University of Amsterdam

MSc Mathematics

Master Thesis

Linear regression testing with *e*-variables

Author:
S. U. Arias

Supervisor:
dr. T.A.L Erven

Examination date:
February 9, 2026

Daily supervisor:
prof. dr. P.D. Grünwald

Korteweg-de Vries Institute for Mathematics

Centrum voor Wiskunde en Informatica

Abstract

P-values have been the main tool for designing statistical tests for almost a century. The novel theory of *e*-values offers an alternative approach, providing testing procedures that allow safe optional stopping of data collection, and safety guarantees for data-dependent significance levels. We systematically compare two types of *e*-variables that have been proposed for linear regression testing: the Model-X (MX) and the Simple Linear Regression (S-LR) *e*-variable. These *e*-variables differ fundamentally in their design principles. The MX *e*-variable was originally developed for conditional independence testing and relies on knowledge of part of the underlying distribution of the *covariates*. In contrast, the S-LR *e*-variable was specifically designed for linear regression testing, exploiting linearity and distributional assumptions on the *target* variable.
We analyze and mathematically compare these *e*-variables in a setting in which both methods are simultaneously applicable. A key instance of this setting is provided by randomized controlled trials. For simple alternative hypotheses, we prove that, asymptotically as the sample size tends to infinity, MX never outperforms S-LR in terms of *e*-power. However, we show that MX can provide robust tests with theoretical safety guarantees of great practical relevance. Surprisingly, we find that in a generalized ANCOVA setting, the asymptotic *e*-power difference is very small, of order $O(||\delta||^6)$ as the effect size vector δ tends to zero. All our findings naturally extend to the adaptations of S-LR and MX *e*-variables for testing the full (composite) linear alternative hypothesis, which are directly applicable to linear regression testing in practice.

Title: Linear regression testing with *e*-variables

Author: S. U. Arias, sebastian.arias2@student.uva.nl, 15075281
Supervisor: prof. dr. P.D. Grünwald, dr. T.A.L. Erven
Second Examiner: prof. dr. J. M. Mooij

Examination date: February 9, 2026

Korteweg-de Vries Institute for Mathematics
University of Amsterdam
Science Park 105-107, 1098 XG Amsterdam
`http://kdvi.uva.nl`

Centrum voor Wiskunde en Informatica
Science Park 123, 1098 XG Amsterdam
`http://www.cwi.nl`

Contents

1 Introduction **4**
- 1.1 Statistical hypothesis testing . . . 5
- 1.2 Testing with e-variables . . . 9
- 1.3 Contents, contributions and conclusion . . . 11

2 Mathematical preliminaries **13**
- 2.1 Properties of Multivariate Gaussians . . . 13
- 2.2 Background on e-variables . . . 15
  - 2.2.1 Fixed-sample testing . . . 15
  - 2.2.2 Sequential testing . . . 19
- 2.3 Group theory . . . 22
- 2.4 Markov kernels and Model-X . . . 24
- 2.5 Classical LR testing . . . 27

3 Simple alternatives **32**
- 3.1 Simple LR test . . . 33
  - 3.1.1 Fixed design . . . 33
  - 3.1.2 Random design . . . 35
- 3.2 Model-X test for LR . . . 38
- 3.3 Examples - AN(C)OVA testing . . . 43
  - 3.3.1 ANOVA testing with MX and S-LR . . . 43
  - 3.3.2 Two-factor ANOVA with S-LR . . . 45
  - 3.3.3 Robust ANCOVA with MX . . . 47
- 3.4 The fully specified null . . . 48
- 3.5 E-power comparison . . . 52
- 3.6 Cheat sheet . . . 61

4 Composite alternatives **62**
- 4.1 Introduction: Point vs. Composite Alternatives . . . 62
- 4.2 Group LR test . . . 64
  - 4.2.1 Fixed design . . . 64
  - 4.2.2 Random design . . . 68
- 4.3 Adaptive MX . . . 70
- 4.4 E-power analysis and simulations . . . 71

5 Conclusion **76**

6 Popular summary **78**

1 Introduction

Consider a medical scientist testing the effectiveness of a newly developed drug. Suppose the drug is intended to affect a target quantity e.g. blood pressure. She may conduct a randomized experiment by assigning participants to a treatment or control group and measuring the difference in blood pressure before and after the intervention period.
This experiment can be summarized by the following equation:

$$Y = \beta X + \mu + \varepsilon,$$

where Y denotes the target variable (i.e., the difference in blood pressure before/after intervention), X is a binary variable indicating treatment assignment, μ is the expected value of Y under control, β represents the treatment effect, i.e. difference in expected value of Y under treatment and control, and ε is a 0-mean random variable that captures unobserved effects and measurement errors.
Additional measured variables related to blood pressure, such as age or body mass index, can be included as covariates, e.g. by assuming a linear relation with Y, yielding the extended linear model

$$Y = \beta X + \gamma^T Z + \mu + \varepsilon. \tag{1.1}$$

The existence of a treatment effect corresponds to $\beta \neq 0$, and if a researcher is willing to accept this model, inference reduces to assessing the plausibility of $\beta \neq 0$ given the observed data and all formulated assumptions. The data, in this context, consists of observations $\{(Y_1, X_1, Z_1), \ldots, (Y_n, X_n, Z_n)\}$, where each triplet (Y_i, X_i, Z_i) corresponds to an individual participant. Testing whether the medication has an effect based on observed data via a *linear regression model* such as (1.1) is a classical problem of statistics, and it is also the main topic of this thesis. In particular, we consider new testing approaches based on e-variables, which provide multiple practical advantages over classical linear regression tests, such as the possibility of applying the test sequentially, safety properties under data-driven security levels, and a simple way of combining the results of multiple studies.

The theory of hypothesis testing with e-variables, is very recent, with almost all major results having been developed after 2019, and since then two fundamentally different approaches for linear regression testing have been proposed. The first approach, dubbed Simple Linear Regression (S-LR) is a direct extension of the classical approach, testing whether $\beta = 0$ as above, and heavily relying on an assumption of normally distributed errors ε [Grünwald et al., 2025]. The second rephrases the problem as conditional independence testing and relies instead on a so-called *Model-X* (MX) Assumption - an approach that is quite different from the classical one, but that is eminently suited for

combination with e-variables [Grünwald et al., 2023]. We systematically compare the approaches in settings in which both of them can be used, and for each of the two, we identify situations in which it works better. We find that, while they can exhibit strikingly different behaviour on some examples, they also show surprisingly similar performance on many others.

More precisely, the condition $\beta = 0$ in the model (1.1) can be viewed as a linearized version of the full conditional independence statement:

$$Y \perp X|Z,$$

which may be read as: "given the covariates, the target does not depend on the treatment assignment". While the distinction between linear and full conditional independence is subtle, it is central to the questions addressed in this thesis. Unlike standard regression models, the Model-X framework makes no assumptions on the marginal distribution of the error term. Instead, it relies on knowledge of the conditional distribution $P(X|Z)$ of the treatment assignment given the covariates. Owing to this non-standard modeling assumption, the Model-X approach can give rise to tests with properties that differ substantially from those of classical methods, both in terms of safety guarantees and statistical power, making it particularly worthy of a systematic comparative study. Importantly, the Model-X requirement is satisfied in many practical settings, most notably in randomized controlled trials, where treatment assignment is under the experimenter's control and determined by a known randomization mechanism, such as a coin toss.

Our main contributions include showing that asymptotically as the sample size grows, if the assumptions needed for both approaches hold, then S-LR yields more e-powerful tests than MX, where e-power represents the e-variable theory analog of classical statistical power. This result does not come as a big surprise, as the first approach makes arguably stronger assumptions on the underlying distribution of the data. Nevertheless, we do describe a special setting in which MX dominates on finite samples. As our most surprising result, we show that in a very general setting, which includes randomized controlled trials such as in model (1.1), the two methods yield tests with very similar asymptotic e-power.

The remainder of this chapter provides background on statistical hypothesis testing, a brief historical overview, and an introduction to e-variable theory within the modern statistical framework. Section 1.3 outlines the structure of the thesis, describes our contributions in more depth, and summarizes the main conclusions.

1.1 Statistical hypothesis testing

Abstractly, a (statistical) *hypothesis* is a set of probability distributions defined over a fixed measurable space $(\mathcal{D}, \mathcal{F}^{\mathcal{D}})$ - which represents possible observations of the data, and

is called the *data space*.
For example, a scientist may hypothesize that a particular treatment has no effect on the expected value of a target random variable Y. To investigate this hypothesis, she draws one participant from the population of interest, administers the treatment, and subsequently observes the value of Y. Assuming that the untreated population's mean value of Y is known, and equals μ, her hypothesis can mathematically be expressed as:

$$\mathcal{H} = \{P : P \in \mathcal{M}_1, \mathbb{E}_P[Y] = \mu\},$$

were $\mathcal{M}_1$ represents the set of all probability measures over the data space.

A hypothesis is called *simple*, or a *point hypothesis* if it contains a single distribution, and otherwise it is *composite*.
The theory of (statistical) hypothesis testing formalizes the task of drawing general conclusions from finite samples and of quantifying the risks involved.

The Neyman–Pearson (NP) approach, named after Jerzy Neyman and Egon Pearson, is the standard framework for statistical hypothesis testing and adopts a decision-theoretic perspective [Neyman and Pearson, 1933]. They define two hypotheses for each testing problem: the null $\mathcal{H}_0$ and the alternative hypothesis $\mathcal{H}_1$. A (statistical) *test* is merely a decision rule, which is formalized as a measurable function mapping the observed data to a binary decision. The decision is either to reject the null and accept the alternative hypothesis (output 1), or accept the null and reject the alternative (output 0). There are two types of errors one can make when testing this way:

- Type-1 error: reject $\mathcal{H}_0$ when it is true
- Type-2 error: not reject $\mathcal{H}_0$ when it is false

Controlling the type-1 error probability is considered a safety guarantee, and is given primary importance when designing tests. This motivates the definition of level-α tests.

Definition 1.1.1. (Level-α tests)
A statistical test for a null hypothesis $\mathcal{H}_0$ is said to be of level-α if the probability of rejecting the null is upper-bounded by α under all $P \in \mathcal{H}_0$.
Furthermore, the value α is called the significance (or security) level of the test.

The *power* of a statistical test ϕ under distribution Q is simply the probability of rejecting the null hypothesis if the data is generated by Q, i.e. the value $Q(\{\phi = 1\})$. The power of a test under $Q \notin \mathcal{H}_0$ and the type-2 error probability under Q are complementary:

$$\mathrm{pow}_Q = Q(\{\phi = 1\}) = 1 - Q(\{\phi = 0\}).$$

According to the NP framework, for a fixed $\alpha \in (0, 1)$, the best level-α test is one which maximizes power uniformly (simultaneously) under all distributions in $\mathcal{H}_1$. Such a test is called *uniformly most powerful* (UMP). UMP tests exist only in highly structured special cases, such as having point alternatives or when both null and alternative are

subsets of a one-parameter parametric family satisfying the monotone likelihood ratio property, which in turn includes the one-dimensional exponential families [Casella and Berger, 2024]. In most realistic statistical models, they provably do not exist, and practitioners must rely on other criteria, such as worst-case-optimality [Casella and Berger, 2024].

Ronald Fisher, another famous statistician, proposed a different view of statistical testing [Fisher, 1934]. He argued that the output of a statistical test ought to be interpreted as a measure of observed evidence against the null hypothesis, and considered the alternative hypothesis irrelevant. An important mathematical tool which stems from his work is the p-variable.

Definition 1.1.2. (p-variables)
A random variable U is said to be a p-variable for $\mathcal{H}_0$ if for all $P \in \mathcal{H}_0$, the distribution of U under P is stochastically larger than the uniform distribution. In other words:

$$P(U \leq c) \leq c$$

holds for all $c \in [0, 1]$ and $P \in \mathcal{H}_0$.

P-variables are generally called p-values, but in this thesis we reserve the term "variable" to indicate the random nature of the object, and "value" whenever we refer to a specific realization of the random variable.
An immediate consequence of the definition is that for any p-variable U for $\mathcal{H}_0$:

$$\phi_\alpha := 1_{[0,\alpha]}(U)$$

is a level-α test for $\mathcal{H}_0$.

The two frameworks can, therefore, be connected. From a NP perspective, p-variables may be used as tools for constructing tests, but the final outcome is an accept–reject decision, with no distinction made between p-values that fall on the same side of the threshold α.
In modern practice, however, the p-value is often reported alongside the accept-reject decision, and in accordance with Fisher's views, regarded as a measure of evidence against the null. This practice departs from the NP approach, and treating the p-value as evidence has been repeatedly criticized across statistics [Lindley, 1957, Berger and Sellke, 1987].
Another downside of the modern framework is that it can incentivize *p-hacking*. Researchers might feel the urge to sample additional data points and re-test after observing a p-value which is slightly above α - and thus is almost, but not quite small enough to reject the null hypothesis. Another common hack is that of choosing α post-hoc - after observing the p-value.
If either of these hacks is used, α no longer represents an upper bound on the type-1 error probability, and the safety guarantees of the test are lost.

Example 1.1.3. (p-hacking by sequential testing)
The most extreme example of the first kind of p-hacking would be to use a fixed-sample test and re-test after every new observation. We will use the classical z-test to illustrate how safety guarantees can be lost. The simple null hypothesis is that the data are generated by a standard normal distribution, while under the alternative the data are still normal with variance 1, but the mean can be any real value. The test statistic for n i.i.d. observations $(Y_1, ..., Y_n)$ is their mean value $\overline{Y_n}$. This statistic can be converted into a p-variable for testing, but one can also work with it directly. To get the level-α version of the z-test for n data points, one merely has to reject the null whenever:

$$|\sqrt{n\overline{Y_n}}| \geq q(1 - \frac{\alpha}{2}),$$

and not reject otherwise. Here q is the quantile function of the standard normal distribution, and thus makes the rejection probability at each fixed n exactly α when the null is true. By Kolmogorov's law of the iterated logarithm [Billingsley, 2017]:

$$\limsup_{n\to\infty} |\sqrt{n\overline{Y_n}}| = \infty$$

holds almost surely under the null hypothesis. Therefore, although for each fixed n the rejection probability is α, if we re-test sequentially at every sample size, the null gets rejected with probability one, even when it is true.

In practice, researchers would never re-test infinitely many times, but performing the test twice is already enough to inflate the type-1 error probability, and make it surpass α.

Now, let us address the second p-hack. If the security level α is chosen after observing the realization of the p-variable U, i.e. is data-dependent, in the most severe case one can reject the null hypothesis always, by simply taking a function α which satisfies $\alpha(U) \geq U$.

To remedy the second issue, we could try to modify the p-variable definition, to also allow data-dependent security levels, as shown by Grünwald [2024], Koning [2024].
The p-variable requirement can equivalently be rewritten as:

$$P(U \leq \alpha) \leq \alpha \ \forall P \in \mathcal{H}_0 \iff \mathbb{E}_P[1_{\{U\leq\alpha\}}\frac{1}{\alpha}] \leq 1 \ \forall P \in \mathcal{H}_0, \tag{1.2}$$

for all $\alpha \in (0, 1)$. The right-hand side (RHS) formulation of the safety requirement can be imposed even if we consider α a random variable. So, let us now impose a stricter safety requirement on U – that the RHS of (1.2) holds for all non-negative measurable functions of the data α. It is not hard to see that the maximal value over α is achieved exactly for the random variable $\alpha = U$, and that the new safety requirement on U amounts to:

$$\mathbb{E}_P\left[\frac{1}{U}\right] \leq 1 \ \forall P \in \mathcal{H}_0.$$

Definition 1.1.4. (e-variables)
A non-negative random variable S is called an e-variable for the hypothesis $\mathcal{H}_0$ if:

$$\mathbb{E}_P[S] \leq 1$$

for all $P \in \mathcal{H}_0$.

With this new terminology, we can state that a p-variable U for $\mathcal{H}_0$ satisfies (1.2) for any data-dependent α if and only if U^{-1} is an e-variable. On the other hand, by Markov's inequality, the inverse S^{-1} of any e-variable S for $\mathcal{H}_0$ is a p-variable for $\mathcal{H}_0$:

$$P(S^{-1} \leq \alpha) = P(S \geq \frac{1}{\alpha}) \leq \alpha \mathbb{E}_P[S] \leq \alpha.$$

Therefore, there is a one-to-one correspondence between the set of e-variables for $\mathcal{H}_0$, and a subset of p-variables for $\mathcal{H}_0$, which provide the safety guarantee (1.2) for all data-dependent choices of α, as opposed to only fixed constants $\alpha \in (0, 1)$.

It is important to realize that classical p-variables, which are used for testing in practice, do not provide this stronger safety guarantee. They are generally designed to take values in $[0, 1]$, meaning that the inverse of any such p-variable U can only be an e-variable if

$$U = U^{-1} = 1$$

holds a.s. under all distributions in the null. But any such p-variable is clearly useless for testing.

E-variables thus provide an answer to the problem of data dependent significance levels. If we agree that the RHS of (1.2) is a meaningful generalization of our original safety requirement for data-dependent values α, then we merely need to restrict the set of p-variables we consider for designing tests to those whose inverses are e-variables. Since our tests are based on thresholding p-variables:

$$\phi_\alpha = 1_{[0,\alpha]}(U) = 1_{[\frac{1}{\alpha},\infty]}(U^{-1}),$$

we could equivalently say that, if we want safety guarantees for data-dependent significance levels, we should use tests which are based on thresholding e-variables instead of (general) p-variables.

We will return to the other issue with p-variables, which we encountered in Example 1.1.3 in the next section.

1.2 Testing with e-variables

The majority of e-variable research has approached hypothesis testing from a decision-theoretic perspective, in the spirit of Neyman and Pearson. The focus is on creating

level-α tests, which provide type-1 error guarantees. The null hypothesis has the same classical interpretation as in the NP framework, simply providing a set of distributions of the data, under which the rejection probability of a test has to be controlled if the test is to be considered safe. The alternative hypothesis, on the other hand, serves exclusively as a guideline for choosing useful e-variables. The output of a test is a decision to reject or not to reject the null, while the alternative is never considered rejected or accepted.

In order to grasp all the benefits of e-variable theory it is necessary to differentiate between *fixed-sample* and *sequential* statistical tests [1] . So far, we have only discussed fixed-sample tests, which were the focus of Neyman, Pearson and Fisher. In a fixed-sample test, the sample size n is predetermined, and n data points are observed and analyzed at once. The data space $(\mathcal{D}, \mathcal{F}^{\mathcal{D}})$ is the set of all possible values of the n data points, endowed with a suitable σ-algebra. As we have seen before, a *fixed-sample test* is simply a measurable function mapping the data to $\{0, 1\}$.
We can also model a testing scenario where the data points are observed sequentially, one-by-one. Then, the data space $\mathcal{D}$ contains the possible outcomes of a whole countable sequence of data points $(D_1, D_2, ...)$, and is additionally equipped with the data filtration $(\mathcal{F}_n^{\mathcal{D}})_{n\in\mathbb{N}}$, where $\mathcal{F}_n^{\mathcal{D}}$ is the σ-algebra generated by the first n data points. A *sequential test* is then any stopping time τ, adapted to the data filtration. This means that stopping at time $i \in \mathbb{N}$ can be interpreted as rejecting the null hypothesis based on the first i data points, after which in practice data sampling would be halted. To simplify our definitions and results, we always assume that infinitely many data points are available when testing sequentially, i.e. a researcher can, if she wants, keep sampling data indefinitely. With this interpretation, a sequential test τ rejects the null whenever $\tau < \infty$, and does not reject only when $\tau = \infty$. Therefore, τ can be converted into a fixed-sample test:

$$1_{\{\tau<\infty\}},$$

and fixed-sample testing definitions of type-1 errors and level-α tests can also be applied to sequential tests.

The first form of p-hacking, discussed in 1.1.3, is a consequence of using a fixed-sample test in a sequential way, for which the test does not provide safety guarantees by design. E-variables can be used to create sequential tests which allow the user to stop or proceed with data-collection based on the data observed so far, without sacrificing safety guarantees. Thus, they can be regarded as a solution to this form of p-hacking as well.

As a final note on e-variables, one can easily check that multiplying two independent e-variables for the same null leads to a new e-variable. Researchers can, therefore, easily combine the results of two experiments which share the same null hypothesis (i.e. research question) and are based on independent data. Combining the outcomes this way is possible even if the second experiment would not have happened at all were it

[1] Sequential testing was pioneered by Abraham Wald, whose work was highly influential in the development of e-variable theory [Wald and Wolfowitz, 1948].

not for the results of the first [Ter Schure and Grünwald, 2019], which is often the case in scientific practice; with p-variables, this is impossible to achieve without destroying type-1 error guarantees.

All statistical tests described in this thesis are e-variable based. We are primarily interested in linear regression testing [Agresti, 2015], as one of the most widely used statistical tests in practice. Our hope is to spark the reader's curiosity about e-variables, by showing their importance for contemporary statistical practice and describing several methods for creating useful e-variables in the context of linear regression testing.

The following section describes the contents of the remainder of the thesis. A major part of our results are related to measuring and comparing the expected performance – or usefulness of our e-variables. The measure of usefulness is formalized by e-power. A GRO e-variable is one which achieves optimal e-power against a given point alternative, while GROW/REGROW refers to versions of optimal e-power for composite alternatives. All of these concepts will be defined in detail later on.

1.3 Contents, contributions and conclusion

In Chapter 2, we present the necessary mathematical background. We begin by reviewing basic properties of multivariate Gaussian distributions and related concepts, including exponential families and the Kullback–Leibler divergence. We then provide a general introduction to e-variable theory, outlining its key definitions and results, such as the GRO e-variable and its construction. This is followed by a brief section on group theory, which plays a central role in the development of e-variable-based linear regression tests for composite alternatives, as well as an overview of the Model-X framework for conditional independence testing. The chapter is concluded with a recap of classical linear regression testing.

Our own contributions begin in Chapter 3, with an analysis of the random design version of the Simple Linear Regression (S-LR) e-variable, which was originally introduced and described in fixed design by Grünwald et al. [2025]. We show that the optimality properties from fixed design transfer directly to random design under a suitable choice of the null hypothesis. Next, we analyze the MX e-variable in the context of LR testing, and provide practical examples in which S-LR and MX e-variables have very different properties. We go on to describe the practical relevance and harshness of effectively testing the intersection of the S-LR and Model-X (MX) null hypotheses, and of the even smaller, Fully Specified Linear Regression (FSLR) null. This analysis leads us to a special setting, in which the MX e-variable coincides with the sequential Growth Rate Optimal (GRO) e-variable for testing the FSLR null. The chapter is concluded with a theoretical e-power comparison of S-LR and MX e-variables, which contains our most important results. We find that, asymptotically as the sample size tends to infinity, S-LR reaches the highest possible benchmark, achieving the same asymptotic relative power

(ARP) as the GRO e-variable for the FSLR null (Theorem 3.5.6). As a consequence, asymptotically S-LR never performs worse than MX (in terms of ARP), but we show that MX can dominate on finite samples (Proposition 3.5.7). A deeper analysis for the ANCOVA setting is conducted (Theorem 3.5.8), and shows that the asymptotic e-power difference between the two types of e-variables is of order $O(||\delta||^6)$ as the effect size vector δ tends to zero. At the end of this chapter we provide a one-page "cheat-sheet", containing the most important conclusions and formulas of the chapter.

All our results in Chapter 3 were in the context of a simple (point) alternatives — a case which is unrealistic but carries essentially all the difficult work: once it has been achieved the extension to composite alternatives is straightforward. Thus, in Chapter 4 we describe the construction of composite alternative adaptations of the MX and S-LR e-variables from Chapter 3. The adaptations are called Adaptive MX (AMX) and the Group Linear Regression (G-LR) e-variables, and were originally described by Pérez-Ortiz et al. [2024], Lindon et al. [2024] and Grünwald et al. [2023]. We show that GROW/REGROW optimality and safe sequential testing properties transfer directly to random design, and empirically verify the e-power similarities of G-LR and AMX e-variables in the ANCOVA setting.

We end in Chapter 5 with a thorough description of our most important conclusions and suggestions for future work, while a short summary of the most notable findings is given right away.
The S-LR e-variable, despite of not being built on any assumptions about the distribution of the covariates, achieves the same asymptotic relative power as the GRO e-variable for the FSLR null (Theorem 3.5.6). The latter e-variable does have knowledge of the true covariate distribution incorporated by design, but it can rarely be analytically described and used in practice. Thus, the price a researcher pays for not having, or not being able to incorporate (partial) knowledge of the true covariate distribution is at most of order $o(n)$, where n is the sample size, which is remarkable, since the e-power itself typically increases linearly in n. As a special consequence, we also know that asymptotically S-LR is always at least as e-powerful as MX. This difference however, is almost negligible in a described setting, which includes randomized controlled trials with ANCOVA designs (Theorem 3.5.8). On a robust ANCOVA example (see Section 3.3), we demonstrate that the MX e-variable can provide, practically very important, theoretical safety guarantees which do not necessarily hold for S-LR. Importantly, all conclusions about asymptotic relative powers of S-LR and MX transfer directly to the composite alternative extensions - G-LR and AMX, making the findings practically relevant.

2 Mathematical preliminaries

This chapter covers the mathematical background necessary to follow the remainder of the thesis.

Outline:
Section 2.1 introduces some important properties of multivariate Gaussians, and their central role in linear regression models. Furthermore, we discuss the Kullback-Leibler divergence, exponential families and sufficient statistics, which are all essential concepts in the process of finding optimal e-variables and will be used throughout Chapters 3 and 4.
In Section 2.2 we lay out some of the most important findings in e-variable theory, including the main results on the existence and construction of optimal e-variables, which are used to define both the S-LR e-variable in Section 3.1, and the MX e-variable in Section 2.4. We further discuss sequential testing, and two methods for adapting point-alternative tests to composite-alternative tests, which are used in Chapter 4.
The short Section 2.3 gives fundamental group-theory definitions needed to follow the construction of the Group Linear Regression e-variable in Section 4.2.
Section 2.4 introduces the Markov kernel notation and presents the MX framework for conditional independence testing, including the MX e-variable, its properties, and the main results of Grünwald et al. [2023].
Finally, Section 2.5 gives a quick overview of classical linear regression testing, including ordinary-least-squares (OLS) estimators, fundamental results such as the Gauss-Markov theorem, and the classical F-test for linear regression. The primary purpose of this section is comparative: it can be used to contrast the classical framework with the e-value theory introduced later.

Important note:
To ensure regular behavior of random variables and processes, throughout this thesis the data space $(\mathcal{D}, \mathcal{F}^{\mathcal{D}})$ is assumed to be a standard Borel space.

2.1 Properties of Multivariate Gaussians

The multivariate Gaussian distribution will be one of the main mathematical objects used throughout this thesis, as it is used to model error terms in classical linear regression. We lay out some of its important properties and notation in this section.
We use $\mathcal{N}(\mu, \Sigma)$ to denote the multivariate Gaussian distribution with mean μ and

covariance matrix Σ, which we invariably assume to be regular. The probability density function (PDF) of $\mathcal{N}(\mu, \Sigma)$ is given by:

$$\mathcal{N}(y^n|\mu, \Sigma) = \frac{1}{\sqrt{2\pi|\Sigma|}} \exp\left(-\frac{1}{2}(y^n - \mu)^T \Sigma^{-1}(y^n - \mu)\right).$$

The notation we use for the distribution itself and its PDF is similar, but this should not lead to confusion as the density always has an additional input.

The KL-divergence is an information-theoretic quantity, traditionally used to represent the expected number of extra bits required when encoding data generated by distribution Q, using a code optimized for distribution P. It plays a central role in statistical testing, appearing in fundamental theorems such as the Chernoff-Stein lemma [Cover and Thomas, 1991]. The generalized, measure theoretic definition of the KL-divergence between two σ-finite measures Q and P is:

$$D_{\mathrm{KL}}(Q||P) := \begin{cases} \mathbb{E}_Q\left[\ln \frac{dQ}{dP}\right], & \text{if } Q \ll P \\ \infty, & \text{if } Q \not\ll P \end{cases}$$

where $Q \ll P$ means that Q is absolutely continuous w.r.t. P. As we will see later on, the KL-divergence is a central concept in e-variable theory. The KL-divergence between two n-dimensional multivariate Gaussians is given by [Cover and Thomas, 1991]:

$$\begin{aligned} D_{\mathrm{KL}}(\mathcal{N}(\mu_1, \Sigma_1)||\mathcal{N}(\mu_2, \Sigma_2)) &= \int \mathcal{N}(y^n|\mu_1, \Sigma_1) \ln \frac{\mathcal{N}(y^n|\mu_1, \Sigma_1)}{\mathcal{N}(y^n|\mu_2, \Sigma_2)} dy^n \\ &= \frac{1}{2}\left(\ln \det\left(\frac{\Sigma_2}{\Sigma_1}\right) + \mathrm{Tr}\left(\frac{\Sigma_1}{\Sigma_2}\right) + (\mu_1 - \mu_2)^T \Sigma_2^{-1}(\mu_1 - \mu_2) - n\right). \quad (2.1) \end{aligned}$$

The fraction $\frac{A}{B}$ of two matrices in the formula can be interpreted as either $B^{-1}A$ or AB^{-1}, since the trace and determinant of the product do not depend on the order of multiplication. We are particularly interested in Gaussian families of the following form:

$$\{\mathcal{N}(w^n\eta, \sigma^2 I_n) : \eta \in \mathbb{R}^{d_w}, \sigma^2 > 0\}, \quad (2.2)$$

where w^n is a fixed $n \times d_w$ matrix of full column rank, as they are used to represent the null and alternative hypotheses in linear regression testing. The density can be rewritten as:

$$\mathcal{N}(y^n|w^n\eta, \sigma^2 I_n) = c(\eta, \sigma^2) \exp\left(-\frac{1}{2\sigma^2}||y^n||^2 + \frac{\eta^T}{\sigma^2} w^{n^T} y^n\right).$$

This specifically means, (2.2) is an *exponential family* [Casella and Berger, 2024], with a *sufficient statistic*:

$$T(Y^n) = (Y^{n^T} Y^n, w^{n^T} Y^n).$$

Definition 2.1.1. (Exponential families)
A hypothesis $\mathcal{H}$ over the data space $\mathcal{D}$ is said to be an exponential family if there exists a σ-finite dominating measure μ, index set Θ and functions:

$$\begin{aligned} q &: \mathcal{D} \to [0, \infty), \\ A &: \Theta \to \mathbb{R}, \\ T &: \mathcal{D} \to \mathbb{R}^k, \\ \eta &: \Theta \to \mathbb{R}^k, \end{aligned}$$

such that q and T are measurable,

$$\mathcal{H} = \{P_\theta : \theta \in \Theta\},$$

and the μ-density of P_θ is given by:

$$\frac{dP_\theta}{d\mu}(d) = q(d) \exp(\eta(\theta)^T T(d) - A(\theta)).$$

The function $T(d)$ in this decomposition is called the sufficient statistic. The family is said to regular if the image of η is an open subset of $\mathbb{R}^k$.

Generally, a sufficient statistic is a measurable function of the data that captures all information in the sample relevant to the parameter of the data-generating distribution. Once the sufficient statistic is known, the rest of the data provides no additional information about the parameter. More formally, the conditional distribution:

$$P_\theta(D|T(D))$$

does not depend on the parameter θ.
As we will see in Proposition 3.1.2, having an exponential family null hypothesis greatly simplifies the problem of finding an effective e-variable for testing point alternatives, while sufficient statistics are crucial in the development of e-variables for linear regression testing with composite alternatives, which is described in Chapter 4.

2.2 Background on e-variables

2.2.1 Fixed-sample testing

All measures and random variables in this section are defined over a fixed measurable space $(\mathcal{D}, \mathcal{F}^{\mathcal{D}})$.
A simple corollary of Markov's inequality is that for any e-variable S for $\mathcal{H}_0$:

$$\phi_\alpha := 1_{[\frac{1}{\alpha}, \infty]}(S)$$

is a level-α fixed-sample test for $\mathcal{H}_0$. All fixed-sample tests of interest for us have this, thresholding form, meaning that the test is completely determined by the choice of the

e-variable S.
The constant 1 is an e-variable for any null hypothesis, but is completely useless for testing. It would be useful to have some notion of optimality, according to which one can look for the "best" e-variable for testing $\mathcal{H}_0$ against $\mathcal{H}_1$. What should the optimality criterion be?
To simplify the question, let us first assume a point alternative hypothesis.
A good e-variable for testing $\mathcal{H}_0$ against $\{Q\}$ should attain large values with high probability under Q. It makes sense to look for e-variables which maximize:

$$\mathbb{E}_Q[u(S)]$$

among all e-variables S, and for some fixed non-decreasing function u.
Grünwald et al. [2024], Koolen and Grünwald [2022] argue that a good choice for u is the logarithm if our goal is to test hypotheses sequentially, using products of e-variables. As we will see in the following subsection, (very often) this is exactly what we want to do. This motivates the definition of the Growth Rate Optimal e-variable.
Let $\mathcal{E}(\mathcal{H}_0)$ be the set of all e-variables for $\mathcal{H}_0$, and $\mathcal{M}_{\leq 1}(\mathcal{D})$ the set of all sub-probability measures on $(\mathcal{D}, \mathcal{F}^{\mathcal{D}})$.

Definition 2.2.1. (e-power and GRO e-variable)
The e-power of e-variable S for testing $\mathcal{H}_0$ against $\{Q\}$ is:

$$\mathbb{E}_Q[\ln S].$$

A Growth Rate Optimal (GRO) e-variable is any e-variable S^* which satisfies:

$$\mathbb{E}_Q\left[\ln \frac{S}{S^*}\right] \leq 0$$

for all $S \in \mathcal{E}(\mathcal{H}_0)$.

The GRO e-variable is also known as the *numeraire*.

Definition 2.2.2. (The effective null hypothesis)
The set:

$$\mathcal{H}_0^{\text{eff}} := \{P \in \mathcal{M}_{\leq 1}(\mathcal{D}) : \mathbb{E}_P[S] \leq 1 \ \forall S \in \mathcal{E}(\mathcal{H}_0)\}$$

is called the effective null hypothesis.

The effective null is the largest set of subprobability measures to which we can extend the original null hypothesis without losing any e-variables.
We say that the alternative Q is absolutely continuous w.r.t. the null:

$$Q \ll \mathcal{H}_0$$

if

$$P(A) = 0 \ \forall P \in \mathcal{H}_0 \Rightarrow Q(A) = 0$$

holds for all $A \in \mathcal{F}^{\mathcal{D}}$. This is a very mild regularity condition, and it will be satisfied in all cases of interest for us.
Note that if the opposite holds, the testing problem is trivial, and one can easily design an e-variable with infinite e-power, e.g.:

$$S = 1 + \infty 1_A$$

where A is any set on which $Q(A) > 0$, and $P(A) = 0 \; \forall P \in \mathcal{H}_0$.
We say that measures P and Q are *equivalent* if they are absolutely continuous w.r.t each other.

Proposition 2.2.3. (The GRO e-variable always exists)
Let $\mathcal{H}_0$ be any collection of probability distributions, and Q a fixed distribution over $\mathcal{D}$. There exists a Q-a.s. unique e-variable S^* for $\mathcal{H}_0$ such that:

$$\mathbb{E}_Q \left[\ln \frac{S}{S^*} \right] \leq 0$$

for all $S \in \mathcal{E}(\mathcal{H}_0)$.

Theorem 2.2.4. (The GRO and the RIPr)
Whenever $Q \ll \mathcal{H}_0$, there exists a unique subprobability distribution $P^* \in \mathcal{H}_0^{\text{eff}}$, equivalent to Q, such that the GRO e-variable S^* satisfies:

$$S^* = \frac{dQ}{dP^*},$$

Q-a.s.

P^* from Theorem 2.2.4 is called the Reverse Information Projection (RIPr) of Q onto $\mathcal{H}_0$.

Theorem 2.2.5. (Finding the GRO and RIPr)
In the setting of the previous theorem, if $\frac{dQ}{dP}$ is an e-variable for some $P \in \mathcal{H}_0^{\text{eff}}$, then $\frac{dQ}{dP}$ is a version of the GRO e-variable. Furthermore, if P is equivalent to Q, then P is the RIPr of Q onto $\mathcal{H}_0$.
Finally, the following duality holds:

$$\mathbb{E}_Q[\ln S^*] = \sup_{S \in \mathcal{E}(\mathcal{H}_0)} \mathbb{E}_Q[\ln S] = \inf_{P \in \mathcal{H}_0^{\text{eff}}} D_{\text{KL}}(Q, P) = D_{\text{KL}}(Q, P^*). \tag{2.3}$$

Theorem 2.2.5 tells us that the GRO e-variable can be identified by finding the minimizer of the KL-divergence between the alternative, and elements of the effective null hypothesis. Although the optimization goal is quite clear, it is in general hard to describe the effective null hypothesis, and even harder to optimize over it. The problem greatly simplifies if we know the minimizer P^* is an element of the null hypothesis itself. This idea will be used extensively throughout Chapter 3.

The statements of Theorems 2.2.4 and 2.2.5 were first proposed by Grünwald et al. [2024] for a more restricted setting, and proved in full generality by Larsson et al. [2025], along with Proposition 2.2.3.

Definition 2.2.6. (The simple case)
We say that the testing problem is in the "simple case" whenever the alternative is simple - $\{Q\}$, and the RIPr of Q onto $\mathcal{H}_0$ is an element of $\mathcal{H}_0$.

A classical example of a GRO e-variable is the simple likelihood ratio. Assume the null hypothesis is simple: $\mathcal{H}_0 = \{P\}$, and that Q and P are equivalent measures. Then $S := \frac{dQ}{dP}$ is an e-variable because:

$$\mathbb{E}_P\left[\frac{dQ}{dP}\right] = Q(\mathcal{D}) = 1.$$

Furthermore, since $P \in \mathcal{H}_0 \subset \mathcal{H}_0^{\text{eff}}$, Theorem 2.2.5 implies that P must be the RIPr of Q onto $\{P\}$, and that $\frac{dQ}{dP}$ is the GRO e-variable.

The simplest interesting example of a GRO e-variable for a composite null can be found in the one-sample t-test setting.

Example 2.2.7. (One sample t-test)
Under the null hypothesis, the data consists of n i.i.d. Gaussians with mean 0 and unknown variance:

$$\mathcal{H}_0 = \{\mathcal{N}(0, \sigma^2 I_n) : \sigma^2 > 0\}.$$

The point alternative is:

$$\{Q\} := \{\mathcal{N}(\mu_1 1_n, \sigma_1^2 I_n)\},$$

for some fixed $\mu_1 \in \mathbb{R}$ and $\sigma_1^2 > 0$. Grünwald et al. [2025] show that the described testing problem is in the simple case. Finding the RIPr then amounts to solving a simple convex optimization problem:

$$\begin{aligned}\arg\min_{\sigma^2} D_{\text{KL}}(\mathcal{N}(\mu_1 1_n, \sigma_1^2 I_n)||\mathcal{N}(0, \sigma^2 I_n)) &= \arg\min_{\sigma^2} n D_{\text{KL}}(\mathcal{N}(\mu_1, \sigma_1^2)||\mathcal{N}(0, \sigma^2)) \\ &= \arg\min_{\sigma^2} \frac{n}{2}(\ln(\frac{\sigma^2}{\sigma_1^2}) + \frac{\sigma_1^2 + \mu_1^2}{\sigma^2} - 1) \\ &= \sigma_1^2 + \mu_1^2.\end{aligned}$$

Therefore, (2.3) implies the RIPr is the distribution:

$$P^* = \mathcal{N}(0, \mu_1^2 + \sigma_1^2)$$

and the GRO e-variable:

$$S(Y^n) = \frac{dQ}{dP^*}(Y^n) = \frac{\mathcal{N}(Y^n|\mu_1 1_n, \sigma_1^2 I_n)}{\mathcal{N}(Y^n|0, (\mu_1^2 + \sigma_1^2) I_n)},$$

where $Y^n := (Y_1, ..., Y_n)$ represents the sampled data.

2.2.2 Sequential testing

In this subsection, we focus on describing e-variable based sequential tests for i.i.d. data, and two standard methods for generalizing simple-alternative sequential tests to composite alternatives.
Let $\mathcal{H}_{0|1}$ be a null hypothesis for the distribution of a single data point - represented by the random variable D. We use P^n to denote the n-fold product of the distribution P. Then:

$$\mathcal{H}_{0|n} := \{P^n : P \in \mathcal{H}_0\}$$

is the extension of $\mathcal{H}_{0|1}$ to n i.i.d. data points $D^n = (D_1, ..., D_n)$. If $S_1(D)$ is an e-variable for $\mathcal{H}_{0|1}$, it is easy to see that:

$$S_n(D^n) := \prod_{i=1}^{n} S_1(D_i)$$

is an e-variable for $\mathcal{H}_{0|n}$.

Definition 2.2.8. (The sequential GRO e-variable)
Let $S_1^{\mathrm{GRO}}(D)$ be the GRO e-variable for testing $\mathcal{H}_{0|1}$ against $\{Q\}$. Then:

$$S_n^{\text{seq-GRO}}(D^n) := \prod_{i=1}^{n} S_1^{\mathrm{GRO}}(D_i)$$

is called the sequential GRO e-variable.

A question that naturally arises now is, under which conditions can we say that the sequential GRO e-variable is equal to the GRO.

Proposition 2.2.9. (The relation between sequential GRO and GRO)
Assume $Q \ll \mathcal{H}_{0|1}$, $Q^n \ll \mathcal{H}_{0|n}$, and the RIPr of Q onto $\mathcal{H}_{0|1}$ is an element of the null hypothesis (the simple case). Then, the sequential GRO and the GRO e-variable for $\mathcal{H}_{0|n}$ against $\{Q^n\}$ coincide Q^n-a.s.

Proof. Let $\frac{dQ}{dP^*}(D)$ be the GRO e-variable for $n = 1$, and P^* the RIPr on Q onto $\mathcal{H}_{0|1}$. Then, $\frac{dQ^n}{dP^{*n}}(D^n) = \prod_{i=1}^{n} \frac{dQ}{dP^*}(D_i)$ is an e-variable. P^{*^n} is an element of $\mathcal{H}_{0|n}$ because $P^* \in \mathcal{H}_{0|1}$. Thus, Theorem 2.2.5 implies $\prod_{i=1}^{n} \frac{dQ}{dP^*}(D_i)$ is the GRO e-variable. □

We can go further and define $\mathcal{H}_{0|\infty}$ using countably infinite products, to represent a hypothesis for the whole (i.i.d.) data process.

Definition 2.2.10. (Test supermartingale)
A discrete time stochastic process $(S_n)_{n\geq 1}$ is a test (super) martingale for a null hypothesis $\mathcal{H}_0$ if it is a non-negative (super) martingale with respect to the data filtration, with expectation at most one at time $n = 1$ under all distributions in $\mathcal{H}_0$.

It is not hard to see that $(S_n^{\text{seq-GRO}})_{n\in\mathbb{N}}$ is a test supermartingale.
Test (super) martingales are the heart of e-variable theory, and the reason will become obvious after the following proposition.

Proposition 2.2.11. (Ville's inequality)
Let $(S_n)_n$ be a test supermartingale for $\mathcal{H}_0$. Then:

$$P(\{\exists n \in \mathbb{N} : S_n \geq \frac{1}{\alpha}\}) \leq \alpha$$

for any $P \in \mathcal{H}_0$.

Ville's inequality [Ville, 1939] can be thought of as a generalization of Markov's inequality to non-negative supermartingales. It immediately implies that the *greedy stopping time*:

$$\tau_\alpha := \min\{n \in \mathbb{N} : S_n \geq \frac{1}{\alpha}\} \tag{2.4}$$

is a level-α sequential test for $\mathcal{H}_0$. This holds because:

$$\{\tau_\alpha < \infty\} = \{\exists n \in \mathbb{N} : S_n \geq \frac{1}{\alpha}\},$$

and thus, the probability of ever rejecting the null with τ_α, at any finite time n, is at most α under the null hypothesis.

As we have seen in the p-hacking by re-testing Example 1.1.3, general sequences of p-variables do not offer type-1 error guarantees of this kind when used sequentially. Ville's inequality (when it holds) is precisely the reason why one can not break safety guarantees of an e-variable based test, or equivalently, a test based on thresholding a p-variable whose inverse is an e-variable, using this form of p-hacking.

So far, we have described e-variables for testing composite nulls against simple alternatives in fixed-sample and sequential settings. The two main methods for extending simple alternative, e-variable-based, tests to composite alternatives are the *prequential plug-in* method and the method of *mixtures* [Ramdas et al., 2023, Grünwald, 2007].
Let us discuss the prequential plug-in method first. Assume that for each fixed distribution $Q \in \mathcal{H}_{1|1}$, we know how to find a "good" e-variable S_Q for testing $\mathcal{H}_{0|1}$ against $\{Q\}$, e.g. the GRO e-variable. Furthermore, given n data points D^n, we have a way of estimating the most plausible alternative distribution $\hat{Q}_n(D^n) \in \mathcal{H}_{1|1}$. This could be done, for example, by maximum likelihood estimation. Then, as long as there are no measurability issues (which can be avoided under mild additional assumptions), the stochastic process defined recursively by:

$$S_n(D^n) := S_{n-1}(D^{n-1}) \cdot S_{\hat{Q}_{n-1}(D^{n-1})}(D_n), \tag{2.5}$$

$$S_1(D_1) := S_{Q_0}(D_1) \tag{2.6}$$

is a test supermartingale for $\mathcal{H}_{0|\infty}$. Q_0 in the definition represents some fixed starting distribution. Thus, the prequential plug-in method yields a simple modification of the sequential GRO e-variable, which allows the process to learn the most likely true alternative distribution over time.

The method of mixtures, on the other hand, works by fixing a Bayesian prior distribution $\pi(\theta)$ over the the parameter space Θ of the (parametric) alternative hypothesis, and using a mixture of e-variables S_θ for $\mathcal{H}_{0|n}$:

$$S_\pi(D^n) := \int_\Theta S_\theta(D^n)\pi(\theta)d\theta \tag{2.7}$$

A common choice for S_θ is the GRO e-variable for $\mathcal{H}_{0|n}$ against $\{Q_\theta^n\}$.
The method of mixtures can be used for fixed-sample or sequential testing alike. Note that if $(S_{\theta,n}(D^n))_{n\in\mathbb{N}}$ is a test supermartingale for each θ, then under mild regularity conditions:

$$(\int_\Theta S_{n,\theta}(D^n)\pi(\theta)d\theta)_{n\in\mathbb{N}}$$

is also a test supermartingale [Shafer et al., 2011].

The described methods are not as different as they might seem at first sight. For example, consider testing a point null hypothesis $\{p\}$ against a composite alternative $\{q_\theta : \theta \in \Theta\}$, both represented by positive densities w.r.t some common measure, and extended to n i.i.d. data points as usual. The GRO e-variable for testing against q_θ is then given by the simple likelihood ratio:

$$S_{n,\theta}(D^n) = \frac{q_\theta(D^n)}{p(D^n)} = \prod_{i=1}^n \frac{q_\theta(D_i)}{p(D_i)},$$

and the method of mixtures suggests using the e-variable

$$\int_\Theta \prod_{i=1}^n \frac{q_\theta(D_i)}{p(D_i)}\pi(\theta)d\theta = \frac{\int_\Theta \prod_{i=1}^n q_\theta(D_i)\pi(\theta)d\theta}{\prod_{i=1}^n p(D_i)} = \prod_{i=1}^n \frac{\hat{q}_i(D_i)}{p(D_i)}, \tag{2.8}$$

where the density $\hat{q}_i$ depends on D^{i-1} and is given by the Bayesian predictive distribution

$$\hat{q}_i(D_i) = \int_\Theta q_\theta(D_i)\pi(\theta|D^{i-1})d\theta,$$

with the Bayesian posterior density

$$\pi(\theta|D^{i-1}) := \frac{q_\theta(D^{i-1})\pi(\theta)}{\int_\Theta q_\theta(D^{i-1})\pi(\theta)d\theta}.$$

From the right-hand side of (2.8), it is clear that, in this setting, the method of mixtures is merely a mild extension of of the prequential plug-in method, where the rule for

choosing new alternatives $\hat{Q}$ in each step is determined by the prior π. The resulting $\hat{Q}$'s being posterior predictive distributions, they then lie in the convex hull of $\mathcal{H}_{1|1}$, whereas in the original formulation they were restricted to lie in $\mathcal{H}_{1|1}$. In Section 4.3 we show that sometimes it is possible to use a combination of the two methods which inherits useful properties from both.

As a final note, instead of using tricks to extend point alternative tests to composite alternatives, one could, more fundamentally, come up with an entirely new definition of optimal e-variables in this setting. The GROW and REGROW optimality criteria have been proposed by Grünwald et al. [2024] to do just that. We will get back to them in Chapter 4, where we describe e-variables for linear regression testing with composite alternatives.

2.3 Group theory

Group theory plays a central role in the design of e-variables for linear regression testing with composite alternatives. We devote this small section to recapping some fundamental definitions. The section is not needed for understanding Chapter 3 (which contains most of our contributions), and may be skipped if the reader is primarily interested in point alternatives.

Definition 2.3.1. (Group)
Let G be non-empty set, and $\cdot$ a binary operation on G. We say that $(G, \cdot)$ is a group if:

- $g \cdot h \in G \ \forall g, h \in G$ (closure)
- $g \cdot (h \cdot l) = (g \cdot h) \cdot l \ \forall g, h, l \in G$ (associativity)
- $\exists e \in G \ \forall g \in G \ eg = ge = g$ (neutral element)
- $\forall g \in G \ \exists g^{-1} \in G \ gg^{-1} = g^{-1}g = e$ (inverses).

In statistics, groups are used to represent transformations which act on the data and distributions of the data.
For example, imagine a one sample t-test setting with n data points, and a given effect size δ. Under the null hypothesis the data points are i.i.d. Gaussians with mean 0 and an unknown variance. Under the alternative, the mean is instead proportional to the standard deviation, with a proportionality constant δ. Precisely, the hypotheses are:

$$\mathcal{H}_{0|n} = \{\mathcal{N}(0, \sigma^2 I_n) : \sigma^2 > 0\},$$
$$\mathcal{H}_{1|n} = \{\mathcal{N}(\sigma\delta 1_n, \sigma^2 I_n) : \sigma^2 > 0\}.$$

The set $(0, \infty)$ with standard multiplication forms the so called *scale group* [Lehmann and Romano, 2005]. Informally, the alternative hypothesis can be viewed as the set

of distributions that we get by "acting" on the distribution $\mathcal{N}(\delta 1_n, I_n)$ with the scale group:

$$\sigma \cdot \mathcal{N}(\delta 1_n, I_n) = \mathcal{N}(\sigma \delta 1_n, \sigma^2 I_n).$$

Similarly, we get the null by acting on $\mathcal{N}(0, I_n)$. This is an example of a testing problem where the null and alternative hypotheses can each be represented as an "orbit" [Lehmann and Romano, 2005] of a single distribution. Under regularity assumptions, Pérez-Ortiz et al. [2024] have shown how to find GROW and REGROW (see Chapter 4) e-variables for testing problems of this kind, which, as already indicated by the t-test example, frequently occur in practice.
Let us now formalize the meaning of group actions and orbits.

Definition 2.3.2. (Group action on data)
The *action* of group $(G, \cdot)$ on the data is a function $A : G \times \mathcal{D} \to \mathcal{D}$ which satisfies:

$$A(e, d) = d,$$
$$A(h, A(g, d)) = A(h \cdot g, d)$$

for all $g, h \in G$, and $d \in \mathcal{D}$. To simplify notation, instead of $A(g, d)$ we use $g \cdot d$.

The requirements on the action ensure that it is well aligned with the group operation.

Definition 2.3.3. (Data orbits)
The *orbit* of an element $d \in \mathcal{D}$ is defined as the set:

$$O(d) = \{g \cdot d : g \in G\}.$$

Group actions and orbits can be defined analogously for distributions of the data:

$$g \cdot P(D) := P(g \cdot D),$$
$$O(P(Y)) := \{P(g \cdot D) : g \in G\}.$$

This is exactly how the scale group acted on Gaussian distributions in the previous example.

It is useful to notice that a group action generates an equivalence relation where each orbit represents an equivalence class. Furthermore, the group action over the data immediately induces a group action over any observed statistic $S(d)$ of the data:

$$g \cdot S(d) := S(g \cdot d).$$

Invariant statistics provide the natural gateway through which group-theoretic ideas enter statistical inference and hypothesis testing.

Definition 2.3.4. (Maximally invariant statistic)
Statistic S is invariant w.r.t. group G if for any $d \in \mathcal{D}, g \in G$:

$$g \cdot S(d) = S(d).$$

S is a maximally invariant statistic if for any data points $d, \tilde{d}$:

$$S(d) = S(\tilde{d}) \iff \tilde{d} \in O(d).$$

A simple, yet very useful consequence of the definition is that the distribution of any G-invariant statistic $S(D)$ is the same under any two distributions P and Q of the data, which belong to the same distribution orbit:

$$P(S(D)) = Q(S(g \cdot D)) = Q(S(D)),$$

for some $g \in G$. Specifically, it means that the probability-density function of an invariant statistic depends only on the distribution orbit of the data-generating distribution. This fact is one of the basic ideas behind the results of Pérez-Ortiz et al. [2024], and proves to be very useful for the development of GROW and REGROW e-variables for testing $\mathcal{H}_0$ against $\mathcal{H}_1$, when the hypotheses are two distribution orbits of the same group. As we will see in Chapter 4, linear regression testing almost fits this description.

2.4 Markov kernels and Model-X

Throughout this thesis, any set defined as:

$$\{P(D) : \text{some restriction}\}$$

should be understood as the set of all probability distributions over the data space $(\mathcal{D}, \mathcal{F}^{\mathcal{D}})$ which satisfy the restriction.
We use the convenient notation $P(A, B)$ for the joint probability distribution of random variables (or vectors) A and B, and $P(A)$ for the marginal distribution of A. The underlying assumption then is that $\mathcal{D} = \mathcal{A} \times \mathcal{B}$ is a product space with the product σ-algebra, A taking values in $\mathcal{A}$, and B in $\mathcal{B}$. The object $P(A|B)$ is a Markov kernel, i.e. a regular conditional probability, which is well defined for any joint probability $P(A, B)$ due to the standard Borel space assumption on the data space. We use $\otimes$ for the usual product operation on Markov kernels [Forré and Mooij, 2024], so specifically:

$$P(A, B) = P(A|B) \otimes P(B)$$

holds. If we have two probabilities $P(A, B) \ll Q(A, B)$, then:

$$\frac{dP(A, B)}{dQ(A, B)}(a, b) \tag{2.9}$$

is the Radon-Nikodym derivative of $P(A, B)$ and $Q(A, B)$ evaluated at a fixed point (a, b). If the derivative (2.9) is evaluated at a random point (A, B), then we simply use:

$$\frac{dP(A, B)}{dQ(A, B)}(A, B).$$

Finally, whenever $P(A|B) \ll Q(A|B)$,

$$\frac{dP(A|B)}{dQ(A|B)}$$

denotes the Doob-Radon-Nikodym derivative of the two kernels [Forré and Mooij, 2024]. With this notation set, we are ready to delve into the Model-X (MX) framework.

Given data from three random vectors Y, X and Z, testing the conditional independence null:

$$\mathcal{H}_0 := \{P(Y, X, Z) : Y \perp_P X|Z\},$$

or the corresponding n i.i.d. data point version, turns out to be a very hard problem. In a famous result, Shah and Peters [2020] prove that if a null hypothesis contains all distributions which satisfy the conditional independence requirement and admit a Lebesgue density, then any level-α test for that null has at most α power under all absolutely continuous alternatives. Therefore, in order to test conditional independence effectively, strong assumptions are necessary.
Making the Model-X assumption [Candès et al., 2018, Grünwald et al., 2023] means assuming:

- i.i.d. data
- knowledge of the true conditional distribution of X given Z, which we denote $K(X|Z)$, and call the *propensity kernel*.

The classical example of this setting is a randomized controlled trial (RCT). The variable Y could represent a target variable, measured for each participant, which the researchers hope to influence using a treatment. X encodes whether a participant is given the real treatment or a placebo, and Z could be a medical history of the participant. In this case, a common propensity kernel is a simple Bernoulli distribution $K(X|Z) = K(X) \sim B(0.5)$, meaning that participants are divided into the experimental and control group by a fair coin toss. Testing $Y \perp X|Z$ is then equivalent to testing whether the treatment has an effect on the target variable, after controlling the influence of the participant's medical histories.

As shown by Grünwald et al. [2023], the MX assumption allows us to create useful e-variables for testing the conditional independence $Y \perp X|Z$. Here, we restate their results in the language of Markov kernels, which we keep using in the coming sections. Formally, the Model-X null hypothesis for n data points $D^n = (Y^n, X^n, Z^n)$ with kernel $K(X|Z)$ is given by

$$\mathcal{H}^{\mathrm{MX}}_{0|n,K(X|Z)} = \{P(D^n) : D_i \text{ are i.i.d. under } P, P(Y, X, Z) = K(X|Z) \otimes P(Y, Z)\} \quad (2.10)$$

The notation on the left hand side will be motivated in the beginning of Chapter 3, where it will arise as a special case of a general notational convention.
Let $Q(Y, X, Z)$ be any probability distribution over a single data point which satisfies the MX assumption, i.e. comes from

$$\mathcal{H}^{\mathrm{MX}}_{1|n,K(X|Z)} = \{P(D^n) : D_i \text{ are i.i.d. under } P, \\ P(Y, X, Z) = P(Y|X, Z) \otimes K(X|Z) \otimes P(Z)\}. \quad (2.11)$$

Then, as shown by Grünwald et al. [2023], the GRO e-variable for testing $\mathcal{H}^{\mathrm{MX}}_{0|1,K(X|Z)}$ against $\{\mathrm{Q(Y,X,Z)}\}$ is given by:

$$S_1^{\mathrm{MX}} = \frac{dQ(Y,X,Z)}{d(K(X|Z) \otimes Q(Y,Z))} = \frac{dQ(X|Y,Z)}{dK(X|Z)}. \tag{2.12}$$

This object is well-defined because $Q(Y,X,Z) \ll Q(X|Z) \otimes Q(Y,Z)$ holds for any distribution Q. From the rightmost expression it is easy to show the e-variable property; let P be an element of the null. Then:

$$\begin{aligned}
\mathbb{E}_{P(Y,X,Z)}[S_1^{\mathrm{MX}}] &= \mathbb{E}_{P(Y,Z)}[\mathbb{E}_{K(X|Z)}[S_1^{\mathrm{MX}}]] \\
&= \mathbb{E}_{P(Y,Z)}\left[\int_x \frac{dQ(X|Y,Z)}{dK(X|Z)}(Y,x,Z)dK(X|Z=z)(x)\right] \\
&= \mathbb{E}_{(y,z)\sim P(Y,Z)}\left[\int_x dQ(X|Y=y,Z=z)(x)\right] \\
&= \mathbb{E}_{P(Y,Z)}[1] = 1.
\end{aligned}$$

In fact, we have shown something even stronger, namely that S_1^{MX} is a (Y,Z)-*conditional* e-variable:

Definition 2.4.1. (Conditional and strong conditional e-variables)
We say that $S(A,B)$ is a B-conditional e-variable for a null hypothesis $\mathcal{H}_0$ over the data $D = (A,B)$, if for every $P \in \mathcal{H}_0$:

$$\mathbb{E}_P[S(A,B)|B] \leq 1,$$

$P(B)$-a.s. Equivalently, for any $P \in \mathcal{H}_0$, there exists a measurable set $\mathcal{B}' \subseteq \mathcal{B}$ (which may depend on P) such that $P(B \in \mathcal{B}') = 1$ and

$$\mathbb{E}_{P(A|B=b)}[S(A,b)] \leq 1 \ \forall b \in \mathcal{B}'.$$

$S(A,B)$ is a strong B-conditional e-variable if:

$$\mathbb{E}_{P(A|B=b)}[S(A,b)] \leq 1 \ \forall b \in \mathcal{B}, P \in \mathcal{H}_0.$$

Clearly, any strong conditional e-variable is a conditional e-variable, and any conditional e-variable is an ordinary e-variable for the same null. The distinction between strong conditional and conditional e-variables will be important for extending fixed-design results to random design in Chapter 3.

Now, let us get back to S_1^{MX}. Its GRO property follows directly from Theorem 2.2.5. Furthermore, the GRO e-variable for $\mathcal{H}^{\mathrm{MX}}_{0|n,K(X|Z)}$ against the n-fold product distribution $\{Q(Y,X,Z)^n\}$ is simply the product e-variable:

$$S_n^{\mathrm{MX}}(Y^n,X^n,Z^n) = \prod_{i=1}^n S_1^{\mathrm{MX}}(Y_i,X_i,Z_i).$$

The last claim follows from Proposition 2.2.9, as $Q \ll \mathcal{H}_0^{\mathrm{MX}}$ and $Q^n \ll \mathcal{H}_{0|n}^{\mathrm{MX}}$ clearly hold since $Q(Y, X, Z) \ll K(X|Z) \otimes Q(Y, Z)$, and the same is true for their n-fold products.

Finally, the e-power of S_1^{MX} is given by a well known information-theoretic measure of conditional dependence: the *(conditional) mutual information* of Y and X given Z, denoted by $I(Y; X|Z)$ [Cover and Thomas, 1991].

$$\begin{aligned}\mathbb{E}_Q\left[\ln S_1^{\mathrm{MX}}\right] &= \mathbb{E}_Q\left[\ln \frac{dQ(Y, X, Z)}{d(K(X|Z) \otimes Q(Y, Z))}\right] \\ &= D_{\mathrm{KL}}(Q(Y, X, Z)||Q(X|Z) \otimes Q(Y, Z)) \\ &= I(Y; X|Z).\end{aligned}$$

The e-power of S_n^{MX} is then clearly $nI(Y; X|Z)$.

2.5 Classical LR testing

In this section we define the linear regression testing problem and analyze it from the fixed and random design perspectives. We end the section with a description of several important results in classical LR testing theory.

In general regression problems, the goal is to predict a target variable Y using several predictor variables summarized in a vector W. It is common to measure the approximation error using the L^2 norm. In that case, assuming (Y, W) have well-defined second moments, the conditional expectation:

$$\mathbb{E}[Y|W]$$

is the best predictor of Y among all measurable functions of W. More precisely:

$$\mathbb{E}[(Y - \mathbb{E}[Y|W])^2] \leq \mathbb{E}[(Y - f(W))^2]$$

for any measurable function f, with equality if and only if $\mathbb{E}[Y|W] = f(W)$ a.s. In practice, with finite amounts of data, searching for a good approximation of the conditional expectation among all measurable functions is a difficult task. Thus, it is often simplified by assuming a linear relation between the target and the predictors:

$$\mathbb{E}[Y|W] = \eta^T W$$

for some vector of unknown coefficients $\eta \in \mathbb{R}^{d_w}$. Here d_w is the dimensionality of W. Finding the optimal coefficient is now a very simple convex optimization task:

$$\begin{aligned}\frac{d}{d\eta}\mathbb{E}[(Y - \eta^T W)^2] = 0 &\iff \mathbb{E}[(Y - \eta^T W)W^T] = 0 \\ &\iff \mathbb{E}[WY] = \mathbb{E}[WW^T]\eta.\end{aligned}$$

The optimizer is unique whenever $\mathbb{E}[WW^T]$ is of full rank. This is the case if and only if the vector of covariates is a linearly independent random vector. In other words, when the following implication holds:

$$a^T W = 0 \text{ a.s. } \Rightarrow a = 0.$$

Of course, this analysis was purely theoretical, as in practice one does not have access to the true values $\mathbb{E}[WW^T]$ and $\mathbb{E}[WY]$. Instead, a practitioner might have a sample $(Y^n, W^n) \in \mathbb{R}^{n\times(1+d_w)}$ of n i.i.d. data points available. The most commonly used estimator for η is the Ordinary Least Squares (OLS) estimator:

$$\begin{aligned}\hat{\eta}_{\text{OLS}}(Y^n, W^n) &:= (W^{n^T} W^n)^{-1} W^{n^T} Y^n \\ &= \left(\frac{1}{n}\sum_{i=1}^n W_i W_i^T\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^n W_i Y_i\right).\end{aligned}$$

By the law of large numbers, the OLS estimator converges almost surely to the theoretically optimal η that we previously found, whenever (Y, W) has a finite second moment and W is linearly independent.

So far in our analysis, we have modeled the covariates as random objects. This is called a random design setting. While developing statistical tests for regression problems, or proving theorems about estimators, it is often useful to temporarily assume a fixed-design, i.e. forget about the randomness of the covariates and consider them given, deterministic values. One may think of fixed design models as representing their random design counterparts after conditioning on the event $\{W^n = w^n\}$. The fixed matrix of covariate values w^n is called the *design matrix*. As we will see, the conditional interpretation makes it easy to generalize optimality properties of e-variables from fixed to random design settings. Therefore, all theorems in this section, along with the definition of the linear regression testing problem will be given assuming a fixed design.

We first review the celebrated Gauss-Markov theorem, which shows that the OLS estimator is optimal in terms of variance minimization, within the class of linear and unbiased estimators of η.

Theorem 2.5.1. (Gauss-Markov BLUE)
Assume the data satisfy the model:

$$Y^n = w^n \eta + \varepsilon^n$$

with some random error term ε^n and fixed design matrix w^n. If the Gauss-Markov assumptions:

- $\mathbb{E}[\varepsilon_i] = 0$ for all i
- $Var(\varepsilon_i) = \sigma^2$ for all i

- $\mathrm{Cov}(\varepsilon_i, \varepsilon_j) = 0$ for $i \neq j$

hold and w^n has full column rank, then the OLS estimator:

$$\hat{\eta}_{\mathrm{OLS}} = ({w^n}^T w^n)^{-1} {w^n}^T Y^n$$

is, the Best Linear Unbiased Estimator (BLUE) of η.
In other words, it minimizes $\mathrm{Cov}(\hat{\eta})$, in Loewner order among all estimators $\hat{\eta}$ of η which are measurable functions of (w^n, Y^n), linear in Y^n and unbiased.

The OLS is a clear favorite among linear unbiased estimators, but if we drop one of those assumptions, there do exist estimators which are arguably better. For example, *shrinkage* estimators such as ridge regression and the James Stein estimator [Berger, 2013] introduce bias, but can substantially decrease the variance, leading to a more favorable bias/variance trade-off.
Nevertheless, the OLS estimator has another useful feature. If we assume the model

$$Y^n = w^n \eta + \varepsilon^n$$

holds with an error vector ε^n consisting of i.i.d. 0-mean Gaussians, then calculating the distribution of $\hat{\eta}_{\mathrm{OLS}}$ becomes feasible. The whole distribution of an estimator is more informative than merely having a point estimate, and knowing it often represents the first step in the process of designing a statistical test.
Assuming Gaussian errors is, thus, mathematically very convenient. A common real-world justification is that, whenever the error term can be represented as a sum of smaller independent influences, by the Central Limit Theorem the error distribution tends to a Gaussian. In general, this assumption can lead to completely inadequate models and catastrophic mistakes. A well known example is the 2008 financial crisis. Gaussians were used to model the correlation structure between mortgage defaults, and lead to a great underestimation of simultaneous default probability, contributing to a global disaster. On the other hand, many common statistical tests which assume Gaussian errors, such as the ANOVA, proved to be robust in settings with misspecified error distributions [Glass et al., 1972].

We now move from estimation to testing, where we keep the Gaussian error assumption. The goal of LR testing is to assess whether $\beta = 0$ holds in the linear model

$$Y = \beta^T X + \gamma^T Z + \varepsilon, \varepsilon \sim \mathcal{N}(0, \sigma^2).$$

If $\beta = 0$, it simply means that X is not a useful linear predictor of Y on top of the predictor Z. The statistician usually has n i.i.d. data points at her disposal, so the model takes the form

$$Y^n = X^n \beta + Z^n \gamma + \varepsilon^n, \varepsilon^n \sim \mathcal{N}(0, \sigma^2 I_n).$$

We give an exact formulation of the random-design null and alternative LR hypotheses in Chapter 3 (3.1). But for now, let us go back to fixed design, and see what the classical

theory has to offer.

The hypotheses in a (fixed-design) linear regression test are collections of probability distributions of the random vector Y^n. They are given by

$$\mathcal{H}_{0|z^n} := \{\mathcal{N}(z^n\gamma, \sigma^2 I_n) : \gamma \in \mathbb{R}^{d_z}, \sigma^2 > 0\}, \tag{2.13}$$

$$\mathcal{H}_{1|x^n,z^n} := \{\mathcal{N}(x^n\beta + z^n\gamma, \sigma^2 I_n) : \beta \in \mathbb{R}^{d_x}, \gamma \in \mathbb{R}^{d_z}, \sigma^2 > 0\}, \tag{2.14}$$

with $w^n = (x^n, z^n)$ a predefined design matrix with full column rank. The full rank assumption is important because it ensures that every distribution is uniquely identified by the parameters $(\beta, \gamma, \sigma^2)$.
We call β the parameter of interest, γ, σ^2 nuisance parameters, and $\delta := \frac{\beta}{\sigma}$ the effect size vector. Sometimes we use $\eta := [\beta^T, \gamma^T]^T$ for the concatenation of β and γ. Finally, $[\{a^n\}]$ will represent the linear subspace of $\mathbb{R}^n$ spanned by the columns of the $n \times d_a$ matrix a^n.
The null hypothesis is a restriction of the alternative, where $\beta = 0$ holds. The testing problem at hand is merely the question of whether the mean-value parameter of a multivariate Gaussian lies in $[\{z^n\}]$ or in the larger space $[\{(x^n, z^n)\}]$.
This may seem like a very restricted setting, but many widely used statistical tests, such as various versions of the analysis of variance, analysis of covariance, and correlation tests can be formulated as linear regression tests of this kind, for specific choices of the design matrix.

Proposition 2.5.2. (Estimator distributions)
Assume the data Y^n is generated by the distribution from $\mathcal{H}^{\mathrm{LR}}_{1|n,x^n,z^n}$ with full rank (x^n, z^n), and parameters $(\beta, \gamma, \sigma^2)$. Then:

$$\hat{\eta}_{\mathrm{OLS}} \sim \mathcal{N}(\eta, \sigma^2 (w^{n^T} w^n)^{-1}),$$

$$\hat{\beta}_{\mathrm{OLS}} \sim \mathcal{N}(\beta, \sigma^2 (x^{n^T} (I_n - P_{z^n}) x^n)^{-1}),$$

$$\hat{\sigma}^2_{\mathrm{OLS}} = \frac{||(I_n - P_{w^n})Y^n||^2}{n - d_w} \sim \frac{\sigma^2}{n - d_w}\chi^2(n - d_w),$$

where P_{a^n} is the orthogonal projection matrix onto $[\{a^n\}]$, and $\hat{\beta}_{\mathrm{OLS}}$ simply represents the coordinates of $\hat{\eta}_{\mathrm{OLS}}$ which belong to β.

The proof amounts to carefully applying linear transformations to Gaussians, and using block-matrix inversion results. It can be found in Agresti [2015]. The previous proposition is the workhorse of the classical linear regression test (the F-test):

Corollary 2.5.3. (F-test for linear regression)
Let $F(a, b, c)$ denote the non-central F-distribution with degrees of freedom (a, b) and

non-centrality parameter c.
Under the same assumptions as the previous proposition:

$$\hat{F} := \frac{||(I_n - P_{z^n})x^n\hat{\beta}_{\text{OLS}}||^2}{d_x\hat{\sigma}^2_{\text{OLS}}} \sim F(d_x, n - d_w, \delta^T {x^n}^T (I_n - P_{z^n})x^n\delta).$$

If F_{cdf} is the cumulative distribution function of the (central) F-distribution $F(d_x, n - d_w, 0)$, then:

$$U_{LR} := F_{\text{cdf}}(\hat{F})$$

is a p-variable for $\mathcal{H}_{0|z^n}$.

There are other, conceptually distinct paths which also lead to the classical linear regression test described in Corollary 2.5.3. For instance, the resulting test coincides with the *likelihood ratio test* for this problem [Agresti, 2015], and it can also be derived from a group-theoretic perspective, by seeking tests that satisfy suitable invariance properties [Lehmann and Romano, 2005].

It is not hard to see that U_{LR} is an *exact* p-variable, i.e. it has a uniform distribution on $[0, 1]$ under all data-distributions in the null hypothesis. The inverse U_{LR}^{-1} is clearly not an e-variable for the null, meaning that p-hacking of the forms described in Chapter 1 leads to a loss of safety guarantees. In the following chapter, we describe and compare two e-variables for LR testing.

3 Simple alternatives

In this chapter, we focus on LR testing of composite nulls against simple (point) alternatives $\{Q\}$. We call these *oracle tests*, because having a simple alternative is like gaining information from an oracle that tells you: "If the true data-generating distribution is not in the null, then it must be exactly Q." This setting is artificial and in itself not very useful in practice. But as we have seen in Section 2.2, with e-variables extending oracle tests (for composite nulls) to composite vs composite tests is usually much easier than designing the oracle test in the first place. It can simply be done by mixing over e-variables for each point alternative or using the prequential plug-in method. We now address the real difficulty!

Outline:
In Sections 3.1 and 3.2 we introduce S-LR and MX e-variables for LR testing and discuss the practical importance of testing the intersection of their null hypotheses, illustrated with examples in Section 3.3. In Section 3.4 we further describe the difficulty of finding the GRO e-variable for testing this intersection. Our main theoretical contributions are Theorems 3.5.6 and 3.5.8 in Section 3.5, which compare the performance of S-LR and MX e-variables in terms of asymptotic relative power and clarify their relationship in a special setting that is particularly relevant for randomized controlled trials.

Conventions:
Throughout the chapter, $\mathcal{H}_1$ will never formally represent the simple alternative hypothesis for which we optimize e-variables. Instead, it will be a set of probability distributions from which we choose the point alternatives $\{Q\}$. $\mathcal{H}_1$ will often be an extension of the null hypothesis, allowing additional parameters or dropping restrictions, so we call it the *extended model*.
Nevertheless, we do occasionally address $\mathcal{H}_0$ and $\mathcal{H}_1$ together as "the hypotheses", but only in contexts where this may not lead to confusion.
In order to distinguish the many different collections of probability distributions that we talk about, we introduce the following notation:

$$\mathcal{H}^{\text{name}}_{i|n,K}.$$

Index $i \in \{0, 1\}$ tells us whether the collection represents the null hypothesis or the extended model, respectively. n is the number of data points, K may be specified or not, and it represents either a propensity kernel $K(X|Z)$ or a joint distribution $K(X, Z)$ over the covariates. The "name" will be one of "LR", "MX","SPLR" or "FSLR" and the exact meaning of each will be explained in due time.

As a further addition to this rule, when working in fixed design with a given design matrix $w^n = (x^n, z^n)$ we use the notation from the previous chapter: $\mathcal{H}_{0|z^n}$ and $\mathcal{H}_{1|x^n,z^n}$.

3.1 Simple LR test

The focus of this section is on designing an e-variable-based oracle test for linear regression specifically.
The Simple Linear Regression (S-LR) test derives its name from the fact that the fixed-design version of the testing problem falls within the simple case (see Definition 2.2.6). As we have seen with the t-test, this makes the optimization problem for finding the GRO e-variable very simple and analytically solvable. GRO e-variables rarely have this form, but sufficient conditions have been established for exponential family nulls [Grünwald et al., 2025].

3.1.1 Fixed design

We start this section by assuming a fixed design. For the time being, we will treat the sample size n and covariate matrices (x^n, z^n) as fixed objects. Our hypotheses of interest are as in Chapter 2 (2.13).
We will, as we did in the classical setting, assume that $w^n := (x^n, z^n)$ has full column rank while presenting initial results. The full rank assumption is useful as it allows us to encode the hypotheses as parametric families without ambiguity, but as we will see is not necessary for finding the GRO e-variable.

Definition 3.1.1. (The S-LR e-variable)
For a given alternative distribution $Q \in \mathcal{H}_{1|x^n,z^n}$ with density $q(\cdot|x^n, z^n)$, we define

$$S_n^{\text{S-LR}}(Y^n) := \frac{q(Y^n|x^n, z^n)}{p^*(Y^n|z^n)},$$

where p^* is the reverse information projection of $q(\cdot|x^n, z^n)$ onto $\mathcal{H}_{0|z^n}$.

$S_n^{\text{S-LR}}$ is, by construction the GRO e-variable for testing $\mathcal{H}_{0|z^n}$ against $\{Q\}$. Grünwald et al. [2025] show that fixed-design linear regression testing belongs to the simple case. In other words, $p^*(\cdot|z^n)$ is the density function of an element of the null $\mathcal{H}_{0|z^n}$ itself. Knowing this allows us to explicitly find the RIPr, GRO e-variable, and to calculate the e-power (see Section 2.2).

Proposition 3.1.2. (S-LR - finding RIPr, GRO e-var and e-power)
Assume (x^n, z^n) has full column rank. Let the density $q(\cdot|x^n, z^n)$ be given by parameters $(\beta_1, \gamma_1, \sigma_1^2)$. Then, the RIPr is an element of the null, and its density is

$p^* = \mathcal{N}(y^n | z^n \gamma_0, \sigma_0^2 I_n)$ with parameters:

$$\gamma_0 = \gamma_1 + (z^{n^T} z^n)^{-1} z^{n^T} x^n \beta_1,$$
$$\sigma_0^2 = \sigma_1^2 + \frac{||(I_n - P_{z^n}) x^n \beta_1||^2}{n}.$$

The GRO e-variable is given by:

$$S_n^{\text{S-LR}} = \frac{\mathcal{N}(Y^n | x^n \beta_1 + z^n \gamma_1, \sigma_1^2 I_n)}{\mathcal{N}(Y^n | z^n \gamma_0, \sigma_0^2 I_n)},$$

with e-power:

$$\mathbb{E}_Q[\ln S_n^{\text{S-LR}}] = \frac{n}{2} \ln(1 + \delta_1^T \frac{x^{n^T} (I_n - P_{z^n}) x^n}{n} \delta_1).$$

Proof. The design matrix is of full rank, so the RIPr is an element of the null, as shown by Grünwald et al. [2025]. To find the KL minimizer among the distributions in the null we use the KL divergence formula for general Gaussians (Chapter 2, (2.1)), which yields

$$\frac{2}{n} D_{\text{KL}}(\mathcal{N}(x^n \beta_1 + z^n \gamma_1, \sigma_1^2 I_n) || \mathcal{N}(z^n \gamma_0, \sigma_0^2 I_n)) = \ln \left(\frac{\sigma_0^2}{\sigma_1^2} \right) + \frac{\sigma_1^2 + \frac{||x^n \beta_1 + z^n \gamma_1 - z^n \gamma_0||^2}{n}}{\sigma_0^2} - 1.$$

Minimizing the KL can be done in two simple steps. First minimizing over γ_0 using ordinary least squares and then minimizing the remaining convex function in σ_0^2. The expressions for the minimizers and e-power follow immediately.

□

In general, the GRO e-variable for testing a null $\mathcal{H}_0$ against an alternative $\{Q\}$ depends only on the distributions contained in the hypotheses, not on the choice of parameterization (if there is one). This allows us to easily generalize the previous proposition to design matrices which do not have a full column rank. We now outline how to do this. Suppose then (x^n, z^n) has linearly dependent columns. We need to perform a *backward basis selection procedure*. This amounts to removing every column from w^n which can be represented as a linear combination of columns on its right-hand side. The elimination also includes the rightmost column in case it is a zero vector. We get a new design matrix $\tilde{w}^n = (\tilde{x}^n, \tilde{z}^n)$ which has full rank, and the property that $[\{z^n\}] = [\{\tilde{z}^n\}]$ and $[\{w^n\}] = [\{\tilde{w}^n\}]$. This specifically means that:

$$\mathcal{H}_{1|x^n, z^n} = \mathcal{H}_{1|\tilde{x}^n, \tilde{z}^n}; \ \mathcal{H}_{0|z^n} = \mathcal{H}_{0|\tilde{z}^n}.$$

Thus, we have merely simplified the parameterization of the hypotheses by removing unnecessary parameters, but did not change the hypotheses themselves.
We can find the GRO e-variable for testing $\mathcal{H}_{0|z^n}$ against any $\{Q\} \subset \mathcal{H}_{1|x^n, z^n}$ simply by using Proposition 3.1.2 on the newly parameterized hypotheses.
It is in fact possible to avoid this procedure and find the GRO e-variable in a single step using the following corollary.

Corollary 3.1.3. (S-LR without full rank requirement)
The statement of Proposition 3.1.2 holds for arbitrary design matrices with the adapted formulas:

$$\begin{aligned}\gamma_0 &= \gamma_1 + (z^n)^+ x^n \beta_1,\\ P_{z^n} &= z^n (z^n)^+,\\ \sigma_0^2 &= \frac{||(I_n - P_{z^n}) x^n \beta_1||^2}{n},\end{aligned}$$

where $(z^n)^+$ is the Moore-Penrose inverse of z^n.

Proof. Note that $z^n(z^n)^+$ is the projection matrix onto the column space of z^n by definition of the Moore-Penrose inverse. This choice of γ_0 ensures that $z^n\gamma_0 = z^n\gamma_1 + P_{z^n}x^n\beta_1$ even when z^n is not of full column rank. Therefore, (γ_0, σ_0^2) is a possibly non-unique minimizer of the KL optimization objective from the proof of Proposition 3.1.2. If we reparameterize the hypotheses using the backward basis selection procedure, the KL minimization objective remains the same, so we know from Proposition 3.1.2 that the minimizer actually must be unique, and that it must be the RIPr. This means that $\mathcal{N}(z^n\gamma_0, \sigma_0^2 I_n)$ is (a representation of) the RIPr. □

Assuming a fixed design greatly simplifies the analysis; the distribution of covariates can be disregarded completely, and restrictions such as "w^n must have full rank" can easily be imposed. However, not all LR tests admit a fixed design version. As we will see in the following section, conditional independence testing with Model-X can be directly repurposed for LR testing. This test relies heavily on the random nature of the covariates and would become completely useless if we were to assume the covariates fixed. Comparing the performance of S-LR and MX will be one of our main tasks, and thus we have to work in the random design framework, where both testing principles can be analysed.

3.1.2 Random design

This subsection is devoted to explaining how we can take our findings in fixed design and transfer them to the random design setting. We consider our covariates random, and we have to adapt our hypotheses accordingly.
Using the notation $D_i = (Y_i, X_i, Z_i)$, they take the following form:

$$\mathcal{H}_{0|n}^{\mathrm{LR}} = \{P(D^n) : P(D^n) = \mathcal{N}(Z^n\gamma, \sigma^2 I_n) \otimes P(X^n, Z^n), \gamma \in \mathbb{R}^{d_z}, \sigma^2 > 0\}, \tag{3.1}$$

$$\mathcal{H}_{1|n}^{\mathrm{LR}} = \{P(D^n) : P(D^n) = \mathcal{N}(X^n\beta + Z^n\gamma, \sigma^2 I_n) \otimes P(X^n, Z^n), \\ (\beta, \gamma) \in \mathbb{R}^{d_x} \times \mathbb{R}^{d_z}, \sigma^2 > 0\}.$$

Any marginal distribution of (X^n, Z^n) is allowed in our hypotheses. The covariates need not be independent over time or identically distributed. But the conditional distribution of $Y^n|X^n, Z^n$ must satisfy the linear model with probability 1 over the covariates. Some questions we need to address at this point are:

- Can we use an analog of the Simple Linear Regression test in this setting?
- Will we get an e-variable for the new null, and if we do, is it the GRO?
- Can we test the null this way against any element of the extended model, or do we have to impose additional restrictions?

As we will see, the S-LR test will carry most of its properties into random design. Before stating the following proposition, let us extend the definition of the $S_n^{\text{S-LR}}$ e-variable.

Definition 3.1.4. (S-LR e-variable for random design)
For a given alternative $Q = \mathcal{N}(X^n\beta_1 + Z^n\gamma_1, \sigma_1^2) \otimes Q(X^n, Z^n)$ the random design S-LR e-variable is:

$$S_n^{\text{S-LR}}(Y^n, X^n, Z^n) := \frac{\mathcal{N}(Y^n | X^n\beta_1 + Z^n\gamma_1, \sigma_1^2 I_n)}{\mathcal{N}(Z^n\gamma_0, \sigma_0^2 I_n)},$$

where:

$$\gamma_0 = \gamma_1 + (Z^n)^+ X^n \beta_1,$$
$$\sigma_0^2 = \sigma_1^2 + \frac{||(I_n - P_{Z^n})X^n\beta_1||^2}{n}.$$

The parameters γ_0, σ_0^2 are random variables now, and we will occasionally stress this fact with notation $\gamma_0(X^n, Z^n), \sigma_0^2(X^n, Z^n)$.

Proposition 3.1.5. (S-LR test for random design)
$S_n^{\text{S-LR}}$ is the GRO e-variable for testing $\mathcal{H}_{0|n}^{\text{LR}}$ against any $\{Q\} \subset \mathcal{H}_{1|n}^{\text{LR}}$.
The RIPr of Q onto $\mathcal{H}_{0|n}^{\text{LR}}$ is the distribution:

$$P^* := \mathcal{N}(Z^n\gamma_0(X^n, Z^n), \sigma_0^2(X^n, Z^n) I_n) \otimes Q(X^n, Z^n).$$

And the e-power is:

$$\mathbb{E}_Q[\ln(S_n^{\text{S-LR}})] = \mathbb{E}_{Q(X^n, Z^n)}\left[\frac{n}{2}\ln(1 + \delta_1^T \frac{X^{n^T}(I_n - P_{Z^n})X^n}{n}\delta_1)\right].$$

Proof. The first thing we need to check is that $S_n^{\text{S-LR}}$ is a non-negative measurable function of the data. If this is not satisfied, it can not possibly be an e-variable. Non-negativity clearly holds. The Borel measurability of the function $A \to A^+$ follows since the Moore–Penrose inverse admits a representation via the singular value decomposition, whose components are measurable. This immediately implies that $\gamma_0(X^n, Z^n)$ and $\sigma_0^2(X^n, Z^n)$ are Borel measurable and thus, so is $S_n^{\text{S-LR}}$.
Secondly, for any fixed value (x^n, z^n) and any $P \in \mathcal{H}_{0|n}^{\text{LR}}$ we get:

$$\mathbb{E}_{P(Y^n | X^n = x^n, Z^n = z^n)}[S_n^{\text{S-LR}}(Y^n, x^n, z^n)] \leq 1 \tag{3.2}$$

by the e-variable property of $S_n^{\text{S-LR}}$ in fixed-design. Therefore, $S_n^{\text{S-LR}}$ is a strong (X^n, Z^n)-conditional e-variable for $\mathcal{H}_{0|n}^{\text{LR}}$ (Chapter 2, (2.4.1)). In fact, every e-variable for $\mathcal{H}_{0|n}^{\text{LR}}$

must be a strong (X^n, Z^n)-conditional e-variable, because it has to satisfy the e-variable property for distributions P with Dirac $\delta_{(x^n,z^n)}$ distributed covariates.
It is easy to see that $Q \ll \mathcal{H}_{0|n}^{\text{LR}}$ holds, simply by picking any distribution in the null with the same covariate distribution as Q. Thus, Theorem 2.2.5 implies the GRO e-variable is the unique e-variable which can be represented as a likelihood ratio between the alternative and an equivalent element in the effective null. We will use this fact to show that $S_n^{\text{S-LR}}$ is (a version of) the GRO e-variable. Define:

$$P^* := \mathcal{N}(Z^n\gamma_0(X^n, Z^n), \sigma_0^2(X^n, Z^n)I_n) \otimes Q(X^n, Z^n)$$

and the σ-finite measure

$$\nu := \lambda^n \otimes Q(X^n, Z^n)$$

Both Q and P^* are absolutely continuous w.r.t ν, and

$$\frac{dQ}{dP*} = \frac{\frac{dQ}{d\nu}}{\frac{dP^*}{d\nu}} = \frac{\mathcal{N}(Y^n|X^n\beta_1 + Z^n\gamma_1, \sigma_1^2 I_n)}{\mathcal{N}(Y^n|Z^n\gamma_0(X^n, Z^n), \sigma_0^2(X^n, Z^n)I_n)} = S_n^{\text{S-LR}}.$$

If we can show that P^* is an element of $\mathcal{H}_{0|n}^{\text{LR}^{\text{eff}}}$, then $S_n^{\text{S-LR}}$ must be the GRO e-variable, and P^* the RIPr.
By definition P^* is an element of the effective null, if all e-variables for $\mathcal{H}_{0|n}^{\text{LR}}$ have expectation at most 1 under P^*, i.e. are also e-variables for $\{P^*\}$. So, let us take some e-variable $S \in \mathcal{E}(\mathcal{H}_{0|n}^{\text{LR}})$.
We know S is a strong (X^n, Z^n)-conditional e-variable for $\mathcal{H}_0^{\text{LR}}$, i.e. satisfies (3.2) for any P in the null and any fixed (x^n, z^n). Therefore, for any choice of $(x^n, z^n, \gamma, \sigma^2)$, the following must hold:

$$f(x^n, z^n, \gamma, \sigma^2) := \int_{\mathbb{R}^n} S(y^n, x^n, z^n)\mathcal{N}(y^n|z^n\gamma, \sigma^2 I_n)d\lambda^n(y^n) \leq 1.$$

Specifically, the inequality also holds when we set $\gamma = \gamma_0(x^n, z^n)$ and $\sigma^2 = \sigma_0^2(x^n, z^n)$. But then

$$\mathbb{E}_{P^*(Y^n|X^n=x^n,Z^n=z^n)}[S] = f(x^n, z^n, \gamma_0(x^n, z^n), \sigma_0^2(x^n, z^n)) \leq 1, \forall x^n, z^n,$$

meaning that S is a strong (X^n, Z^n)-conditional e-variable for $\{P^*\}$.
The e-power formula follows immediately from the e-power in fixed design.

□

It is very important to notice that unlike in fixed design, now the RIPr is not an element of the null, i.e. this is not the simple case. This is not hard to see - in the null, the coefficients γ_0 and σ_0^2 are not allowed to be functions of (X^n, Z^n), they must be constant values. Since we are not in the simple case, we would generally not expect to be able to find analytic expressions for the RIPr and GRO e-variable. In this example it was possible, because we were able to solve the simpler, fixed-design version of the problem

first, and then transfer the conclusions back to random design.

The same procedure could be followed in a more general regression setting to show that a set of fixed design GRO e-variables (one for each design matrix) naturally yields a random design GRO e-variable for the null hypothesis with unrestricted covariate distributions. We leave this for future work.

Another important observation is that $S_n^{\text{S-LR}}$ does not have a product form like the sequential GRO e-variable (see Definition 2.2.8). The process $(S_n^{\text{S-LR}})_{n\in\mathbb{N}}$ in fact does not form a test supermartingale, and does not provide type-1 error guarantees when used sequentially. In other words, the stopping time $\tau := \min\{n \in \mathbb{N} : S_n^{\text{S-LR}} \geq \frac{1}{\alpha}\}$ is not a level-α sequential test for $\mathcal{H}_{0|n}^{\text{LR}}$. This fact might seem very disappointing at first, as sequential testing is one the main advantages of e-variable theory. However, quite incredibly, the composite alternative version of the test (see Section 4.2) will yield a test supermartingale (under a coarser filtration), and allow sequential testing with the greedy stopping time.

Before we end this section, we want to mention a different, but also very natural way of defining the random design hypotheses for LR testing. Instead of allowing all distributions over the covariates (X^n, Z^n), we could have required the rows of the design matrix to be i.i.d. Then for all P in the hypotheses, we would have $P(X^n, Z^n) = P(X, Z)^n$, but $P(X, Z)$ could still be any distribution. The null hypothesis defined this way is a subset of $\mathcal{H}_{0|n}^{\text{LR}}$, which means that $S_n^{\text{S-LR}}$ would still be an e-variable, but it is not clear whether it would be the GRO for that null.
The proof that we used in the previous proposition does not work anymore. We are not free to choose Dirac marginal distributions $\delta_{(x^n,z^n)}$ for elements in the null, except for the special case when all rows of (x^n, z^n) are exactly the same - because of the i.i.d. assumption.
As a consequence, it can be shown that in the i.i.d. setting, all e-variables for the null are necessarily (X^n, Z^n)-conditional e-variables, but not necessarily *strong* conditional! This distinction makes it impossible to use the same arguments to show that P^* is an element of the effective null. Thus, it is possible that a more e-powerful e-variable than $S_n^{\text{S-LR}}$ exists when i.i.d. covariates are assumed.

3.2 Model-X test for LR

In this section our goal will be the same as in the previous one, namely to describe an e-variable which can effectively be used for LR testing with point alternatives.
However, this time the approach will be very different. Instead of describing how to design an e-variable for our exact problem at hand, we will try to recycle the existing, MX e-variable (see Section 2.4). Thus, the focus of the section is not on the construction of a new e-variable, but instead on arguing why it makes sense to use the MX e-variable in this setting, and what benefits and detriments it might bring.

Recall that the MX e-variable is used for testing conditional independence $Y \perp X|Z$. The null hypothesis and extended model are as given in Chapter 2, (2.10) and (2.11).
Our motivation for using the MX e-variable in LR testing is two-fold.
First, whenever the covariates are sampled i.i.d. and the true propensity kernel $K(X|Z)$ is known, we get a new LR test "for free". This allows us to pick the more convenient test (S-LR or MX based), either based on power, safety guarantees against model misspecification or other relevant properties.
Secondly, in clinical trials the propensity kernel is in fact known, the data are i.i.d., and the most commonly used statistical analyses are linear regression tests (often in disguise). This is true for the t-test, (multiple-factor) ANOVA, ANCOVA, and even some repeated measurements ANOVA variants (which are outside of the scope of this thesis). Therefore, simultaneously satisfying the MX assumption, and additionally imposing a linear model on the target happens often in practice.

We start by explaining how testing the conditional independence (CI) $Y \perp X|Z$ can be viewed as a test of the general regression question: "Is (X, Z) a better predictor of Y than Z alone?". Then, we explore some assumptions under which it makes sense to use CI testing for linear regression testing, and give examples of situations when this is not the case. We proceed with a careful formalization of our new testing setting, and end the section with Proposition 3.2.4, which gives an upper bound on the e-power of the MX e-variable in the context of LR testing. This finding, together with Proposition 3.1.5 plays a crucial role in comparing the e-powers of S-LR and MX e-variables in Section 3.5.

For the moment, let us forget about the MX test specifically and assume we have a conditonal independence test which works without any assumptions on the distributions of (Y, X, Z). Also, we drop the linearity requirement completely and assume the general regression model:

$$Y = f(X, Z) + \varepsilon,$$

where f can be any measurable function, and ε is a 0-mean error term, independent of (X, Z).
From the assumptions on the error, we know $f(X, Z)$ is a version of the conditional expectation $\mathbb{E}[Y|X, Z]$ under the data-generating distribution. If $Y \perp X|Z$, then $\mathbb{E}[Y|X, Z] = \mathbb{E}[Y|Z]$, is a.s. equal to some measurable function of Z. On the other hand, if we start by assuming $\mathbb{E}[Y|X, Z] = \mathbb{E}[Y|Z]$ a.s., then $Y \perp X|Z$ immediately holds since we assumed $\varepsilon \perp (X, Z)$. Therefore, under the general regression model with an independent error we have:

$$\begin{aligned} Y \perp X|Z &\iff \mathbb{E}[Y|X, Z] = \mathbb{E}[Y|Z] \text{ a.s.} \\ &\iff \mathbb{E}[Y|X, Z] \text{ has a } \sigma(Z) \text{ measurable version.} \end{aligned}$$

Furthermore, we can decompose the minimal mean squared approximation error as fol-

lows:

$$\begin{aligned}\mathbb{E}[(Y - \mathbb{E}[Y|Z])^2] &= \mathbb{E}[(Y - \mathbb{E}[Y|X,Z] + \mathbb{E}[Y|X,Z] - \mathbb{E}[Y|Z])^2] \\ &= \mathbb{E}[(Y - \mathbb{E}[Y|X,Z])^2] + \mathbb{E}[(\mathbb{E}[Y|X,Z] - \mathbb{E}[Y|Z])^2].\end{aligned}$$

This implies, (X, Z) and Z are equally good predictors of Y if and only if $\mathbb{E}[Y|X,Z] = \mathbb{E}[Y|Z]$-a.s., or equivalently if and only if $Y \perp X|Z$.
Now, assume the conditional expectation is in fact linear in (X, Z). How does the second formulation of independence now relate to the value of β?
We have:

$$\begin{aligned}\mathbb{E}[Y|X,Z] = \mathbb{E}[Y|Z] \iff \beta^T X + \gamma^T Z = \beta^T \mathbb{E}[X|Z] + \gamma^T Z \\ \iff \beta^T (X - \mathbb{E}[X|Z]) = 0,\end{aligned}$$

where all equalities should be interpreted almost surely under the data-generating distribution. We have proved the following proposition:

Proposition 3.2.1. Assuming the model

$$Y = \beta^T X + \gamma^T Z + \varepsilon,$$

with 0-mean independent error term ε:

- $\beta = 0 \Rightarrow Y \perp X|Z$,
- $Y \perp X|Z$ and $X - \mathbb{E}[X|Z]$ is linearly independent $\Rightarrow \beta = 0$.

The first implication is very important, as it tells us that testing conditional independence instead of $\beta = 0$ is safe, i.e. does not lead to a loss of type-1 error guarantees.
The implication $Y \perp X|Z \Rightarrow \beta = 0$ does not hold (in general). For example, imagine a setting in which X is a deteriministic function of Z. The conditional independence then clearly holds for any choice of β. Equivalently,

$$\beta \neq 0 \not\Rightarrow Y \not\perp X|Z.$$

When formulated this way, it is clear that there can be testing scenarios in which $\beta \neq 0$, but conditional independence holds. Whenever this is the case, the rejection probability of any level-α conditional independence test will be at most α.
Thus, using a hypothetical conditional independence test, which works without any additional assumptions, would never break safety guarantees, but could lead to a substantial loss of power when used for testing $\beta = 0$ in our linear model.

Now, we want to completely formalize the previous conclusion for LR testing with the MX e-variable, in settings where on top of linearity, the propensity kernel $K(X|Z)$ is known, and the covariates are i.i.d. We do so by defining a new null hypothesis and extended model, both of which will depend on the specified propensity kernel.

Definition 3.2.2. (The specified propensity LR hypotheses)
We define a new null hypothesis and extended model as the intersection of their MX and S-LR counterparts:

$$\mathcal{H}^{\text{SPLR}}_{0|n,K(X|Z)} := \mathcal{H}^{\text{LR}}_{0|n} \cap \mathcal{H}^{\text{MX}}_{0|n,K(X|Z)},$$
$$\mathcal{H}^{\text{SPLR}}_{1|n,K(X|Z)} := \mathcal{H}^{\text{LR}}_{1|n} \cap \mathcal{H}^{\text{MX}}_{1|n,K(X|Z)}.$$

The null contains all distributions $P(D^n)$ over the data under which the linear model $Y^n|X^n, Z^n \sim \mathcal{N}(Z^n\gamma, \sigma^2 I_n)$ holds $P(X^n, Z^n)$-a.s., the covariates (X_i, Z_i) are i.i.d., and have a fixed propensity kernel $P(X|Z) = K(X|Z)$, $P(Z)$-a.s.
Note that there is no need to specifically impose that $P(X|Y, Z) = K(X|Z)$ holds $P(Y, Z)$-a.s. anymore, since $Y \perp_P X|Z$ is implied by the linearity requirement under the null.

Any distribution P_0 in the SPLR null is specified by the two values (γ_0, σ_0^2) and the marginal distribution $P_0(Z)$. We can think of the new null as a reduction of the original LR null hypothesis after incorporating additional information; that the covariates are i.i.d. and that the true propensity kernel is $K(X|Z)$.

It is important to realize that reducing a null hypothesis by removing distributions which are certainly not the data-generators (i.e. including true knowledge) does not lead to a loss of safety guarantees, but can lead to a stronger GRO e-variable.
To clearly explain what we mean by not losing safety guarantees, we have to slightly rephrase our current definition of safety, i.e. level-α tests (Chapter 1, (1.1.1)). Let us redefine a level-α fixed-sample test ϕ for $\mathcal{H}_0$ as one which makes the following implication true:
"If the data-generating distribution P is in $\mathcal{H}_0$, then the rejection probability of ϕ under P is at most α".
This definition might seem equivalent to the original, but a small difference does exist. Assume there exists a set of distributions $\mathcal{I} \subset \mathcal{H}_0$ which are certainly not the data-generators. Any level-α test for $\mathcal{H}_0 \setminus \mathcal{I}$ is also a level-α test for $\mathcal{H}_0$ according to the new definition. Because, if the data generating distribution is in the null, by definition of I, it is necessarily in $\mathcal{H}_0 \setminus \mathcal{I}$.
Another way to make the argument for decreasing the null is by simply noticing that there is no practical advantage from providing safety guarantees under distributions which are certainly not the data generators. Therefore, we may take all such distributions out of the null hypothesis.
Furthermore, the GRO e-variable for $\mathcal{H}_0$ is still an e-variable for $\mathcal{H}_0 \setminus \mathcal{I}$, but is not necessarily the GRO for the reduced null. Thus, reducing the null by incorporating true knowledge can only lead to better GRO e-variables and tests.

Analogously, we are not interested in optimizing e-power under distributions which can not possibly be the data generators, so we reduce the extended model as well.
For any Q in the extended model - which is uniquely specified by $(\beta_1, \gamma_1, \sigma_1^2, Q(Z))$,

following the results in Section 2.4, the MX e-variable (for testing $\mathcal{H}^{\mathrm{MX}}_{0|n,K(X|Z)}$ against $\{Q\}$) takes the following form:

$$S_n^{\mathrm{MX}}(D^n) = \prod_{i=1}^{n} \frac{\mathcal{N}(Y_i|\beta_1^T X_i + \gamma_1^T Z_i, \sigma_1^2)}{\mathbb{E}_{K(X|Z_i)}[\mathcal{N}(Y_i|\beta_1^T X + \gamma_1^T Z_i, \sigma_1^2)]}.$$

Whenever linearity, i.i.d. covariates and a propensity kernel are specified, one can use either the S-LR or the MX e-variable for testing, as they are both e-variables for the intersection of their nulls, but importantly, neither is necessarily the GRO e-variable for the intersection. Unfortunately, finding the GRO e-variable for testing $\mathcal{H}^{\mathrm{SPLR}}_{0|n,K(X|Z)}$ against relevant point alternatives is very often impossible, as the optimization problem described in Theorem 2.2.5 is not tractable. The MX and S-LR settings represent rare exceptions to this rule.

As a consequence, comparing the e-power and safety guarantees (under model misspecification) of the MX and S-LR e-variables becomes important. It can provide valuable information for choosing the better e-variable for specific use-cases. As a first step in this direction, we provide an upper bound on the e-power of the MX e-variable in the context of LR testing.
The derivation makes use of another important information theoretic quantity:

Definition 3.2.3. (Differential entropy)
The (differential) entropy of a random variable A with Lebesgue density $p(a)$ is given by:

$$-\int \ln p(a) p(a) da,$$

where the integral is taken over the support of p.

The following proposition is proved using a classical result, which states that Gaussians $\mathcal{N}(\mu, \Sigma)$ uniquely maximize the differential entropy among all (Lebesgue) absolutely continuous random variables on $\mathbb{R}^n$ with fixed covariance matrix Σ [Cover and Thomas, 1991].

Proposition 3.2.4. (MX-LR e-power upper bound and maximizer)
Let $\{Q(Y, X, Z)\} \subset \mathcal{H}^{\mathrm{SPLR}}_{1|n,K(X|Z)}$ with parameters $(\beta_1, \gamma_1, \sigma_1^2)$. Assume $\beta_1^T X$ has a finite second moment under Q. Then, the conditional mutual information, and thus the e-power of S_1^{MX} is upper bounded by:

$$I(Y; X|Z) \leq \frac{1}{2} \ln(1 + \delta_1^T \mathrm{Cov}(X - \mathbb{E}[X|Z])\delta_1).$$

The bound is achieved if and only if $\beta_1^T X|Z = z \sim \mathcal{N}(\mu(z), \sigma^2)$ with $Q(Z)$-a.s. constant variance σ^2.

Proof. The mutual information can be represented as a difference of two conditional entropies [Cover and Thomas, 1991]:

$$I(Y;X|Z) = h(Y|Z) - h(Y|X,Z).$$

The second is always equal to the Gaussian entropy $\frac{1}{2}\ln(2e\pi\sigma_1^2)$ under any $Q \in \mathcal{H}^{\text{SPLR}}_{1|n,K(X|Z)}$ because the model on Y is specified. The first can be bounded using the fact that Gaussians maximize entropy and Jensen's inequality as follows:

$$\begin{aligned} h(Y|Z) = \mathbb{E}_{z\sim Q(Z)}[h(Y|Z=z)] &\leq \mathbb{E}\left[\frac{1}{2}\ln(2e\pi(\sigma_1^2 + \beta_1^T \operatorname{Cov}(X|Z=z)\beta_1))\right] \\ &\leq \frac{1}{2}\ln(2e\pi(\sigma_1^2 + \beta_1^T \mathbb{E}[\operatorname{Cov}(X|Z=z)]\beta_1)) \\ &= \frac{1}{2}\ln(2e\pi(\sigma_1^2 + \beta_1^T \operatorname{Cov}(X - \mathbb{E}[X|Z])\beta_1)). \end{aligned}$$

All expectations are taken over $z \sim Q(Z)$, but we omit this for notational simplicity. Also, note that after conditioning on $\{Z = z\}$, the distribution of Y is absolutely continuous w.r.t. the Lebesgue measure, allowing us to use the Gaussian bounds.
To get a perfect equality, first, on a set of $Q(Z)$-probability-1 $h(Y|Z=z)$ has to be maximized among all distributions with variance $\operatorname{Var}(\beta_1^T X|Z=z) + \sigma_1^2$. This maximum is achieved iff $Y|Z=z$ is Gaussian with the same variance, which is equivalent to saying that $\beta_1^T X|Z=z$ is Gaussian. Jensen's inequality is an equality iff the integrated variable is a.s. constant, thus the requirement that $\operatorname{Var}(\beta_1^T X|Z=z)$ is $Q(Z)$-a.s. constant. □

3.3 Examples - AN(C)OVA testing

The MX and S-LR tests are based on very different assumptions, and it is not hard to find settings in which one test can be used, and performs very well, while the other performs very poorly, i.e. either does not even provide type-1 error guarantees, or has trivially low power. The easiest examples would be situations in which a linear dependence of Y on (X, Z) holds, but no information about the propensity kernel is available, or vice versa. In this section we show that there exist more subtle, yet practically relevant experimental designs in which, both linearity and the propensity kernel are known, but nevertheless, only one of the tests can be used effectively.
To warm up, we start with a simple setting where either method can be used.

3.3.1 ANOVA testing with MX and S-LR

The *analysis of variance* test - ANOVA (also known as *one-way* or *one-factor* ANOVA) can be used in the following experimental setting.
A total of n data points $\{(Y_1, X_1), \ldots, (Y_n, X_n)\}$, each corresponding to a participant in an RCT are available. The participants were selected independently from a population of interest, and divided into k experimental groups by sampling the group index from a

uniform distribution on $\{1, \ldots, k\}$. Each group is given a different treatment, and the variable Y represents the difference in a target quantity (e.g., blood pressure), before and after receiving the treatment. The goal is to test whether the expected value of Y varies across the groups.

This question can be completely encoded into a linear regression testing problem in the following way: the covariate of interest X is the group-indicator vector of length $k-1$, taking values in $\{e_1^{k-1}, e_2^{k-1} \ldots, e_{k-1}^{k-1}, e_0^{k-1}\}$, i.e. in the canonical basis of $\mathbb{R}^{k-1}$ or the zero-vector which we denoted as e_0^{k-1}. Each vector X_i uniquely determines the experimental group assigned to participant i. The group corresponding to the zero-vector is called the *reference group*, and for simplicity we assume it is the final group, with index k. The remaining covariates Z contain only an intercept. The measurement errors are as usual, assumed to be Gaussian, yielding the linear model:

$$Y = \beta^T X + \gamma + \varepsilon, X \perp \varepsilon \sim \mathcal{N}(0, \sigma^2)$$

for a single outcome. This encoding scheme is called *reference coding* [Agresti, 2015], and it is used in most practical implementations of the classical ANOVA test.
In an actual RCT, the reference group is usually the control group, i.e. the group that is given no treatment or a placebo, and the ANOVA test verifies whether at least one of the treatments has an actual effect - different than that of the placebo.

Now, it is easy to check that:

$$\mathbb{E}[Y|X] = \text{const} \iff \beta^T X = \text{const} \iff \beta^T X = 0 \iff \beta = 0,$$

where the first three equalities should be interpreted almost surely.
γ represents the mean value of the reference group, while β_j represents the difference in expected value between the group with vector e_j^{k-1} and the reference. For any fixed alternative distribution Q with parameters $(\beta_1, \gamma_1, \sigma_1^2)$, we can use the S-LR e-variable:

$$S_n^{\text{S-LR}} = \frac{\mathcal{N}(Y^n | X^n \beta_1 + 1_n \gamma_1, \sigma_1^2 I_n)}{\mathcal{N}(Y^n | 1_n \gamma_0(X^n, 1_n), \sigma_0^2(X^n, 1_n) I_n)},$$

where γ_0 and σ_0^2 are as in Definition 3.1.4, and 1_n is an n-dimensional vector of ones. Finally,

$$\phi_\alpha := 1_{[\frac{1}{\alpha}, \infty]}(S_n^{\text{S-LR}})$$

gives us the Simple-LR version of the ANOVA test.

On the other hand, since the data are i.i.d. and the propensity kernel $K(X|Z) = K(X)$ is exactly known, we can also use the MX e-variable, which now instead takes form

$$S_n^{\text{MX}} = \prod_{i=1}^{n} \frac{\mathcal{N}(Y_i | \beta_1^T X_i + \gamma_1, \sigma_1^2)}{\frac{1}{k} \sum_{j=0}^{k-1} \mathcal{N}(Y_i | \beta_1^T e_j^{k-1} + \gamma_1, \sigma_1^2)},$$

to perform the ANOVA test.

As always, the MX e-variable has a product form, yielding a test supermartingale and allowing sequential testing. The S-LR e-variable, generally does not produce a test supermartingale and does not provide type-1 error guarantees when used sequentially. In this specific setting however, that issue can be resolved.
Assume that instead of randomly assigning participants to experimental groups, we divide them into clusters of size k, making sure that no two participants within the same cluster end up in the same experimental group. This can be done, e.g. by randomly permuting the participants inside the clusters and then assigning them to experimental groups $1, 2, ..., k$ in this order. If we follow this procedure, using the final group as reference, the design matrix after measuring the results of $n = mk$ participants is a block matrix consisting of m repetitions of the matrix:

$$(x^k, z^k) = \begin{bmatrix} I_{k-1} & 1_{k-1} \\ 0_{k-1}^T & 1 \end{bmatrix}.$$

Using $s_{\beta_1} := 1_{k-1}^T \beta_1$, we get parameters:

$$\gamma_0(x^n, z^n) = \gamma_1 + \frac{s_{\beta_1}}{k},$$
$$\sigma_0^2(x^n, z^n) = \sigma_1^2 + \frac{||\beta_1||^2}{k} - (\frac{s_{\beta_1}}{k})^2,$$

which depend exclusively on $(\beta_1, \gamma_1, \sigma_1^2)$ and the number of experimental groups k. Therefore, in this case, the S-LR e-variable can be rewritten as:

$$S_n^{\text{S-LR}} = \prod_{i=1}^{m} \frac{\mathcal{N}(Y_i^k | \tilde{\beta}_1 + 1_k \gamma_1, \sigma_1^2 I_k)}{\mathcal{N}(Y_i^k | 1_k \gamma_0, \sigma_0^2 I_k)},$$

where Y_i^k represents the k outcomes of the i-th cluster of participants, and $\tilde{\beta}_1 = [\beta_1^T, 0]^T$. Thus, as long as the data points arrive in clusters of size k, we can think of Y as a single k-dimensional data point, and the S-LR e-variable yields a test supermartingale under this new data filtration.

The same trick can not be used for general LR testing, as it relies on the block structure of the design matrix, which appears only in highly structured settings, such as this one. Now, let us explore two settings of great practical importance in which S-LR and MX behave very differently.

3.3.2 Two-factor ANOVA with S-LR

A natural extension of the analysis of variance test is the two-factor (two-way) ANOVA. This time, the experimental design contains two categorical variables of interest (factors) instead of one. Abstractly, we could say the group indices are given by the product set:

$\{1, \ldots, k_1\} \times \{1, \ldots, k_2\}$. In general, researchers are interested in testing three distinct statistical questions: the existence of two *main effects* - one for each of the factors, and the existence of an *interaction effect*. To explain the meaning of these effects we will use the following notation:

- $\mu_{i,j}$ - expected value of Y in group (i, j),
- $\mu_{i,\cdot} := \frac{1}{k_2} \sum_{j=1}^{k_2} \mu_{i,j}$,
- $\mu_{\cdot,j} := \frac{1}{k_1} \sum_{i=1}^{k_1} \mu_{i,j}$,
- $\mu_{\cdot,\cdot} = \frac{1}{k_1 k_2} \sum_{i,j} \mu_{i,j}$.

Testing the first main effect means testing the hypothesis

$$H_0 : \mu_{1,\cdot} = \mu_{2,\cdot} = \ldots, \mu_{k_1,\cdot},$$

and it is completely equivalent to the one-way ANOVA testing problem which we would get by merging all groups along the second factor. Testing the second main effect is analogous. The novel and interesting part of this test is the interaction effect. We say that there is no interaction between the two factors if for all (i, j)

$$\begin{aligned} \mu_{i,j} &= \mu_{i,\cdot} + \mu_{\cdot,j} - \mu_{\cdot,\cdot} \\ &= (\mu_{i,\cdot} - \mu_{\cdot,\cdot}) + (\mu_{\cdot,j} - \mu_{\cdot,\cdot}) + \mu_{\cdot,\cdot}. \end{aligned}$$

It helps to think of $\mu_{i,j} - \mu_{\cdot,\cdot}$ as the effect that being placed in group (i, j) has on the target variable Y. The previous equation says that there is no interaction if the effect of being in group (i, j), is equal to the sum of the marginal effects of being in group i along the first factor and j along the second, for all (i, j). This is precisely what we want to test now.

To formalize the interaction question completely, we can again write it as a linear regression testing problem. The procedure is similar to that of the one-way ANOVA, but notationally much more cumbersome, so we leave the details out ; more information can be found in the book by Casella and Berger [2024]. We are left with the linear model:

$$Y = \beta_I \odot X^I + \gamma_1^T X^1 + \gamma_2^T X^2 + \gamma_0 + \varepsilon, X^{\text{all}} \perp \varepsilon \sim \mathcal{N}(0, \sigma^2) \tag{3.3}$$

for a single outcome. Here $X^{\text{all}} := (X^1, X^2, X^I)$. X^1 and X^2 are group-indicator vectors for each factor separately, they encode the main effects and are designed the same way as in the one-way ANOVA, with reference groups k_1 and k_2. For notational convenience, X^I is the $(k_1 - 1) \times (k_2 - 1)$ matrix defined as:

$$X^I = X^1 X^{2^T}$$

instead of a vector. It is used to represent the interaction effect, which is present if and only if the matrix of parameters β_I, which has the same shape as X^I, has a non-zero entry. $\odot$ denotes a pointwise product of matrices. This way, testing the interaction

effect simplifies into performing a linear regression test with the variable of interest X^I, and remaining covariates $(X^1, X^2, 1)$.

Therefore, it is possible to test the existence of an interaction effect in a two-way ANOVA using the S-LR e-variable. However, testing the interaction effect with the MX e-variable would be futile even though the propensity kernel $K(X^I|X^1, X^2)$ is known. The reason is that X_I is a deterministic function of (X_1, X_2), making the conditional independence $Y \perp X^I|X^1, X^2$ always true. Thus, the MX based test would reject with at most α probability for any value of β_I.

3.3.3 Robust ANCOVA with MX

Our final example is a twist on the classical *analysis of covariance* (ANCOVA). The ANCOVA test is another generalization of the one-way ANOVA.
The setting is exactly the same as before, except that the covariates Z now consist of an intercept term and other measured variables, which are assumed to be related to the target Y. In this framework, unlike the one-way ANOVA, researchers take into account possible differences in the baseline value of Y for different participants, based on the values of their covariates Z. The usual linear model with an additive error:

$$Y^n = X^n\beta + Z^n\gamma + \varepsilon^n \tag{3.4}$$

is used for n outcomes. The propensity kernel $K(X|Z) = K(X)$ is again the uniform distribution over the group indicator vectors (X), and describes the usual RCT setting, where participants are assigned into experimental groups in a random-uniform way, independent of their properties such as medical histories, which are encoded in Z.

The standard assumption in practice, and which we made in the previous two examples is that the errors are i.i.d. Gaussians, independent of all the covariates. In practice, this assumption can break in many different ways. The error terms could be dependent of each other $\varepsilon_i \not\perp \varepsilon_j$, e.g. if the same researcher, who tends to overestimate, made the measurements for participants i and j. They could depend on the covariates Z, e.g. it is known that a portable glucometer has a larger measurement error when the patient's skin temperature is low. Specifically, this dependence can lead to *non-linear* relations between Y and Z. Further, the error could also depend on the group allocation X in some way, e.g. it could be *heteroscedastic*, meaning that the variance of the error differs across experimental groups. Finally, even if all other assumptions hold, the errors could generally follow non-Gaussian distributions.
The S-LR based test does not provide theoretical type-1 error guarantees in any of the mentioned settings (in general), i.e. even when $\beta = 0$ holds in (3.4), the test

$$\phi_\alpha = 1_{[\frac{1}{\alpha},\infty]}(S_n^{\text{S-LR}})$$

might reject with more than α probability. On the other hand, the MX based test does

provide theoretical safety guarantees in some of these cases.

To make the claim precise, we need to take a step back and reconsider the fundamental assumptions of MX-based testing. Assume that a (possibly non-i.i.d.) distribution $P(D^n)$ satisfies a *generalized MX condition*:

$$P(X_i|Y_i, Z_i, D^{i-1}) = K(X|Z) \tag{3.5}$$

for all $i \leq n$, and a fixed kernel $K(X|Z)$. Then:

$$\begin{aligned}\mathbb{E}_P[S_n^{\mathrm{MX}}|Y_n, Z_n, D^{n-1}] &= S_{n-1}^{\mathrm{MX}}\mathbb{E}_P\left[\frac{dQ(X|Y,Z)}{dK(X|Z)}(Y_n, X_n, Z_n)|Y_n, Z_n, D^{n-1}\right] \\ &= S_{n-1}^{\mathrm{MX}}.\end{aligned}$$

The second equality follows immediately from (3.5). Thus, the MX e-variable S_n^{MX} yields a test supermartingale under all distributions of the data which satisfy the generalized MX condition, even if the data are not i.i.d. under them. The i.i.d. data assumption, which is a part of the "ordinary" MX setting described in Section 2.4 can, therefore, be relaxed.

Getting back to our example, we need to find conditions under which the requirement (3.5) holds in the ANCOVA model. The group allocation of participant i in most RCT designs is determined by external randomization, such as a coin or dice toss or random number generation, and thus is clearly independent of all previous data D^{i-1} and covariates Z_i. So $X_i \perp Z_i, D^{i-1}$ is satisfied. If on top of that

$$X_i \perp \varepsilon_i|Z_i, D^{i-1} \tag{3.6}$$

holds, then using standard Markov kernel rules [Forré and Mooij, 2024] we get:

$$X_i \perp \varepsilon_i, Z_i, D^{i-1}. \tag{3.7}$$

It is not hard to see that whenever $\beta = 0$, (3.7) implies that the generalized MX requirement $X_i \perp Y_i, Z_i, D^{i-1}$ indeed holds in the model (3.4). Therefore, as long as (3.6) is true, the MX e-variable can safely be used for ANCOVA testing. This specifically means that any joint distribution of (Z^n, ε^n) may be allowed, as long as $X^n \perp (Z^n, \varepsilon^n)$, and the MX based ANCOVA test will not lose safety guarantees.

Thus, except under heteroscedasticity, the MX e-variable provides theoretical safety guarantees under the model misspecification examples we gave earlier. This example shows that MX might generally serve as a good alternative to the S-LR e-variable in settings where robustness and safety guarantees are of primary importance.

3.4 The fully specified null

The main goal of this section is to describe the problems related to finding the GRO e-variable for testing the newly defined Fully Specified Linear Regression (FSLR) null

hypothesis against relevant point alternatives. At the end of the section we mark out a special setting in which the MX e-variable coincides with the sequential GRO e-variable for the (FSLR) null.

As we have argued in Section 3.2, if i.i.d. data, the true kernel $K(X|Z)$ and linearity are known or assumed to be true, then using either MX or S-LR is a waste of (informational) resources. We should instead strive to test the intersection of the two nulls: $\mathcal{H}_{0|n}^{\text{SPLR}}$. The issue with this approach is that, in almost all practical settings we are not be able to analytically express the GRO e-variable. To illustrate this fact, we will consider an even simpler testing scenario - Fully Specified Linear Regression (FSLR), where i.i.d. covariates, linear dependency of Y on (X, Z) and knowledge of the full distribution $K(X, Z)$ of the covariates are assumed.

Definition 3.4.1. (FSLR null and extended model)

$$\mathcal{H}_{0|n,K(X,Z)}^{\text{FSLR}} = \{P(D^n) : D_i \text{ are i.i.d., } D \sim \mathcal{N}(\gamma^T Z, \sigma^2) \otimes K(X, Z), \gamma \in \mathbb{R}^{d_z}, \sigma^2 > 0\},$$

$$\mathcal{H}_{1|n,K(X,Z)}^{\text{FSLR}} = \{P(D^n) : D_i \text{ are i.i.d., } D \sim \mathcal{N}(\beta^T X + \gamma^T Z, \sigma^2) \otimes K(X, Z), \\ (\beta, \gamma) \in \mathbb{R}^{d_x+d_z}, \sigma^2 > 0\}.$$

Let us describe the problem of finding the RIPr for the fully specified null in more detail. We work in the $n = 1$ case, as then we can say a bit more.

All distributions in the null and the extended model are absolutely continuous w.r.t. the σ-finite measure $\nu := \lambda \otimes K(X, Z)$, where λ is the Lebesgue measure on $\mathcal{B}(\mathbb{R})$. The alternative distribution Q can be represented by its density wrt ν

$$\frac{dQ}{d\nu}(y, x, z) := q(y|x, z) = \mathcal{N}(y|\beta_1^T x + \gamma_1^T z, \sigma_1^2),$$

and similarly, the distributions in the null with:

$$p_{\gamma,\sigma^2}(y|z) = \mathcal{N}(y|\gamma^T z, \sigma^2),$$

for varying parameters γ and σ^2.
Thus, the null and extended model are parametric families of probability distributions which are all absolutely continuous w.r.t. the same measure ν. This simplifies the problem of finding the RIPr. According to Grünwald et al. [2024], if

$$\inf_{\pi} \mathbb{E}_Q \left[\ln\left(\frac{q(Y|X,Z)}{\int p_{\gamma,\sigma^2}(Y|Z)d\pi(\gamma,\sigma^2)}\right)\right] = \mathbb{E}_Q \left[\ln\left(\frac{q(Y|X,Z)}{\int p_{\gamma,\sigma^2}(Y|Z)d\pi^*(\gamma,\sigma^2)}\right)\right] < \infty, \quad (3.8)$$

where the infimum is taken over all probability distributions (priors) π on $\mathcal{B}(\mathbb{R}^{d_z} \times (0, \infty))$, then $\int p_{\gamma,\sigma^2}(y|z)d\pi^*(\gamma, \sigma^2)$ is the ν-density of the RIPr of Q onto the null. The RIPr

can, therefore, be identified with the minimizing distribution π^* over (γ, σ^2).

When π^* happens to be a Dirac measure, we are in the simple case (see Definition 2.2.6), and finding it reduces to a finite-dimensional convex optimization task, which can easily be solved analytically.
In Proposition 3.4.4, we present a slightly more complicated, *semi-simple* setting, in which we are also able to pinpoint π^*, although it is not necessarily Dirac.
In full generality however, π^* is the solution to the described infinite-dimensional convex optimization problem (3.8), and it depends on $(K(X,Z), \beta_1, \gamma_1, \sigma_1^2)$ in ways that we are often unable to describe.

Let us now remember our most important tool in the search for optimal e-variables - the fact that the GRO is the only e-variable which can be represented as a likelihood-ratio $\frac{dQ}{dP}$ of the alternative and an element of the effective null, and additionally, that if P and Q are equivalent, P must be the RIPr (Theorem 2.2.5).
Note that, although we use conditional notation, $q(y|x,z)$ and $p(y|z)$ are the actual density functions of Q and some $P \in \mathcal{H}^{\text{FSLR}}_{0|1,K(X,Z)}$ w.r.t. ν. Thus, if we ever find an e-variable of the form

$$\frac{q(Y|X,Z)}{\int p_{\gamma,\sigma^2}(Y|Z)d\pi(\gamma,\sigma^2)},$$

we immediately know it has to be the GRO, and $\int p_{\gamma,\sigma^2}(Y|Z)d\pi(\gamma,\sigma^2)$ the density of the RIPr. The e-variable requirement states

$$\mathbb{E}_{P'}\left[\frac{q(Y|X,Z)}{\int p_{\gamma,\sigma^2}(Y|Z)d\pi(\gamma,\sigma^2)}\right] \leq 1$$

for all $P' \in \mathcal{H}^{\text{FSLR}}_{0|1,K(X,Z)}$. Under the null hypothesis $Y \perp X|Z$ always holds, and the previous inequality decomposes into

$$\mathbb{E}_{P'(Y,Z)}\left[\frac{\mathbb{E}_{K(X|Z)}[q(Y|X,Z)]}{\int p_{\gamma,\sigma^2}(Y|Z)d\pi(\gamma,\sigma^2)}\right] \leq 1.$$

Now, let us define a new density

$$q'(y|z) := \mathbb{E}_{K(X|Z=z)}[q(y|X,z)].$$

A trivial consequence is that

$$\mathbb{E}_{P'}\left[\frac{q(Y|X,Z)}{\int p_{\gamma,\sigma^2}(Y|Z)d\pi(\gamma,\sigma^2)}\right] \leq 1 \iff \mathbb{E}_{P'}\left[\frac{q'(Y|Z)}{\int p_{\gamma,\sigma^2}(Y|Z)d\pi(\gamma,\sigma^2)}\right] \leq 1.$$

The random variable on the left is an e-variable for $\mathcal{H}^{\text{FSLR}}_{0|1,K(X,Z)}$ if and only if the r.v. on the right is, which implies that the RIPrs of the original alternative Q and the new distribution Q', defined by $q' := \frac{dQ'}{d\nu}$, onto $\mathcal{H}^{\text{FSLR}}_{0|1,K(X,Z)}$ coincide.
Now, notice that Q' is exactly the RIPr of Q onto the MX null hypothesis $\mathcal{H}^{\text{MX}}_{0|1,K(X|Z)}$.
We have proved the following proposition:

Proposition 3.4.2. (The double projection)
Finding the RIPr of $Q \in \mathcal{H}^{\mathrm{FSLR}}_{1|1,K(X,Z)}$ onto $\mathcal{H}^{\mathrm{FSLR}}_{0|1,K(X,Z)}$ can be done in two steps:

- Project Q onto the larger $\mathcal{H}^{\mathrm{MX}}_{0|1,K(X|Z)}$ to obtain the RIPr Q',
- Project Q' onto $\mathcal{H}^{\mathrm{FSLR}}_{0|1,K(X,Z)}$.

If Q' turns out to be an element of $\mathcal{H}^{\mathrm{FSLR^{eff}}}_{0|1,K(X,Z)}$, the second projection is not needed, and the RIPr and GRO e-variable for testing the FSLR and MX nulls coincide. The following proposition gives a sufficient condition for this, "semi-simple" case.

Definition 3.4.3. (The intercept condition)
We say that distribution $K(A)$ satisfies the intercept condition if there exists a vector $c \in \mathbb{R}^{d_a}$ such that

$$c^T A = 1$$

holds $K(A)$-a.s.

Proposition 3.4.4. (MX and FSLR: the semi-simple case)
Assume $Q \in \mathcal{H}^{\mathrm{FSLR}}_{1|1,K(X,Z)}$ with $\lambda \otimes K(X,Z)$-density $\mathcal{N}(y|\beta_1^T x + \gamma_1^T z, \sigma_1^2)$. Furthermore, $K(X,Z) = K(X) \otimes K(Z)$, and the intercept condition holds for $K(Z)$. Then,

$$\mathbb{E}_{K(X|Z=z)}[\mathcal{N}(y|\beta_1^T X + \gamma_1^T z, \sigma_1^2)] = \mathbb{E}_{K(X)}[\mathcal{N}(y|\beta_1^T X + \gamma_1^T z, \sigma_1^2)] \tag{3.9}$$

is the $\lambda \otimes K(X,Z)$-density of the RIPr of Q onto $\mathcal{H}^{\mathrm{MX}}_{0|1,K(X|Z)}$ and onto $\mathcal{H}^{\mathrm{FSLR}}_{0|1,K(X,Z)}$.

Proof. (3.9) is clearly the density of the RIPr of Q onto the corresponding MX null. To show the second statement, it is enough to express (3.9) as a mixture of densities from the FSLR null, because any such mixture is certainly an element of the effective null $\mathcal{H}^{\mathrm{FSLR^{eff}}}_{0|1,K(X,Z)}$. The statement then follows by the double-projection idea from Proposition 3.4.2, and the fact that projecting an element of the effective null onto the null always yields the same distribution.
Let c be a vector such that $c^T Z = 1$ holds $K(Z)$-a.s. Then,

$$\beta_1^T X + \gamma_1^T Z = ((\beta_1^T X) \cdot c + \gamma_1)^T Z =: \gamma_0(X)^T Z.$$

We can define $\tilde{\pi}(\gamma)$ as the push-forward distribution of $\gamma_0(X)$:

$$\tilde{\pi}(\gamma \in A) := K(\gamma_0(X) \in A)$$

for all $A \in \mathcal{B}(\mathbb{R}^{d_z})$. Then, $\pi^* := \tilde{\pi} \otimes \delta_{\sigma_1^2}$ satisfies

$$\mathbb{E}_{K(X)}[\mathcal{N}(y|\beta_1^T X + \gamma_1^T z, \sigma_1^2)] = \int \mathcal{N}(y|\gamma^T z, \sigma^2) d\pi^*(\gamma, \sigma^2).$$

□

A concise example of the semi-simple case is given when $X \sim B(0.5)$ has a Bernoulli distribution, and $Z = 1$ only contains an intercept, i.e. $K(X, Z) = B(0.5) \otimes 1$. Then, Proposition 3.4.4 implies that

$$\frac{1}{2}\mathcal{N}(y|\gamma_1, \sigma_1^2) + \frac{1}{2}\mathcal{N}(y|\beta_1 + \gamma_1, \sigma_1^2)$$

is the $\lambda \otimes B(0.5) \otimes 1$ density of the RIPr of distribution $Q(Y, X, Z) = \mathcal{N}(\beta_1 X + \gamma_1, \sigma_1^2) \otimes B(0.5) \otimes 1$, onto $\mathcal{H}^{\text{FSLR}}_{0|1,B(0.5)\otimes 1}$.
This is not the simple case, as the prior $\pi^*(\gamma, \sigma^2)$ is not a Dirac measure, but instead a uniform mixture of two Diracs, one centered at (γ_1, σ_1^2) and the other at $(\gamma_1 + \beta_1, \sigma_1^2)$. But nevertheless, we are able to express π^* exactly, because it is the semi-simple case, and finding the RIPr onto the FSLR null hypothesis reduces to finding the RIPr onto the MX null – which we know how to do.

The $X \perp Z$ assumption in Proposition 3.4.4 is necessary because the mixing distribution π^* could otherwise depend on z. It is also important to notice that the statement works for $n = 1$ data points, and it can not be directly generalized to $n > 1$.
Nevertheless, we get an interesting corollary.

Corollary 3.4.5. (MX as sequential GRO for FSLR)
Under the assumptions of the previous proposition, the MX e-variable

$$S_n^{\text{MX}}(D^n) = \prod_{i=1}^{n} \frac{\mathcal{N}(Y_i|\beta_1^T X_i + \gamma_1^T Z_i, \sigma_1^2)}{\mathbb{E}_{K(X)}[\mathcal{N}(Y_i|\beta_1^T X + \gamma_1^T Z_i, \sigma_1^2)]}$$

is the sequential-GRO (see Definition 2.2.8) e-variable for testing $\mathcal{H}^{\text{FSLR}}_{0|n,K(X,Z)}$ against $\{Q^n\}$.

This is a direct consequence of the fact that for $n = 1$, the GRO e-variables for MX and FSLR coincide, as shown in Proposition 3.4.4.

Corollary 3.4.5 is particularly useful, as it provides a meaningful interpretation of the, otherwise unexpected, statement of Theorem 3.5.8.

3.5 E-power comparison

Our goal in this section is to describe how S-LR and MX e-variables compare in terms of e-power when a linear dependence of Y on (X, Z), i.i.d. covariates and a known true propensity kernel $K(X|Z)$ are all assumed, i.e. when both e-variables can be applied. As we have mentioned in Sections 3.2 and 3.4, one should ideally construct an e-variable which takes into account all available information. A way to do so would be to base our tests on the GRO e-variable for testing the SPLR null hypothesis against relevant alternatives (see Section 3.2). The SPLR null explicitly uses all our assumptions, but we

are generally unable to find the corresponding GRO e-variable. Comparing the e-powers of the available e-variables thus gains practical importance, as it provides relevant information which can be used, along with safety guarantees and other useful properties, to select the better e-variable in a specific testing setting.

Theorems 3.5.6 and 3.5.8, and Proposition 3.5.7 represent the most important theoretical contributions of this thesis. Theorem 3.5.6 states that asymptotically, as the sample size tends to infinity, the S-LR e-variable always has at least as much e-power as the MX e-variable. In fact, the asymptotic e-power difference is generally positive, except in a special setting which we describe in detail, where it is zero. Proposition 3.5.7 serves as a counter-example for the statement of Theorem 3.5.6 on finite samples, describing a setting of practical importance in which MX outperforms S-LR. Finally, Theorem 3.5.8 describes the behavior of the asymptotic e-power difference in a special setting which is relevant for randomized controlled trials.

We start by establishing a benchmark for the e-powers of both e-variables. Then we proceed to compare a measure of their asymptotic e-powers directly.

Definition 3.5.1. (The benchmark)
Let $K(X, Z)$ be any distribution over the covariates with finite second moments, and $Q \in \mathcal{H}^{\text{FSLR}}_{1|1,K(X,Z)}$. We define our e-power benchmark as:

$$\mathbf{B}_{K(X,Z)} := \min_{P \in \mathcal{H}^{\text{FSLR}}_{0|1,K(X,Z)}} D_{\text{KL}}(Q||P).$$

Definition 3.5.2. (The best linear predictor)
Let (X, Z) be a random vector with a finite second moment. We use

$$L_{X|Z} := \mathbb{E}[XZ^T]\mathbb{E}[ZZ^T]^+$$

to denote the matrix representation of the best linear predictor of X with Z. Here, as in Section 3.1, A^+ denotes the Moore-Penrose inverse of matrix A.

A short proof showing that $\mathbb{E}[(X - L_{X|Z}Z)^2]$ is the minimal L^2 approximation error achievable with linear functions of Z can be found in Agresti [2015].

Proposition 3.5.3. (Value of the benchmark)
The e-power benchmark $\mathbf{B}_{K(X,Z)}$ is well defined, and the minimum is attained at $P_0(Y, X, Z) = \mathcal{N}(\gamma_0^T Z, \sigma_0^2) \otimes K(X, Z)$ with parameters:

$$\gamma_0 = \gamma_1 + \mathbb{E}[ZZ]^+\mathbb{E}[ZX^T]\beta_1 = \gamma_1 + L^T_{X|Z}\beta_1,$$
$$\sigma_0^2 = \sigma_1^2 + \mathbb{E}[(\beta_1^T(X - L_{X|Z}Z))^2],$$

where the expectations are taken w.r.t. $K(X, Z)$.
The minimal value is:

$$\mathbf{B}_{K(X,Z)} = D_{\text{KL}}(Q||P_0) = \frac{1}{2}\ln(1 + \mathbb{E}[(\delta_1^T(X - L_{X|Z}Z))^2]).$$

Further, $n\mathbf{B}_{K(X,Z)}$ is an upper bound on the e-powers of the GRO e-variables for all nulls MX, LR, SPLR and FSLR (with corresponding propensity kernel and covariate distribution) against $\{Q^n\}$, for any number of data points n.

Proof. Using equation (2.1) from Chapter 2, we get

$$\begin{aligned} D_{\mathrm{KL}}(Q||P) &= \mathbb{E}_{Q(X,Z)} D_{\mathrm{KL}}(\mathcal{N}(y|\beta_1^T x + \gamma_1^T z, \sigma_1^2)||\mathcal{N}(y|\gamma^T z, \sigma^2)) \\ &= \frac{1}{2}\mathbb{E}_{Q(X,Z)}\left[\ln(\frac{\sigma^2}{\sigma_1^2}) - 1 + \frac{\sigma_1^2}{\sigma^2} + \frac{(\beta_1^T x + \gamma_1^T z - \gamma^T z)^2}{\sigma^2}\right] \\ &= \frac{1}{2}\left[\ln(\frac{\sigma^2}{\sigma_1^2}) - 1 + \frac{\sigma_1^2}{\sigma^2} + \frac{\mathbb{E}_{(X,Z)}(\beta_1^T x + \gamma_1^T z - \gamma^T z)^2}{\sigma^2}\right]. \end{aligned}$$

As we have seen before, we can minimize over γ first by solving a simple least squares problem, we get the minimizer in γ:

$$\gamma_0 = \gamma_1 + \mathbb{E}[ZZ^T]^+\mathbb{E}[ZX^T]\beta_1.$$

Then, minimizing over σ^2 yields

$$\sigma_0^2 = \sigma_1^2 + \mathbb{E}[(\beta_1^T(X - L_{X|Z}Z))^2].$$

For the final statement we simply need to notice that

$$nD_{\mathrm{KL}}(Q||P_0) = D_{\mathrm{KL}}(Q^n||P_0^n).$$

Since P_0^n is an element of all: $\mathcal{H}^{\mathrm{MX}}_{0|n,Q(X|Z)}$, $\mathcal{H}^{\mathrm{LR}}_{0|n}$, $\mathcal{H}^{\mathrm{SPLR}}_{0|n,Q(X|Z)}$ and $\mathcal{H}^{\mathrm{FSLR}}_{0|n,Q(X,Z)}$, the uper bound holds for all their GRO e-variables by the duality from Theorem 2.2.5. □

Definition 3.5.4. (Asymptotic relative power)
Let $(S_n)_{n\in\mathbb{N}}$ be a sequence of e-variables each defined on n i.i.d. data points. The asymptotic relative power (ARP) of $(S_n)_n$ under distribution Q is defined as

$$\mathrm{ARP}(S_n, Q) := \lim_n \mathbb{E}_{Q^n}\left[\frac{\ln(S_n)}{n}\right],$$

whenever the limit exists.

Sometimes we will simplify the notation further by putting the name of a null hypothesis in the first argument. This should be understood as the ARP of the GRO e-variable for the specified null and alternative.

It is important to keep in mind that the ARP is a very crude measure of performance of an e-variable. Firstly, it only depends on the asymptotic behavior, and thus might favor e-variables whose performance dominates in settings with unrealistically large sample sizes, but are suboptimal on small samples. The second, and possibly more important reason, is that all differences in growth of order $o(n)$ become invisible. Hao and Grünwald

[2024] compare e-powers of several e-variables, including the GRO and sequential GRO, but also others such as the conditional e-variable and universal inference (which are outside of the scope of this thesis). In their setting, all e-variables were designed for testing the same null against the same alternative, and they showed that the differences in e-power are indeed very often of the order $o(n)$.
For example, the GRO e-variable S_n^{GRO} (which is the most e-powerful by definition) could be logarithmically more e-powerful than another e-variable S_n:

$$\mathbb{E}_{Q^n}[\ln S_n^{\text{GRO}}] \approx \mathbb{E}_{Q^n}[\ln S_n] + \ln(n).$$

With ARP alone as the measure of performance, we could not distinguish the superiority of the GRO in this case.
Luckily for us, we are comparing very different e-variables. $S_n^{\text{S-LR}}$ and S_n^{MX} were originally designed with different null hypotheses in mind, and an analysis of ARP will in fact be enough to distinguish the more powerful among them in a wide range of situations.

In the following theorem, we show that in terms of asymptotic relative power, $S_n^{\text{S-LR}}$ reaches the benchmark, and thus performs as well as the GRO e-variable for the Fully Specified Linear Regression null (3.4.1). As we have seen in Section 3.4, in general we do not know how this e-variable looks, but we do know that it exists, and is a.s. unique under the alternative (Theorem 2.2.4). On the other hand, the MX e-variable does not reach the benchmark, except for a very narrow collection of alternatives.
Before showing the results, we need a few definitions.

Definition 3.5.5. (e-power analysis conditions)
A distribution $Q(Y, X, Z) = \mathcal{N}(\beta_1^T X + \gamma_1^T Z, \sigma_1^2) \otimes Q(X, Z)$ is said to satisfy:

1. "*The second moment condition*" if the vector (X, Z) has a finite second moment under $Q(X, Z)$.

2. "*The conditional Gaussian condition*" if $Q(\beta_1^T X | Z = z) \sim \mathcal{N}(\mu(z), \sigma^2)$ $Q(Z)$-a.s. for some measurable function μ and constant σ^2.

3. "*The linear conditional Gaussian condition*" if it satisfies (2) with a linear mean function $\mu(z) = Lz$.

Theorem 3.5.6. (S-LR, MX and FSLR ARPs)
Let $Q(Y, X, Z) = \mathcal{N}(\beta_1^T X + \gamma_1^T Z, \sigma_1^2) \otimes Q(X, Z)$ be the fixed alternative, and assume it satisfies the second moment condition. Then, the asymptotic relative powers of S-LR and MX e-variables are:

$$\text{ARP}(\text{S-LR}, Q) = \lim_n \mathbb{E}_{Q^n}\left[\frac{\ln(S_n^{\text{S-LR}})}{n}\right] = \frac{1}{2}\ln(1 + \mathbb{E}[(\delta_1^T(X - L_{X|Z}Z))^2]) = \mathbf{B}_{Q(X,Z)},$$

$$\text{ARP}(\text{MX}, Q) = I(Y; X|Z) \leq \frac{1}{2}\ln(1 + \mathbb{E}[(\delta_1^T(X - \mathbb{E}[X|Z]))^2]).$$

Furthermore,

$$\mathrm{ARP}(\text{S-LR}, Q) = \mathrm{ARP}(\text{FSLR}, Q) \geq \mathrm{ARP}(\text{MX}, Q)$$

holds for any such distribution Q.
The ARPs of all three e-variables are exactly equal if and only if Q satisfies the linear conditional Gaussian condition.

Proof. The MX ARP statement follows from Proposition 3.2.4. Now, we show that the ARP of S-LR is at least as big as the benchmark. The statement follows from Proposition 3.1.5 and Fatou's lemma, which we use to take the limit inside the expectation:

$$\begin{aligned}
&\lim_n \mathbb{E}_Q\left[\frac{\ln(S_n^{\text{S-LR}})}{n}\right] = \lim_n \mathbb{E}_{Q(X,Z)}\left[\frac{1}{2}\ln\left(1+\delta_1^T\frac{X^{n^T}(I_n - P_{Z^n})X^n}{n}\delta_1\right)\right] \\
&\geq \mathbb{E}_{Q(X,Z)}\left[\lim_n \frac{1}{2}\ln\left(1+\delta_1^T\frac{X^{n^T}(I_n - P_{Z^n})X^n}{n}\delta_1\right)\right] \\
&= \mathbb{E}_{Q(X,Z)}\left[\lim_n \frac{1}{2}\ln\left(1+\delta_1^T\frac{X^{n^T}(I_n - Z^n(Z^{n^T}Z^n)^+Z^{n^T})X^n}{n}\delta_1\right)\right] \\
&= \mathbb{E}_{Q(X,Z)}\left[\frac{1}{2}\ln(1+\delta_1^T(\mathbb{E}[XX^T] - \mathbb{E}[XZ^T]\mathbb{E}[ZZ^T]^+\mathbb{E}[ZX^T])\delta_1)\right] \\
&= \frac{1}{2}\ln(1+\delta_1^T(\mathbb{E}[XX^T] - \mathbb{E}[XZ^T]\mathbb{E}[ZZ^T]^+\mathbb{E}[ZX^T])\delta_1) \\
&= \frac{1}{2}\ln(1+\mathbb{E}[(\delta_1^T(X - L_{X|Z}Z))^2]) \\
&= \mathbf{B}_{Q(X,Z)}.
\end{aligned}$$

To calculate the almost sure limit, we used the law of large numbers, finite second moments of (X, Z) and a result by Rakočević [1997], which states that $A_n \to A$ implies $(A_n)^+ \to A^+$ in Frobenius norm, whenever the sequence $(A_n)_{n\in\mathbb{N}}$ has a constant rank, equal to that of A from some $n_0 \in \mathbb{N}$ onwards. This assumption is almost surely true in our case, as with $Q(Z^\infty)$-probability 1, the matrix Z^n must eventually have rank equal to the number of linearly independent components of the random vector Z [Van der Vaart, 2000, Chapter 5]. From that point onwards, the rank of $Z^{n^T}Z^n$ coincides with that of $\mathbb{E}[ZZ^T]$.
Furthermore,

$$\mathrm{ARP}(\text{S-LR}, Q) \leq \mathrm{ARP}(\text{FSLR}, Q) \leq \mathbf{B}_{Q(X,Z)}.$$

The first inequality holds because the fully specified null is a subset of the linear regression null, implying that the GRO e-variable of the latter is at least as e-powerful as that of the former. The second holds by Proposition 3.5.3.
Thus, the GRO e-variable for the FSLR null also reaches the benchmark asymptotically. To get equal e-power with MX, we need first $\mathbb{E}[X|Z] = L_{X|Z}Z$ $Q(Z)$-a.s. And second, the mutual information $I(Y; X|Z)$ needs to be exactly equal to the upper bound, which,

as we have seen in Proposition 3.2.4, only happens if $\beta_1^T X|Z=z$ is Gaussian with fixed variance $Q(Z)$-a.s. □

The theorem shows an interesting fact. In terms of asymptotic relative power, S-LR reaches the highest possible benchmark - same performance as the GRO e-variable for the FSLR null. In other words, when it comes to asymptotic relative power, knowing the true distribution of the covariates does not help us to create better e-variables. Furthermore, it shows that the ARP of the MX e-variable is always lower or equal to that of the S-LR e-variable, and that the gap can appear for exactly two reasons:

1. The value $\mathbb{E}[(\delta_1^T(X-\mathbb{E}[X|Z]))^2]$ is smaller than $\mathbb{E}[(\delta_1^T(X-L_{X|Z}Z))^2]$, i.e. $\mathbb{E}[\delta_1^T X|Z]$ is a non-linear function of Z.

2. The actual conditional mutual information $I(Y;X|Z)$ is small compared to the maximal value it could achieve under fixed $\mathrm{Cov}(X-\mathbb{E}[X|Z])$; which is $\frac{1}{2}\ln(1+\mathbb{E}[(\delta_1^T(X-\mathbb{E}[X|Z]))^2])$.

If neither of these two conditions hold then there is no gap. In fact, with finite samples there even exists a class of alternative distributions under which MX has strictly larger non-asymptotic e-power than S-LR.

Proposition 3.5.7. (sufficient condition for MX dominance)
Let $Q(Y,X,Z)=\mathcal{N}(\beta_1^T X+\gamma_1^T Z,\sigma_1^2)\otimes Q(X,Z)$. Assume Q satisfies the second moment, and linear conditional Gaussian conditions, and that Z is not $Q(Z)$-a.s. equal to 0.
Let $\{Q^n\}$ be the fixed alternative against which we test. Then, S_n^{MX} is strictly more e-powerful than $S_n^{\text{S-LR}}$.

Proof. The finite sample e-powers of S-LR and MX e-variables are given by (Propositions 3.1.5 and 3.2.4):

$$\mathbb{E}_Q[\ln(S_n^{\text{S-LR}})]=\mathbb{E}_{Q(X^n,Z^n)}\left[\frac{n}{2}\ln\left(1+\delta_1^T\frac{X^{n^T}(I_n-P_{Z^n})X^n}{n}\delta_1\right)\right],$$
$$\mathbb{E}_Q[\ln(S_n^{\mathrm{MX}})]=nI(Y;X|Z).$$

On top of that, finite second moments of (X,Z) and Jensen's inequality ensure the following upper bounds

$$\mathbb{E}_Q[\ln(S_n^{\text{S-LR}})]\leq\frac{n}{2}\ln\left(1+\delta_1^T\frac{\mathbb{E}_{Q(X^n,Z^n)}[X^{n^T}(I_n-P_{Z^n})X^n]}{n}\delta_1\right),$$
$$\mathbb{E}_Q[\ln(S_n^{\mathrm{MX}})]=nI(Y;X|Z)\leq\frac{n}{2}\ln(1+\delta_1^T\mathrm{Cov}(X-\mathbb{E}[X|Z])\delta_1).$$

For MX, the upper bound is actually reached because of the linear conditional Gaussian assumption.

For S-LR, the upper bound becomes

$$\frac{n}{2}\ln\left(1+\left(1-\frac{\mathbb{E}[r(Z^n)]}{n}\right)\delta_1^T\mathrm{Cov}(X-\mathbb{E}[X|Z])\delta_1\right) < nI(Y;X|Z),$$

where $r(Z^n)$ is the rank of Z^n. The expected value of the rank is larger than 0 because Z is not a.s. 0. Thus, the MX e-variable is more e-powerful.
The bound is proved as follows:

$$\begin{aligned}(X^n\delta_1)^T(I_n-P_{Z^n})(X^n\delta_1) &= \sum_{i,j}(\delta_1^TX_i)(\delta_1^TX_j)(I_n-P_{Z^n})_{i,j}\\ &= \sum_{i=1}^n(\delta_1^TX_i)^2(I_n-P_{Z^n})_{i,i}+\sum_{i\neq j}(\delta_1^TX_i)(\delta_1^TX_j)(I_n-P_{Z^n})_{i,j}.\end{aligned}$$

Applying the conditional expectation $\mathbb{E}[\cdot|Z^n]$ to the RHS, and using $X_i \perp X_j|Z^n$ whenever $i\neq j$, we get

$$\begin{aligned}&\sum_{i=1}^n(\mathrm{Var}(\delta_1^TX_i|Z_i)+\mathbb{E}[\delta_1^TX_i|Z_i]^2)(I_n-P_{Z^n})_{i,i}+\sum_{i\neq j}\mathbb{E}[\delta_1^TX_i|Z_i]\mathbb{E}[\delta_1^TX_j|Z_j](I_n-P_{Z^n})_{i,j}\\ &=\sum_{i=1}^n(I_n-P_{Z^n})_{i,i}Var(\delta_1^TX_i|Z_i)+\mathbb{E}[X^n\delta_1|Z^n]^T(I_n-P_{Z^n})\mathbb{E}[X^n\delta_1|Z^n].\end{aligned}$$

Now, using that the conditional variance is assumed to be independent of Z, and that the conditional expectation is linear in Z, the final expression becomes

$$\sum_{i=1}^n(I_n-P_{Z^n})_{i,i}\mathrm{Var}(\delta_1^TX-\mathbb{E}[\delta_1^TX|Z]) = (n-r(Z^n))\mathrm{Var}(\delta_1^TX-\mathbb{E}[\delta_1^TX|Z]).$$

□

Importantly, all assumptions of the previous proposition are satisfied if Z contains an intercept term, and the remaining covariates in (X,Z) form a non-degenerate multivariate Gaussian. This example shows that MX dominance is not tied to trivial corner cases. Instead, the MX e-variable can be more e-powerful than S-LR on finite samples in practically relevant settings.

Now, we move to the special case where Z contains an intercept as above, and the covariates X and Z are independent, but not necessarily Gaussian. This setting is practically important as it often appears in clinical trials. On top of that, in Corollary 3.4.5 we have seen that the MX e-variable has a very specific interpretation in this case - it is the sequential GRO e-variable for testing the Fully Specified LR null hypothesis. Therefore, we would expect it to perform relatively well compared to the GRO e-variable for the FSLR null, and consequently also compared to the (weaker) S-LR e-variable.

Theorem 3.5.8. (e-power analysis of MX as sequential GRO for FSLR)
Let $Q(Y, X, Z) = \mathcal{N}(\beta_1^T X + \gamma_1^T Z, \sigma_1^2) \otimes Q(X) \otimes Q(Z)$. Assume that Q satisfies the finite second moment condition. Furthermore, $Q(Z)$ satisfies the intercept condition (3.4.3), and $\mathbb{E}_Q[(\delta_1^T X)^6] < \infty$. Then, the difference in ARP between the S-LR and MX e-variables scales as $O(||\delta_1||^6)$ when the effect size vector δ_1 tends to 0.

Proof. Due to the assumed independence and linearity of Y in (X, Z), conditioning on $Z = z$ only translates the expected value of Y and has no effect on X, implying that $I(Y; X|Z) = I(Y; X|Z = 0)$.
Furthermore, using the invariance property of the mutual information w.r.t. bijective transformations, we get

$$I(Y; X|Z = 0) = I(\beta_1^T X + \varepsilon; \beta_1^T X) = I(\delta_1^T X + \varepsilon'; \delta_1^T X),$$

where $\varepsilon' \sim \mathcal{N}(0, 1)$.
Let us define the function $g(\gamma) := I(\sqrt{\gamma}\delta_1^T X + \varepsilon'; \delta_1^T X)$ on $[0, \infty)$. The parameter γ is called the signal-to-noise ratio. As described by Guo et al. [2011], g is differentiable infinitely many times at all points $\gamma > 0$, and the first six derivatives can be continuously extended to $\gamma = 0$ as a consequence of our assumption $E|\delta_1^T X|^6 < \infty$. To simplify notation, we further define the i-th conditional central moment

$$M_i(\gamma) := \mathbb{E}[(\delta_1^T X - \mathbb{E}[\delta_1^T X|\sqrt{\gamma}\delta_1^T X + \varepsilon'])^i|\sqrt{\gamma}\delta_1^T X + \varepsilon']$$

for $i \leq 6$.
The first three derivatives of g are then given by

$$g^{(n)}(\gamma) = \frac{-1^{n+1}}{2}\mathbb{E}[M_2^n(\gamma)] = \frac{-1^{n+1}}{2}\mathbb{E}[Var(\delta^T X|\sqrt{\gamma}\delta^T X + \varepsilon')^n]$$

for $n = 1, 2$, and

$$g'''(\gamma) = \mathbb{E}[M_2^3(\gamma) - \frac{1}{2}M_3^2(\gamma)].$$

Thus, by a Taylor series expansion we have

$$\begin{aligned} I(Y; X|Z = 0) = g(1) &= g(0) + g'(0) + \frac{g''(0)}{2} + \frac{g'''(\xi)}{6} \\ &= \frac{1}{2}\delta_1^T \mathrm{Cov}(X)\delta_1 - \frac{(\delta_1^T \mathrm{Cov}(X)\delta_1)^2}{4} + \frac{g'''(\xi)}{6}, \end{aligned}$$

for some $\xi \in (0,1)$. All that is left is to bound the error term:

$$\begin{aligned}
|g'''(\gamma)| &\leq \mathbb{E}[M_2^3(\gamma)] + \frac{1}{2}\mathbb{E}[|M_3(\gamma)|^2] \leq \frac{3}{2}\mathbb{E}[M_6(\gamma)] \\
&= \frac{3}{2}\mathbb{E}[(\delta_1^T X - \mathbb{E}[\delta_1^T X|\sqrt{\gamma}\delta_1^T X + \varepsilon'])^6] \\
&\leq \frac{3}{2}2^5\mathbb{E}[(\delta_1^T X)^6 + (\mathbb{E}[\delta_1^T X|\sqrt{\gamma}\delta_1^T X + \varepsilon'])^6] \\
&\leq 48\mathbb{E}[(\delta_1^T X)^6 + \mathbb{E}[(\delta_1^T X)^6|\sqrt{\gamma}\delta_1^T X + \varepsilon']] \\
&= 96\mathbb{E}[(\delta_1^T X)^6] = O(||\delta_1||^6).
\end{aligned}$$

Finally, the difference between the S-LR and MX ARPs is then also $O(||\delta_1||^6)$. The ARP of the S-LR e-variable in this setting is given by

$$\frac{1}{2}\ln(1 + \delta_1^T \mathrm{Cov}(X)\delta_1).$$

This is a direct consequence of Theorem 3.5.6, and the fact that $L_{X|Z}Z = \mathbb{E}[X]$. The latter holds simply because of the independence $X \perp_Q Z$ and the intercept condition. The first two terms in $I(Y, X|Z = 0)$ and in the Taylor series expansion of $\frac{1}{2}\ln(1 + \delta_1^T \mathrm{Cov}(X)\delta_1)$ coincide, while the remaining part of both expressions converges to 0 at the rate $O(||\delta_1||^6)$. □

The most important representative of the setting of Theorem 3.5.8 is ANCOVA testing (see Section 3.3). Our findings suggest that for small effect size vectors δ_1 and large samples, the MX and S-LR e-variables should have a similar performance on this task. In Chapter 4, after introducing methods for composite alternatives, we empirically assess this conjecture.

3.6 Cheat sheet

The Model-X and Simple Linear Regression e-variables (S_n^{MX} and $S_n^{\text{S-LR}}$) are the growth-rate-optimal (GRO) e-variables for testing their nulls $\mathcal{H}_{0|n,K(X|Z)}^{\text{MX}}$ and $\mathcal{H}_{0|n}^{\text{LR}}$ against point alternatives from their extended models $\mathcal{H}_{1|n,K(X|Z)}^{\text{MX}}$ and $\mathcal{H}_{1|n}^{\text{LR}}$.
They are both suboptimal e-variables for testing the intersection of their nulls, denoted by $\mathcal{H}_{0|n,K(X|Z)}^{\text{SPLR}}$. The Specified Propensity Linear Regression null is used to represent a testing scenario in which all: linearity of Y in (X, Z), i.i.d. covariates, and a known propensity kernel are assumed. This is a practically important setting because it represents the information which is usually available (or assumed) in clinical trials.
When testing against a point alternative $\{Q^n\} \subset \mathcal{H}_{1|n,K(X|Z)}^{\text{SPLR}}$ of the form

$$Q(Y, X, Z) = \mathcal{N}(\beta_1^T X + \gamma_1^T Z, \sigma_1^2) \otimes K(X|Z) \otimes Q(Z),$$

the MX and S-LR e-variables are given by:

$$S_n^{\text{MX}} = \prod_{i=1}^{n} \frac{\mathcal{N}(Y_i|\beta_1^T X_i + \gamma_1^T Z_i, \sigma_1^2)}{\mathbb{E}_{K(X|Z_i)}[\mathcal{N}(Y_i|\beta_1^T X + \gamma_1^T Z_i, \sigma_1^2)]},$$
$$S_n^{\text{S-LR}} = \frac{\mathcal{N}(Y^n|X^n\beta_1 + Z^n\gamma_1, \sigma_1^2 I_n)}{\mathcal{N}(Y^n|Z^n\gamma_0(X^n, Z^n), \sigma_0^2(X^n, Z^n) I_n)},$$

with

$$\gamma_0(X^n, Z^n) = \gamma_1 + (Z^n)^+ X^n \beta_1,$$
$$\sigma_0^2(X^n, Z^n) = \sigma_1^2 + \frac{||(I_n - P_{Z^n})X^n\beta_1||^2}{n}.$$

In general, we can not find the GRO e-variable for testing the SPLR null against relevant alternatives directly, and have no better option (to the best of our knowledge) than using MX or S-LR.
Even if we assume complete knowledge of the distribution of covariates $K(X, Z)$, which is encoded in the Fully Specified Linear Regression null - $\mathcal{H}_{0|n,K(X,Z)}^{\text{FSLR}}$, we are still unable to find the GRO e-variable. Nevertheless, analyzing the FSLR null gives us a way to compare the asymptotic relative power (ARP) of the two currently available e-variables. The S-LR e-variable has the same ARP as the GRO e-variable for the FSLR null, meaning that in terms of ARP, knowledge of the true distribution of covariates can not help us to get a more (e-)powerful e-variable. The MX e-variable has a strictly lower ARP in almost all testing scenarios. On finite samples however, there do exists settings in which MX dominates S-LR.
Whenever $X \perp Z$ and Z contains an intercept - such as in the ANCOVA test, the MX e-variable coincides with the sequential GRO e-variable for testing the FSLR null, and the difference in ARP between the S-LR and MX e-variables is of the order $O(||\delta_1||^6)$. This means that for small effect sizes and large samples, the e-power difference is negligible.

4 Composite alternatives

In this chapter, we show how one can adapt the S-LR and MX e-variables from Chapter 3, and make them useful for LR testing in practice, where the alternative hypothesis is composite.

Outline:
We start with a detailed description of the construction of GROW and REGROW (see Definition 4.1.1) e-variables for fixed-design LR testing, based on Pérez-Ortiz et al. [2024], and show that their optimality properties are directly transferred to random design. We go on to describe a method for extending the MX e-variable to composite alternatives, and finish the chapter with a simulation study, which empirically assesses the e-powers of the constructed test supermartingales in an ANCOVA setting.

4.1 Introduction: Point vs. Composite Alternatives

Before diving into the details, let us shortly discuss the relation and main differences between testing point and composite alternative hypotheses with e-variables.
The e-variable property depends exclusively on the null hypothesis, meaning that the set of e-variables is invariant w.r.t. the choice and cardinality of the alternative. Under a point alternative $\{Q\}$, the (Q-a.s.) unique "best" e-variable for our purposes is the GRO e-variable (see Section 2.2). Finding closed-form expressions for the RIPr and GRO e-variable is impossible in general, but nevertheless, we have a clear definition of what it means for an e-variable to be "the best" for testing $\mathcal{H}_0$ against $\{Q\}$. The same definition no longer makes sense if the alternative hypothesis is composite, as the GRO e-variables for different choices of Q generally differ. Although we do not necessarily have a precise optimality definition, it is clear that we would like to find e-variables which have a similar e-power to the (oracle) GRO e-variable for Q, under all $Q \in \mathcal{H}_1$. There are multiple ways to pursue this goal.

Approach 1: Estimating or Mixing The first idea is to completely avoid making a new optimality definition for composite alternatives, and instead use a combination of optimal e-variables for relevant point alternatives.
For example, we can use the data to pinpoint a single distribution $\hat{Q}$ from the alternative hypothesis which is most likely to be the true generator of the data. Then, simply use the $\hat{Q}$-GRO e-variable for testing. This approach can be dangerous, as using the same

data twice, once for estimating $\hat{Q}$ and again for testing, can lead to a loss of type-1 error guarantees. The issue can be remedied, e.g. by separating the dataset into a part for estimating the point alternative and a part for testing. A similar, but often more practical approach, is to use the prequential plug-in method (Chapter 2, (2.5)), which allows a sequential estimation of $\hat{Q}$ and testing as new data points arrive.
Another way to go about this problem is to mix the collection of optimal e-variables for all distributions in the alternative hypothesis using the method of mixtures (Chapter 2, (2.7)).

The process of finding a good e-variable for testing a composite null against a composite alternative can thus often be divided into two steps. First, solve the oracle testing version of the problem, i.e. find the GRO e-variable for testing $\mathcal{H}_0$ against point alternatives $\{Q\}$ from $\mathcal{H}_1$. Then, combine those e-variables either by estimating the true alternative over time, or by taking mixtures.

Approach 2: Redefine best The second option is to come up with a new definition of "best" e-variable for testing composite alternative hypotheses. Two game-theoretic definitions were given by Grünwald et al. [2024]:

Definition 4.1.1. (GROW and REGROW optimality criteria)
A growth rate optimal in the worst case (GROW) e-variable for testing $\mathcal{H}_0$ against $\mathcal{H}_1$ is defined as any e-variable $S^* \in \mathcal{E}(\mathcal{H}_0)$ which satisfies

$$\mathbb{E}_Q[\ln S^*] = \max_{S \in \mathcal{E}(\mathcal{H}_0)} \inf_{Q \in \mathcal{H}_1} \mathbb{E}_Q[\ln S].$$

Similarly, the relative growth rate optimal in the worst case (REGROW) e-variable is one that satisfies

$$\mathbb{E}_Q[\ln S^*] = \max_{S \in \mathcal{E}(\mathcal{H}_0)} \inf_{Q \in \mathcal{H}_1} (\mathbb{E}_Q[\ln S] - \mathbb{E}_Q[\ln S_Q^{\mathrm{GRO}}]),$$

where S_Q^{GRO} is the GRO e-variable for testing $\mathcal{H}_0$ against $\{Q\}$.

A GROW e-variable maximizes the worst-case e-power over the alternative, while the REGROW e-variable minimizes the worst-case gap between its own e-power and that of an oracle.

The "mixture/estimating" vs "redefine best" approaches are not necessarily incompatible: Grünwald et al. [2024] show that the GROW and REGROW e-variables can often be re-interpreted in terms of putting a prior on the alternative hypothesis and then taking the GRO e-variable relative to the Bayesian marginal distribution corresponding to that prior. If one allows improper priors, the e-variables shown in Section 4.2 can be arrived at in such a way as well. The details of this approach are discussed by Pérez-Ortiz et al. [2024].

In practice it is not always clear which optimization criterion GROW or REGROW one should settle for. Luckily, whereas in general the two criteria can lead to very different e-variables [Grünwald et al., 2024] in the setting in which we will apply this criterion, it turns out that both approaches are equivalent: in Section 4.2 we design the counterparts to the Simple LR e-variable for composite alternatives using the GROW and REGROW criteria, and we will see that the resulting e-variables coincide.
On the other hand, the composite alternative version of the MX e-variable described in Section 4.3 is formed by combining optimal e-variables for point alternatives, using a method that, in a sense, combines the prequential plug-in method and the method of mixtures.

4.2 Group LR test

The results of this section heavily rely on group theory. We refer to Section 2.3 for the necessary background.

4.2.1 Fixed design

The Group LR test (G-LR) can be thought of as an extension of the Simple LR test for composite alternatives. The name "Group" comes from the importance of several theorems connecting group theory and statistics [Pérez-Ortiz et al., 2024, Lindon et al., 2024] in the design of the e-variable. As usual, we start by giving fixed-design results, so let us fix a full column rank design matrix (x^n, z^n) right away.

Definition 4.2.1. (The location-scale group and the actions)
The group $G_{z^n} = (\mathbb{R}_+ \times \mathbb{R}^{d_z}, \cdot)$ with operation

$$(s_1, g_1) \cdot (s_2, g_2) = (s_1 s_2, s_1 g_2 + g_1)$$

is called the location-scale (LS) group. We further endow G_{z^n} with group actions over observed data, and over distributions of the data

$$(s, g) \cdot y^n := s y^n + z^n g,$$

$$(s, g) \cdot Q(Y^n) := Q(sY^n + z^n g).$$

Specifically, for distributions from the linear model $\mathcal{H}_{1|x^n,z^n}$ (Chapter 2, (2.13)) we get

$$(s, g) \cdot \mathcal{N}(x^n \beta + z^n \gamma, \sigma^2 I_n) = \mathcal{N}(x^n s\beta + z^n (s\gamma + g), s^2 \sigma^2 I_n).$$

The described actions are two different ways of thinking about the same transformation - on the level of the observed data, or on the level of the underlying distributions.

By Definition 2.3.3, the data and distribution orbits of the location-scale group are given by

$$O(y^n) := \{(s,g) \cdot y^n : (s,g) \in G\},$$
$$O(Q) := \{Q(sY^n + z^n g) : (s,g) \in G\}.$$

Specifically, the distribution orbit of $\mathcal{N}(x^n\beta_1 + z^n\gamma_1, \sigma_1^2 I_n)$ is given by

$$\begin{aligned}\mathcal{H}_{1|x^n,z^n,\delta_1} &:= \{(s,g) \cdot \mathcal{N}(x^n\delta_1\sigma_1 + z^n\gamma_1, \sigma_1^2 I_n) : (s,g) \in G\} \\ &= \{\mathcal{N}(x^n\delta_1\sigma + z^n\gamma, \sigma^2 I_n) : \sigma^2 > 0, \gamma \in \mathbb{R}^{d_z}\}.\end{aligned}$$

It is easy to see that the distribution orbit of any element of $\mathcal{H}_{1|x^n,z^n}$ depends exclusively on the effect size vector δ_1, so we may also refer to it as the *distribution orbit of* δ_1.

A maximally invariant statistic (see Definition 2.3.4) w.r.t. the G_{z^n} group is given by [Pérez-Ortiz et al., 2024]

$$m_n(y^n) = \frac{(I_n - P_{z^n})y^n}{||(I_n - P_{z^n})y^n||}, \tag{4.1}$$

where P_{z^n} is the projection matrix onto the column space of z^n.

The following proposition shows the deep connection between the location-scale group and optimal e-variables for LR testing in the GROW and REGROW sense.

Proposition 4.2.2. (Testing one Gaussian orbit against another)
Suppose that P_n and Q_n are non-degenerate multivariate Gaussian distributions of the data Y^n. Let p^{m_n} and q^{m_n} be the corresponding density functions of some maximally invariant statistic of the data, e.g. (4.1). Then both the GROW and REGROW e-variables for testing G_{z^n} orbits:

$$O(P_n) \text{ vs } O(Q_n)$$

are given by the likelihood ratio:

$$S_n(Y^n) := \frac{q^{m_n}(m_n(Y^n))}{p^{m_n}(m_n(Y^n))}.$$

Furthermore, if z^∞ is fixed, and P_∞, Q_∞ are distributions of the whole Gaussian process $(Y_n)_{n=1}^\infty$ with non-degenerate finite-dimensional distributions, then $(S_n(Y^n))_{n=1}^\infty$ is a test supermartingale under the filtration $\mathcal{F}_M := (\sigma(m_i(Y^i) : i \leq n))_n$ generated by the maximally invariant statistic.

The proposition was proved for a far more general setting than Gaussian orbit testing by Pérez-Ortiz et al. [2024].
Choosing $P_n = \mathcal{N}(0, I_n)$ and $Q_n = \mathcal{N}(x^n\delta_1, I_n)$ in Proposition 4.2.2 immediately gives us the GROW and REGROW e-variable for testing the linear null $\mathcal{H}_{0|z^n}$ against a specified orbit $\mathcal{H}_{1|x^n,z^n,\delta_1}$ in the alternative. The density function of the maximally invariant

statistic (4.1) under Gaussian data is very hard to work with [Pérez-Ortiz et al., 2024, Lindon et al., 2024], but it nevertheless brings us one step closer to being able to (effectively) test the linear null against the entire composite alternative.

Quite incredibly, and unlike the stochastic process generated by the S-LR e-variable $(S_n^{\text{S-LR}})_n$, the sequence of e-variables

$$(\frac{q(m_n(Y^n))}{p(m_n(Y^n))})_n$$

from the previous proposition, does form a test supermartingale. The supermartingale property holds only under a coarser filtration, generated by the maximally invariant statistic, as opposed to the whole data, but nevertheless, this allows safe sequential testing with the usual, greedy stopping time

$$\tau = \min\{n \in \mathbb{N} : \frac{q(m_n(Y^n))}{p(m_n(Y^n))} \geq \frac{1}{\alpha}\}.$$

This is possible because the stopping rule only uses the data through the maximally invariant statistic, or more precisely, because τ is a stopping time under $\mathcal{F}_M$ as well [Pérez-Ortiz et al., 2024, Appendix C].

To implement a composite alternative LR test in practice, there are two questions that still need to be answered:

- Can we find a different maximally invariant statistic with a possibly simpler probability density function?
- How to extend the findings of the previous proposition to get a useful e-variable for testing the null against the full linear alternative, instead of the orbit corresponding to one particular δ?

To simplify the first problem, Pérez-Ortiz et al. [2024] show that under mild assumptions - which are satisfied in our setting, the statements of their theorems, and thus also of Proposition 4.2.2 still hold if instead of working with the original data, one summarizes it using a sufficient statistic (see Section 2.1). Thus, instead of working with the raw data Y^n, we may find a convenient sufficient statistic $T^n(Y^n)$ which induces a simpler maximally invariant statistic. In case of linear regression testing, it is not very hard to do so. We can use the sufficient statistic based on the ordinary least squares estimators [Casella and Berger, 2024]:

$$T^n(Y^n) := (\hat{\beta}_{\text{OLS}}(Y^n), \hat{\gamma}_{\text{OLS}}(Y^n), ||(I_n - P_{w^n})Y^n||^2),$$

where $w^n = (x^n, z^n)$. The estimator of the effect size vector used in classical LR testing is

$$\hat{\delta}(Y^n) = \frac{\hat{\beta}_{\text{OLS}}}{\hat{\sigma}} = \frac{\hat{\beta}_{\text{OLS}}(Y^n)}{\sqrt{\frac{||(I_n - P_{w^n})Y^n||^2}{n - d_w}}},$$

and a short proof of the fact that $\hat{\delta}$ is a maximally invariant statistic of $T^n(Y^n)$ is given by Lindon et al. [2024].

Furthermore, if $Y^n \sim \mathcal{N}(x^n\beta_1 + z^n\gamma_1, \sigma_1^2 I_n)$, then

$$\hat{\beta}_{\text{OLS}} \sim \mathcal{N}(\beta_1, \sigma_1^2(x^{n^T}(I_n - P_{z^n})x^n)^{-1}) = \sigma_1\mathcal{N}(\delta_1, (x^{n^T}(I_n - P_{z^n})x^n)^{-1}),$$
$$\hat{\sigma}^2 \sim \frac{\sigma_1^2}{n - d_w}\chi^2(n - d_w),$$

and importantly, the estimators are independent [Casella and Berger, 2024]. Thus, the maximally invariant estimator is a ratio of a Gaussian and the square-root of an independent (and properly scaled) χ^2-distributed variable. This by definition means it has a non-central t-distribution:

$$\hat{\delta}(Y^n) \sim t^{\text{nc}}_{n-d_w}(\delta_1, (x^{n^T}(I_n - P_{z^n})x^n)^{-1}).$$

Therefore, the simple likelihood ratio

$$S_n^{\text{GO}} := \frac{t^{\text{nc}}_{n-d_w}(\hat{\delta}(Y^n)|\delta_1, (x^{n^T}(I_n - P_{z^n})x^n)^{-1})}{t^{\text{nc}}_{n-d_w}(\hat{\delta}(Y^n)|\delta_0, (x^{n^T}(I_n - P_{z^n})x^n)^{-1})} \tag{4.2}$$

is the GROW and REGROW e-variable for testing Gaussian orbits $\mathcal{H}_{1|x^n,z^n,\delta_0}$ against $\mathcal{H}_{1|x^n,z^n,\delta_1}$. The non-central t-density cannot be evaluated via a closed formula, but it can be numerically approximated with high accuracy. Thus, the described e-variable can easily be computed and used for (sequential) testing in practice.

To test the LR null against the full alternative, Pérez-Ortiz et al. [2024] propose to first mix over the unknown effect size vector δ. So, let us fix a prior distribution $p(\delta) = \mathcal{N}(\delta|0, \Lambda^{-1})$, and marginalize the parameter out

$$\int \mathcal{N}(y^n|x^n\sigma_1\delta + z^n\gamma_1, \sigma_1^2 I_n)p(\delta)d\delta = \mathcal{N}(y^n|z^n\gamma_1, \sigma_1^2(I_n + x^n\Lambda^{-1}x^{n^T})).$$

This leads to a new alternative hypothesis

$$\begin{aligned}\mathcal{H}_{1|x^n,z^n,\Lambda^{-1}} &:= \{\mathcal{N}(z^n\gamma, \sigma^2(I_n + x^n\Lambda^{-1}x^{n^T}) : \sigma^2 > 0, \gamma \in \mathbb{R}^{d_z}\} \\ &= \{(s, g) \cdot \mathcal{N}(0, I_n + x^n\Lambda^{-1}x^{n^T}) : (s, g) \in G\}.\end{aligned}$$

The benefit of choosing a Gaussian prior is now clear; $\mathcal{H}_{1|x^n,z^n,\Lambda^{-1}}$ is again a distribution orbit of a Gaussian, and allows us to reuse Proposition 4.2.2. The same statistic $T^n(Y^n)$ is still sufficient for the alternative. $\hat{\delta}(Y^n)$ is maximally invariant as a function of T^n, but now has a central multivariate t-distribution [Lindon et al., 2024]

$$\hat{\delta}(Y^n) \sim t_{n-d_w}(0, \Lambda^{-1} + (x^{n^T}(I_n - P_{z^n})x^n)^{-1}).$$

Finally, the likelihood ratio

$$S_n^{\text{G-LR}}(Y^n) := \frac{t_{n-d_w}(\hat{\delta}(Y^n)|0, \Lambda^{-1} + (x^{n^T}(I_n - P_{z^n})x^n)^{-1})}{t_{n-d_w}(\hat{\delta}(Y^n)|0, (x^{n^T}(I_n - P_{z^n})x^n)^{-1})} \tag{4.3}$$

is the GROW and REGROW e-variable for testing $\mathcal{H}_{0|z^n}$ against $\mathcal{H}_{1|x^n,z^n,\Lambda^{-1}}$, and we call it the Group LR e-variable. Just like in the first case of Gaussian orbit testing, the sequence $(S_n^{\text{G-LR}})_{n=1}^{\infty}$ is a test supermartingale under the filtration induced by the maximally invariant statistic $\hat{\delta}(Y^n)$, and allows safe sequential testing.

4.2.2 Random design

In the previous subsection, either a single full column rank covariance matrix (x^n, z^n) or a whole sequence $(x^n, z^n)_{n \geq d_w}$ were given upfront. This subsection is devoted to showing that both safety and GROW/REGROW optimality conclusions still hold for S_n^{GO} once we switch to random design.

To show this, we must first generalize the current results to unrestricted design matrices. Without full rank assumptions, the location-scale group does not necessarily act *freely* on the data y^n or on the sufficient statistics $S^n(y^n)$. This is a fundamental assumption in Theorem 2 by Pérez-Ortiz et al. [2024], on which Proposition 4.2.2 relies.
The degenerate rank issue can easily be solved using the same trick as in Section 3.1. The e-variable property and GROW/ REGROW properties do not depend on how the hypotheses are parameterized, but only on the distributions they contain. If a degenerate design matrix (x^n, z^n) is given, simply perform a backward basis selection procedure on the columns to get the full rank matrix $(\tilde{x}^n, \tilde{z}^n)$. Then, the following holds:

$$\mathcal{H}_{1|x^n,z^n,\delta} = \mathcal{H}_{1|\tilde{x}^n,\tilde{z}^n,\tilde{\delta}},$$

where $\tilde{\delta}$ is the unique vector such that $x^n\delta = \tilde{x}^n\tilde{\delta} + \tilde{z}^n\tilde{\gamma}$ holds for some $\tilde{\gamma}$. We can reuse the results of this section to conclude that the GROW/REGROW e-variable for testing two specified Gaussian orbits:

$$\mathcal{H}_{1|x^n,z^n,\delta_0} \text{ vs } \mathcal{H}_{1|x^n,z^n,\delta_1}$$

with arbitrary design matrix (x^n, z^n) is given by the likelihood ratio

$$\frac{t^{\text{nc}}_{n-d_{\tilde{w}}}(\hat{\delta}(Y^n)|\tilde{\delta}_1, (\tilde{x}^{n^T}(I_n - P_{\tilde{z}^n})\tilde{x}^n)^{-1})}{t^{\text{nc}}_{n-d_{\tilde{w}}}(\hat{\delta}(Y^n)|\tilde{\delta}_0, (\tilde{x}^{n^T}(I_n - P_{\tilde{z}^n})\tilde{x}^n)^{-1})}.$$

The transformation $(x^n, z^n, \delta_1) \to (\tilde{\delta}_1, (\tilde{x}^{n^T}(I_n - P_{\tilde{z}^n})\tilde{x}^n))$ is measurable since the basis selection procedure can be represented with a series of matrix multiplications, and the remaining functions are continuous. This allows a simple extension of the results to random design.

Definition 4.2.3. (e-variable for random design Gaussian orbit testing)

$$S_n^{\mathrm{GO}}(Y^n, X^n, Z^n) := \frac{t^{\mathrm{nc}}_{n-d_{\tilde{w}}}(\hat{\delta}(Y^n)|\tilde{\delta}_1, (\tilde{X}^{n^T}(I_n - P_{\tilde{Z}^n})\tilde{X}^n)^{-1})}{t^{\mathrm{nc}}_{n-d_{\tilde{w}}}(\hat{\delta}(Y^n)|\tilde{\delta}_0, (\tilde{X}^{n^T}(I_n - P_{\tilde{Z}^n})\tilde{X}^n)^{-1})}.$$

This ratio is considered equal to 1 whenever all columns of X^n lie in the space spanned by Z^n, i.e. when $\tilde{X}^n$ is an empty matrix.

The natural extension of LR distribution orbits to random design is:

$$\mathcal{H}^{\mathrm{LR}}_{1|n,\delta} = \{P(D^n) : P(Y^n, X^n, Z^n) = \mathcal{N}(X^n\sigma\delta + Z^n\gamma, \sigma^2 I_n) \otimes P(X^n, Z^n), \\ \sigma^2 > 0, \gamma \in \mathbb{R}^{d_z}\}.$$

S_n^{GO} is clearly a strong (X^n, Z^n)-conditional e-variable (see Definition 2.4.1) for $\mathcal{H}^{\mathrm{LR}}_{1|n,\delta_0}$. The statement follows directly from the e-variable property in fixed design.
Furthermore, fixed design GROW and REGROW properties immediately imply the same in random design

$$\begin{aligned}
&\sup_{S\in\mathcal{E}(\mathcal{H}^{\mathrm{LR}}_{0|n})} \inf_{Q\in\mathcal{H}^{\mathrm{LR}}_{1|n,\delta_1}} \mathbb{E}_Q[\ln S] = \sup_{S\in\mathcal{E}(\mathcal{H}^{\mathrm{LR}}_{0|n})} \inf_{Q\in\mathcal{H}^{\mathrm{LR}}_{1|n,\delta_1}} \mathbb{E}_{Q(X^n,Z^n)}[\mathbb{E}_{Q(Y^n|X^n,Z^n)}[\ln S]] \\
&= \sup_{S\in\mathcal{E}(\mathcal{H}^{\mathrm{LR}}_{0|n})} \inf_{\gamma,\sigma^2} \inf_{Q(X^n,Z^n)} \mathbb{E}_{Q(X^n,Z^n)}[\mathbb{E}_{Q_{\gamma,\sigma^2}(Y^n|X^n,Z^n)}[\ln S]] \\
&= \sup_{S\in\mathcal{E}(\mathcal{H}^{\mathrm{LR}}_{0|n})} \inf_{\gamma,\sigma^2} \inf_{(x^n,z^n)} \mathbb{E}_{Q_{\gamma,\sigma^2}(Y^n|x^n,z^n)}[\ln S] \\
&= \sup_{S\in\mathcal{E}(\mathcal{H}^{\mathrm{LR}}_{0|n})} \inf_{(x^n,z^n)} \inf_{\gamma,\sigma^2} \mathbb{E}_{Q_{\gamma,\sigma^2}(Y^n|x^n,z^n)}[\ln S] \\
&\leq \inf_{(x^n,z^n)} \inf_{\gamma,\sigma^2} \mathbb{E}_{Q_{\gamma,\sigma^2}(Y^n|x^n,z^n)}[\ln S^{\mathrm{GO}}_n] \\
&= \inf_{Q\in\mathcal{H}^{\mathrm{LR}}_{1|n,\delta_1}} \mathbb{E}_Q[\ln S^{\mathrm{GO}}_n].
\end{aligned}$$

The third equality holds because an infimum over averages can always be achieved by putting all weight on a single point. The inequality is a consequence of the fixed design GROW property of S_n^{GO}. The same argument works for the REGROW objective.

Finally, if $\tau(D^\infty)$ denotes the usual greedy stopping time (2.4) with security level α, and based on S_n^{GO}, then

$$P(\tau(Y^\infty, X^\infty, Z^\infty) < \infty) = \mathbb{E}_{(x^\infty,z^\infty)\sim P(X^\infty,Z^\infty)}[P(\tau(Y^\infty, x^\infty, z^\infty) < \infty)] \leq \alpha$$

for any $P \in \mathcal{H}^{\mathrm{LR}}_{1|\infty,\delta_0}$, directly from the fixed-design guarantees.

Although unconditional optimality may be theoretically appealing, these extensions are of limited practical relevance. The number of available data points in classical LR settings is much larger than the number of covariates, and the probability of observing a

full rank design matrix is generally close to one. Whenever this is the case, researchers may be satisfied having an e-variable which is GROW/REGROW optimal conditionally on the event that the design matrix has full rank, while ignoring its behavior in the highly unprobable alternative setting.

It is not as easy to generalize the G-LR e-variable (4.3) to random design in a way which would secure the unconditional GROW/REGROW property, but as we have just argued, this is not really necessary for practical purposes. Furthermore, safe sequential testing in random design can be secured by simply setting the value of the e-variable to one whenever the design matrix is not of full rank.

4.3 Adaptive MX

In Chapter 2 (equations (2.5) and (2.7)), we introduced the prequential plug-in method and the method of mixtures as general tools for extending e-variables designed for point alternatives to composite alternatives. Both approaches can be applied directly to the MX e-variable, yielding test supermartingales suitable for (sequential) testing of the full linear alternative hypothesis.
Each method, however, has practical limitations. The prequential plug-in approach often performs poorly on small samples due to inaccurate parameter estimation, while the method of mixtures typically requires numerical integration, which can be computationally expensive.

To address these issues, we adopt a third approach, referred to as Adaptive MX, which combines key ideas from both the prequential plug-in method and the method of mixtures. This technique has previously been used in settings without covariates by Turner et al. [2024] and discussed in the MX framework by Grünwald et al. [2023].

The hybrid approach resolves the small-sample instability of the prequential plug-in method by introducing a prior distribution $p(\beta, \gamma, \sigma^2)$ over the model parameters, in the spirit of the method of mixtures. Rather than mixing e-variables at each step, the prior is used to define a posterior parameter distribution

$$p(\beta, \gamma, \sigma^2 | d^n) := \frac{[\prod_{i=1}^n \mathcal{N}(y_i | \beta^T x_i + \gamma^T z_i, \sigma^2)] p(\beta, \gamma, \sigma^2)}{\int [\prod_{i=1}^n \mathcal{N}(y_i | \beta^T x_i + \gamma^T z_i, \sigma^2)] p(\beta, \gamma, \sigma^2) d(\beta, \gamma, \sigma^2)},$$

and a Bayesian predictive distribution

$$f_n(y_n | x_n, z_n, d^{n-1}) \quad := \quad \int \mathcal{N}(y_n | \beta^T x_n \; + \; \gamma^T z_n, \sigma^2) p(\beta, \gamma, \sigma^2 | d^{n-1}) d(\beta, \gamma, \sigma^2), \quad (4.4)$$

where $d^n := (y^n, x^n, z^n)$ is the observed data. The predictive distribution $f_n(\cdot|\cdot, \cdot, d^{n-1})$ yields a data-dependent Markov kernel $F_n(Y|X, Z)$, which is interpreted as an estimate of the data-generating conditional distribution:

$$\hat{Q}_n(Y|X, Z) := F_n(Y|X, Z).$$

The GRO e-variable (Chapter 2, (2.12)) for testing $\mathcal{H}^{\mathrm{MX}}_{0|1,K(X|Z)}$ against the point alternative

$$\{F_n(Y|X,Z) \otimes K(X|Z) \otimes Q(Z)\}$$

for any marginal distribution $Q(Z)$ is given by

$$\frac{f_n(Y|X,Z,d^{n-1})}{\mathbb{E}_{K(X'|Z)}[f_n(Y|X',Z,d^{n-1})]}. \tag{4.5}$$

Following the prequential plug-in approach, (4.5) is used as a multiplicative increment to form a test supermartingale, and the full process is given by

$$S^{\mathrm{AMX}}_n(D^n) := S^{\mathrm{AMX}}_{n-1}(D^{n-1}) \cdot \frac{f_n(Y_n|X_n,Z_n,D^{n-1})}{\mathbb{E}_{K(X|Z_n)}[f_n(Y_n|X,Z_n,D^{n-1})]},$$
$$S^{\mathrm{AMX}}_0 := 1.$$

For simple choices of the prior, predictive distributions f_n admit a closed-form expression. As a consequence, the approach stabilizes inference in small samples, while avoiding the additional numerical integration over parameters $(\beta, \gamma, \sigma^2)$ typically required by the method of mixtures.

It is easy to check that $(S^{\mathrm{AMX}}_n)_n$ is in fact a test supermartingale w.r.t. the data filtration. Indeed, for any $P \in \mathcal{H}^{\mathrm{MX}}_{0|n,K(X|Z)}$:

$$\begin{aligned}
&\mathbb{E}_P[S^{\mathrm{AMX}}_n|D^{n-1}] = \mathbb{E}_P\left[\mathbb{E}_P[S^{\mathrm{AMX}}_n|Y_n,Z_n,D^{n-1}]|D^{n-1}\right] \\
&= S^{\mathrm{AMX}}_{n-1} \cdot \mathbb{E}_P\left[\mathbb{E}_P\left[\frac{f_n(Y_n|X_n,Z_n,D^{n-1})}{\mathbb{E}_{K(X|Z_n)}\left[f_n(Y_n|X,Z_n,D^{n-1})\right]}|Y_n,Z_n,D^{n-1}\right]|D^{n-1}\right] \\
&= S^{\mathrm{AMX}}_{n-1} \cdot \mathbb{E}_P[1|D^{n-1}] \\
&= S^{\mathrm{AMX}}_{n-1}.
\end{aligned}$$

The third equality follows simply because $X_n \perp Y_n, D^{n-1}|Z_n$ holds under the null. The exact same reasoning was used in the robust ANCOVA example, where we introduced the generalized MX condition (Chapter 3, (3.5)).

4.4 E-power analysis and simulations

In this section, we compare the e-powers of Adaptive MX and G-LR e-variables to each other, and to their corresponding point-alternative versions on both a theoretical and empirical level.
We start with a theoretical comparison of the asymptotic relative power (ARP) of the original S-LR and MX e-variables, to their composite alternative adaptations G-LR and AMX.

Proposition 4.4.1. (No ARP loss compared to oracles)
Under the assumptions of Theorem 3.5.6, the ARP of the G-LR e-variable for any choice of Λ^{-1} is the same as that of S-LR and equals

$$\frac{1}{2}\ln(1+\mathbb{E}[(\delta_1^T(X-L_{X|Z}Z))^2]).$$

Under any $Q \in \mathcal{H}^{\mathrm{MX}}_{1|\infty,K(X|Z)}$, the ARP of Adaptive MX with prior p and Bayesian predictive distribution f_n as in (4.4) is the same as that of the oracle MX e-variable, i.e. equals $I(Y;X|Z)$ whenever:

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{Q(D^{n-1})}[D_{\mathrm{KL}}(\mathcal{N}(y|\beta_1^T x+\gamma_1^T z,\sigma_1^2)||f_n(y|x,z,D^{n-1}))]\to 0, \tag{4.6}$$

and $\mathcal{N}(y|\beta_1^T x+\gamma_1^T z,\sigma_1^2)$ is the true Lebesgue-density of the kernel $Q(Y|X,Z)$.

The statement for the G-LR e-variable is proved by Lindon et al. [2024], and for AMX by Grünwald et al. [2023].
The convergence requirement (4.6) is in fact satisfied when p is the *Gaussian-inverse-gamma* prior, which is commonly used in Bayesian linear regression [Bernardo et al., 1994].

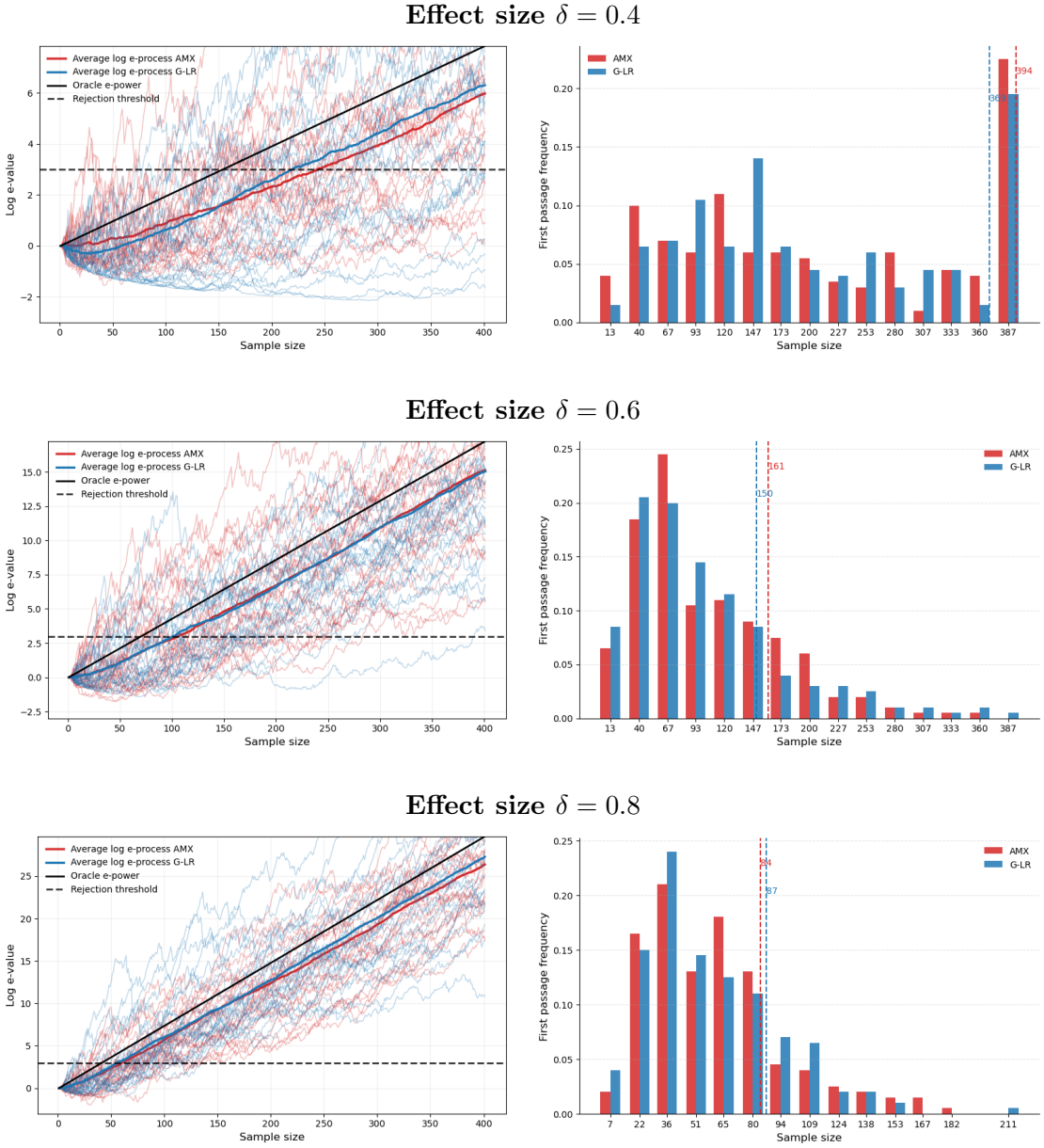
As we have discussed in Section 3.5, the ARP is a fairly imprecise measure of the practical usefulness of an e-variable. To solidify our theoretical findings, we conduct an e-power simulation study in a simple ANCOVA setting. X is a binary random variable indicating one of two experimental groups, the covariates Z consist of an intercept term and an additional normally distributed variable $Z_1 \sim \mathcal{N}(0,1)$. The target variable Y satisfies the linear model:

$$Y = \delta X + \gamma^T Z + \varepsilon, \varepsilon \sim \mathcal{N}(0,\sigma^2), \tag{4.7}$$

where $\gamma^T = [1,1]$, $\sigma^2 = 1$, and X, Z, ε are all independent. We use the G-LR and AMX e-variables to test the usual linear null hypothesis $\delta = 0$, against the composite linear alternative. For each of the six settings (2 methods and 3 effect sizes), a total of 200 independent datasets, each containing 400 samples were generated and used for testing. Figure 4.1 shows the results of the simulations. The oracle power (black line) represents the e-power of the point-alternative (oracle) MX e-variable, which "knows" the true parameters. The theoretical e-powers of the oracle MX are S-LR e-variables are practically identical in this setting and under the considered effect sizes, so the black line can be interpreted as the optimal growth rate that both AMX and G-LR can only hope to achieve. The average sample paths (thick curves) represent estimates of G-LR and AMX e-powers over time.

The plots on the LHS of Figure 4.1 show that e-power curves of AMX and G-LR quickly become parallel with that of the oracle, supporting the results of Proposition 4.4.1 in this setting. Furthermore, AMX and G-LR exhibit a very similar empirical growth-rate, which aligns with the results of Theorem 3.5.6. The similarity of the two methods

can additionally be observed by a more practically relevant measure - the empirical distribution of the greedy stopping time (Chapter 2, 2.4), depicted on the RHS of Figure 4.1. Given the strong similarity between the empirical stopping time distributions, we conclude that, in practice, a comparable number of samples is required to reject the null hypothesis under either method. This is particularly evident from the closeness of the 80% quantiles of the two distributions. In classical hypothesis testing, this quantile typically corresponds to the power level used by practitioners when determining the required sample size. In our setting, the near equality of the 80% stopping time quantiles can therefore be interpreted as indicating that both methods require approximately the same sample size to reject the null hypothesis under the greedy stopping rule.



(a) E-powers (b) Stopping times

Figure 4.1: Comparison of G-LR and AMX e-variables across different effect sizes. Each panel shows on the left the theoretical e-power of the oracle (black line), individual sample paths (thin curves), the average sample path (thick curves), and the rejection threshold $\ln(\frac{1}{0.05})$ (dashed line). The plots on the right show the empirical distribution of the greedy stopping time and the 80% quantile (dashed line). Rows correspond to effect sizes $\delta = 0.4, 0.6, 0.8$.

Given that it is hard to prefer one or the other e-variable in this setting based on (e-) power alone, we turn our attention to safety guarantees under model misspecification. As demonstrated in Section 3.3, the (oracle) MX e-variable yields a robust ANCOVA test with explicit theoretical guarantees: it remains valid even when the error terms exhibit arbitrary dependence on the covariates Z over time and across observations. By construction, the AMX e-variable inherits this robustness.
For S-LR and G-LR however, such guarantees are not immediate. Under the linear null $\mathcal{H}_{0|n}^{\mathrm{LR}}$ the errors are specifically assumed to be independent Gaussians conditional on (X^n, Z^n). It remains an open problem to verify how much bigger the effective null $\mathcal{H}_{0|n}^{\mathrm{LR}^{\mathrm{eff}}}$ is, i.e. under which kinds of model violations the type-1 error guarantees continue to hold for all corresponding e-variables, and in particular for S-LR and G-LR.

Finally, we conclude that based on our theoretical findings concerning asymptotic e-power (Theorem 3.5.8) and safety guarantees (Chapter 3, (3.5)), and our empirical examination of finite-sample behavior, the AMX e-variable is the preferable choice over G-LR in the ANCOVA and similar testing scenarios, while it is likely that in many general settings, with arbitrary dependencies between the covariates X and Z, G-LR has a significant e-power advantage (Theorem 3.5.6).

5 Conclusion

Our main goal in this thesis was to provide a careful description and comparison of the S-LR and MX e-variables for linear regression testing, as well as their counterparts for testing composite alternatives. To enable a meaningful comparison, we first extended the existing fixed-design S-LR e-variable to a random-design setting. We showed that, for an appropriate choice of the random-design null hypothesis $\mathcal{H}_{0|n}^{\mathrm{LR}}$ —namely, one that imposes no restrictions on the covariate distributions, the S-LR e-variable inherits the GRO property from its fixed-design counterpart. An analogous phenomenon was observed for the GROW/REGROW e-variables for Gaussian orbit testing; fixed-design GROW/REGROW optimality carries over directly to the random-design setting. This transferability follows from the fact that the relevant optimality criteria are defined in terms of expectations, and similar conclusions are likely to hold for more general regression models beyond linear regression.

We identified experimental settings in which both e-variables are applicable, yet one performs well while the other is essentially useless. For instance, the MX e-variable has trivial power for interaction testing in a two-factor ANOVA model, whereas the S-LR e-variable loses its theoretical safety guarantees under certain misspecifications of the ANCOVA model. In contrast, the MX e-variable provides theoretical type-1 error guarantees under a wide range of model misspecifications (robust ANCOVA example in Section 3.3).

We further discussed that, in randomized controlled trials, researchers often possess structural knowledge such as linearity, i.i.d. covariates, and the true propensity kernel. Ideally, one would use the GRO e-variable corresponding to a null hypothesis that fully incorporates this information. However, such a construction is generally infeasible, which motivates our comparison of the e-power of the S-LR and MX e-variables. To the best of our knowledge, these represent the only practical options, requiring a choice between two suboptimal procedures. We also briefly considered linear regression testing under a fully specified covariate distribution. In this setting, the GRO e-variable is strictly more e-powerful than either S-LR or MX. Remarkably, we showed that the S-LR e-variable, despite making no assumptions about the covariate distribution, achieves the same asymptotic relative power (ARP) as the GRO e-variable for the FSLR null. Consequently, the cost of not knowing or not being able to incorporate partial or full information about the true covariate distribution is at most of order $o(n)$, where n denotes the sample size. This result also implies that S-LR is never inferior to MX in terms of ARP.
In finite samples, however, the comparison is more subtle. We demonstrated the exis-

tence of practically relevant scenarios in which the MX e-variable is strictly superior. In the context of randomized controlled trials, we showed that, within a generalized ANCOVA framework, the ARP difference between the two e-variables is of order $O(\|\delta\|^6)$, where δ denotes the effect size vector. A possible explanation for the remarkably strong performance by MX is that in the generalized ANCOVA setting, the MX e-variable coincides with the sequential GRO e-variable for the FSLR null hypothesis. This fact comes as a consequence of an, unusual, double-projection property — for $n = 1$, finding the RIPr of a relevant alternative distribution onto the FSLR null hypothesis can be done in two steps, by first projecting onto the larger, MX null hypothesis, and then projecting the MX-RIPr onto the FSLR null.
The small ARP gap suggests that in practically relevant settings, the two e-variables may have a very similar performance. We empirically supported this conclusion through simulation studies on the AMX and G-LR e-variables. Taken together, these results indicate that, for randomized controlled trials with ANCOVA designs, the MX approach may be preferable: the power difference relative to S-LR is small, and the stronger safety guarantees of MX can be highly advantageous.

As a direction for future research, a rigorous theoretical analysis of the safety guarantees of S-LR and G-LR e-variables under model misspecification—such as non-Gaussian errors or nonlinear dependence structures—would be of substantial practical value. Furthermore, it would be natural to move beyond strict linear regression to more flexible settings such as *linear mixed models*. These models are widely used in clinical trials and allow for dependent data arising from, for example, repeated measurements, thereby offering greater flexibility and practical relevance than the framework considered here.
Another potential extension involves generalized linear models, such as the practically highly important *logistic regression*. It is known that it is exceedingly difficult to implement the corresponding generalization of the S-LR e-variable for logistic-regression, as the reverse information projection (Theorem 2.2.5) takes on a very complicated form Lardy [2021]. However, the construction of the MX e-variable is no more difficult than for the standard linear case treated in this thesis [Grünwald et al., 2023]. This suggests that it is worth extending our investigation to generalized linear regression and see whether we can identify situations in which, there too, both e-variables behave very similarly, and it is therefore intrinsically harmless to use the simpler MX e-variable.

6 Popular summary

Consider a scientist who wants to test the null hypothesis

$$\mathcal{H}_0 : \text{the treatment has no effect,}$$

against the alternative hypothesis

$$\mathcal{H}_1 : \text{the treatment is effective.}$$

Classical statistical tests for answering this question were developed in the 1920s, and many of them are still used today in their original forms. Classical tests crucially rely on conditions such as prespecifying the number of participants in the study, and performing the test only after all the data has been collected. If these conditions are not satisfied, the safety properties of the tests are often lost. This means that a test which usually makes a false positive decision (rejecting the null when it is in fact true) 5% of the time, could actually be making it much more often. In practice such a false positive finding might result in an ineffective drug entering the market, or the development of a scientific theory based on inexistent effects. In the long-run it leads to reduced reproducibility of scientific research and generally, a waste of resources. The conditions on which classical tests are based were not deemed too stringent at the time of their first development, but that is no longer the case. Today, sequential tests - which accumulate evidence against the null hypothesis while the data is being collected, and allow the researcher to stop acquiring data once sufficient evidence has been observed, are more practically useful. The theory of e-variables, which has been rapidly developing since 2019, provides a theoretical framework for creating such sequential tests.
In this thesis we give a detailed mathematical description of two e-variables called: Simple Linear Regression (S-LR) and Model-X (MX) that have been proposed for linear regression testing - one of the most widely used statistical tests in practice. We first describe what e-variables are and how they can be used for hypothesis testing, and then compare the performance of the induced S-LR and MX tests in different settings. We show that S-LR tends to have a larger e-power, meaning that it can detect existing treatment effects quicker, while MX is more robust, and provides a wider range of practically important safety guarantees against false-positive findings. We specifically examine the analysis of covariance setting - which is often used to model clinical trials, and mathematically prove that in this configuration S-LR and MX have very similar asymptotic (as the sample size tends to infinity) e-power. In fact, we precisely quantify the discrepancy in asymptotic e-power as being of order $O(|\delta|^6)$, when the magnitude of the treatment effect δ converges to 0. In practice, the treatment effects are small, and our result suggests that, in this setting, S-LR and MX have a very similar performance.

This outcome is very optimistic for the future development of sequential tests, because versions of the MX e-variable can be used for testing in more complicated, generalized linear regression settings, and our findings suggest that maybe there too, it will be useful.

Bibliography

Alan Agresti. *Foundations of Linear and Generalized Linear Models*. John Wiley & Sons, 2015.

James O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer, 2013.

James O Berger and Thomas Sellke. Testing a point null hypothesis: The irreconcilability of p values and evidence. *Journal of the American statistical Association*, 82(397):112–122, 1987.

José M. Bernardo, Adrian FM Smith, and Mark Berliner. *Bayesian Theory*. John Wiley & Sons, 1994.

Patrick Billingsley. *Probability and measure*. John Wiley & Sons, 2017.

Emmanuel Candès, Yingying Fan, Lucas Janson, and Jinchi Lv. Panning for gold: 'model-X' knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society B*, 80:551–577, 2018. doi: 10.1111/rssb.12265. URL `https://rss.onlinelibrary.wiley.com/doi/epdf/10.1111/rssb.12265`.

George Casella and Roger L. Berger. *Statistical Inference*. Chapman and Hall, 2nd edition, 2024.

T.M. Cover and J.A. Thomas. *Elements of Information Theory*. Wiley-Interscience, New York, 1991.

Ronald Aylmer Fisher. *Statistical methods for research workers*. 1934.

Patrick Forré and Joris M. Mooij. A mathematical introduction to causality, 2024.

Gene V Glass, Percy D Peckham, and James R Sanders. Consequences of failure to meet assumptions underlying the fixed effects analyses of variance and covariance. *Review of educational research*, 42(3):237–288, 1972.

P. Grünwald, T. Lardy, Y. Hao, S. Bar-Lev, and M. De Jong. Optimal e-values for exponential families: the simple case. In *Information Theory, Probability and Statistical Learning — A Festschrift in Honor of Andrew Barron*. Springer, 2025. To Appear.

Peter Grünwald. Beyond neyman-pearson: e-values enable hypothesis testing with a data-driven alpha. *Proceedings National Academy of Sciences of the USA (PNAS)*, 121(39), 2024.

Peter D. Grünwald. *The Minimum Description Length Principle*. MIT Press, 2007.

Peter D. Grünwald, Rianne de Heide, and Wouter M. Koolen. Safe testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(5):1091–1128, 2024. With Discussion.

Peter Grünwald, Alexander Henzi, and Tyron Lardy. Anytime-valid tests of conditional independence under model-x. *Journal of the American Statistical Association*, 2023.

Dongning Guo, Yihong Wu, Shlomo S Shitz, and Sergio Verdú. Estimation in Gaussian noise: Properties of the minimum mean-square error. *IEEE Transactions on Information Theory*, 57(4):2371–2385, 2011.

Yunda Hao and Peter Grünwald. E-values for exponential families: the general case, 2024. URL `https://arxiv.org/abs/2409.11134`.

Nick Koning. Continuous testing: unifying tests and e-values. *arXiv preprint 2409.05654*, 2024.

W. M. Koolen and P. Grünwald. Log-optimal anytime-valid e-values. *International Journal of Approximate Reasoning*, 141:69–82, 2022. Festschrift for G. Shafer's 75th Birthday.

Tyron Lardy. E-values for hypothesis testing wih covariates: Construction and application to regression models. Master's thesis, Leiden University, 2021.

Martin Larsson, Aaditya Ramdas, and Johannes Ruf. The numeraire e-variable and reverse information projection. *Annals of Statistics*, 53(3):1015–1–43, 2025.

Erich L. Lehmann and Joseph P. Romano. *Testing Statistical Hypotheses*. Springer New York, 2005.

Dennis V Lindley. A statistical paradox. *Biometrika*, 44(1/2):187–192, 1957.

Michael Lindon, Dae Woong Ham, Martin Tingley, and Iavor Bojinov. Anytime-valid linear models and regression adjusted causal inference in randomized experiments. *arXiv preprint arXiv:2210.08589*, 2024.

Jerzy Neyman and Egon Sharpe Pearson. Ix. on the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 231(694-706): 289–337, 1933.

Muriel Felipe Pérez-Ortiz, Tyron Lardy, Rianne de Heide, and Peter D. Grünwald. E-statistics, group invariance and anytime-valid testing. *Ann. Statist.*, 52(4):1410–1432, 2024. ISSN 0090-5364. doi: 10.1214/24-AOS2394.

Vladimir Rakočević. On continuity of the Moore-Penrose and Drazin inverses. 1997. URL `https://api.semanticscholar.org/CorpusID:54176864`.

Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statist. Sci.*, 38(4):576–601, 2023. ISSN 0883-4237. doi: 10.1214/23-STS894.

Glenn Shafer, Alexander Shen, Nikolai Vereshchagin, and Vladimir Vovk. Test martingales, Bayes factors and p-values. *Statistical Science*, 26(1):84–101, 2011.

Rajen D. Shah and Jonas Peters. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3), 2020. doi: 10.1214/19-aos1857. URL `https://doi.org/10.1214%2F19-aos1857`.

Judith Ter Schure and Peter Grünwald. Accumulation bias in meta-analysis: the need to consider time in error control [version 1; peer review: 2 approved]. *F1000Research*, 8:962, 2019. URL `https://doi.org/10.12688/f1000research.19375.1`.

Rosanne Turner, Alexander Ly, and Peter Grünwald. Generic e-variables for exact sequential k-sample tests that allow for optional stopping. *Statistical Planning and Inference*, 230:106116, 2024. doi: https://doi.org/10.1016/j.jspi.2023.106116.

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

Jean Ville. *Etude critique de la notion de collectif*. Gauthier-Villars, 1939.

Abraham Wald and Jacob Wolfowitz. Optimum character of the sequential probability ratio test. *The Annals of Mathematical Statistics*, pages 326–339, 1948.