E-values for Hypothesis Testing wih Covariates: Construction & Application to Regression Models

Lardy, T.

MATHEMATISCH INSTITUUT
UNIVERSITEIT LEIDEN

MASTER'S THESIS

E-values for Hypothesis Testing with Covariates

Construction & Application to Regression Models

Author:
T.D. Lardy

Thesis Supervisor:
Prof.dr. P.D. Grünwald

Examination Date: June 23, 2021

E-values for Hypothesis Testing with Covariates

T.D. Lardy

Mathematical Institute, Leiden University
Niels Bohrweg 1, 2333 CA Leiden, The Netherlands

Examination Date: June 23, 2021

Abstract

The E-value is a notion of evidence that — unlike the p-value — allows for the seamless combination of results from multiple tests of a single hypothesis. Statistical inferences drawn from the combination of E-values are valid even when the decision to perform one of the tests was based on the results of another test. However, the theory surrounding E-values has not yet been fully developed. In particular, little is known about using E-values for hypothesis tests that involve covariates. This thesis describes two methods of constructing E-values and shows that they can be applied to tests with covariates. The first method is based on the principles of invariant testing and is shown to lead to the GROW (Growth-Rate-Optimal-in-Worst-Case) E-value for a hypothesis test in a linear regression model. The second method uses an algorithm to approximate the GRO E-value. It is shown that this approximation can be normalised to become an E-value. Preliminary simulations are done for a hypothesis test in a logistic regression model.

Acknowledgements

First and foremost, I would like to express my most profound appreciation to my supervisor, Peter Grünwald. His immense knowledge, ingenuity, and enthusiasm have been a driving factor in the completion of this thesis. My sincere gratitude also extends to Muriel Perez for his invaluable insights and clever suggestions to further my research. Additionally, I am thankful to the CWI and all members of the machine learning group for providing a friendly and stimulating environment to work on my thesis. Lastly, I am forever indebted to my parents for their wise counsel and unwavering support.

Contents

Chapter 1

Introduction

Nearly a century since its introduction, the p-value has become *the* standard by which scientific conclusions are judged [1]. Yet, it has been recognised for decades that the p-value has mathematical and even ethical shortcomings [2, 3]. The result is that the p-value is often misinterpreted and misused, jeopardising the integrity of scientific research [4]. An example is when researchers engage in *optional continuation*, i.e. when they base their decision whether or not to collect more data on side information, such as previous results or the amount of funding. While this practice is ubiquitous in science, it invalidates conclusions drawn from the p-value.

For such reasons, an alternative to the p-value has recently been gaining traction: the E-value [5, 6]. The E-value is a test statistic that improves the *interpretability* and *flexibility* of classical hypothesis testing methods. It can be construed as a simple bet against the null hypothesis [7] and any derived conclusions are valid even under optional continuation [5]. Moreover, E-values from different studies can effortlessly be combined by multiplying them, as multiplication is an operation under which they remain valid [5, 6]. This improves the simplicity and validity of meta-analyses, as current techniques are prone to various biases [8].

However, the widespread adoption of the E-value is greatly hindered by its current lack of adaptability. There are simply many hypothesis testing problems for which the theory of testing based on E-values has not yet been fully developed. In particular, little is currently known about E-values for problems involving side information, or *covariates*. This is a significant problem for fields like pharmaceutics, where studies on the effect of drugs on a patients health usually involve information, such as the patient's age, prior afflictions, etc. Failing to account for such information might weaken the results of the study.

The primary aim of this thesis is to help further develop the theory of E-values for such problems. Two methods are discussed: one for obtaining precise E-values and one for approximating E-values. The first method is based on results by Perez et al. [9], who adapt the established theory of invariant testing [10, 11] to the theory of E-values. It will be shown that

this method applies to a hypothesis testing problem in a linear regression model that involves covariates. The resulting E-value is, in a sense, the best E-value one can use for this particular problem.

The second method revolves around an approximation algorithm by Li [12], which almost directly leads to an E-value. It will be shown that this E-value can be used for a hypothesis test in a logistic regression model, which also involves covariates. Preliminary simulations are done to investigate this E-value, but due to time considerations they are very limited. Consequently, a lot of research into E-values for this hypothesis testing problem remains to be done. Before further discussing these results, the remainder of this Chapter is dedicated to formally introducing the E-value.

1.1 E-values

One of the ultimate goals of science is to gain true knowledge about the universe. This process often involves formulating a so-called 'null hypothesis' and collecting and analysing data to quiz this hypothesis. Conclusions derived from the data might represent universal truths, but they could also be observed by chance. To help discriminate between the two, researchers turn to the theory of hypothesis testing. In this branch of statistics, the null hypothesis $\mathcal{H}_0$ is considered to be a set of possible probability distributions of some random variable $\mathbf{Y}$, and the collected data $\mathbf{y}$ is deemed to be a realisation of this random variable. Given the observed realisation, it is assessed how likely it is that the distribution of $\mathbf{Y}$ is an element of $\mathcal{H}_0$, often relative to an alternative hypothesis $\mathcal{H}_1$. In the framework considered here, this assessment is based on the E-test statistic.

Definition 1.1. An E-test statistic for testing a set of probability distributions $\mathcal{H}_0$ is a non-negative random variable $S := s(\mathbf{Y})$ such that

$$\text{for all } Q \in \mathcal{H}_0 : \mathbb{E}_{\mathbf{Y}\sim Q}[s(\mathbf{Y})] \leq 1.$$

That is, the expected value of any E-test statistic under the null hypothesis is smaller than 1. Therefore, a large E-test statistic ($\gg 1$) offers evidence against the null hypothesis and the larger it is, the more evidence it provides. E-test statistics will be referred to as E-variables and their realisation on specific data as E-values.

Example. Consider a clinical trial designed to determine whether medicine ABC is effective at treating high blood pressure. To that end, a patient's blood pressure is measured before and after being treated with ABC. The

difference between the two measurements, denoted by $\mathbf{X}$, is reported. Regardless of the effect of ABC, one cannot expect the measurements to be the same, because of regular fluctuations and the influence of things like measurement errors. Therefore, the two measurements should be considered to be independent random variables. If ABC does not affect the blood pressure, the distribution of the two measurements should have the same mean, and so the distribution of X should be symmetric around 0, i.e.

$$\mathcal{H}_0 = \{Q : Q \text{ symmetric around } 0\}.$$

If ABC does affect the blood pressure, the mean of the two measurements should differ, so one might take

$$\mathcal{H}_1 = \{Q : Q \text{ not symmetric around } 0\}.$$

One possible E-variable to use for this testing problem is the so-called Efron-De la Pena E-variable [13] given by

$$s_\lambda(\mathbf{X}) = e^{\lambda \mathbf{X} - \frac{\lambda^2 \mathbf{X}^2}{2}},$$

where $\lambda \in \mathbb{R}$ is a free parameter. To see that this is indeed an E-variable, $\mathbf{X}$ is written as $\mathbf{X} = \mathbf{Z} \cdot \mathbf{B}$, where $\mathbf{Z} = |\mathbf{X}|$ and $\mathbf{B} = \text{sign}(\mathbf{X})$. Under any $Q_0 \in \mathcal{H}_0$, it must hold for all $z \in \mathbb{R}^+$ that $Q_0(\mathbf{B} = 1 | \mathbf{Z} = z) = \frac{1}{2}$. Therefore,

$$\begin{aligned}
\mathbb{E}_{\mathbf{X} \sim Q_0}\left[s_\lambda(\mathbf{X})\right] &= \mathbb{E}_{\mathbf{Z} \sim Q_0} \mathbb{E}_{\mathbf{B} \sim Q_0 | \mathbf{Z}}\left[s_\lambda(\mathbf{BZ})\right] \\
&= \mathbb{E}_{\mathbf{Z} \sim Q_0} \mathbb{E}_{\mathbf{B} \sim Q_0 | \mathbf{Z}}\left[e^{\lambda \mathbf{BZ} - \frac{\lambda^2 (\mathbf{BZ})^2}{2}}\right] \\
&= \mathbb{E}_{\mathbf{Z} \sim Q_0}\left[\frac{1}{2} e^{-\lambda \mathbf{Z} - \frac{\lambda^2 (\mathbf{BZ})^2}{2}} + \frac{1}{2} e^{\lambda \mathbf{Z} - \frac{\lambda^2 (\mathbf{BZ})^2}{2}}\right] \\
&\leq 1,
\end{aligned}$$

where the last step follows from the cosh-inequality. The statistic $s_\lambda(\mathbf{X})$ can be extended to data involving multiple people by simply taking the product.

The rest of this section will discuss the favourable properties of E-variables, how they can be interpreted, and how they are related to the established theory of hypothesis testing.

1.1.1 Betting strategy analogy

The most natural interpretation of E-variables is as the outcome of a *betting strategy* against the null hypothesis [7]. Imagine that there is a banker that

sells gambling tickets for €1. He allows people to buy as many tickets as they like, even fractional amounts (e.g. 3.71 tickets). After seeing some data y, such as the outcome of a sports match, each of these tickets is worth a payoff of $€S := s(y)$, for some function s. It is common sense to assume that the banker picked S such that he does not expect to lose any money. Accordingly, under the null hypothesis of any rational player, it should hold that $\mathbb{E}[S] \leq 1$, i.e. S is an E-variable. If one buys tickets regardless and wins a lot of money, this can be considered to be evidence against the null hypothesis. After all, one bet against the null hypothesis and won. The larger s is, the more money one wins and the more evidential value it offers. From this betting interpretation, one can immediately see that combining E-values by multiplication is a valid operation corresponding to reinvesting one's profit, see Section 1.3. However, even if a lot of money was won, one should keep in the back of their mind that the possibility exists that the outcome was a fluke. This uncertainty is fully embraced by the betting interpretation of the E-variable, whereas classical hypothesis testing methods often tend to shy away from ambiguity. Instead, definite statements are typically made stating whether a finding is significant or not, which leads to misleading conclusions [14].

1.1.2 Relation to p-values

More often than not, these conclusions are based on the p-value. It is, therefore, useful to discuss the connection between E-values and p-values before going into more depth about the former. Informally, a p-value is the probability under a particular model that a statistic of the data would be equal to or more extreme than the observed value [1]. The formal definition is slightly more general:

Definition 1.2. A p-value is a random variable P such that for all $\epsilon \in [0, 1]$ and $Q \in \mathcal{H}_0$

$$Q(P \leq \epsilon) \leq \epsilon.$$

Note that both the random variable and its realisation will be referred to as p-value, while the former will be denoted with upper-case letters and the latter with lower-case letters. Roughly, one can state that E-values and p-values are linked by taking reciprocals. In one direction, this is precise: the reciprocal of any E-variable is a p-value [5]. The opposite is not precisely true, but Vovk and Wang [6] show that the reciprocal of a p-value can, after a correction, be seen as an E-variable. The following two Propositions make these statements mathematically rigorous.

Proposition 1.1. *For any E-variable S, the random variable $1/S$ is a p-value.*

Proof. Let S be an E-variable for testing $\mathcal{H}_0$. By Markov's inequality, for all $Q \in \mathcal{H}_0$

$$Q\left(\frac{1}{S} \leq \epsilon\right) = Q\left(S \geq \frac{1}{\epsilon}\right) \leq \epsilon \mathbb{E}_Q[S] \leq \epsilon.$$

□

Proposition 1.2 (Vovk and Wang)**.** *Let $\kappa \in (0,1)$ arbitrarily and consider the function $f_\kappa : \mathbb{R} \to \mathbb{R}$ defined by $p \mapsto \kappa p^{\kappa-1}$. For any p-value P, the random variable $f_\kappa(P)$ is an E-variable.*

That the reciprocal of any E-variable is a p-value follows immediately. By considering that for small values of κ, $f_\kappa(P)$ can roughly be seen as the reciprocal of P, it also follows that the reciprocal of a p-value can be seen as an E-variable after a correction. Moreover, Vovk and Wang [6] show that taking reciprocals is essentially the only way to transform E-variables into p-values. At the same time, the functions f_κ are only a subset of the rich family of functions that transform p-values into E-variables. From this relation, it can directly be seen that E-values generally give a lot less evidence than p-values. For example, transforming a significant p-value of 0.05 with $\kappa = \frac{1}{2}$ gives an E-value of $f_{\frac{1}{2}}(0.05) \approx 2.24$. This value is not a lot larger than 1 and transforming it back to a p-value gives a value of roughly 0.45, which is a lot larger than the initial value of 0.05. Hence, evidential value was lost by transforming the p-value into an E-value.

1.1.3 Relation to Bayes factors

Another statistic that is frequently used in hypothesis testing is the Bayes Factor [15]. To explore its relation to E-variables, it is helpful to specify the hypothesis testing problem in more detail. To this end, consider a family of probability distributions $\mathcal{P}$, that is parametrised by the parameter set Θ, i.e.

$$\mathcal{P} = \{P_\theta : \theta \in \Theta\}.$$

The P_θ's are distributions defined on the same sample space, and their density functions are denoted by p_θ. It is then assumed that the hypotheses are of the form

$$\mathcal{H}_0 = \{P_\theta : \theta \in \Theta_0\} \text{ and } \mathcal{H}_1 = \{P_\theta : \theta \in \Theta_1\},$$

where $\Theta_0 \subseteq \Theta$ and $\Theta_1 \subseteq \Theta \setminus \Theta_0$.

Bayesians equip both hypotheses $\mathcal{H}_i (i \in \{0,1\})$ with a *prior* W_i. This prior is essentially a distribution on Θ_i, which represents some pre-existing belief about how likely the distributions within the hypothesis are, assuming this hypothesis is true. After all, even if the conclusion is that the distribution of $\mathbf{Y}$ is an element of Θ_i, only one of its elements can be the actual distribution. The set of all possible priors will be denoted by $\mathcal{W}(\Theta_i)$. The Bayes marginal distribution P_{W_i} represents how likely the data is under hypothesis $\mathcal{H}_i$ given the prior

$$p_{W_i}(\mathbf{Y}) = \int_{\Theta_i} p_\theta(\mathbf{Y}) dW_i(\theta).$$

The Bayes Factor $\mathrm{BF}(\mathbf{Y})$ is then defined as

$$\mathrm{BF}(\mathbf{Y}) = \frac{p_{W_1}(\mathbf{Y})}{p_{W_0}(\mathbf{Y})}. \tag{1.1}$$

It represents how many times more likely the data is under the alternative hypothesis than under the null hypothesis, given the prior belief about the individual hypotheses. Therefore, similar to E-variables, the larger the Bayes Factor, the more evidence it constitutes against the null hypothesis. It should then be no surprise that there is a simple condition under which Bayes Factors are E-variables.

Proposition 1.3. *If $\mathcal{H}_0$ is a singleton, the Bayes Factor is an E-variable.*

Proof. Since $\mathcal{H}_0$ only consists of one distribution, say P_0, the prior W_0 must put all weight on this distribution, i.e. $P_{W_0} = P_0$. Therefore, for any prior W_1, it holds that

$$\mathbb{E}_{\mathbf{Y} \sim P_0} [\mathrm{BF}(\mathbf{Y})] = \int p_0(y) \frac{p_{W_1}(y)}{p_0(y)} dy = \int p_{W_1}(y) dy = 1.$$

Since P_0 is the only distribution in $\mathcal{H}_0$, the Bayes Factor satisfies the conditions of Definition 1.1 and is thus an E-variable. □

In general, Bayes Factors for more complicated, composite, null hypotheses are not E-variables. Conversely, E-variables need not necessarily be of the form (1.1) either. However, Grünwald et al. show that the class of E-variables that, in some sense, provide the most substantial evidence *are* Bayes Factors. A notion of strength of E-variables needs to be introduced to make this precise.

1.2 The GRO(W) Criterion

So far, E-variables have been fully characterised by their expectation under the null hypothesis. This characterisation ensures that if $\mathcal{H}_0$ is true, it is unlikely to find evidence against it. However, this alone is not enough to guarantee that an E-variable is useful. For example, the constant $s(\mathbf{Y}) := 1$ satisfies the condition to be an E-variable but will never offer any evidence against $\mathcal{H}_0$. Therefore, another critical aspect is to consider what happens if $\mathcal{H}_1$ is true. In this case, it is desirable to gather as much evidence as possible against $\mathcal{H}_0$. Since the size of the E-variable corresponds to the amount of evidence, the most useful E-variables are those that grow large as fast as possible if $\mathcal{H}_1$ is true.

To formalise this property, Grünwald et al. [5] introduce the concept of GROW (Growth-Rate-Optimal-in-Worst-case) E-variables. This concept is based on requiring that an E-variable gives maximal evidence if $\mathcal{H}_1$ is true, even for the distribution in $\mathcal{H}_1$ that looks most like $\mathcal{H}_0$.

Definition 1.3. The GROW E-variable S^* is defined as the E-variable that achieves (if it exists):

$$\sup_{S \in \mathcal{E}(\Theta_0)} \min_{\theta_1 \in \Theta_1} \mathrm{E}_{P_{\theta_1}}[\log S]. \tag{1.2}$$

Here, $\mathcal{E}(\Theta_0)$ denotes the set of all possible E-variables on $\mathbf{Y}$ for Θ_0.

Grünwald et al. give the full reasoning behind using the logarithm as opposed to any other increasing function [5, Section 3.1]. However, it is straightforward to see why optimising over $\mathrm{E}[S]$ itself is not a good idea: one might end up with an E-variable that is equal to zero with a positive probability and large otherwise. If such an E-variable is used to combine evidence from multiple experiments by multiplying (see Section 1.3), there is a positive probability of losing *all* the gathered evidence. Taking the logarithm prevents this by infinitely penalising such E-variables.

The worst-case aspect of Definition 1.2 is not always desirable. In the presence of a prior on the alternative, it may not make sense to look at all elements of the alternative hypothesis separately. Instead, the strength of an E-variable should be defined relative to the prior. In such cases, the definition of GROW is adapted to GRO, as worst-case behaviour should no longer be considered.

Definition 1.4. Let $W_1 \in \mathcal{W}(\Theta_1)$ be a prior on $\mathcal{H}_1$. The GRO E-variable $S^*_{W_1}$ with respect to W_1 is the E-variable that achieves (if it exists):

$$\sup_{S \in \mathcal{E}(\Theta_0)} \mathbb{E}_{P_{W_1}}[\log S]. \tag{1.3}$$

Note that if S in (1.3) is of the form of a Bayes Factor, then the expectation of the logarithm reduces to the Kullback-Leibler divergence. That is, if $S = \frac{p_{W_1}(\mathbf{Y})}{p_{W_0}(\mathbf{Y})}$, then

$$\mathbb{E}_{P_{W_1}}[\log S] = \int_{\mathbf{y}} p_{W_1}(\mathbf{y}) \log\left(\frac{p_{W_1}(\mathbf{y})}{p_{W_0}(\mathbf{y})}\right) d\mathbf{y} = D(P_{W_1} \| P_{W_0}).$$

From results by Li [12, Lemma 4.1], such a Bayes Factor is an E-variable if and only if P_{W_0} is given by the so-called *Reverse Information Projection* (RIPr) of P_{W_1} on $\{P_W : W \in \mathcal{W}(\Theta_0)\}$.

Definition 1.5. The Reverse Information Projection of P_{W_1} on $\{P_W : W \in \mathcal{W}'\}$ is defined as the distribution that achieves

$$\inf_{W \in \mathcal{W}'} D(P_{W_1} \| P_W). \tag{1.4}$$

Li [12] shows that a measure achieving (1.4) exists under no further conditions, although in general it may be a sub-distribution that integrates to less than 1. Depending on how complex $\mathcal{P}$ is, it may or may not be feasible to find an explicit expression for this RIPr. For cases in which it is not, Li [12] demonstrates a way to approximate the RIPr. Such an approximation might be useful because the main theorem proved by Grünwald et al. [5] shows that the RIPr leads to the GRO E-variable relative to W_1.

Theorem 1.4 (Grünwald et al.)**.** *Let $W_1 \in \mathcal{W}(\Theta_1)$, such that the infimum $\inf_{W_0 \in \mathcal{W}(\Theta_0)} D(P_{W_1} \| P_{W_0})$ is finite and achieved by some prior W_0^o, and such that for all $\theta \in \Theta_0$, P_θ is absolutely continuous with respect to P_{W_1}, then the GRO E-variable takes the form*

$$S^*_{W_1} := \frac{p_{W_1}(\mathbf{Y})}{p_{W_0^o}(\mathbf{Y})}.$$

1.3 Merging E-values

A final step in the general discussion of E-values is to examine how researchers can combine the evidential value of multiple E-values. Suppose that they have collected independent data samples $\mathbf{Y}_{(1)}, \ldots \mathbf{Y}_{(k)}$ and have respectively calculated E-values $S_{(1)} := s_{(1)}(\mathbf{Y}_{(1)}), \ldots, S_{(k)} := s_{(k)}(\mathbf{Y}_{(k)})$ for the same testing problem. Of course, the researchers would like to combine their results, or *merge* the E-values. To this end, Grünwald et al. [5] show that regardless of any intricate dependencies between the experiments, the product $\prod_{i=1}^{k} S_{(i)}$ is again an E-value. Similarly, Vovk and

Wang [6] show that the arithmetic mean $\sum_{i=1}^{k} S_{(i)}/k$ is also an E-value. Moreover, they show that there are certain conditions for both of these methods, under which they lead to better E-values than any other possible merging methods.

These results hold even in the most extreme cases thinkable. Think of a researcher that first collects a single batch of data and calculates the first E-value. Depending on this E-value and perhaps other conditions (e.g. the amount of funding), the researcher decides whether or not to collect another batch of data. This process continues until the researcher decides to stop, either because the final E-value is pleasing or for any other reason. Such complex dependencies between the experiments would make it extremely hard to reach valid conclusions in the classical hypothesis testing approach. On the other hand, it is perfectly safe to make statistical inferences from the product of the E-variables.

To make this intuitive, consider the betting analogy from Section 1.1.1. Imagine that one first invests €1 and gets a payoff of €$s_{(1)}$, which is invested in buying new tickets. After observing the new payoff per ticket, €$s_{(2)}$, the total capital is €$s_{(1)} \cdot s_{(2)}$. This is repeated k times, after which it is decided to stop for whatever reason. The accumulated capital will be €$\prod_{i=1}^{k} s_{(k)}$. Since the individual bets were not expected to be profitable, it should also not be expected that reinvesting the money from the bets is profitable, i.e. $\mathbb{E}\left[\prod_{i=1}^{k} s_{(k)}\right] \leq 1$. This will be true regardless of the decision rule used to continue or stop betting.

1.4 Main Contributions and Conclusions

In Chapter 2, the theory of invariant testing [10, 11] is discussed. Roughly, this theory states that statistical inferences should exhibit certain invariance properties. For example, suppose that the data is given as the realisation of a Gaussian random variable $\mathbf{Y} \sim \mathcal{N}(\mu, \sigma^2)$ for some $\mu \in \mathbb{R}$, $\sigma^2 \in \mathbb{R}_{>0}$. A t-test is concerned with whether the *effect size* $\frac{\mu}{\sigma}$ is equal to 0 or equal to some predefined threshold $\delta \in \mathbb{R}_{>0}$. That is, the hypotheses are given by

$$\mathcal{H}_0 = \{\mathcal{N}(0, \sigma^2) : \sigma^2 \in \mathbb{R}_{>0}\}$$

and

$$\mathcal{H}_1 = \left\{\mathcal{N}(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 \in \mathbb{R}_{>0}, \frac{\mu}{\sigma} = \delta\right\}.$$

For such a test, the principle of invariance states that it should not matter whether the data is given by $\mathbf{Y}$ or $c\mathbf{Y}$ for any $c \in \mathbb{R}_{>0}$, since scaling the

random variable does not change the effect size:

$$\frac{\mathbb{E}[\mathbf{Y}]}{\sqrt{\text{Var}[\mathbf{Y}]}} = \frac{c\mathbb{E}[\mathbf{Y}]}{\sqrt{c^2\text{Var}[\mathbf{Y}]}} = \frac{\mathbb{E}[c\mathbf{Y}]}{\sqrt{\text{Var}[c\mathbf{Y}]}}.$$

The statistic that the test is based on is therefore not allowed to change when the data is scaled. An obvious choice is then to use the statistic $t(\mathbf{Y}) := \frac{\mathbf{Y}}{|\mathbf{Y}|}$. Conveniently, it is not hard to verify that under all elements of the null hypothesis, $t(\mathbf{Y})$ has the same distribution (see e.g. [5, Appendix E] for a proof). Analogous to the proof of Proposition 1.3 it follows that

$$\frac{p_{W_1}^{[T]}(t(\mathbf{Y}))}{p_0^{[T]}(t(\mathbf{Y}))}$$

is an E-value for any prior W_1 on the alternative. Here, $p_x^{[T]}$ is used to denote the density of $t(\mathbf{Y})$ when $\mathbf{Y}$ has distribution P_x. It has been shown that the infimum of this E-value over all possible priors is even GROW [5, Theorem 3].

The t-test is only a specific example of a hypothesis test to which the principle of invariant testing can be applied. For general hypothesis tests, it is often the case that there is a group of transformations that do not change the relevant properties of the data. It will be shown that the existence of such a group leads to a straightforward method of obtaining E-variables, analogous to the example above. The difficulty is then in determining whether these E-variables are GROW in general. To this end, a result by Perez et al. [9] will be presented, which gives a number of conditions that taken together are sufficient for this to be the case. The most notable of these conditions is amenability of the aforementioned group [16, 17]. Amenability is a group theoretic property that has a large number of equivalent formulations, which makes amenable groups particularly pleasant to work with. As the main contribution of this chapter, it will be shown in Theorem 2.10 that all the required conditions, including amenability, are met to find the GROW E-variable for a hypothesis test in a linear regression model. This result is essentially a generalisation of the example above involving covariates.

Next, Chapter 3 outlines an approximation algorithm by Li [12], which can be used to approximate the RIPr based on any prior on the alternative for most hypothesis testing problems. An upper bound is found on the expected value of the likelihood ratio of the prior and approximation of its RIPr under the null. Dividing the likelihood ratio by this upper bound

gives an E-variable, which can be seen as an approximation of the GRO E-variable. This follows from the fact that it converges to the GRO E-variable when the number of iterations of the approximation algorithm increases. It is demonstrated that this E-variable can be applied to a hypothesis test in a logistic regression model, which involves covariates. Preliminary simulations are performed to analyse whether the established upper bound is sharp and how well this E-variable performs for this particular hypothesis test.

For this second part, the approximation of the GRO E-variable is compared to a randomisation-based E-variable. The latter owes its name to the fact that it is only an E-variable if the assumption is made that the covariate whose effect is tested is a Bernoulli random variable. While this might seem like a contrived assumption, it arises frequently in practice. One example is in clinical trials, where the effect of a certain drug is tested by dividing the participants in the trial in two groups. The first group is given the treatment, while the second group acts as control group and is given a placebo. The assignment of patients to the groups is often randomised, so that the covariate corresponding to treatment or control is a Bernoulli random variable. While the (approximation of) the earlier mentioned GRO E-variable (that does not rely on randomisation) is still an E-variable in this setting, it might no longer be GRO due to the extra assumption. The main benefit of the randomisation-based E-variable is that it is very easy to compute, whereas the approximation algorithm turns out to be computationally demanding.

Due to the time constraints of this project, it was not possible to fix all the complications that arose during the simulations. One consequence is that the number of data points used in the simulations was very limited $(n \leq 10)$. Another is that all simulations were performed with a degenerate prior on the alternative, while it is generally desirable to be able to work with more generic, non-informative priors. Nevertheless, the results indicated that the established upper bound is not sharp. Since the likelihood ratio is divided by this upper bound, this means that the approximation of the GRO E-variable is unnecessarily weakened. Furthermore, the results suggested that the randomisation-based E-variable outperformed the approximation of the GRO E-variable, even before normalising the latter. So while it is shown that it is possible to find the GROW E-variable for the linear regression model, a lot of future research is needed on E-variables for the logistic regression model.

Chapter 2

E-variables by Invariant Testing

It is a broadly accepted principle that statistical inferences should exhibit certain invariance properties [10, 11]. For example, if the test data is listed in kilometres, the conclusion should be the same as when the data would have been given in miles. The chosen unit of measurement is only one of many arbitrarily chosen notions, which should not impact the inference. Mathematically, this is described by invariance under a suitable group of transformations, which will be made precise in this chapter. Furthermore, it will be shown that this theory leads to a generic method of generating E-variables. This method will be made concrete by constructing an E-variable for linear regression.

The required background on group theory is outlined in Appendix A. This includes a discussion of basic group-theoretic concepts, the Haar measure [18], and amenability [16, 17].

2.1 Background

Consider a statistical model with Lebesgue measurable sample space $\mathcal{Y} \subseteq \mathbb{R}^n$, parameter space Θ, and $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$ a class of distributions on $\mathcal{Y}$. It is assumed that each P_θ with $\theta \in \Theta$ has density p_θ with respect to the Lebesgue measure. Furthermore, $\mathbf{Y}$ will denote a random variable with probability distribution P_θ for some $\theta \in \Theta$. The hypotheses with regards to its distribution are given by

$$\mathcal{H}_0 = \{P_{\theta'} : \theta' \in \Theta_0\} \text{ and } \mathcal{H}_1 = \{P_{\theta'} : \theta' \in \Theta_1\},$$

for subsets $\Theta_0 \subset \Theta$ and $\Theta_1 \subset \Theta \setminus \Theta_0$. Finally, let f be a transformation of $\mathcal{Y}$. For example, if each entry of $\mathbf{Y}$ corresponds to a certain distance in kilometres, $f(\mathbf{y}) = 0.62\mathbf{y}$ is the transformation that changes the unit to miles. All such transformations considered in this section will be assumed to be one-to-one, onto and differentiable. The following definitions introduce a notion of what it means for the transformation f to leave the testing problem invariant.

Definition 2.1. The parameter space Θ is *invariant* under f if, for every $\theta \in \Theta$, there exists a unique $\theta^* \in \Theta$ such that

$$\mathbf{Y} \sim P_\theta \Rightarrow f(\mathbf{Y}) \sim P_{\theta^*}.$$

In such situations, θ^* will be denoted by $\overline{f}(\theta)$.

Definition 2.2. The problem of testing $\mathcal{H}_0$ against $\mathcal{H}_1$ is *invariant* under f if Θ is invariant under f and

$$\overline{f}(\Theta_0) = \Theta_0 \text{ and } \overline{f}(\Theta_1) = \Theta_1.$$

For many testing problems there is not just a single transformation that leaves the problem invariant, but a whole class of transformations. For example, suppose that $Y_1, \ldots, Y_n$ are independent and have distributions $P_{\theta_1}, \ldots, P_{\theta_n}$. It can quite easily be verified that the problem of testing the hypothesis $\theta_1 = \cdots = \theta_n$ versus the hypothesis that the parameters are not all equal is invariant under all permutations of the data points. The numbering of these variables is, after all, an arbitrary notion.

Suppose $\mathcal{C}$ is such a class of transformations of $\mathcal{Y}$, under which the problem of testing $\mathcal{H}_0$ against $\mathcal{H}_1$ is invariant. It should be clear that the problem is also invariant under the composition of elements of $\mathcal{C}$. Similarly, the problem is invariant under all inverses of the transformations in $\mathcal{C}$ too. These inverses exist because the transformations are assumed to be one-to-one and onto. Therefore, $\mathcal{C}$ can always safely be extended to a group. To avoid introducing extra notation for this extension, $(\mathcal{C}, \circ)$ is henceforth simply assumed to be a group. Furthermore, it is assumed that $\mathcal{C}$ can be indexed by some set G, i.e. $\mathcal{C} = \{f_g : g \in G\}$. Consequently, for all $g_1, g_2 \in G$ there exists a $g' \in G$ such that $f_{g_1} \circ f_{g_2} = f_{g'}$. Therefore, an operation $\cdot$ can be defined on G as $g_1 \cdot g_2 = g'$. Then $(G, \cdot)$ is a group since it inherits the group properties from $\mathcal{C}$. Considering G instead of $\mathcal{C}$ simplifies matters because G can be considered to act on both $\mathcal{Y}$ and Θ in the following ways

$$g\mathbf{y} = f_g(\mathbf{y}) \text{ and } g\theta = \overline{f}_g(\theta).$$

Suitably, G will be referred to as the *invariance group*.

2.2 Invariant E-variable

In an invariant testing problem, the relevant structure of $\mathbf{Y}$ and $g\mathbf{Y}$ is equivalent for all $g \in G$. The invariance principle thus states that the outcome

of any test should be the same for input $\mathbf{Y}$ and $g\mathbf{Y}$. It is therefore natural to use a test based on an *invariant statistic*. A statistic is simply a random variable that is defined as a function of the data. Invariance is a property of this function.

Definition 2.3. A function $t(\mathbf{y})$ is said to be *invariant* with respect to G if for all $\mathbf{y} \in \mathcal{Y}$ and $g \in G$,

$$t(g\mathbf{y}) = t(\mathbf{y}). \tag{2.1}$$

This definition leads to a straightforward condition, under which the null hypothesis is simple for data coarsened to an invariant statistic and therefore leads to an E-variable.

Theorem 2.1. *Let $T := t(\mathbf{Y})$ be an invariant statistic. If G acts transitively on Θ_0, then T has the same distribution $P_0^{[T]}$ under all $\theta \in \Theta_0$. Moreover, under this condition the following quantity is an E-variable:*

$$s_{inv}^{W[\theta]}(T) := \frac{p_{W[\theta]}^{[T]}(T)}{p_0^{[T]}(T)} = \frac{\int_{\theta \in \Theta_1} p_\theta^{[T]}(T) dW[\theta]}{p_0^{[T]}(T)}, \tag{2.2}$$

where $W[\theta] \in \mathcal{W}(\Theta_1)$.

Proof. Let $\theta, \theta' \in \Theta_0$ arbitrarily. Since G acts transitively on Θ_0, there exists a $g^* \in G$ such that $g^*\theta = \theta'$. With a slight abuse of notation, it then holds that

$$\begin{aligned} P_\theta^{[T]}(T \in A) &= P_\theta(t(\mathbf{Y}) \in A) \\ &= P_\theta(t(g^*\mathbf{Y}) \in A) \\ &= P_{g^*\theta}(t(\mathbf{Y}) \in A) \\ &= P_{\theta'}^{[T]}(T \in A). \end{aligned}$$

In the second step it is used that, if $\mathbf{Y} \sim P_\theta$, then $g^*\mathbf{Y} = f_{g^*}(\mathbf{Y})$ has distribution $P_{f_{g^*}(\theta)} = P_{g^*\theta}$. This concludes the first part of the statement. The second part follows directly from the fact that the null hypothesis is simple under T. □

In theory, Theorem 2.1 provides an unambiguous method of obtaining an E-value: evaluate an invariant function in the data and calculate the likelihood ratio. In practice, it is still not clear which invariant function should be used, or what the distribution of the resulting invariant statistic is. Before being able to address these points, the concept of *maximal invariants* needs to be discussed.

Definition 2.4. A function $m(\mathbf{y})$ is said to be *maximally invariant* with respect to G if it is invariant and if for $\mathbf{y}_1, \mathbf{y}_2 \in \mathcal{Y}$

$$m(\mathbf{y}_1) = m(\mathbf{y}_2) \Rightarrow \mathbf{y}_1 = g\mathbf{y}_2, \tag{2.3}$$

for some $g \in G$.

The difference between invariants and maximal invariants is best seen by considering how they act on the *orbits* of G. The orbit of an element $\mathbf{y} \in \mathcal{Y}$ under G is the set of all elements that $\mathbf{y}$ can be transformed into by G, i.e.

$$G\mathbf{y} = \{g\mathbf{y} | g \in G\}.$$

Together, all the different orbits of points in $\mathcal{Y}$ under G form a partition of $\mathcal{Y}$. The distinction is then that a function is invariant if it is constant on orbits, while a function is maximally invariant if it is constant on orbits and it assigns a different value to each orbit. To see that maximally invariant functions exist, consider the following systematic way of obtaining one: select a unique point in each orbit and map all points to the uniquely selected point in their respective orbit.

One of the reasons why maximal invariants are useful is given by the following well-known result, which states that any maximal invariant fully characterises all invariant functions.

Proposition 2.2. *If $m(\mathbf{y})$ is a maximal invariant, then $t(\mathbf{y})$ is invariant if and only if it is a function of $m(\mathbf{y})$.*

Proof. Assume that $t(\mathbf{y})$ is invariant and that $m(\mathbf{y}_1) = m(\mathbf{y}_2)$ for some $\mathbf{y}_1, \mathbf{y}_2 \in \mathcal{Y}$. By definition, there exists a $g \in G$ such that $\mathbf{y}_1 = g\mathbf{y}_2$. Then

$$t(\mathbf{y}_1) = t(g\mathbf{y}_2) = t(\mathbf{y}_2),$$

so that if $m(\mathbf{y})$ is known, $t(\mathbf{y})$ can be deduced as well. In other words, $t(\mathbf{y})$ is a function of $m(\mathbf{y})$.

For the other way around, assume that $t(\mathbf{y}) = \rho(m(\mathbf{y}))$ for some function ρ. Then for all $g \in G$

$$t(g\mathbf{y}) = \rho(m(g\mathbf{y})) = \rho(m(\mathbf{y})) = t(\mathbf{y}).$$

So if $t(\mathbf{y})$ is a function of $m(\mathbf{y})$, it is indeed invariant. □

Since any maximal invariant fully determines all invariant functions, they can be considered to somehow contain more information than invariant functions. Intuitively, $s_{inv}^{W[\Theta]}$ should therefore be strongest when evaluated on maximally invariant statistics and it should not matter which

maximal invariant is chosen. Under certain regularity conditions, this can be made precise.

The first step in showing this is given by Andersson's [19] version of a Theorem by Wijsman [20]. It roughly states that all maximally invariant statistics lead to the same E-variable and that it is not necessary to know the distribution of M to calculate $s_{\text{inv}}^{W[\theta]}(M)$.

Theorem 2.3 (Andersson)**.** *Let $M := m(\mathbf{Y})$ be a maximally invariant statistic. If G acts transitively on $\mathcal{H}_0$, G is a σ-locally compact Hausdorff group, and the action of G on $\mathcal{Y}$ is proper, then for any $\theta_0 \in \Theta_0$ and prior $W[\theta]$ on Θ_1, it holds that*

$$s_{inv}^{W[\theta]}(M) = \frac{\int_G p_{W[\theta]}(gY)|\chi_g| d\nu_l(g)}{\int_G p_{\theta_0}(gY)|\chi_g| d\nu_l(g)}, \tag{2.4}$$

where χ_g is the Jacobian determinant of f_g and ν_l is a left Haar-measure on G.

The proof of this Theorem is omitted, as it requires technical background and buildup, which is irrelevant to the present discussion. The details are given by e.g. Andersson [19] or Eaton [21]. The main takeaway here is that the E-variable based on a maximal invariant can be computed without knowledge of its distribution and it is independent of the choice of maximal invariant. The latter is seen because the right hand side of Equation (2.4) does not depend on $m(\mathbf{y})$.

The second and final step is given by Perez et al. [9], who prove that there are conditions under which the E-variable $s_{\text{inv}}^{W[\theta]}(M)$ based on a maximally invariant statistic M is GROW.

Theorem 2.4 (Perez et al.)**.** *Let $M := m(\mathbf{Y})$ be a maximal invariant under G and suppose that the following conditions*[1] *hold:*

1. *The problem of testing $\mathcal{H}_0$ against $\mathcal{H}_1$ is invariant under G,*
2. *G acts transitively on $\mathcal{H}_0$ and $\mathcal{H}_1$, so that the distribution of M can be written as $P_0^{[M]}$ and $P_1^{[M]}$ respectively,*
3. *There exists a function $\phi : \mathcal{Y} \to G$, such that for all $\mathbf{y} \in \mathcal{Y}$ and $g \in G$, $\phi(g\mathbf{y}) = g\phi(\mathbf{y})$ and such that the map $\mathbf{y} \mapsto (\phi(\mathbf{y}), m(\mathbf{y}))$ is a homeomorphism,*

[1] The context considered here differs slightly from the one considered by Perez et al. The assumptions of the Theorem have been adapted to reflect this.

4. *For all* $\theta_1 \in \Theta_1$ *and* $\theta_0 \in \Theta_0$, *there exists an* $\epsilon > 0$ *such that*[2]

$$\mathbb{E}_{h \sim P_{\theta_1}^{[\phi(\mathbf{Y})]}} \left[\left| \log \left(\frac{p_{\theta_1}^{[\phi(\mathbf{Y})]}(h)}{p_{\theta_0}^{[\phi(\mathbf{Y})]}(h)} \right) \right|^{1+\epsilon} \right] < \infty$$

and

$$\mathbb{E}_{m \sim P_1^{[M]}} \left[\left| \log \left(\frac{p_1^{[M]}(m)}{p_0^{[M]}(m)} \right) \right|^{1+\epsilon} \right] < \infty.$$

5. $D\left(P_1^{[M]} \| P_0^{[M]}\right) < \infty$,

6. *G is a* σ*-compact and locally compact Hausdorff topological group,*

7. *G is amenable,*

then the E-variable $s_{inv}^{W[\theta]}(M)$ *is independent of the prior* $W[\theta]$ *and is GROW with respect to* Y.

It may seem unnatural that amenability plays a role in determining whether the invariant E-variable is GROW. However, amenability has previously been shown to be a necessary condition for the Hunt-Stein theorem on the existence of minimax invariant tests [22]. Theorem 2.4 can therefore be seen as an analogue to known results in the classical setting. The established theory of invariance can thus be adapted to the framework of E-variables. Of course, this is only useful if there are settings that satisfy all the required conditions. To demonstrate that this is the case, the next section will show how the theory discussed so far can be applied to hypothesis testing in linear regression models.

[2]This need not hold for all ϕ as in condition 3, only for one particular choice.

2.3 GROW E-variable for Linear Regression

Suppose that the data consists of n observations: $\{y_i, \mathbf{x}_i\}_{i=1}^n$. Each y_i is a scalar *response/dependent variable/regressand* and each $\mathbf{x}_i \in \mathbb{R}^{p-1}$ is a vector of *covariates/independent variables/regressors*. The linear regression model considered here assumes a linear relation between response and covariates:

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \dots \beta_p x_{i,p-1} + \epsilon_i, \tag{2.5}$$

where $\boldsymbol{\beta} \in \mathbb{R}^p$ is a vector of (unknown) coefficients and ϵ_i is unobserved noise. To simplify the notation, each $\mathbf{x}_i$ is padded with a zeroth entry equal to 1 and X is used to denote the matrix with row-vectors $\mathbf{x}_i$, so that Eq. (2.5) reduces to

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{2.6}$$

Adding a constant entry to the covariates allows for the assumption that the expected value of the noise is zero. After all, if it is not zero, this can simply be compensated for by β_0. On top of that, it is often assumed that the noise is i.i.d. Gaussian.

A testing problem that frequently arises is concerned with whether the impact of some covariate (say covariate j) on the response crosses a certain threshold. To this end, it is tested whether the *effect size* β_j/σ is equal to 0 or larger than some $\delta \in \mathbb{R}_{\geq 0}$. Since the interesting properties of hypothesis tests are defined for the worst case, this is equivalent to testing whether β_j/σ is equal to 0 or equal to δ. To do so, it is useful to rewrite Equation (2.6) to:

$$\mathbf{y} = X_{-j}\boldsymbol{\beta}_{-j} + X_j\beta_j + \boldsymbol{\epsilon}, \tag{2.7}$$

where X_j is the j-th column of X and X_{-j} denotes X without the j-th column (and similar for $\boldsymbol{\beta}$). It is assumed that X_{-j} is full column rank. Considering $\boldsymbol{\beta}_{-j}$ and β_j as two separate parameters, this leads to the statistical model $\mathcal{Y} = \mathbb{R}^n$, $\Theta = \mathbb{R}^{p-1} \times \mathbb{R} \times \mathbb{R}^+$ and

$$\mathcal{P} = \{P_{\boldsymbol{\beta}_{-j},\beta_j,\sigma} : (\boldsymbol{\beta}_{-j}, \beta_j, \sigma) \in \Theta\}.$$

Here, $P_{\boldsymbol{\beta}_{-j},\beta_j,\sigma}$ represents the normal distribution with mean $X_{-j}\boldsymbol{\beta}_{-j} + X_j\beta_j$ and variance $\sigma^2 I_n$. The null hypothesis $\mathcal{H}_0$ and alternative hypothesis $\mathcal{H}_1$ are given by

$$\mathcal{H}_0 = \left\{ P_{\boldsymbol{\beta}_{-j},\beta_j,\sigma} \,\middle|\, (\boldsymbol{\beta}_{-j}, \beta_j, \sigma) \in \Theta_0 = \left\{ (\boldsymbol{\beta}'_{-j}, \beta'_j, \sigma') \in \Theta \,\middle|\, \frac{\beta'_j}{\sigma'} = 0 \right\} \right\} \tag{2.8}$$

and

$$\mathcal{H}_1 = \left\{ P_{\boldsymbol{\beta}_{-j},\beta_j,\sigma} \,\middle|\, (\boldsymbol{\beta}_{-j},\beta_j,\sigma) \in \Theta_1 = \left\{ (\boldsymbol{\beta}'_{-j},\beta'_j,\sigma') \in \Theta \,\middle|\, \frac{\beta'_j}{\sigma'} = \delta \right\} \right\}. \quad (2.9)$$

Example (t-test). If there are no covariates ($p = 1$), (2.6) reduces to

$$\mathbf{y} = \beta + \boldsymbol{\epsilon}.$$

In this case, it is thus assumed that n i.i.d. observations of the same normal distribution are observed. There are only two parameters: the mean β and variance σ^2 of this normal distribution. Testing $\mathcal{H}_0$ versus $\mathcal{H}_1$ as in (2.8) and (2.9) comes down to testing whether the effect size β/σ is equal to zero or to some specified value, which is commonly referred to as a t-test. This example is continued throughout this section, but solely for clarity: all results in this simplified setting have previously been discussed by Grünwald et al. [5, Section 4.3].

2.3.1 Invariance group

To obtain an expression for the invariant E-variable of this testing problem, the invariance group will first be examined. No novelty is claimed in this section, as this has previously been described by e.g. Eaton [21, Section 3.4] and Kariya [23].

Since the testing problem is concerned with whether the j-th coefficient is equal to zero, it should not matter whether the data is scaled or if the influence of other coefficients is altered. In mathematical terms, such transformations are contained in the class

$$\mathcal{C} = \{ f_{\alpha_0,\boldsymbol{\alpha}} | \alpha_0 \in \mathbb{R}_{>0}, \boldsymbol{\alpha} \in \mathbb{R}^q \},$$

where $q := p - 1$ and $f_{\alpha_0,\boldsymbol{\alpha}}(\mathbf{y}) = \alpha_0 \mathbf{y} + X_{-j}\boldsymbol{\alpha}$.

Proposition 2.5. *The problem of testing $\mathcal{H}_0$ (2.8) versus $\mathcal{H}_1$ (2.9) is invariant under $\mathcal{C}$.*

Proof. Let $\mathbf{Y} \sim P_{\boldsymbol{\beta}_{-j},\beta_j,\sigma}$. For arbitrary $\alpha_0 \in \mathbb{R}_{>0}$ and $\boldsymbol{\alpha} \in \mathbb{R}^q$, it holds that

$$f_{\alpha_0,\boldsymbol{\alpha}}(\mathbf{Y}) = \alpha_0 \mathbf{Y} + X_{-j}\boldsymbol{\alpha} \sim P_{\alpha_0\boldsymbol{\beta}_{-j}+\boldsymbol{\alpha},\alpha_0\beta_j,\alpha_0\sigma},$$

i.e. $\overline{f}_{\alpha_0,\boldsymbol{\alpha}}(\boldsymbol{\beta}_{-j},\beta_j,\sigma) = (\alpha_0\boldsymbol{\beta}_{-j} + \boldsymbol{\alpha}, \alpha_0\beta_j, \alpha_0\sigma)$. Since $\alpha_0 \in \mathbb{R}_{>0}$, it is true that $\frac{\alpha_0\beta_j}{\alpha_0\sigma} = \frac{\beta_j}{\sigma}$. Therefore, for $i \in \{0,1\}$,

$$(\boldsymbol{\beta}_{-j},\beta_j,\sigma) \in \Theta_i \Leftrightarrow \overline{f}_{\alpha_0,\boldsymbol{\alpha}}(\boldsymbol{\beta}_{-j},\beta_j,\sigma) \in \Theta_i.$$

It follows that $\overline{f}_{\alpha_0,\boldsymbol{\alpha}}(\Theta_j) = \Theta_j$, so that the testing problem is indeed invariant under $\mathcal{C}$. □

Note that the class $\mathcal{C}$ itself contains compositions and inverses, because for $\alpha_0, \gamma_0 \in \mathbb{R}_{>0}$ and $\boldsymbol{\alpha}, \boldsymbol{\gamma} \in \mathbb{R}^q$, it holds that

$$\begin{aligned} f_{\gamma_0,\boldsymbol{\gamma}}(f_{\alpha_0,\boldsymbol{\alpha}}(y)) &= f_{\gamma_0,\boldsymbol{\gamma}}(\alpha_0 y + X_{-j}\boldsymbol{\alpha}) \\ &= \gamma_0\alpha_0 y + X_{-j}(\gamma_0\boldsymbol{\alpha} + \boldsymbol{\gamma}) \\ &= f_{\gamma_0\alpha_0,\gamma_0\boldsymbol{\alpha}+\boldsymbol{\gamma}}(y) \end{aligned}$$

and

$$\begin{aligned} f_{\frac{1}{\alpha_0},-\frac{1}{\alpha_0}\boldsymbol{\alpha}}(f_{\alpha_0,\boldsymbol{\alpha}}(y)) &= \frac{1}{\alpha_0}(\alpha_0 y + X_{-j}\boldsymbol{\alpha}) - \frac{1}{\alpha_0}X_{-j}\boldsymbol{\alpha} \\ &= y. \end{aligned}$$

Therefore, the invariance group is simply the index set

$$G = \{(\alpha_0, \boldsymbol{\alpha}) : \alpha_0 \in \mathbb{R}_{>0}, \boldsymbol{\alpha} \in \mathbb{R}^q\} = \mathbb{R}_{>0} \times \mathbb{R}^q$$

together with the operation $(\gamma_0, \boldsymbol{\gamma}) \cdot (\alpha_0, \boldsymbol{\alpha}) = (\gamma_0\alpha_0, \gamma_0\boldsymbol{\alpha} + \boldsymbol{\gamma})$.

Example (t-test, continued)**.** If $Y \sim \mathcal{N}(\beta, \sigma^2)$ and $c \in \mathbb{R}_{>0}$, then $cY \sim \mathcal{N}(c\beta, c^2\sigma^2)$. The effect size of both of these distributions is equal, i.e. $(c\beta)/(\sqrt{c^2\sigma^2}) = \beta/\sigma$. Since this is the quantity that is being tested, scaling the data does not change the properties of the data which are relevant to the test. The same does not hold for just any transformation. For example, the addition of a scalar does change the relevant properties, since $(c + \mathbf{Y}) \sim \mathcal{N}(\beta + c, \sigma^2)$ and $(\beta + c)/\sigma \neq \beta/\sigma$. The natural class of transformations that leave the testing problem invariant is therefore given by

$$\mathcal{C} = \{f_{\alpha_0} | \alpha_0 \in \mathbb{R}_{>0}\},$$

where $f_{\alpha_0}(\mathbf{y}) = \alpha_0\mathbf{y}$. Composing two elements of $\mathcal{C}$ again gives an element of $\mathcal{C}$:

$$f_{\gamma_0}(f_{\alpha_0}(\mathbf{y})) = \gamma_0\alpha_0\mathbf{y} = f_{\gamma_0\alpha_0}(\mathbf{y}).$$

Similarly, $\mathcal{C}$ contains inverses too and therefore $(\mathcal{C}, \circ)$ is a group. The invariance group is then simply the index set $\mathbb{R}_{>0}$ with group operation multiplication.

It is shown in Appendix B.1 that G is a σ-compact and locally compact Hausdorff topological group and that it is amenable too, i.e. conditions 6 and 7 of Theorem 2.4 are satisfied. Additionally, the following lemma shows that conditions 2 of Theorem 2.4 is satisfied too.

Lemma 2.6. *The action of G on $\mathcal{H}_0$ and $\mathcal{H}_1$ is transitive.*

Proof. Let $(\boldsymbol{\beta}_{-j}, \beta_j, \sigma), (\boldsymbol{\beta}'_{-j}, \beta'_j, \sigma') \in \mathcal{H}_1$ arbitrarily. Then $\left(\frac{\sigma'}{\sigma}, \boldsymbol{\beta}'_{-j} - \frac{\sigma'}{\sigma}\boldsymbol{\beta}_{-j}\right) \in G$ and it holds that

$$\left(\frac{\sigma'}{\sigma}, \boldsymbol{\beta}'_{-j} - \frac{\sigma'}{\sigma}\boldsymbol{\beta}_{-j}\right)(\boldsymbol{\beta}_{-j}, \beta_j, \sigma) = (\boldsymbol{\beta}'_{-j}, \frac{\sigma'}{\sigma}\beta_j, \sigma') = (\boldsymbol{\beta}'_{-j}, \beta'_j, \sigma'),$$

where it is used that $\beta_j/\sigma = \delta$ and $\delta\sigma' = \beta'_j$. Analogously, G works transitively on $\mathcal{H}_0$ too. □

2.3.2 Maximal invariant

Now that the invariance group is determined, the next step is to show that a maximal invariant under this group exists. The maximal invariant described for this purpose has previously been studied by e.g. Bhowmik [24] and Kariya [23].

Consider the matrix $N = I_n - X_{-j}(X_{-j}^T X_{-j})^{-1} X_{-j}^T$. This matrix is well-defined, as the existence of the inverse $(X_{-j}^T X_{-j})^{-1}$ follows from the assumption that X_{-j} has full column rank. Next, let A be a $k \times n$ matrix such that $AA^T = I_k$ and $A^T A = N$, where $k := n - q$. The existence of such a matrix can be shown as follows: $X_{-j}(X_{-j}^T X_{-j})^{-1} X_{-j}^T$ represents the standard projection onto the column space of X_{-j}. Therefore, N represents the projection onto the orthogonal complement of the column space of X_{-j}. Any matrix, whose rows form an orthonormal basis for this orthogonal complement then automatically satisfies the restrictions for A. Finally, denote $M^* := m^*(\mathbf{Y})$, where $m^* : \mathcal{Y} \to \mathbb{R}^k$ is the function that maps[3] $\mathbf{y} \mapsto \frac{A\mathbf{y}}{\|A\mathbf{y}\|}$.

Proposition 2.7. *The function $m^*(\mathbf{y})$ is maximally invariant with respect to G.*

Proof. It will first be shown that m^* is invariant. To that end, let $\mathbf{y} \in \mathcal{Y}$ and $(\alpha_0, \boldsymbol{\alpha}) \in G$ arbitrarily. Then it holds that

$$\begin{aligned} A((\alpha_0, \boldsymbol{\alpha})\mathbf{y}) &= A(\alpha_0 \mathbf{y} + X_{-j}\boldsymbol{\alpha}) \\ &= A\alpha_0 \mathbf{y} + AX_{-j}\boldsymbol{\alpha} \\ &= A\alpha_0 \mathbf{y} + ANX_{-j}\boldsymbol{\alpha} \\ &= \alpha_0 A\mathbf{y}, \end{aligned}$$

[3]The function m^* would actually only be well defined if the sample space was $\mathbb{R}^n \setminus \{\mathbf{0}\}$. However, $\mathbf{0}$ has measure 0 under all distributions $P_\theta, \theta \in \Theta$, so this technicality is simply ignored.

where it is used that $A = AN$ and $NX_{-j} = 0$. Then also

$$\begin{aligned} m^*((\alpha_0, \boldsymbol{\alpha})(\mathbf{y})) &= \frac{\alpha_0 A\mathbf{y}}{\|\alpha_0 A\mathbf{y}\|} \\ &= \frac{A\mathbf{y}}{\|A\mathbf{y}\|} = m^*(\mathbf{y}). \end{aligned}$$

Secondly, to show that m^* is maximally invariant, let $\mathbf{y}_1, \mathbf{y}_2 \in \mathcal{Y}$ arbitrarily such that $m^*(\mathbf{y}_1) = m^*(\mathbf{y}_2)$. Then it is true that

$$\begin{aligned} & A\left(\frac{\mathbf{y}_1}{\|A\mathbf{y}_1\|} - \frac{\mathbf{y}_2}{\|A\mathbf{y}_2\|}\right) = 0 \\ & \Rightarrow \frac{\mathbf{y}_1}{\|A\mathbf{y}_1\|} - \frac{\mathbf{y}_2}{\|A\mathbf{y}_2\|} \in \mathrm{Nul}(A) \subseteq \mathrm{Nul}(N). \end{aligned}$$

The nullspace of A is a subset (they are in fact equal) of the nullspace of N because for $\mathbf{x} \in \mathcal{Y}$: if $A\mathbf{x} = 0$, then $N\mathbf{x} = A^T A\mathbf{x} = A^T\mathbf{0} = \mathbf{0}$. This means that $\frac{\mathbf{y}_1}{\|A\mathbf{y}_1\|} - \frac{\mathbf{y}_2}{\|A\mathbf{y}_2\|}$ is orthogonal to the orthogonal space of $\mathrm{col}(X_{-j})$, which is exactly the case if it is an element of $\mathrm{col}(X_{-j})$. Therefore, there is an $\boldsymbol{\alpha} \in \mathbb{R}^q$ such that

$$\begin{aligned} \frac{\mathbf{y}_1}{\|A\mathbf{y}_1\|} - \frac{\mathbf{y}_2}{\|A\mathbf{y}_2\|} &= X_{-j}\boldsymbol{\alpha} \\ \frac{\mathbf{y}_1}{\|A\mathbf{y}_1\|} &= \frac{\mathbf{y}_2}{\|A\mathbf{y}_2\|} + X_{-j}\boldsymbol{\alpha} \\ \mathbf{y}_1 &= \mathbf{y}_2 \frac{\|A\mathbf{y}_1\|}{\|A\mathbf{y}_2\|} + X_{-j}\boldsymbol{\alpha} \\ \mathbf{y}_1 &= \left(\frac{\|A\mathbf{y}_1\|}{\|A\mathbf{y}_2\|}, \boldsymbol{\alpha}\right)\mathbf{y}_2. \end{aligned}$$

Since $\left(\frac{\|A\mathbf{y}_1\|}{\|A\mathbf{y}_2\|}, \boldsymbol{\alpha}\right) \in G$, m^* is maximally invariant. □

Example (t-test, continued)**.** Without covariates, the matrix X_{-j} is not well-defined. This can be sidestepped by defining it as a vector of n zeroes. The column space of X_{-j} is then the set containing only the origin, i.e. $\{\mathbf{0}\}$. The matrix N represents the projection onto the orthogonal complement of $\{\mathbf{0}\}$. This orthogonal complement is given by all of $\mathbb{R}^n$ so that N is equal to the identity matrix I_n. Any matrix whose rows form an orthonormal basis of $\mathbb{R}^n$ is, therefore, a valid choice for A. For simplicity, let A be the identity matrix. The maximally invariant function is then given by $m^*(\mathbf{y}) = \mathbf{y}/\|\mathbf{y}\|$. This is a variation of the maximal invariant used by Grünwald et al. [5], who use the mapping $m'(\mathbf{y}) = \mathbf{y}/|y_1|$. The intuition

behind both invariants is the same: by dividing, the influence of the scale parameter is nullified. That is, for $\mathbf{Y} \sim \mathcal{N}(\boldsymbol{\beta}, \sigma^2)$, the distribution of both $m^*(\mathbf{Y})$ and $m'(\mathbf{Y})$ no longer depends on σ, but only on the effect size β/σ. The latter is, of course, precisely the quantity that is being tested!

To compute the likelihood ratio for M^*, Bhowmik [24] shows that its density function is given by

$$p^{[M^*]}_{\boldsymbol{\beta}_{-j},\beta_j,\sigma}(\mathbf{m}) = \frac{1}{2}\Gamma\left(\frac{k}{2}\right)\pi^{-\frac{k}{2}}e^{c(\beta_j/\sigma)}\left[{}_1F_1\left(\frac{k}{2},\frac{1}{2},\frac{a^2(\mathbf{m},\beta_j/\sigma)}{2}\right)\right.$$
$$\left. + \sqrt{2}a(\mathbf{m},\beta_j/\sigma)\,\frac{\Gamma((1+k)/2)}{\Gamma(k/2)}{}_1F_1\left(\frac{1+k}{2},\frac{3}{2},\frac{a^2(\mathbf{m},\beta_j/\sigma)}{2}\right)\right],$$

where $\mathbf{m}$ is a unit vector,

$$a\left(\mathbf{m},\frac{\beta_j}{\sigma}\right) = \frac{\beta_j}{\sigma}\mathbf{m}^T A X_j, \quad c\left(\frac{\beta_j}{\sigma}\right) = -\frac{1}{2}\left(\frac{\beta_j}{\sigma}\right)^2 X_j^T N X_j$$

and ${}_1F_1$ is the confluent hypergeometric function. Under the null hypothesis this reduces to the uniform distribution over the k-dimensional unit sphere. With this knowledge, it can be checked that condition 5 of Theorem 2.4 holds.

Lemma 2.8. $D\left(P_1^{[M^*]} \| P_0^{[M^*]}\right) < \infty$.

Proof. It can be seen that for every unit vector $\mathbf{m}$

$$\log\left(\frac{p_1^{[M^*]}(\mathbf{m})}{P_0^{[M^*]}(\mathbf{m})}\right) = \log(e^{c(\delta)}\left[{}_1F_1\left(\frac{k}{2},\frac{1}{2},\frac{a^2(\mathbf{m},\delta)}{2}\right)\right.$$
$$\left. + \sqrt{2}a(\mathbf{m},\delta)\,\frac{\Gamma((1+k)/2)}{\Gamma(k/2)}{}_1F_1\left(\frac{1+k}{2},\frac{3}{2},\frac{a^2(\mathbf{m},\delta)}{2}\right)\right]).$$

The confluent hypergeometric function is given by

$${}_1F_1(a,b,z) = 1 + \frac{a}{b}z + \frac{a(a+1)}{b(b+1)}\frac{z^2}{2} + \cdots = \sum_{k=0}^{\infty}\frac{\Gamma(a+k)\Gamma(b)}{\Gamma(a)\Gamma(b+k)}\frac{z^k}{k!}.$$

The ratio of the $k+1$st and kth term of this summation is given by

$$\frac{\frac{\Gamma(a+k+1)\Gamma(b)}{\Gamma(a)\Gamma(b+k+1)}\frac{z^{k+1}}{(k+1)!}}{\frac{\Gamma(a+k)\Gamma(b)}{\Gamma(a)\Gamma(b+k)}\frac{z^k}{k!}} = \frac{a+k}{b+k}\frac{z}{k+1},$$

so by the ratio test, ${}_1F_1(a,b,z)$ is finite for all finite a, b and z. Furthermore, it can be seen that for positive inputs, ${}_1F_1(a,b,z)$ is increasing in z. Also, by the Cauchy-Schwarz inequality, it holds that

$$|a(\mathbf{m},\delta)| = \delta|\mathbf{m}^T A X_j| \leq \delta\|\mathbf{m}\|\|AX_j\| = \delta\|AX_j\|.$$

Putting these things together, it follows that $\log\left(\frac{p_1^{[M^*]}(\mathbf{m})}{p_0^{[M^*]}(\mathbf{m})}\right)$ is bounded from above. Furthermore, the support of $P_1^{[M]}$ is the k-dimensional unit sphere, so that the integral over its support is finite. Therefore,

$$D\left(P_1^{[M]}\|P_0^{[M]}\right) = \int \log\left(\frac{p_1^{[M]}(\mathbf{m})}{P_0^{[M]}(\mathbf{m})}\right) dP_1^{[M]} < \infty.$$

□

Now, let $B := (X_{-j}^T X_{-j})^{-1} X_{-j}^T$ and suppose that for some unknown $\mathbf{y} \in \mathcal{Y}$, the triplet

$$(\|A\mathbf{y}\|, B\mathbf{y}, m^*(\mathbf{y}))$$

is known. With this info at hand, multiplying $m^*(\mathbf{y})$ by $\|A\mathbf{y}\|$ gives the vector $A\mathbf{y}$. In turn, multiplication by A^T and addition of $X_{-j}B\mathbf{y}$ grants the previously unknown $\mathbf{y}$:

$$\begin{aligned} A^T A\mathbf{y} + X_{-j}B\mathbf{y} &= (A^T A + X_{-j}B)\mathbf{y} \\ &= (I_n - X_{-j}(X_{-j}^T X_{-j})^{-1}X_{-j}^T + X_{-j}(X_{-j}^T X_{-j})^{-1}X_{-j}^T)\mathbf{y} \\ &= \mathbf{y}. \end{aligned}$$

Note that $(\|A\mathbf{y}\|, B\mathbf{y}) \in \mathbb{R}_{>0} \times \mathbb{R}^q = G$ and that all involved operations are continuous. Assuming that A and X_{-j} are given, the map given by $\mathbf{y} \mapsto (\|A\mathbf{y}\|, B\mathbf{y}, m^*(\mathbf{y}))$ is a homeomorphism between $\mathcal{Y}$ and $G \times \mathbb{R}^k$. To see that condition 3 of Theorem 2.4 is then satisfied, let $\phi : \mathcal{Y} \mapsto G$ denote the function that maps $\mathbf{y} \mapsto (\|A\mathbf{y}\|, B\mathbf{y})$. It then indeed holds for $(\alpha_0, \boldsymbol{\alpha}) \in G$ and $\mathbf{y} \in \mathcal{Y}$,

$$\begin{aligned} \phi((\alpha_0,\boldsymbol{\alpha})\mathbf{y}) &= \phi(\alpha_0\mathbf{y} + X_{-j}\boldsymbol{\alpha}) \\ &= (\|A\alpha_0\mathbf{y} + AX_{-j}\boldsymbol{\alpha}\|, B(\alpha_0\mathbf{y} + X_{-j}\boldsymbol{\alpha})) \\ &= (\alpha_0\|A\mathbf{y}\|, \alpha_0 B\mathbf{y} + \boldsymbol{\alpha}) \\ &= (\alpha_0,\boldsymbol{\alpha})\phi(\mathbf{y}), \end{aligned}$$

Notice that ϕ depends on the particular choice of A so that there exists such a function for each possible choice.

Furthermore, if $\mathbf{Y} \sim P_{\beta_{-j},\beta_j,\sigma}$, then

$$A\mathbf{Y} \sim \mathcal{N}(AX_j\beta_j, \sigma^2 I_k) \text{ and } B\mathbf{Y} \sim \mathcal{N}(\beta_{-j} + BX_j\beta_j, \sigma^2(X_{-j}^T X_{-j})^{-1}).$$

The covariance of these jointly normally distributed random variables is given by

$$\begin{aligned}
\operatorname{cov}(A\mathbf{Y}, B\mathbf{Y}) &= \mathbb{E}\left[(A\mathbf{Y} - \mathbb{E}[A\mathbf{Y}])(B\mathbf{Y} - \mathbb{E}[B\mathbf{Y}])^T\right] \\
&= A\mathbb{E}\left[(\mathbf{Y} - \mathbb{E}[\mathbf{Y}])(\mathbf{Y} - \mathbb{E}[\mathbf{Y}])^T\right] B^T \\
&= \sigma^2 AB^T \\
&= \sigma^2 A\left(I_n - X_{-j}(X_{-j}^T X_{-j})^{-1} X_{-j}^T\right) X_{-j}(X_{-j}^T X_{-j})^{-1} \\
&= 0,
\end{aligned}$$

where it is used that $A = A(I_n - X_{-j}(X_{-j}^T X_{-j})^{-1} X_{-j}^T)$. It follows that $A\mathbf{Y}$ and $B\mathbf{Y}$ are independent. Additionally, $\|A\mathbf{Y}\|^2$ is the sum of squares of independent Gaussians and thus follows a scaled non-central chi-square distribution with k degrees of freedom and non-centrality parameter

$$\lambda(\beta_j) = \frac{1}{\sigma^2}\sum_{i=1}^{k}(AX_j\beta_j)_i^2.$$

Under the null hypothesis, this reduces to a scaled chi-square distribution with k degrees of freedom. The density of $\|A\mathbf{Y}\|$ can then be found using a simple change of variables. Now, it can be shown that condition 4 of Theorem 2.4 holds.

Lemma 2.9. *For all $\theta_1 \in \Theta_1$ and $\theta_0 \in \Theta_0$, there exists an $\epsilon > 0$ such that*

$$\mathbb{E}_{h\sim P_{\theta_1}^{[\phi(\mathbf{Y})]}}\left[\left|\log\left(\frac{p_{\theta_1}^{[\phi(\mathbf{Y})]}(h)}{p_{\theta_0}^{[\phi(\mathbf{Y})]}(h)}\right)\right|^{1+\epsilon}\right] < \infty$$

and

$$\mathbb{E}_{m\sim P_1^{[M^*]}}\left[\left|\log\left(\frac{p_1^{[M^*]}(m)}{p_0^{[M^*]}(m)}\right)\right|^{1+\epsilon}\right] < \infty,$$

Proof. The proof is given in Appendix B.1. □

2.3.3 GROW E-variable

Finally, all components can be put together to find the best E-variable for this testing problem.

Theorem 2.10. *The GROW E-variable S^* for testing $\mathcal{H}_0$ against $\mathcal{H}_1$ exists and satisfies*

$$S^* = \frac{p_1^{[M^*]}(M^*)}{p_0^{[M^*]}(M^*)} \tag{2.10}$$

Proof. The regularity conditions shown in Appendix B.1 together with Lemma 2.6, 2.8, and 2.9 show that the conditions of Theorem 2.4 are satisfied. The statement then follows directly. □

One problem with the representation of the E-variable in the form of (2.10) is that the density of the maximal invariant does not have an appealing expression. In the case without covariates, Grünwald et al. [5, Section 4.3] avoid this problem by showing that the density of the statistic can be rewritten to a straightforward integral. The following Proposition demonstrates how Theorem 2.3 can be applied to generalise this result. This idea is not novel, as a more general result has previously been proved by Berger et al. [25, Theorem 2.1].

Proposition 2.11. *The GROW E-variable S^* satisfies*

$$S^* = \frac{\int_{\theta_1 \in \Theta_1} p_{\theta_1}(y) \frac{1}{\sigma(\theta_1)} d\theta_1}{\int_{\theta_0 \in \Theta_0} p_{\theta_0}(y) \frac{1}{\sigma(\theta_0)} d\theta_0}. \tag{2.11}$$

Proof. It is shown in Appendix B.1 that action of G on $\mathcal{Y}$ is proper, so by Theorem 2.3 it holds for arbitrary $\theta_0 \in \Theta_0$ and $\theta_1 \in \Theta_1$ that

$$\begin{aligned} \frac{p_1^{[M^*]}(M^*)}{p_0^{[M^*]}(M^*)} &= \frac{\int_G p_{\theta_1}(gY)|\chi_g| d\nu_l(g)}{\int_G p_{\theta_0}(gY)|\chi_g| d\nu_l(g)} \\ &= \frac{\int_{(\alpha_0, \boldsymbol{\alpha}) \in G} p_{\theta_1}((\alpha_0, \boldsymbol{\alpha})Y)|\chi_{\alpha_0, \boldsymbol{\alpha}}| d\nu_l(\alpha_0, \boldsymbol{\alpha})}{\int_{(\alpha_0, \boldsymbol{\alpha}) \in G} p_{\theta_0}((\alpha_0, \boldsymbol{\alpha})Y)|\chi_{\alpha_0, \boldsymbol{\alpha}}| d\nu_l(\alpha_0, \boldsymbol{\alpha})}. \end{aligned}$$

Recall that $\chi_{\alpha_0, \boldsymbol{\alpha}}$ is the Jacobian determinant of the transformation $f_{\alpha_0, \boldsymbol{\alpha}}$, which maps $\mathbf{y} \mapsto \alpha_0 \mathbf{y} + \boldsymbol{\alpha}$. The Jacobian matrix of this transformation equals $\alpha_0 I_n$, so that $\chi_{\alpha_0, \boldsymbol{\alpha}} = \alpha_0^n$. Furthermore, it is shown in Appendix B.1

that $d\nu_l(\alpha_0, \alpha) = \frac{1}{\alpha_0^{n+1}} d\alpha_0 d\boldsymbol{\alpha}$ is a left Haar measure on G. Substituting this in the previous equation results in

$$\begin{aligned}\frac{p_1^{[M^*]}(M^*)}{p_0^{[M^*]}(M^*)} &= \frac{\int_{(\alpha_0,\boldsymbol{\alpha})\in G} p_{\theta_1}((\alpha_0,\boldsymbol{\alpha})Y)\frac{1}{\alpha_0} d\alpha_0 d\boldsymbol{\alpha}}{\int_{(\alpha_0,\boldsymbol{\alpha})\in G} p_{\theta_0}((\alpha_0,\boldsymbol{\alpha})Y)\frac{1}{\alpha_0} d\alpha_0 d\boldsymbol{\alpha}} \\ &= \frac{\int_{(\alpha_0,\boldsymbol{\alpha})\in G} p_{(\alpha_0,\boldsymbol{\alpha})\theta_1}(Y)\frac{1}{\alpha_0} d\alpha_0 d\boldsymbol{\alpha}}{\int_{(\alpha_0,\boldsymbol{\alpha})\in G} p_{(\alpha_0,\boldsymbol{\alpha})\theta_0}(Y)\frac{1}{\alpha_0} d\alpha_0 d\boldsymbol{\alpha}}.\end{aligned}$$

Next, denote $\theta_1 = (\boldsymbol{\beta}_{-j}, \beta_j, \sigma)$, so that the numerator can be rewritten as

$$\begin{aligned}\int_{(\alpha_0,\boldsymbol{\alpha})\in G} p_{(\alpha_0,\boldsymbol{\alpha})\theta_1}(Y)\frac{1}{\alpha_0} d\alpha_0 d\boldsymbol{\alpha} &= \int_{(\alpha_0,\boldsymbol{\alpha})\in G} p_{\alpha_0\boldsymbol{\beta}_{-j}+\boldsymbol{\alpha},\alpha_0\beta_j,\alpha_0\sigma}(Y)\frac{1}{\alpha_0} d\alpha_0 d\boldsymbol{\alpha} \\ &= \int_{(\sigma',\boldsymbol{\beta}'_{-j})\in G} p_{\boldsymbol{\beta}'_{-j},\sigma'\delta,\sigma'}(y)\frac{1}{\sigma'} d\beta'_{-j} d\sigma' \\ &= \int_{\theta'_1\in\Theta_1} p_{\theta'_1}(y)\frac{1}{\sigma(\theta'_1)} d\theta'_1.\end{aligned}$$

The second equality follows from the change of variables given by

$$(\alpha_0, \boldsymbol{\alpha}) \to (\alpha_0\sigma, \alpha_0\boldsymbol{\beta}_{-j} + \boldsymbol{\alpha}).$$

The Jacobian determinant of this transformation is σ, so that $d\alpha_0 d\boldsymbol{\alpha} = \frac{1}{\sigma} d\boldsymbol{\beta}_{-j} d\sigma'$. The denominator can be rewritten in the same fashion, from which the result follows. □

It has thus been shown that the common hypothesis testing problem discussed in this section satisfies the required conditions for obtaining E-variables via the principles of invariant testing. This illustrates that the theory of E-variables is rich enough to deal with problems involving covariates. However, not all such problems will satisfy the invariance properties assumed in this chapter. In such cases, approximation methods might offer a solution. This will be further discussed in the next chapter.

Chapter 3

E-variables by Approximation

Unlike the kinds of testing problems discussed in Chapter 2, some other testing problems do not naturally show any invariance properties. Suppose that for such a problem without invariance properties, there is a prior on the alternative. Then there might not be an easier way to find the GRO E-variable than to explicitly find an expression for the RIPr, as used in Theorem 1.4. Unfortunately, this may not always be possible using analytical methods and numerical methods might not always be computationally feasible. However, one of the main Theorems by Li describes a way to construct an iterative approximation of the RIPr [12]. The likelihood ratio based on this approximation can be thought of as an approximation of the GRO E-variable, although it is not an E-variable itself. In this chapter, a bound will be given on its expected value under the null, which gives an E-variable after normalisation. Furthermore, the strength of this method will be assessed by applying it to a hypothesis test for logistic regression.

3.1 Approximating the RIPr

As before, let $\mathcal{Y}$ be a sample space, Θ a parameter space and $\mathcal{P}$ a class of distributions indexed by the parameter space. Recall that for an arbitrary probability distribution Q on $\mathcal{Y}$, the RIPr on $\{P_W : W \in \mathcal{W}(\Theta)\}$ is defined as the mixture of distributions that achieves

$$\inf_{W \in \mathcal{W}(\Theta)} D(Q \| P_W).$$

Whether or not it is possible to evaluate this infimum explicitly depends on the exact properties of $\mathcal{P}$ and Q. However, it is always possible to approximate it using finite mixtures of the elements in $\mathcal{P}$. A finite mixture is simply a distribution P_W, where $W \in \mathcal{W}(\Theta)$ only gives positive weight to a finite number of parameters. The following theorem by Li [12] establishes an error bound for using finite mixtures to approximate the RIPr.

Theorem 3.1 (Li). *Choose $\theta_1 \in \Theta$ such that such that $D(Q\|P_{\theta_1})$ is minimised. Denote $Q_1 = P_{\theta_1}$ and define Q_k for $k = 2, 3, \ldots$ iteratively as*

$$Q_k = (1 - \alpha_k)Q_{k-1} + \alpha_k P_{\theta_k},$$

where $\alpha_k \in [0, 1]$ and $\theta_k \in \Theta$ are chosen to minimise $D(Q\|Q_k)$. Then

$$D(Q\|Q_k) \leq \inf_{W \in \mathcal{W}(\Theta)} D(Q\|P_W) + \frac{c^2_{Q,*}\gamma}{k}, \tag{3.1}$$

where $c^2_{Q,*}$ is a constant determined by Q and γ is a constant determined by $\mathcal{P}$.

That is, a greedy approach is taken to construct a finite mixture of elements in $\mathcal{P}$, which achieves a particular error bound for approximating the RIPr. Note that this approximation error might be infinite under some circumstances, but it will be shown in Proposition 3.2 that it is finite if the likelihood ratio between any two distributions in $\mathcal{P}$ is bounded away from zero and infinity. The advantage of such a greedy procedure is that the computational task of estimating one component at a time is much less complicated than estimating a full mixture. Nevertheless, the involved minimisation might be very difficult. This complexity can be somewhat mitigated because, as was realised years later, the details of Li's proof show that instead of optimising over both α_k and θ_k, the fixed sequence $\alpha_j = 2/(j+1)$ can be used too [12, 26]. For simplicity, this choice of weights will be used from here on out.

Besides the complexity of this procedure, another critical aspect is its performance. To this end, the constants in the error bound will be introduced and discussed. The first constant, γ, is determined by the properties of $\mathcal{P}$. In particular, its value depends on an upper bound of the log-ratio between two densities in $\mathcal{P}$.

Definition 3.1. Define

$$\gamma = 4\left[\log(3\sqrt{e}) + a_{\Theta,\mathcal{Y}}\right],$$

where

$$a_{\Theta,\mathcal{Y}} = \sup_{a,b \in \Theta, y \in \mathcal{Y}} \log\left(\frac{p_a(y)}{p_b(y)}\right).$$

The second constant, $c^2_{Q,*}$, is less straightforward and explaining it requires some intermediary constants. First, for a given point $y \in \mathcal{Y}$ and

prior $W \in \mathcal{W}(\Theta)$, the constant $c^2_{y,W}$ is defined such that $c^2_{y,W} - 1$ equals the coefficient of variation of $p_\theta(y)$ with respect to a prior W, i.e.

$$c^2_{y,W} = \frac{\mathbb{E}_{\theta\sim W}\left[p^2_\theta(y)\right]}{\left(\mathbb{E}_{\theta\sim W}\left[p_\theta(y)\right]\right)^2}.$$

Note that by Jensen's inequality, this quantity is always larger than or equal to 1. Then, $c^2_{Q,W}$ is defined as the expected value of $c^2_{y,W}$ with respect to Q:

$$c^2_{Q,W} = \mathbb{E}_{\mathbf{Y}\sim Q}\left[c^2_{\mathbf{Y},W}\right].$$

Now, all things are in place to properly define $c^2_{Q,*}$.

Definition 3.2. For Q as above, denote by W^* the set of all sequences $(W_i)_i$ such that $D(Q\|P_{W_i}) \to \inf_{W\in\mathcal{W}(\Theta)} D(Q\|P_W)$. Then $c_{Q,*}$ is defined as

$$c^2_{Q,*} = \inf_{(W_i)_i\in W^*} \liminf_{i\to\infty} c_{Q,P_{W_i}}$$

While this definition might seem very contrived, it is motivated by the fact that the bound given in (3.1) holds for all $W \in \mathcal{W}(\Theta)$, i.e.

$$D(Q\|Q_k) \leq D(Q\|P_W) + \frac{c^2_{Q,W}\gamma}{k}.$$

The result in (3.1) follows directly by choosing the best sequence $(W_i)_i$, such that

$$D(Q\|P_{W_i}) \to \inf_{W\in\mathcal{W}(\Theta)} D(Q\|P_W).$$

While this concludes all the needed definitions, nothing has been said about the size of the constants so far. None of the constants introduced so far is even necessarily finite. If they are not, the approximation error might be infinitely large. However, imposing bounds on the parameter space is generally sufficient to ensure that $a_{\Theta,\mathcal{Y}}$ is finite. An example of this can be found in Section 3.3. If $a_{\Theta,\mathcal{Y}}$ is bounded, it trivially follows that γ is finite too, and the following Proposition shows that it implies that $c^2_{Q,*}$ is bounded too.

Proposition 3.2. *If $a_{\Theta,\mathcal{Y}} < \infty$, then $c^2_{Q,*} < \infty$.*

Proof. Fix an arbitrary $\theta^* \in \Theta$. By definition, for every $\theta \in \Theta$ and $y \in \mathcal{Y}$ it holds that

$$\log\left(\frac{p_\theta(y)}{p_{\theta^*}(y)}\right) \leq a_{\Theta,\mathcal{Y}},$$

from which it follows that

$$p_\theta(y) \leq e^{a_{\Theta,\mathcal{Y}}} p_{\theta^*}(y) \quad \text{and} \quad p_\theta(y) \geq e^{-a_{\Theta,\mathcal{Y}}} p_{\theta^*}(y).$$

It is therefore true for any $W \in \mathcal{W}(\Theta)$ that

$$\begin{aligned} c^2_{Q,W} &= \mathbb{E}_{\mathbf{Y}\sim Q}\left[\frac{\mathbb{E}_{\theta\sim W}\left[p^2_\theta(\mathbf{Y})\right]}{\mathbb{E}_{\theta\sim W}\left[p_\theta(\mathbf{Y})\right]^2}\right] \\ &\leq \mathbb{E}_{y\sim Q}\left[\frac{\mathbb{E}_{\theta\sim W}\left[p^2_{\theta^*}(\mathbf{Y})e^{2a_{\Theta,\mathcal{Y}}}\right]}{\mathbb{E}_{\theta\sim W}\left[p_{\theta^*}(\mathbf{Y})e^{-a_{\Theta,\mathcal{Y}}}\right]^2}\right] \\ &= e^{4a_{\Theta,\mathcal{Y}}}, \end{aligned}$$

from which it directly follows that $c^2_{Q,*}$ is finite if $a_{\Theta,\mathcal{Y}}$ is. □

It is thus straightforward to tell whether the constants in Theorem 3.1 are finite or not. If they are, the error bound achieved by the greedy approach converges to zero with a rate of $1/k$. This suggests that, in such cases, the RIPr can be adequately approximated.

3.2 Approximating the GRO E-variable

Now that a method for approximating the RIPr of an arbitrary distribution has been outlined, it will be shown how this can be applied to hypothesis testing. Therefore, consider again the problem of testing the null hypothesis $\mathcal{H}_0 = \{P_{\theta'} : \theta' \in \Theta_0\}$ against the alternative hypothesis $\mathcal{H}_1 = \{P_{\theta'} : \theta' \in \Theta_1\}$. Suppose that some knowledge about the alternative hypothesis is encapsulated in the prior $W_1 \in \mathcal{W}(\Theta_1)$. By Theorem 1.4, the GRO E-variable $S^*_{W_1}$ for this problem is given by

$$S^*_{W_1} = \frac{p_{W_1}(\mathbf{Y})}{p^o_{W_0}(\mathbf{Y})},$$

where $P_{W^o_0}$ is the RIPr of P_{W_1} on $\{P_W : W \in \mathcal{W}(\Theta_0)\}$. While this RIPr may not be easily determinable, Theorem 3.1 shows that it is possible to approximate it. The likelihood ratio based on this approximation will itself not be an E-variable because the likelihood ratio of P_{W_1} and another distribution is an E-variable if and only if that other distribution is the RIPr [12, Lemma 4.1]. However, the following Theorem shows that it is possible to normalise the likelihood ratio to become an E-variable.

Theorem 3.3. *Let W_ϵ such that $D(P_{W_1}\|P_{W_\epsilon}) \leq D(P_{W_1}\|P_{W_0^o}) + \epsilon$. It holds that*

$$\sup_{\theta\in\Theta_0} \mathbb{E}_{\mathbf{Y}\sim P_\theta}\left[\frac{p_{W_1}(\mathbf{Y})}{p_{W_\epsilon}(\mathbf{Y})}\right] \leq 1 + \frac{V\log V}{\sqrt{\log V + 1/V - 1}}\sqrt{\epsilon},$$

where $V = e^{a_{\Theta,\mathcal{Y}}}$.

Proof. See Appendix B.2. □

Corollary 3.3.1. *Let $P_{W_0^k}$ be the iterative approximation of the RIPr of P_{W_1} on $\{P_W : W \in \mathcal{W}(\Theta_0)\}$ after k iterations. Then for $\epsilon_k = \frac{c^2_{P_{W_0},*}\gamma}{k}$, the following is an E-variable for testing $\mathcal{H}_0$ against $\mathcal{H}_1$:*

$$S^k_{W_1} = \frac{p_{W_1}(\mathbf{Y})}{p_{W_0^k}(\mathbf{Y})}\frac{1}{1 + \frac{V\log V}{\sqrt{\log V + 1/V - 1}}\sqrt{\epsilon_k}}.$$

As the number of iterations of the approximation algorithm increases, ϵ_k becomes smaller and $S^k_{W_1}$ converges to $S^*_{W_1}$. However, in reality, resources are always finite, and $S^k_{W_1}$ will not equal the GROW E-variable. Therefore, the amount of evidence it offers against the null will not be as large as possible when the alternative is valid. To assess how much this harms the usefulness of the E-variable, it should be compared to other known E-variables. This will be done in the next sections by performing a series of experiments for a hypothesis test in the logistic regression model.

3.3 Logistic Regression Model

Suppose that some dependent variable can only take on two values: yes or no, pass or fail, healthy or sick, etc. An interesting question is how likely it is to observe these values based on some known covariates. For example, a researcher in a clinical trial might ask: what is the probability that a patient becomes ill, given their age, sex, and that they have received a particular type of medicine? Mathematically, the quantity of interest is $\mathbb{P}(\mathbf{Y}|\mathbf{x})$ for a random variable $\mathbf{Y}$ on $\{-1, 1\}$ and a list of covariates $\mathbf{x} \in \mathbb{R}^{p-1}$. In logistic regression, it is assumed that there is a linear relationship between the log-odds and covariates, i.e.

$$\log\left(\frac{\mathbb{P}(\mathbf{Y}=1|\mathbf{x})}{\mathbb{P}(\mathbf{Y}=-1|\mathbf{x})}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_{p-1}x_{p-1}.$$

Here, $\boldsymbol{\beta} \in \mathbb{R}^p$ is a vector of coefficients. Since the two probabilities sum to one, there is only one unknown in this equation, and basic arithmetic gives

$$\mathbb{P}(\mathbf{Y} = y|\mathbf{x}) = \frac{e^{(\beta_0+\beta_1 x_1+\cdots+\beta_p x_p)y}}{e^{\beta_0+\beta_1 x_1+\cdots+\beta_p x_p} + e^{-(\beta_0+\beta_1 x_1+\cdots+\beta_p x_p)}}. \tag{3.2}$$

A zeroth entry equal to 1 is usually included in $\mathbf{x}$, so that Equation (3.2) can be written as

$$\mathbb{P}(\mathbf{Y} = y|\mathbf{x}) = \frac{e^{\mathbf{x}^T\boldsymbol{\beta} y}}{e^{\mathbf{x}^T\boldsymbol{\beta}} + e^{-\mathbf{x}^T\boldsymbol{\beta}}}. \tag{3.3}$$

In this setting, it is natural to test whether a particular covariate (say covariate j) impacts the dependent variable. To this end, $n \in \mathbb{N}$ data points are collected. Here, each data point consists of a realisation of the dependent variable and a vector of covariates. The totality of the data is thus given by $\{y_i, \mathbf{x}_i\}_{i=1}^n$, where $y_i \in \{-1, 1\}$ and $\mathbf{x}_i \in \mathbb{R}^p$. Formally, this leads to the statistical model $(\mathcal{Y}, \Theta, \mathcal{P})$ for

$$\mathcal{Y} = \{-1, 1\}^n,\ \Theta = \mathbb{R}^p \text{ and } \mathcal{P} = \{P_{\boldsymbol{\beta}} : \boldsymbol{\beta} \in \Theta\}.$$

Denoting X for the matrix with rows equal to $\mathbf{x}_i$, the distribution $P_{\boldsymbol{\beta}}$ has conditional probability mass function given by

$$p_{\boldsymbol{\beta}}(\mathbf{y}|X) = \prod_{i=1}^n \frac{e^{\mathbf{x}_i^T\boldsymbol{\beta} y_i}}{e^{\mathbf{x}_i^T\boldsymbol{\beta}} + e^{-\mathbf{x}_i^T\boldsymbol{\beta}}}.$$

The hypotheses are then given by

$$\mathcal{H}_0 = \left\{P_{\boldsymbol{\beta}} \,\middle|\, \boldsymbol{\beta} \in \Theta_0 = \left\{\boldsymbol{\beta} \in \Theta : \boldsymbol{\beta}_j = 0\right\}\right\}$$

and

$$\mathcal{H}_1 = \left\{P_{\boldsymbol{\beta}} \,\middle|\, \boldsymbol{\beta} \in \Theta_1 = \left\{\boldsymbol{\beta} \in \Theta : \boldsymbol{\beta}_j > 0\right\}\right\}.$$

Assuming that a prior $W_1 \in \mathcal{W}(\Theta_1)$ is given, the GRO E-variable is given by the likelihood ratio of P_{W_1} and its RIPr on $\{P_W : W \in \mathcal{W}(\Theta_0)\}$. The latter is found by solving the following minimisation problem:

$$\inf_{W_0 \in \mathcal{W}(\Theta_0)} \frac{\int_{\boldsymbol{\beta} \in \Theta_1} \prod_{i=1}^n \frac{e^{\mathbf{x}_i^T\boldsymbol{\beta} y_i}}{e^{\mathbf{x}_i^T\boldsymbol{\beta}} + e^{-\mathbf{x}_i^T\boldsymbol{\beta}}} w_1(\boldsymbol{\beta}) d\boldsymbol{\beta}}{\int_{\boldsymbol{\beta}' \in \Theta_0} \prod_{i=1}^n \frac{e^{\mathbf{x}_i^T\boldsymbol{\beta}' y_i}}{e^{\mathbf{x}_i^T\boldsymbol{\beta}'} + e^{-\mathbf{x}_i^T\boldsymbol{\beta}'}} w_0(\boldsymbol{\beta}') d\boldsymbol{\beta}'}.$$

One can imagine that it is a daunting task to explicitly find a solution (even though a solution must exist, by the arguments based on Carathéodory's

theorem in Grünwald et al. [5]). Subsequently, it is probably a lot easier to calculate the E-variable $S^k_{W_1}$, as in Corollary 3.3.1. To this end, it should be checked that this E-variable is well-defined. By Proposition 3.2, this is true as long $a_{\Theta,\mathcal{Y}}$ is finite. To make sure that this is the case, all parameters are restricted to the interval $[-B, B]$ for some $B \in \mathbb{R}_{>0}$. Furthermore, it is assumed that the covariates are normalised to this same interval. Then it holds that

$$\begin{aligned}\exp(a_{\Theta,\mathcal{Y}}) &= \sup_{\beta,\gamma\in\Theta,\mathbf{y}\in\mathcal{Y}} \frac{p_\beta(\mathbf{y}|X)}{p_\gamma(\mathbf{y}|X)} = \sup_{\beta,\gamma\in\Theta,\mathbf{y}\in\mathcal{Y}} \prod_{i=1}^n \frac{p_\beta(y_i|\mathbf{x}_i)}{p_\gamma(y_i|\mathbf{x}_i)} \\ &\leq \prod_{i=1}^n \frac{\sup_{\beta\in\Theta} \frac{e^{\mathbf{x}_i^T\beta}}{e^{\mathbf{x}_i^T\beta}+e^{-\mathbf{x}_i^T\beta}}}{\inf_{\gamma\in\Theta} \frac{e^{\mathbf{x}_i^T\gamma}}{e^{\mathbf{x}_i^T\gamma}+e^{-\mathbf{x}_i^T\gamma}}} = \prod_{i=1}^n e^{2\sqrt{pB^2}\|\mathbf{x}_i\|_2} \\ &\leq e^{2Bnp\sqrt{\max\{B^2,1\}}},\end{aligned}$$

where the Cauchy-Schwarz inequality is used together with the fact that $f(c) = \frac{e^c}{e^c+e^{-c}}$ is increasing in c. So,

$$a_{\Theta,\mathcal{Y}} \leq 2Bnp\sqrt{\max\{B^2,1\}}.$$

Since this upper bound is finite, the E-variable $S^k_{W_1}$ is well-defined.

However, it should be noted that restricting the parameters to $[-B, B]$ does not only impose a limit on the log-ratio of the likelihood but also on the maximum and minimum value of $p_\beta(\mathbf{y})$. For example, in the case $n = 1$, the model would be restricted to Bernoulli$(\frac{1}{2})$ for $B \to 0$. Therefore, B should be chosen large enough to ensure that the model does not become too restrictive. One sensible bound is to choose $B \geq 1.5/p$ because then the probability of a single outcome can still be close to 1, i.e.

$$\sup_{\mathbf{x}\in[0,1]^p,\beta\in\Theta,y\in\{-1,1\}} p_\beta(\mathbf{y}|\mathbf{x}) = \frac{e^{Bp}}{e^{Bp}+e^{-Bp}} \geq \frac{e^{1.5}}{e^{1.5}+e^{-1.5}} \approx 0.95.$$

By symmetry, the probability of a single outcome can be close to 0 too.

3.3.1 Randomisation-based E-variable

While it is shown that the GRO E-variable can properly be estimated, it turns out that another, easier, E-variable can be calculated by making an

extra assumption. Namely that the tested covariate is a Bernoulli($\frac{1}{2}$) random variable. Here, the terminology is slightly abused, as a random variable that takes on value -1 with probability $1-p$ and 1 with probability p is referred to as a Bernoulli(p) random variable. To see why this might be a natural assumption, return once more to the example of the clinical trial. A large group of people is gathered with the purpose of testing whether certain medication positively affects the health. To do so, the patients are either treated with the medicine or given a placebo as control. The assignment to treatment or control is often randomised, so that the corresponding covariate is a Bernoulli random variable, whereas the other covariates (age/sex/etc.) can still be anything. With this extra assumption, it is not hard to see that the following is an E-variable:

$$S = \frac{p_{W_1}(y|\mathbf{x})}{\frac{1}{2}p_{W_1}(y|\mathbf{x}_{-j}, -1) + \frac{1}{2}p_{W_1}(y|\mathbf{x}_{-j}, 1)}. \tag{3.4}$$

Here, $(\mathbf{x}_{-j}, x_j)$ denotes the vector of covariates $\mathbf{x}$ with the j-th coordinate set to $x_j \in \{-1, 1\}$. Since the j-th covariate is assumed to be a Bernoulli($\frac{1}{2}$) random variable, for arbitrary $P_0 \in \mathcal{H}_0$,

$$\begin{aligned}
\mathbb{E}[S] &= \mathbb{E}_{\mathbf{X}_j \sim \text{Ber}(1/2)} \mathbb{E}_{\mathbf{Y} \sim P_0} \left[\frac{p_{W_1}(\mathbf{Y}|\mathbf{x}_{-j}, \mathbf{X}_j)}{\frac{1}{2}p_{W_1}(\mathbf{Y}|\mathbf{x}_{-j}, -1) + \frac{1}{2}p_{W_1}(\mathbf{Y}|\mathbf{x}_{-j}, 1)} \right] \\
&= \frac{1}{2}\int_y p_0(y|\mathbf{x}_{-j}, -1) \frac{p_{W_1}(y|\mathbf{x}_{-j}, -1)}{\frac{1}{2}p_{W_1}(y|\mathbf{x}_{-j}, -1) + \frac{1}{2}p_{W_1}(y|\mathbf{x}_{-j}, 1)} dy \\
&+ \frac{1}{2}\int_y p_0(y|\mathbf{x}_{-j}, 1) \frac{p_{W_1}(y|\mathbf{x}_{-j}, 1)}{\frac{1}{2}p_{W_1}(y|\mathbf{x}_{-j}, -1) + \frac{1}{2}p_{W_1}(y|\mathbf{x}_{-j}, 1)} dy \\
&= \int_y p_0(y|\mathbf{x}_{-j}) \frac{\frac{1}{2}p_{W_1}(y|\mathbf{x}_{-j}, -1) + \frac{1}{2}p_{W_1}(y|\mathbf{x}_{-j}, 1)}{\frac{1}{2}p_{W_1}(y|\mathbf{x}_{-j}, -1) + \frac{1}{2}p_{W_1}(y|\mathbf{x}_{-j}, 1)} dy \\
&= 1.
\end{aligned}$$

The second to last equality holds because, under the null hypothesis, y does not depend on x_j, so $p_0(y|\mathbf{x}_{-j}, -1) = p_0(y|\mathbf{x}_{-j}, 1) =: p_0(y|\mathbf{x}_{-j})$. Note that this holds for arbitrary Bernoulli distributions, but the fair coin flip is chosen here because it is simple and easy to interpret. Furthermore, this E-variable can be extended to multiple outcomes by considering each data point separately and multiplying the resulting E-variables. Essentially the same idea as (3.4) was used to obtain the E-variables for 2×2 contingency tables recently presented by Turner et al. [27].

One favourable property of the randomisation-based E-variable is that it is very easy to compute. On top of that, since each data point is considered separately, it is possible to stop collecting data once a satisfactory E-value is found. This practice, known as optional stopping, does not invalidate the conclusions drawn from the resulting E-value even if the study was initially designed to include more data points. De Jong [28] makes this precise for a variation of the linear regression model discussed in Section 2.3. In this setting, he also shows that the randomisation-based E-variable can be adjusted to the case where randomisation is done in bulk at the start of the study, instead of for each patient separately. It can then be enforced that e.g. half of the patients are actually given the treatment. While this certainly improves the flexibility of the randomisation-based E-variable, it is still not as widely applicable as the approximation of the GRO E-variable, which is an E-variable regardless of any extra assumptions on the covariates. However, it should be noted that in this altered setting, the latter might no longer approximate the actual GRO E-variable, since the problem definition slightly changed. This change allows for the definition of E-variables, which were not E-variables in the more abstract setting, such as the randomisation-based E-variable.

3.4 Experiments

This section exemplifies the approximation of the GRO E-variable on simulated instances of the logistic regression model. To this end, Li's algorithm is implemented in a very straightforward manner: assuming that $k-1$ steps of the algorithm have already been run with resulting output $P_{W_0^{k-1}}$, the next iteration uses convex optimisation methods to find

$$\min_{\beta \in [-B,B]^p} D(P_{W_1} \| (1-\alpha_k) P_0^{k-1} + \alpha_k P_\beta).$$

However, we have not been able to prove that this problem is actually convex for the logistic regression model, nor have we found any examples to indicate the contrary. One way to potentially prove this is to show that the Hessian matrix is positive semi-definite. The difficulty in doing so is that the expression of this Hessian is complicated by the mixture of P_β with $P_{W_{k-1}}$.

Furthermore, the implementation is computationally intensive and its time complexity grows exponentially in the number of data points. This follows because the Kullback-Leibler divergence is calculated as a sum over all possible data points. Consequently, the performed experiments

are very basic: the number of data points is small ($n \leq 10$) and the prior on the alternative is restricted to degenerate distributions. In all experiments, the tested covariate is assumed to be a Bernoulli($\frac{1}{2}$) random variable and the covariates other than the intercept and treatment/control are drawn uniformly random from $[-B, B]$.

3.4.1 Approximation error and maximum expectation

Theorem 3.1 gives an idea of how the Kullback-Leibler divergence between the distribution on the alternative and the approximation of its RIPr should decrease with the number of iterations. Similarly, Theorem 3.3 gives a theoretical upper bound on the expected value of the likelihood ratio based on this approximation after k iterations. To compare these theoretical results to behaviour in reality, simulations are done for $n = 5$. For different values of B and priors W_1, the iterative approximation $P_{W_0^k}$ of the RIPr of P_{W_1} is computed ($k \in \{1, \ldots, 100\}$). The Kullback-Leibler divergence $D(P_{W_1} \| P_{W_0^k})$ is calculated and a grid search is performed over Θ_0 to estimate

$$\sup_{\theta \in \Theta_0} \mathbb{E}_{P_\theta}\left[S^k_{W_1}\right] = \sup_{\theta \in \Theta_0} \mathbb{E}_{\mathbf{Y} \sim P_\theta}\left[\frac{p_{W_1}(\mathbf{Y})}{p_{W_0^k}(\mathbf{Y})}\right].$$

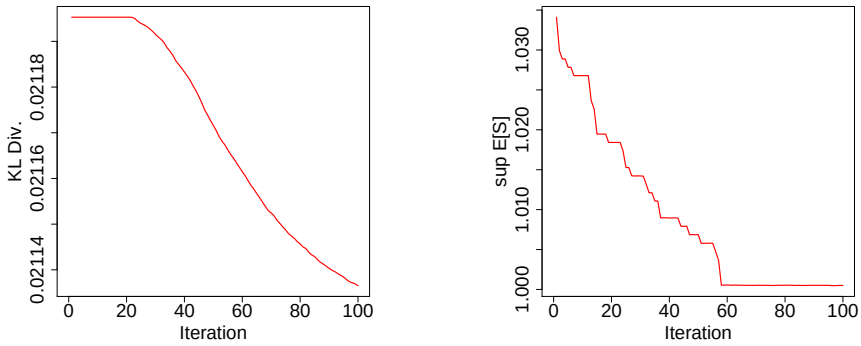
This is averaged over 500 repetitions to smooth out the influence of the covariates. The results are shown in Figure 3.1.

3.4.2 Comparison to randomisation-based E-variable

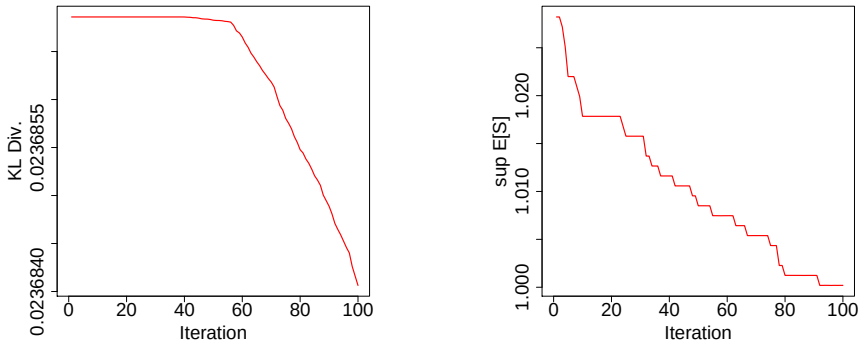
Calculating $S^k_{W_1}$ is much more complicated than calculating the randomisation-based E-variable given by (3.4). The evidence these E-variables give against the null will be compared to check whether this extra effort is worth it. This comparison is made using the likelihood ratio before normalisation (thus giving an unfair advantage to $S^k_{W_1}$). For $n \in \{1, \ldots, 10\}$, $k \in \{1, 10, 20, 30\}$, varying values of B, and different priors W_1 on the alternative, the iterative approximation $P_{W_0^k}$ of the RIPr of P_{W_1} is calculated. The expected value

$$\mathbb{E}_{\mathbf{Y} \sim P_{W_1}}\left[\log\left(\frac{p_{W_1}(\mathbf{Y})}{p_{W_0^k}(\mathbf{Y})}\right)\right]$$

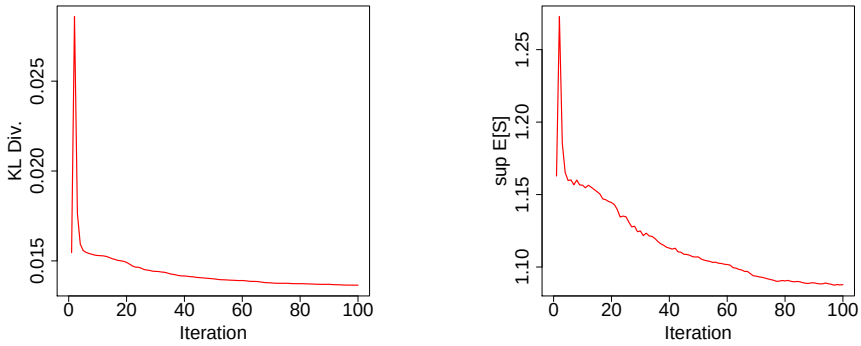
is calculated together with the expected value of the logarithm of (3.4). These values are averaged over 100 repetitions to estimate the expected value under the covariates. The results are shown in Figure 3.2.



(a) *Results for* $B = 0.5$ *and degenerate prior* W_1 *with all weight on the point* $(0.25, 0.25, 0.25)$.

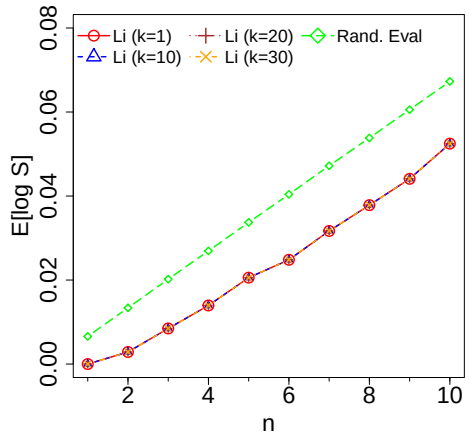


(b) *Results for* $B = 0.5$ *and degenerate prior* W_1 *with all weight on the point* $(0, 0.25, 0.25)$.

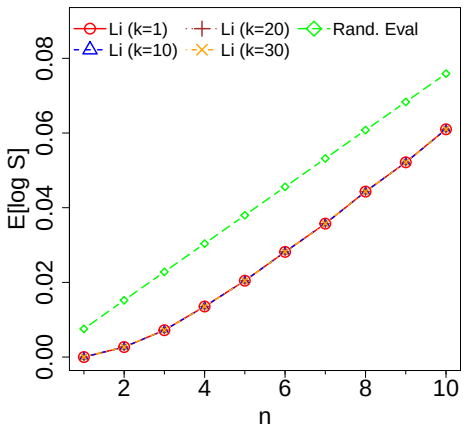


(c) *Results for* $B = 5$ *and degenerate prior* W_1 *with all weight on the point* $(0.25, 0.25, 0.25)$.

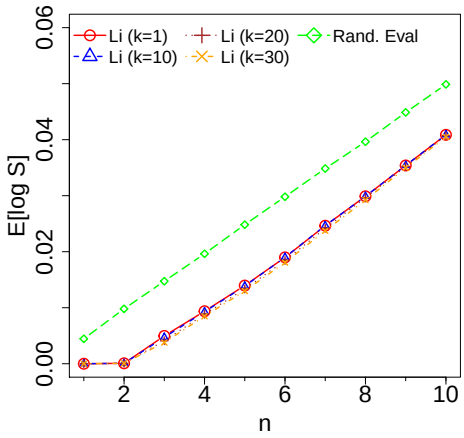
Figure 3.1: *The iterative approximation of the RIPr was calculated for different choices of the prior* W_1 *on the alternative* ($n = 5$). *The Kullback-Leibler divergence of this approximation and* P_{W_1} *was calculated, and the maximum expected value of the resulting likelihood ratio under the null was estimated. These were averaged over* 1000 *repetitions and are plotted against the number of iterations that were run.*



(a) *Results for* $B = 0.5$ *and degenerate prior* W_1 *with all weight on the point* $(0.25, 0.25, 0.25)$.



(b) *Results for* $B = 0.5$ *and degenerate prior* W_1 *with all weight on the point* $(0, 0.25, 0.25)$



(c) *Results for* $B = 5$ *and degenerate prior* W_1 *with all weight on the point* $(0.25, 0.25, 0.25)$

Figure 3.2: *For different choices of the prior* W_1 *on the alternative, the expected value under the alternative of* $\log S^k_{W_1}$ *was calculated for* $k \in \{1, 10, 20, 30\}$. *Additionally, the expected value under the alternative of the logarithm of the randomisation-based* E*-variable given by* (3.4) *was calculated. These values were averaged over* 100 *repetitions to estimate the expected value under the covariates. They are plotted against the number of data points.*

3.5 Discussion

The implementation of Li's algorithm used for the experiments outlined above was very straightforward and hence computationally intensive. As a result, the experiments were very basic and only included a very small number of data points. Future research needs to be done to determine whether it is possible to implement Li's algorithm more efficiently. One potential approach is to find a method of calculating the Kullback-Leibler divergence between two distributions in the logistic regression model, such that the time complexity does not increase exponentially in the number of data points. Another consequence of the straightforward implementation is that all experiments had to be done with an arbitrarily chosen degenerate prior on the alternative. In the majority of cases, it is not realistic to assume that such specific knowledge is available. Rather, it would be appropriate to work with a generic, non-informative prior. Extending the approximation of the GRO E-variable to this setting is straightforward, but the computational problems would get even worse. Nevertheless, the basic implementation allowed for the approximation of the GRO E-variable, albeit in a very limited setting.

The results in Section 3.4.1 suggested that the upper bound in Theorem 3.3 should not be used to normalise the likelihood ratio to an E-variable. This follows because the bound equals

$$1 + \frac{V \log V}{\sqrt{\log V + 1/V - 1}} \sqrt{\epsilon},$$

where $V = e^{2Bnp\sqrt{\max\{B^2,1\}}}$. This is roughly equal to $1 + 8.5 \cdot 10^{146}\sqrt{\epsilon}$ for $n = 5$ and $B = 5$, while the estimated maximum expected value did not exceed 2 in any of the experiments. Dividing the likelihood ratio by this upper bound would therefore make it unnecessarily small. A better approach seems to be to compute the upper bound directly by calculating the expected value under each element of the null by summing over the entire sample space. For all but the smallest number of samples, this seems computationally infeasible but this could presumably be fixed by estimating the expected value using vectors drawn uniformly at random. This would also allow for the use of Li's algorithm without imposing a bound of B on the log-likelihood. While this invalidates the guarantees on the approximation error, future research might conclude through numeric methods that Li's algorithm still leads to a valid approximation of the GRO E-variable. The biggest theoretical question that remains open is whether it is possible to find a tighter theoretic upper bound on the expected value

of the approximation of the GRO E-variable. The interest is then not so much in getting a bound that is practically useful, but rather in getting a better idea of how the approximation really scales with n: maybe it scales linearly rather than exponentially, for example.

Furthermore, it is interesting to note that neither the Kullback-Leibler divergence nor the maximum expectation were found to be strictly decreasing in the number of iterations. This is surprising since it seems reasonable to expect that the approximation error should be decreasing and that a smaller approximation error should give a smaller maximum expectation. One possible explanation is that convex optimisation methods were used to find the approximations of the RIPr, while the involved minimisation was not proven to be convex. Therefore, it might have been the case that the algorithm sometimes wound up in local minima, but more research needs to be done to give a definitive answer as to whether this is indeed the problem.

Finally, it can be seen that the considered E-variables did not constitute substantial evidence against the null hypothesis in any of the experiments in Section 3.4.2. This might be due to the fact that the number of data points was not much larger than the number of covariates. Still, the findings suggested that in the case that the tested covariate is a Bernoulli random variable, the randomisation-based E-variable is to be favoured over the approximation of the GRO E-variable. In all of the experiments, the randomisation-based E-variable offered more evidence against the null, while the approximation of the GRO E-variable was not even normalised yet. This also suggests that the GRO E-variable in the general setting is no longer the GRO E-variable when the tested covariate is assumed to be a Bernoulli random variable. However, the generalisability of these suspicions is limited by the small number of data points. A lot more research remains to be done to determine whether they are actually valid and whether they can be extended to a larger number of data points.

Chapter 4

Conclusion

By discussing and expanding two methods of obtaining E-variables, this thesis added to the general theory of testing with E-variables. Giving explicit applications of these methods offered examples of how to use E-variables for hypothesis testing problems involving covariates. The first method was based on the principles of invariant testing [10, 11] and results by Perez et al. [9]. This approach was shown to give the GROW E-variable for testing whether the effect size of a particular covariate is equal to zero or equal to some predefined threshold in a linear regression model. A potential extension of this result is to determine what happens when a predefined threshold is not known. In such cases, the method can still be used by putting a prior on the threshold and averaging over the corresponding E-variables. The resulting E-variable is still GROW if there are no covariates, as shown by Grünwald et al. [5, Theorem 3]. Further research should conclude whether this generalises to the situation with covariates.

The second method for obtaining E-variables was based on an approximation algorithm by Li [12]. When testing a null hypothesis against an alternative, this algorithm can approximate the RIPr based on a prior on the alternative. It was shown that it is possible to normalise the likelihood ratio of the prior and approximation of its RIPr to become an E-variable. This normalisation used an established upper bound on the expected value of the likelihood ratio under the null hypothesis. Preliminary simulations for a hypothesis test in a logistic regression model indicated that this upper bound is not sharp. Furthermore, the results suggested that for a small number of data points, this E-variable does not provide more evidence than the much easier to calculate randomisation-based E-variable. So while it is shown how to find the GROW E-variable for the linear regression model, a lot of future research is needed on applying E-variables to the logistic regression model.

Bibliography

[1] R. L. Wasserstein and N. A. Lazar, "The ASA's Statement on p-Values: Context, Process, and Purpose," *The American Statistician*, vol. 70, no. 2, pp. 129–133, 2016, [doi:10.1080/00031305.2016.1154108].

[2] E. L. Lehmann, "The Fisher, Neyman-Pearson theories of testing hypotheses: One theory or two?," *Journal of the American Statistical Association*, vol. 88, no. 424, pp. 1242–1249, 1993, [doi:10.1080/01621459.1993.10476404].

[3] S. N. Goodman, "p Values, Hypothesis Tests, and Likelihood: Implications for Epidemiology of a Neglected Historical Debate," *American Journal of Epidemiology*, vol. 137, no. 5, pp. 485–496, 1993, [doi:10.1093/oxfordjournals.aje.a116700].

[4] L. K. John, G. Loewenstein, and D. Prelec, "Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling," *Psychological Science*, vol. 23, no. 5, pp. 524–532, 2012, [doi:10.1177/0956797611430953].

[5] P. Grünwald, R. de Heide, and W. Koolen, "Safe testing," 2020, [arxiv:1906.07801].

[6] V. Vovk and R. Wang, "E-values: Calibration, combination, and applications," 2020, [arxiv:1912.06116].

[7] G. Shafer, "Testing by betting : A strategy for statistical and scientific communication," *Journal of the Royal Statistical Society, Series A*, vol. 184, no. 2, pp. 407–431, 2021, [doi:10.1111/rssa.12647].

[8] J. Ter Schure and P. Grünwald, "Accumulation bias in meta-analysis: The need to consider time in error control," *F1000Research*, vol. 8, no. 962, pp. 1–24, 2019, [doi:10.12688/f1000research.19375.1]. version 1; peer review: 2 approved.

[9] M. F. Perez-Ortiz, R. de Heide, and P. Grünwald, "Optimal e-statistics under group invariance." unpublished, 2021.

[10] A. Davison, *Statistical Models*. Cambridge University Press, 2003, [doi:10.1017/CBO9780511815850].

[11] E. L. Lehmann and J. P. Romano, *Testing Statistical Hypothesis*. New York, NY: Springer, 3 ed., 2005, [doi:10.1007/0-387-27605-X].

[12] Q. Li, *Estimation of Mixture Models*. PhD thesis, Yale University, 1999.

[13] V. H. De La Peña, "A general class of exponential inequalities for martingales and ratios," *Annals of Probability*, vol. 27, no. 1, pp. 537–564, 1999, [doi:10.1214/aop/1022677271].

[14] D. J. Benjamin, J. O. Berger, M. Johannesson, B. A. Nosek, E.-J. Wagenmakers, R. Berk, K. A. Bollen, B. Brembs, L. Brown, and C. Camerer, "Redefine Statistical Significance," *Nature Human Behaviour*, vol. 2, no. 1, pp. 6–10, 2018, [doi:10.1038/s41562-017-0189-z].

[15] J. O. Berger, "Could Fisher, Jeffreys and Neyman have agreed on testing?," *Statistical Science*, vol. 18, no. 1, pp. 1–32, 2003, [doi:10.1214/ss/1056397485].

[16] T. Ceccherini-Silberstein and M. Coornaert, *Cellular automata and groups*, vol. 1. New York, NY: Springer-Verlag, 2012, [doi:10.1007/978-1-4614-1800-9{_}23].

[17] A. Garrido, "An introduction to amenable groups." http://reh.math.uni-duesseldorf.de/~garrido/amenable.pdf, 2013. Lecture Notes.

[18] L. Nachbin and L. Bechtolsheim, *The Haar integral*. Princeton, N.J., Van Nostrand, 1965.

[19] S. Andersson, "Distributions of Maximal Invariants Using Quotient Measures," *The Annals of Statistics*, vol. 10, no. 3, pp. 955–961, 1982, [doi:10.1214/aos/1176345885].

[20] R. Wijsman, "Cross-sections of orbits and their application to densities of maximal invariants," in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics* (L. M. L. Cam and J. Neyman, eds.), University of California Press, December 1967.

[21] M. L. Eaton, *Group Invariance Applications in Statistics*. Hayward, CA: Institute of Mathematical Statistics, 1989.

[22] J. V. Bondar and P. Milnes, "Amenability: A survey for statistical applications of Hunt-Stein and related conditions on groups," *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, vol. 57, no. 1, pp. 103–128, 1981, [doi:10.1007/BF00533716].

[23] T. Kariya, "Locally Robust Tests for Serial Correlation in Least Squares Regression," *Annals of Statistics*, vol. 8, no. 5, pp. 1065–1070, 1980, [doi:10.1214/aos/1176345143].

[24] J. L. Bhowmik, "Constructing Locally Best Invariant Tests of the Linear Regression Model Using the Density Function of a Maximal Invariant," *American Journal of Mathematics and Statistics*, vol. 3, no. 1, pp. 45–52, 2013, [doi:10.5923/j.ajms.20130301.07].

[25] B. J. Berger and L. R. Pericchi, "Bayes Factors and Marginal Distributions in Invariant Situations," *The Indian Journal of Statistics*, vol. 60, no. 3, pp. 307–321, 1998.

[26] W. D. Brinda, *Adaptive Estimation with Gaussian Radial Basis Mixtures*. PhD thesis, Yale University, 2018.

[27] R. Turner, A. Ly, and P. Grünwald, "Safe tests and always-valid confidence intervals for contingency tables and beyond," 2021, [arxiv:2106.02693].

[28] M. De Jong, "Tests of significance for linear regression using e-values," Master's thesis, University of Leiden, 2021. Master's thesis.

[29] S. Banach and A. Tarski, "Sur la décomposition des ensembles de points en parties respectivement congruentes," *Fundamenta Mathematicae*, vol. 6, pp. 244–277, 1924.

[30] A. Das and W. S. Geisler, "A method to integrate and classify normal distributions," 2021, [arxiv:2012.14331].

[31] D. Jones, "Statistical Analysis of Empirical Models Fitted by Optimization," *Biometrika*, vol. 70, no. 1, pp. 67–88, 1983, [doi:10.1093/biomet/70.1.67].

[32] R. A. Wijsman, "Proper Action in Steps, with Application to Density Ratios of Maximal Invariants," *Annals of Statistics*, vol. 13, no. 1, pp. 395–402, 1985, [doi:10.1214/aos/1176346600].

[33] Y. Yang and A. R. Barron, "An asymptotic property of model selection criteria," *IEEE Transactions on Information Theory*, vol. 44, no. 1, pp. 95–116, 1998, [doi:10.1109/18.650993].

Appendix A

Background on Group Theory

In this chapter, relevant concepts and definitions from group theory are described. Most of the material can be found in standard textbooks; citations are given for the more advanced concepts.

Suppose that there is some group G. The nature of the group structure allows for algebraic operations to be performed on G and for G to interact with other spaces. A group G is said to act on some space X if there is a map $\phi : G \times X \to X$, such that for all $x \in X$:

1. $\phi(e, x) = x$, where e is the identity element,
2. $\phi(g, \phi(h, x)) = \phi(g \cdot h, x)$ for all $g, h \in G$.

The action of G on X is usually abbreviated as $\phi(g, x) = gx$.

In addition to algebraic operations, it is often desirable to be able to do some kind of analysis on G. To this end, an extra structure is needed: a topology.

Definition A.1. A topological group is a group G together with a topology, such that the group operation

$$\cdot : G \times G \to G, (g, h) \mapsto gh$$

and inversion map

$$^{-1} : G \to G, g \mapsto g^{-1}$$

are both continuous with respect to this topology.

The following definition classifies a particularly useful type of action for topological groups.

Definition A.2. The action of G on X is called proper if G is a topological group and the map from $G \times X \to X \times X$ given by $(g, x) \to (gx, x)$ is proper.

It has been shown that topological groups that are locally compact and Hausdorff allow for the natural definition of a measure (see e.g. [18, Chapter 2]). This measure can be seen as a generalisation of the Lebesgue measure on Euclidean spaces, because they share a characterising property: invariance under translation. For the Lebesgue measure, this means that the measure of the set $\{a + x : a \in A\}$ is the same as the measure of A, for any Lebesgue-measurable set A and $x \in \mathbb{R}^n$. The so-called Haar measure demonstrates a generalisation of this property. However, since the group operation is not necessarily commutative, there exists both a left- and right-invariant version of this measure.

Definition A.3. A left (resp. right) Haar measure is a countably additive measure μ on the Borel sets of G such that

1. For every Borel set S and $g \in G$,
$$\mu(gS) = \mu(S) \ \ (\text{resp. } \mu(Sg) = \mu(S)),$$
2. For every compact $K \subseteq G$: $\mu(K) < \infty$,
3. For every Borel set S,
$$\mu(S) = \inf\{\mu(U) | U \text{ is open}, S \subseteq U\},$$
4. For every open set $U \subseteq G$,
$$\mu(U) = \sup\{\mu(K) | K \text{ is compact}, K \subseteq U\},$$

The Lebesgue measure is not the only thing that generalises to topological groups. Another is the well-known, counter-intuitive, result by Banach-Tarski [29]. This essentially states that a ball in 3-dimensional space can be deconstructed into a finite number of pieces, which can be reconstructed into two balls identical in size to the first. A direct consequence is that these "paradoxical" sets cannot be measurable by any measure that assigns non-zero measure to the unit ball and is countably additive. As a result, there must be subsets of $\mathbb{R}^3$ which are not Lebesgue measurable. However, this phenomenon does not occur for $\mathbb{R}$ and $\mathbb{R}^2$.

Very similarly, there are paradoxical decompositions that arise in some topological groups, but not in others. The exact nature of these paradoxes will not be discussed here, as it is of little relevance to the rest of the discussion; details can be found in e.g. [16, Chapter 4]. The topological groups on which these paradoxes do not occur, are precisely those groups on which the Haar measure can be extended to be defined on every subset of the group [17]. Such groups are called amenable.

Definition A.4. A locally compact topological group G is amenable if there exists a finitely additive probability measure μ such that for all $A \in \mathcal{P}(G)$ and $g \in G$

$$\mu(gA) = \mu(A).$$

It has been shown that many properties are equivalent to the one used in this definition, see e.g. [22] for a comprehensive list. This makes amenable groups particularly pleasant to work with. Examples of amenable groups include abelian groups, compact groups, and solvable groups.

Appendix B

Proofs

This chapter contains miscellaneous results which were used, but not proved in the main text.

B.1 Proofs for Chapter 2

The notation used here corresponds with that of Section 2.3.

Proof of Lemma 2.9

Proof. **(of Lemma 2.9)** Since $\phi(\mathbf{Y}) = (\|A\mathbf{y}\|, B\mathbf{y})$, denote $\phi_1(\mathbf{Y}) = \|A\mathbf{Y}\|$ and $\phi_2(\mathbf{Y}) = B\mathbf{Y}$. For $\theta_0 = (\beta_{-j}, 0, \sigma) \in \Theta_0$ and $\theta_1 = (\beta'_{-j}, \beta'_j, \sigma') \in \Theta_1$ arbitrarily,

$$p_{\theta_0}^{[\phi_1(\mathbf{Y})]}(h_1) = \frac{2h_1^{k-1}e^{-\frac{h_1^2}{2\sigma^2}}}{(2\sigma^2)^{\frac{k}{2}}\Gamma(\frac{k}{2})}$$

and

$$p_{\theta_1}^{[\phi_1(\mathbf{Y})]}(h_1) = e^{-\frac{\lambda(\beta'_j)}{2}}{}_0F_1\left(;\frac{k}{2};\frac{\lambda(\beta'_j)h_1^2}{4\sigma'^2}\right)\frac{2h_1^{k-1}e^{-\frac{h_1^2}{2\sigma'^2}}}{(2\sigma'^2)^{\frac{k}{2}}\Gamma(\frac{k}{2})},$$

where ${}_0F_1(;a;z) = \sum_{i=0}^{\infty} \frac{\Gamma(a)}{\Gamma(a+i)} \frac{z^i}{i!}$ is the confluent hypergeometric limit function. Assuming that $k > 1$ (i.e. $n > p$), it holds that ${}_0F_1(;\frac{k}{2};z) \leq e^z$, so

$$\begin{aligned}
&\mathbb{E}_{h_1 \sim P_{\theta_1}^{[\phi_1(\mathbf{Y})]}} \left[\left| \log \left(\frac{p_{\theta_1}^{[\phi_1(\mathbf{Y})]}(h_1)}{p_{\theta_0}^{[\phi_1(\mathbf{Y})]}(h_1)} \right) \right|^{1+\epsilon} \right] \\
&\quad = \mathbb{E}_{h_1 \sim P_{\theta_1}^{[\phi_1(\mathbf{Y})]}} \left[\left| -\frac{\lambda(\beta_j')}{2} + h_1^2 \left(\frac{1}{2\sigma^2} - \frac{1}{2\sigma'^2} \right) + k \log \left(\frac{\sigma}{\sigma'} \right) \right. \right. \\
&\quad + \left. \left. \log \left({}_0F_1 \left(;\frac{k}{2}; \frac{\lambda(\beta_j')h_1^2}{4\sigma'^2} \right) \right) \right|^{1+\epsilon} \right] \\
&\quad \leq \mathbb{E}_{h_1 \sim P_{\theta_1}^{[\phi_1(\mathbf{Y})]}} \left[\left| -\frac{\lambda(\beta_j')}{2} + h_1^2 \left(\frac{1}{2\sigma^2} - \frac{1}{2\sigma'^2} \right) + k \log \left(\frac{\sigma}{\sigma'} \right) + \frac{\lambda(\beta_j')h_1^2}{4\sigma'^2} \right|^{1+\epsilon} \right]
\end{aligned}$$

For $\epsilon = 1$, this is equal to

$$\begin{aligned}
&\mathbb{E}_{h_1 \sim P_{\theta_1}^{[\phi_1(\mathbf{Y})]}} \left[\left| -\frac{\lambda(\beta_j')}{2} + k \log \left(\frac{\sigma}{\sigma'} \right) + h_1^2 \left(\frac{1}{2\sigma^2} + \frac{\frac{1}{2}\lambda(\beta_j') - 1}{2\sigma'^2} \right) \right|^2 \right] \\
&\quad = \mathbb{E}_{h_1 \sim P_{\theta_1}^{[\phi_1(\mathbf{Y})]}} \left[\left(-\frac{\lambda(\beta_j')}{2} + k \log \left(\frac{\sigma}{\sigma'} \right) \right)^2 + h_1^4 \left(\frac{1}{2\sigma^2} + \frac{\frac{1}{2}\lambda(\beta_j') - 1}{2\sigma'^2} \right)^2 \right. \\
&\quad \left. + 2h_1^2 \left(-\frac{\lambda(\beta_j')}{2} + k \log \left(\frac{\sigma}{\sigma'} \right) \right) \left(\frac{1}{2\sigma^2} + \frac{\frac{1}{2}\lambda(\beta_j') - 1}{2\sigma'^2} \right) \right] \\
&\quad < \infty,
\end{aligned}$$

where the finity follows from the fact that all involved constants are finite and the first and second moment of a non-central chi-square are too.

Next, denote $C_0 = \beta_{-j}$, $\Sigma_0 = \sigma^2 (X_{-j}^{-1} X_{-j})^{-1}$, $C_1 = \beta_{-j}' + B X_j \beta_j'$ and $\Sigma_1 = \sigma'^2 (X_{-j}^{-1} X_{-j})^{-1}$. Then for $i \in \{0, 1\}$,

$$p_{\theta_i}^{[\phi_2(\mathbf{Y})]}(h_2) = \frac{1}{\sqrt{(2\pi)^q |\Sigma_i|}} e^{-\frac{1}{2}(h_2 - C_i)^T \Sigma_i^{-1} (h_2 - C_i)}.$$

So,

$$\mathbb{E}_{h_2\sim P_{\theta_1}^{[\phi_2(\mathbf{Y})]}}\left[\left|\log\left(\frac{p_{\theta_1}^{[\phi_2(\mathbf{Y})]}(h_2)}{p_{\theta_0}^{[\phi_2(\mathbf{Y})]}(h_2)}\right)\right|^{1+\epsilon}\right]$$
$$= \mathbb{E}_{h_2\sim P_{\theta_1}^{[\phi_2(\mathbf{Y})]}}\left[\left|\frac{1}{2}(h_2-C_0)^T\Sigma_0^{-1}(h_2-C_0)-\frac{1}{2}(h_2-C_1)^T\Sigma_1^{-1}(h_2-C_1)\right.\right.$$
$$\left.\left.+q\log\left(\frac{\sigma}{\sigma'}\right)\right|^{1+\epsilon}\right].$$

Note that for $i \in \{0,1\}$, Σ_i^{-1} is a multiple of $(X_{-j}^T X_{-j})$, so it is positive semi-definite and thus

$$\frac{1}{2}(h_2-C_i)^T\Sigma_i^{-1}(h_2-C_i) \geq 0.$$

Therefore, the product of the first two terms is negative. For the other terms, it can be shown that $(h_2-C_i)^T\Sigma_i^{-1}(h_2-C_i)$ can be written as a weighted sum of finitely many independent chi-square random variables [30, 31]. These all have finite mean and variance, so for $\epsilon = 1$, the quantity of interest is indeed finite.

Putting this together with the independence of $A\mathbf{Y}$ and $B\mathbf{Y}$,

$$\mathbb{E}_{(h_1,h_2)\sim P_{\theta_1}^{[\phi(\mathbf{Y})]}}\left[\left|\log\left(\frac{p_{\theta_1}^{[\phi(\mathbf{Y})]}(h_1,h_2)}{p_{\theta_0}^{[\phi(\mathbf{Y})]}(h_1,h_2)}\right)\right|^2\right]$$
$$= \mathbb{E}_{(h_1,h_2)\sim P_{\theta_1}^{[\phi(\mathbf{Y})]}}\left[\left|\log\left(\frac{p_{\theta_1}^{[\phi_1(\mathbf{Y})]}(h_1)}{p_{\theta_0}^{[\phi_1(\mathbf{Y})]}(h_1)}\right)+\log\left(\frac{p_{\theta_1}^{[\phi_2(\mathbf{Y})]}(h_2)}{p_{\theta_0}^{[\phi_2(\mathbf{Y})]}(h_2)}\right)\right|^2\right]$$
$$\leq \mathbb{E}_{(h_1,h_2)\sim P_{\theta_1}^{[\phi(\mathbf{Y})]}}\left[3\left|\log\left(\frac{p_{\theta_1}^{[\phi_1(\mathbf{Y})]}(h_1)}{p_{\theta_0}^{[\phi_1(\mathbf{Y})]}(h_1)}\right)\right|^2+3\left|\log\left(\frac{p_{\theta_1}^{[\phi_2(\mathbf{Y})]}(h_2)}{p_{\theta_0}^{[\phi_2(\mathbf{Y})]}(h_2)}\right)\right|^2\right]$$
$$< \infty,$$

where it is used that $xy \leq \max\{x^2,y^2\} \leq x^2+y^2$. This concludes the first part of the Lemma. The second part follows from a similar argument to the proof of Lemma 2.8. □

Miscellaneous proofs

None of the results in this subsection are claimed as novel. The proofs rely solely on basic well-known results, and have most likely already been described in some form or fashion.

Lemma B.1. *$G = \mathbb{R}_{>0} \times \mathbb{R}^q$ is a σ-compact and locally compact Hausdorff topological group.*

Proof. The sets $\mathbb{R}_{>0}$ and $\mathbb{R}^q$ are both locally compact, as each point in either of these is contained in a closed and bounded interval, which are compact subsets by the Heine-Borel theorem. As the product of two locally compact Hausdorff spaces, G is itself too locally compact Hausdorff. Similarly, it is well-known that both $R_{>0}$ and $\mathbb{R}^q$ are σ-compact when equipped with the standard topology,. As their product, G is σ-compact too.

Next, recall that the group operation for $(\alpha_0, \boldsymbol{\alpha}), (\gamma_0, \boldsymbol{\gamma}) \in G$ is given by

$$(\gamma_0, \boldsymbol{\gamma}) \cdot (\alpha_0, \boldsymbol{\alpha}) = (\gamma_0 \alpha_0, \gamma_0 \boldsymbol{\alpha} + \boldsymbol{\gamma})$$

and that the inverse mapping is given by

$$(\alpha_0, \boldsymbol{\alpha})^{-1} = \left(\frac{1}{\alpha_0}, -\frac{\boldsymbol{\alpha}}{\alpha_0} \right).$$

Since these maps are both continuous, the group G is a topological group. □

Lemma B.2. *The action of G on $\mathcal{Y}$ is proper.*

Proof. It will be shown that the action of G on $G \times \mathbb{R}^k$ is proper. The result follows from the fact that $\mathcal{Y}$ is shown to be homeomorph to $G \times \mathbb{R}^k$ in Section 2.3.2.

Recall that G acts on $G \times \mathbb{R}^k$ by multiplication on the first coordinate and trivially on the second. The mapping $G \times G \to G \times G$ defined by $(g, g') \mapsto (gg', g')$ is a homeomorphism and therefore proper, so the action of G on G is proper [32, Example 3.1]. Since the trivial action is also proper, it follows that the action of G on $G \times \mathbb{R}^k$ is proper [32, Example 3.2]. □

Lemma B.3. *G is amenable.*

Proof. Consider the subset

$$U = \{(1, \boldsymbol{\alpha}) : \boldsymbol{\alpha} \in \mathbb{R}^q\} \subseteq G.$$

It will first be shown that U is a subgroup. It is evident that U contains the identity element $(1, \mathbf{0})$ of G. Furthermore, for $(1, \boldsymbol{\alpha}) \in U$ arbitrarily,

the inverse $(1, -\boldsymbol{\alpha})$ is also an element of U. Finally, let $(1, \boldsymbol{\gamma}) \in U$ be given arbitrarily too. It holds that

$$(1, \boldsymbol{\gamma}) \cdot (1, \boldsymbol{\alpha}) = (1, \boldsymbol{\alpha} + \boldsymbol{\gamma}) \in U.$$

So U is indeed a subgroup of G.

Next, it will be shown that U contains the commutator $[G, G]$. To this end, let $(\alpha_0, \boldsymbol{\alpha}), (\gamma_0, \boldsymbol{\gamma}) \in G$ arbitrarily. To compute the commutator of these two elements, note that

$$(\gamma_0, \boldsymbol{\gamma}) \cdot (\alpha_0, \boldsymbol{\alpha}) = (\alpha_0\gamma_0, \gamma_0\boldsymbol{\alpha} + \boldsymbol{\gamma})$$

and

$$\begin{aligned}(\gamma_0, \boldsymbol{\gamma})^{-1} \cdot (\alpha_0, \boldsymbol{\alpha})^{-1} &= \left(\frac{1}{\gamma_0}, -\frac{\boldsymbol{\gamma}}{\gamma_0}\right) \cdot \left(\frac{1}{\alpha_0}, -\frac{\boldsymbol{\alpha}}{\alpha_0}\right) \\ &= \left(\frac{1}{\alpha_0\gamma_0}, -\frac{\boldsymbol{\alpha}}{\alpha_0\gamma_0} - \frac{\boldsymbol{\gamma}}{\gamma_0}\right).\end{aligned}$$

Therefore, it is true that

$$\begin{aligned}[(\gamma_0, \boldsymbol{\gamma}), (\alpha_0, \boldsymbol{\alpha})] &= (\gamma_0, \boldsymbol{\gamma})^{-1} \cdot (\alpha_0, \boldsymbol{\alpha})^{-1} \cdot (\gamma_0, \boldsymbol{\gamma}) \cdot (\alpha_0, \boldsymbol{\alpha}) \\ &= \left(\frac{1}{\alpha_0\gamma_0}, -\frac{\boldsymbol{\alpha}}{\alpha_0\gamma_0} - \frac{\boldsymbol{\gamma}}{\gamma_0}\right) \cdot (\alpha_0\gamma_0, \gamma_0\boldsymbol{\alpha} + \boldsymbol{\gamma}) \\ &= \left(1, \frac{\boldsymbol{\alpha}}{\alpha_0} + \frac{\gamma}{\alpha_0\gamma_0} - \frac{\boldsymbol{\alpha}}{\alpha_0\gamma_0} - \frac{\boldsymbol{\gamma}}{\gamma_0}\right) \in U.\end{aligned}$$

Since the commutator of any two elements of G is in U, the commutator of G itself is a subset of U, i.e. $[G, G] \subseteq U$. Therefore, U is a normal subgroup and the quotient group G/U is abelian. Notice that U itself is abelian too and since abelianity is a sufficient condition for amenability [16, Theorem 4.6.1], U and G/U are abelian. This is a sufficient condition for G to be amenable [16, Proposition 4.5.5] □

Lemma B.4. *The measure $d\nu_l(\alpha_0, \boldsymbol{\alpha}) = \frac{1}{\alpha_0^{n+1}} d\alpha_0 d\boldsymbol{\alpha}$ is a left Haar measure on G.*

Proof. (sketch) The statement will be made plausible by showing that ν_l is left-translation invariant.

Let $(\alpha_0, \boldsymbol{\alpha}), (\alpha_0', \boldsymbol{\alpha}') \in G$ arbitrarily. Then

$$\begin{aligned}d\nu_l((\alpha_0', \boldsymbol{\alpha}')(\alpha_0, \boldsymbol{\alpha})) &= \frac{d(\alpha_0\alpha_0')d(\alpha_0'\boldsymbol{\alpha} + \boldsymbol{\alpha}')}{\alpha_0'^{n+1}\alpha_0^{n+1}} \\ &= \frac{\alpha_0'^{n} d\alpha\alpha_0' d\alpha_0}{\alpha_0'^{n+1}\alpha_0^{n+1}} \\ &= \frac{d\alpha_0 d\boldsymbol{\alpha}}{\alpha_0^{n+1}} = d\nu_l((\alpha_0, \boldsymbol{\alpha})).\end{aligned}$$

Here, it is used that $d(\alpha_0\alpha_0') = \alpha_0' d\alpha_0$ and $d(\alpha_0'\boldsymbol{\alpha} + \boldsymbol{\alpha}') = \alpha_0'^n d\boldsymbol{\alpha}$, both of which follow from a simple change of variables. □

B.2 Proofs for Chapter 3

In this Chapter, the notation of Section 3.1 will be used.

Proof. **(of Theorem 3.3)** For $\theta \in \Theta_0$ arbitrarily, it holds that

$$\begin{aligned}\mathbb{E}_{\mathbf{Y}\sim P_\theta}\left[\frac{p_{W_1}(\mathbf{Y})}{p_{W_\epsilon}(\mathbf{Y})}\right] &= \mathbb{E}_{\mathbf{Y}\sim P_{W_1}}\left[\frac{p_\theta(\mathbf{Y})}{p_{W_\epsilon}(\mathbf{Y})}\right] \\ &= \mathbb{E}_{\mathbf{Y}\sim P_{W_1}}\left[\frac{p_\theta(\mathbf{Y})}{p_{W_0^o}(\mathbf{Y})}\frac{p_{W_0^o}(\mathbf{Y})}{p_{W_\epsilon}(\mathbf{Y})}\right].\end{aligned}$$

Define the function

$$g(\eta) := \mathbb{E}_{\mathbf{Y}\sim P_{W_1}}\left[\frac{p_\theta(\mathbf{Y})}{p_{W_0^o}(\mathbf{Y})}\left(\frac{p_{W_0^o}(\mathbf{Y})}{p_{W_\epsilon}(\mathbf{Y})}\right)^\eta\right].$$

By the mean value theorem, there exists an $\eta' \in [0,1]$ such that:

$$\begin{aligned}&\mathbb{E}_{\mathbf{Y}\sim P_\theta}\left[\frac{p_{W_1}(\mathbf{Y})}{p_{W_\epsilon}(\mathbf{Y})}\right] = g(1) = g(0) + \frac{d}{d\eta}g(\eta)|_{\eta=\eta'} \\ &= \mathbb{E}_{\mathbf{Y}\sim P_{W_1}}\left[\frac{p_\theta(\mathbf{Y})}{p_{W_0^o}(\mathbf{Y})}\right] + \mathbb{E}_{\mathbf{Y}\sim P_{W_1}}\left[\frac{p_\theta(\mathbf{Y})}{p_{W_0^o}(\mathbf{Y})}\log\left(\frac{p_{W_0^o}(\mathbf{Y})}{p_{W_\epsilon}(\mathbf{Y})}\right)\left(\frac{p_{W_0^o}(\mathbf{Y})}{p_{W_\epsilon}(\mathbf{Y})}\right)^{\eta'}\right] \\ &\le 1 + \mathbb{E}_{\mathbf{Y}\sim P_{W_1}}\left[\frac{p_\theta(\mathbf{Y})}{p_{W_0^o}(\mathbf{Y})}\left(\frac{p_{W_0^o}(\mathbf{Y})}{p_{W_\epsilon}(\mathbf{Y})}\right)^{\eta'}\log\left(\frac{p_{W_0^o}(\mathbf{Y})}{p_{W_\epsilon}(\mathbf{Y})}\right)\right] \\ &\le 1 + \mathbb{E}_{\mathbf{Y}\sim P_{W_1}}\left[\sup_{\eta''\in(0,1)}\frac{p_\theta(\mathbf{Y})}{p_{W_0^o}(\mathbf{Y})}\left(\frac{p_{W_0^o}(\mathbf{Y})}{p_{W_\epsilon}(\mathbf{Y})}\right)^{\eta''}\log^+\left(\frac{p_{W_0^o}(\mathbf{Y})}{p_{W_\epsilon}(\mathbf{Y})}\right)\right] \\ &\le 1 + V\mathbb{E}_{\mathbf{Y}\sim P_{W_1}}\left[\log^+\left(\frac{p_{W_0^o}(\mathbf{Y})}{p_{W_\epsilon}(\mathbf{Y})}\right)\right] \\ &= 1 + V\int\limits_{y:\frac{p_{W_0^o}(y)}{p_{W_\epsilon}(y)}\ge 1} p_{W_1}(y)\log\left(\frac{p_{W_0^o}(y)}{p_{W_\epsilon}(y)}\right)dy \\ &\le 1 + \frac{V\log V}{\sqrt{\log V + 1/V - 1}}\sqrt{\epsilon},\end{aligned}$$

The final inequality is found using techniques from Lemma 3 and 4 from Yang and Barron [33]. To this end, lower bound $\sqrt{\epsilon}$ as follows:

$$\begin{aligned}
\sqrt{\epsilon} &\geq \sqrt{D(P_{W_1}\|P_{W_\epsilon}) - D(P_{W_1}\|P_{W_0^o})} \\
&= \sqrt{\mathbb{E}_{\mathbf{Y}\sim P_{W_1}}\left[\log \frac{p_{W_0^o}(\mathbf{Y})}{p_{W_\epsilon}(\mathbf{Y})}\right]} \\
&= \sqrt{\mathbb{E}_{\mathbf{Y}\sim P_{W_1}}\left[\log \frac{p_{W_0^o}(\mathbf{Y})}{p_{W_\epsilon}(\mathbf{Y})} + \frac{p_{W_\epsilon}(\mathbf{Y})}{p_{W_0^o}(\mathbf{Y})} - \mathbb{E}_{\mathbf{Y}\sim P_{W_\epsilon}}\left[\frac{p_{W_1}(\mathbf{Y})}{p_{W_0^o}(\mathbf{Y})}\right]\right]} \\
&\geq \sqrt{\mathbb{E}_{\mathbf{Y}\sim p_{W_1}}\left[\log \frac{p_{W_0^o}(\mathbf{Y})}{p_{W_\epsilon}(\mathbf{Y})} + \frac{p_{W_\epsilon}(\mathbf{Y})}{p_{W_0^o}(\mathbf{Y})} - 1\right]} \\
&\geq \mathbb{E}_{\mathbf{Y}\sim P_{W_1}}\left[\sqrt{\log \frac{p_{W_0^o}(\mathbf{Y})}{p_{W_\epsilon}(\mathbf{Y})} + \frac{p_{W_\epsilon}(\mathbf{Y})}{p_{W_0^o}(\mathbf{Y})} - 1}\right],
\end{aligned}$$

where the last step follows from Jensen's inequality. Also note that for all $z \in \mathbb{R}_{>0}$, it holds that $-\log(z) + z - 1 \geq 0$ so that the square root is well-defined. It then can be seen that

$$\begin{aligned}
&\mathbb{E}_{\mathbf{Y}\sim P_{W_1}}\left[\sqrt{\log \frac{P_{W_0^o}(\mathbf{Y})}{P_{W_\epsilon}(\mathbf{Y})} + \frac{P_{W_\epsilon}(\mathbf{Y})}{P_{W_0^o}(\mathbf{Y})} - 1}\right] \\
&\quad= \int_y p_{W_1}(y)\sqrt{\log \frac{p_{W_0^o}(y)}{p_{W_\epsilon}(y)} + \frac{p_{W_\epsilon}(y)}{p_{W_0^o}(y)} - 1}dy \\
&\quad\geq \int\limits_{y:\frac{p_{W_0^o}(y)}{p_{W_\epsilon}(y)}\geq 1} p_{W_1}(y)\sqrt{\log \frac{p_{W_0^o}(y)}{p_{W_\epsilon}(y)} + \frac{p_{W_\epsilon}(y)}{p_{W_0^o}(y)} - 1}dy \\
&\quad\geq \frac{\sqrt{\log V + 1/V - 1}}{\log V} \int\limits_{y:\frac{p_{W_0^o}(y)}{p_{W_\epsilon}(y)}\geq 1} p_{W_1}(y)\log \frac{p_{W_0^o}(y)}{p_{W_\epsilon}(y)}dy.
\end{aligned}$$

The last inequality comes from the fact that the function $\frac{\sqrt{\log x + 1/x - 1}}{\log x}$ is non-negative and decreasing on $(1, \infty)$. Putting everything together gives

the desired inequality:

$$\int\limits_{y:\frac{p_{W_0^o}(y)}{p_{W_\epsilon}(y)}\geq 1} p_{W_1}(y)\log\frac{p_{W_0^o}(y)}{p_{W_\epsilon}(y)} \leq \frac{\log V}{\sqrt{\log V + 1/V - 1}}\sqrt{\epsilon}.$$

□