. A., Smith, S. M. & Nichols, T. E. (2014). Permutation inference for the general linear model. *NeuroImage* **92**, 381–397.

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics* **38**, 894–942.

Zhang, C.-H. & Zhang, S. S. (2013). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **76**, 217–242.

Zhao, S., Witten, D. & Shojaie, A. (2021). In defense of the indefensible: A very naïve approach to high-dimensional inference. *Statistical Science* **36**, 562–577.

Zou, H. (2006). The adaptive lasso ad its oracle properties. *Journal of the American Statistical Association* **101**, 1418–1429.

Zou, H. & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology* **67**, 301–320.