universität wien

MAGISTERARBEIT | MASTER'S THESIS

Titel | Title

A discussion of Universal Inference

verfasst von | submitted by

Lorenz Matz BSc BSc

angestrebter akademischer Grad | in partial fulfilment of the requirements for the degree of

Magister der Sozial- und Wirtschaftswissenschaften (Mag.rer.soc.oec.)

Wien | Vienna, 2025

Studienkennzahl lt. Studienblatt | Degree programme code as it appears on the student record sheet:	UA 066 951
Studienrichtung lt. Studienblatt | Degree programme as it appears on the student record sheet:	Magisterstudium Statistik
Betreut von | Supervisor:	Univ.-Prof. Mag. Dr. Hannes Leeb

Abstract

Universal Inference is a general method for constructing confidence sets and tests in statistical models, proposed by Wasserman et al. (2020). It is based on a combination of the likelihood ratio approach and sample splitting. Unlike inference procedures built on classical asymptotic results or most methods based on resampling, the resulting confidence sets/tests have finite-sample guarantees regarding their coverage probability/significance level and do not require any regularity assumptions on the model or the null hypothesis. However, in models where sensible confidence sets and tests with (finite sample or asymptotic) guarantees are available, the universal confidence sets and tests tend to be conservative in comparison to these. This thesis adds two important points to the ongoing discussion about universal inference: On the one hand, we try to build a mathematically rigorous foundation of this method, which is still missing in the literature. Furthermore, we provide several insights regarding the connection between the universal and the classical likelihood ratio approach. In particular, we compare the split likelihood ratio confidence set which uses the maximum likelihood estimator with the classical likelihood ratio confidence set in terms of subset relations and their average volumes. We find that, in one- and two-dimensional models, the classical likelihood ratio set always constitutes a subset of this specific split likelihood ratio set, and we show that the average volume of the latter is larger than the average volume of the former in multivariate Gaussian models of any dimension. Moreover, we take the Gaussian setting as an example to illustrate why universal confidence sets and tests usually become extremely conservative in high-dimensional models, as was already noticed by other authors.

Zusammenfassung

Universal Inference ist eine allgemeine Methode zur Konstruktion von Konfidenzmengen und Tests in statistischen Modellen, die erstmals in Wasserman et al. (2020) vorgestellt wurde. Sie basiert auf einer Kombination aus dem Likelihood-Quotienten-Ansatz und dem Prinzip der Stichprobenaufteilung. Im Gegensatz zu jenen Inferenzverfahren, die sich aus bekannten asymptotischen Resultaten ableiten, sowie den meisten auf Resampling basierenden Methoden, kontrollieren die daraus resultierenden Konfidenzmengen und Tests die Überdeckungswahrscheinlichkeit bzw. das Signifikanzniveau bereits für endliche Stichprobengrößen. Weiters erfordern sie dafür keinerlei Regularitätsannahmen an das Modell oder an die zu testende Nullhypothese. Jedoch sind in Modellen, in denen sinnvolle Tests und Konfidenzmengen existieren, die ein gefordertes Signifikanzniveau bzw. die geforderte Überdeckungswahrscheinlichkeit (asymptotisch) einhalten, die auf Universal Inference basierenden Konfidenzmengen und Tests im Vergleich zu diesen tendenziell konservativ. Diese Arbeit trägt zwei wichtige Punkte zur laufenden Diskussion über Universal Inference bei: Zum einen wird versucht, eine mathematisch fundierte Grundlage für diese Methode zu schaffen, die in der Literatur bisher noch fehlt. Darüber hinaus werden einige Erkenntnisse zum Zusammenhang zwischen dem Universal Inference- und dem klassischen Likelihood-Quotienten-Ansatz präsentiert. Insbesondere wird die Split-Likelihood-Quotienten-Konfidenzmenge, die den Maximum-Likelihood-Schätzer verwendet, mit der klassischen Likelihood-Quotienten-Konfidenzmenge hinsichtlich Teilmengenrelationen sowie ihrer jeweiligen durchschnittlichen Volumina verglichen. Es zeigt sich, dass die klassische Likelihood-Quotienten-Menge in ein- und zweidimensionalen Modellen stets eine Teilmenge dieser spezifischen Split-Likelihood-Quotienten-Menge ist und dass in multivariaten Normalverteilungsmodellen beliebiger Dimension das durchschnittliche Volumen der letzteren stets größer ist als jenes der ersteren. Außerdem wird das multivariate Gauß'sche Modell zum Anlass dafür genommen, zu erklären, warum die auf Universal Inference basierenden Konfidenzmengen und Tests in hochdimensionalen Modellen üblicherweise sehr konservativ werden, wie bereits von anderen Autoren festgestellt wurde.

Danksagung

*Kein Mensch schafft ein Werk wirklich ganz allein, und so möchte ich an dieser Stelle die Gelegenheit ergreifen, mich bei jenen zu bedanken, ohne die ich weder diese Arbeit zu Ende gebracht, noch eine schöne Studienzeit gehabt hätte: An allererster Stelle bei meinen Eltern Elisa und Karl für viele Jahre der Unterstützung auf allen nur erdenklichen Ebenen - ohne dieser wäre das Verfassen dieser Arbeit so wie überhaupt mein gesamtes Studium auf keinen Fall möglich gewesen. Bei meiner Schwester Larissa, deren Stärke und Durchhaltevermögen mich immer wieder aufs Neue beeindrucken und anspornen. Bei meiner Großmutter Hildrun, die den mit dieser Arbeit verbundenen Schreibprozess mit großem Interesse verfolgt hat und der ich stets von meinen dabei auftretenden Sorgen und Schwierigkeiten erzählen konnte. Bei all den netten Kommiliton*innen, die ich im Laufe der Jahre im Statistikstudium kennengelernt habe; insbesondere bei Andi, Cornelius, Martin, Matthias und Tihomir für den motivierenden Austausch über unsere Arbeiten. Bei Nico und Georg für eine Hilfsbereitschaft, wie ich sie mir an allen Instituten dieser Welt wünschen würde, und bei Lukas für das wichtige Aufrechterhalten meiner Motivation in jener Zeit, als das Schreiben an dieser Arbeit nie zu Ende gehen schien. Bei Gertrude, Illi, Michi und Timmy für jene Ablenkungen, die mir in manchen Momenten wohl mehr bezüglich dieser Arbeit weitergeholfen haben als die Lektüre eines weiteren Papers.*

Zu guter Letzt möchte ich insbesondere noch zwei Menschen meinen Dank aussprechen, ohne die diese Arbeit wohl niemals zustande gekommen wäre: Meinem Betreuer Prof. Dr. Hannes Leeb für den Vorschlag dieses spannenden Themas, das genaue Durchlesen meiner Manuskripte und für jene zahl- und hilfreichen Rückmeldungen sowie Ideen, die diese Arbeit erst zu dem gemacht haben, was sie jetzt ist. Und natürlich meiner Freundin Silvia: Danke, dass du mich trotz einiger langer Schreibnächte und der während der intensiven Schreibphasen häufigen gedanklichen Abwesenheit immer unterstützt hast und vor allem, dass du mich so schätzt, wie ich bin.

Contents

1 Introduction

Statistical inference is one of the most important tasks in the context of data analysis: Given a certain data set, one often seeks to use it for inferring properties of the probability distribution from which the observations were generated. Two of the most frequently used paradigms in statistical inference are *confidence sets for parameters* and *statistical tests of hypotheses*. In both cases, a number $\alpha \in (0, 1)$ is pre-specified by the user (usually to one of the values 0.1, 0.05 or 0.01). A confidence set with *minimum coverage probability* $1 - \alpha$ for some unknown parameter of the underlying distribution is a set computed from the observations according to some specified rule, with the property that, if we would repeatedly sample from the distribution and compute sets with this rule, then, in the long run, at least $(1 - \alpha) * 100\%$ of the sets would contain this unknown parameter. Similarly, a statistical test of a null hypothesis H_0 regarding the data generating distribution with significance level α is a data-driven decision rule that either rejects or does not reject this hypothesis, with the property that, if we repeatedly sample observations from a distribution for which the null hypothesis holds, then, in the long run, the share of decisions that will be wrong in the sense that they reject H_0 will be $\alpha * 100\%$ at most. The mistake of rejecting the null hypothesis although it is in fact true is called *Type I error*.

It is easy to construct confidence sets and statistical tests that satisfy these properties: For example, for any $\alpha \in (0, 1)$, the decision rule 'Never reject H_0' obviously never makes the mistake of rejecting the null hypothesis despite it being true, and therefore, it is a statistical test with significance level α for any $\alpha \in (0, 1)$ and any hypothesis in question. In practice, however, this test will be considered useless - for instance, think about the situation where we observe two groups of cancer patients, one of which has received a certain drug which is not yet approved for the market, while the other has not. Based on some medical outcomes we observe on the patients (the data), we would like to test the null hypothesis that the drug is ineffective against the cancer, versus the alternative that the drug in question increases the chances of the cancer being cured. If we use the decision rule 'Never reject H_0' (for the 'computation' of which we do not need to look at the data at all), we would never reject the hypothesis that the drug is ineffective, even if the observations would indicate the complete opposite. Thus, the real challenge is to construct statistical tests that are *powerful*, which means that if the data generating distribution is in some sense sufficiently far away from the null hypothesis (in the example above, this can be loosely interpreted as saying that the drug increases the chances of being cured by so much that a positive effect can be seen in the observations), then we would like the test to reject the null. In the context of confidence sets, one would like some measure of the size (usually the expected diameter or volume) of the

resulting set not to be unnecessarily large - for example, if we know that a parameter of the data generating distribution must lie in a set Θ before looking at the data, then Θ is a confidence set for this parameter for any confidence level $1 - \alpha$, but usually, this confidence set is not very practical.

The task of finding powerful confidence sets and tests for various parameters and hypotheses of interest has occupied mathematical statisticians for centuries. Numerous criteria for optimality of tests have been discussed (see, for example, Lehmann and Romano, 2005) and a whole asymptotic theory has been built around the issue of finding suitable test statistics and critical values (see, for instance, Van der Vaart, 2007). In general, there exist three general approaches for constructing powerful confidence sets and tests (we will only focus on the latter in the list below since, if a construction principle for tests of certain hypotheses is known, then in most cases, it is easy to determine a corresponding construction principle for confidence sets) :

- In special cases, the *finite sample distribution* of a random variable derived from the data is known under the assumption that the null hypothesis is true and belongs to a well-known family of distributions. For example, in the context of testing the null hypothesis $H_0 : \mu^* = \mu_0$ on the mean parameter μ^* in a univariate Gaussian model with unknown variance, the critical value of a t-test is derived from the (known) distribution of a suitably standardized average.

- Assuming some regularity conditions (see, for example, Van der Vaart, 2007) regarding the statistical model and the null hypothesis, specific data-dependent random variables (such as the likelihood ratio test statistic) are known to have certain *asymptotic distributions* (in the case of the likelihood ratio test statistic, a specific χ^2-distribution) if the null hypothesis is true. This fact can be used to construct tests with *asymptotic* significance level α, which means that the probability of the test making a type I error is not necessarily less than or equal to α for a given sample size n, but approaches α (or, is guaranteed to be smaller than α) as the sample size increases. Some examples which use this construction principle are the Wald-, Score- and Likelihood ratio test.

- In recent decades, due to the increase in computing power, *resampling methods* such as the Bootstrap, the Jackknife or Permutation tests (for testing the hypothesis that two different samples are generated by the same distribution) have become popular as well. Bootstrap and Jackknife produce new samples based on the observations originally given and try to approximate the true distribution of a certain statistic by the empirical distribution of it based on these samples. The permutation test, on the other hand, uses the fact that if the null hypothesis is true, the two samples may be pooled, then randomly split into two groups again, and the distribution of the resulting two groups should be similar to the ones originally observed. While Bootstrap and Jackknife approaches in general only lead to tests with asymptotic significance level α, the permutation test controls the type I error probability exactly at the desired level.

All three of these approaches have been studied thoroughly in the literature. However, for certain statistical models and hypotheses, all of them fail to result in valid construction principles for statistical tests or confidence sets, in the sense that, for a given $\alpha \in (0,1)$, they do not yield reasonably powerful tests with significance level α and reasonably sized confidence sets with coverage probability $1-\alpha$. To illustrate this, we will introduce the reader to the one-dimensional, homoscedastic Gaussian mixture model with two components (general Gaussian mixture models will be discussed in Subsection 3.4.3). This is a relatively simple and also in some fields frequently proposed statistical model: Assume that we collect observations from n individuals and that there exists two groups A and B which each of the individuals are randomly and independently assigned to with some probability (which may be 0 or 1). Given this assignment, the observation for an individual is generated either from normal distribution A or normal distribution B, both of which have the same (possibly unknown) variance $\sigma^2 > 0$, but may have different (unknown) means μ_A and μ_B. Suppose we do not know for any of the individuals which group they have been assigned to. Then, the observations are independently and identically (i.i.d.) distributed according to the probability density function

$$f_{p,\mu_A,\mu_B,\sigma^2}(x) = p\,\phi\Big(\frac{x-\mu_A}{\sigma}\Big) + (1-p)\,\phi\Big(\frac{x-\mu_B}{\sigma}\Big),$$

where $p \in [0,1]$ and where ϕ denotes the density function of the standard normal distribution. An interesting question in the context of this statistical model is whether the observations are generated from two different distributions or really just one. In the context of mixture models, the latter case is referred to as the *homogenous* one, and it is easy to see that in this specific model, it is correct if and only if $p \in \{0,1\}$ or if $\mu_1 = \mu_2$. Formally, we may want to test the null hypothesis

$$H_0 : p^* \in \{0,1\} \text{ or } \mu_A^* = \mu_B^*$$

against the alternative that

$$H_1 : p^* \in (0,1) \text{ and } \mu_A^* \neq \mu_B^*,$$

with p^*, μ_A^*, μ_B^* being the true parameters of the model. We will now argue that all the general approaches for construction mentioned in the list above fail to construct a powerful test for this specific model and null hypothesis:

- In the literature, no pivotal quantity (under the null) that leads to a powerful test seems to be known.
- The usual regularity assumptions do not hold: Under the null hypothesis, the model is not identifiable (for example, $\mu_A^* = \mu \neq \mu_B^*$, $p^* = 1$ leads to the same model as $\mu_A^* \neq \mu = \mu_B^*$, $p^* = 0$), the true parameter may lie on the boundary of the parameter space and the Fisher-Information matrix may be singular (see McLachlan and Rathnayake, 2014).

- To our knowledge, no Bootstrap or Jackknife approach with provable type I error guarantees for this testing problem seems to be known. McLachlan (1987) suggested a method that revolves around generating bootstrap samples of the log-likelihood ratio statistic from a (single) normal distribution with parameters $\mu = \hat{\mu}_0$ and $\sigma = \hat{\sigma}_0$, where $\hat{\mu}_0$, $\hat{\sigma}_0$ are the maximum likelihood estimators for μ_A^* and σ^* under the null (with p^* set to 1). While this approach seems to be popular in applications, we were unable to find any proof in the literature that the resulting test indeed has significance level α. On the other hand, the permutation test principle cannot be applied to this problem as it does not revolve around the comparison of two samples.

It should be noted here that in fact, there exists a powerful test for homogeneity in this model with asymptotic significance level $\alpha \in (0, 1)$, namely the EM-test suggested by Chen and Li (2009). The design of this test cleverly combines the expectation-maximization algorithm with the use of penalty functions. However, this is a very specific construction principle, in the sense that it cannot be easily transferred to other models and hypotheses. Even adapting the EM-test from the univariate to a multivariate Gaussian homoscedastic mixture model is not trivial at all.

Exactly these kinds of statistical hypotheses and models, which are difficult or impossible to tackle with the three general construction principles stated in the list above, were the main motivation for Larry Wasserman, Aaditya Ramdas and Sivaraman Balakrishnan to propose a new general approach to constructing tests and confidence sets: They introduced it to the public under the name *universal inference* in 2020, in an article (Wasserman et al., 2020) with the same title. This method, which will be explained in detail in Chapter 2 of this thesis, has two advantages compared to the other general approaches:

- From a technical point of view, one needs to be able to define the likelihood function of the considered model in order to use this method (this is not possible in some settings, see, for instance, Dalmasso et al., 2024) and, for the testing problem, one also needs to be able to compute or approximate the maximum likelihood estimator under the null. However, regarding the validity of the resulting tests and confidence sets, *no regularity assumptions on the model or the null hypothesis are required*. In other words, universal inference gives statistical tests which are guaranteed to have significance level α as well as confidence sets that will have minimum coverage probability $1 - \alpha$, irrespective of how 'irregular' the underlying model might behave.

- These statements are true not only in an asymptotic sense, but for finite samples as well, i.e. a test constructed with this principle has significance level α for any given sample size $n \in \mathbb{N}$.

The universal inference method essentially boils down to a combination of sample splitting with the classical likelihood ratio approach. All of the confidence sets and tests

constructed with the universal inference principle are based on statistics which, on the one hand, are non-negative and, on the other hand, have their expectation bounded by 1 under the null hypothesis. In recent years, inference based on such random variables, which are called *e-variables* (see Vovk and Wang, 2021), has attracted a lot of attention in the statistical community, especially in the context of anytime valid inference, which can be summarized as the task of designing procedures that control the type I error or non-coverage probability uniformly over a whole data process, not just for a fixed sample size (see, for example, Ramdas et al., 2023; Grünwald, 2024b). The universal approach can also be applied to anytime inference problems (such as constructing confidence sequences and test supermartingales, see Section 7 of Wasserman et al., 2020; Hao and Grünwald, 2024), but for this thesis, we will focus on the classical inference setting where we neither consider optional stopping nor optional continuation (see Grünwald, 2024b).

The critical value of any confidence set or test constructed with universal inference is derived via Markov's inequality. Since Markov's inequality usually corresponds to a rough upper bound of the upper tail probabilities of a random variable, this implies two additional features of universal inference: On the one hand, in many cases, the actual type I error probability of universal tests will be significantly lower than the nominal significance level α and the actual coverage probability of universal confidence sets will be considerably larger than the nominal $1 - \alpha$. On the other hand, in general, if for certain models and null hypotheses, other (possibly only asymptotically) valid and non-trivial tests and confidence sets are available, these will typically be less conservative than their universal inference counterparts, meaning that the corresponding tests will be more powerful and the corresponding confidence sets will have smaller average diameter/area. One could say that this additional conservativeness of universal inference is the price one pays for its universal applicability - more details on this can be found in Subsection 2.1.3. The actual performance of these methods also crucially depends on the choice of the ratio according to which the data is split into two subsets (see Section 3.3).

Universal inference is similar to other construction principles for tests and confidence sets which are built on sample splitting, too. To name a few, Robins and Van der Vaart (2006) propose a splitting method for constructing adaptive nonparametric confidence sets, Lundborg et al. (2024) use sample splitting in the context of testing the significance of a variable X for predicting variable Y given covariates Z, Cai et al. (2024) employ it for a test of independence in high-dimensional models, Zhang and Shao (2024) in the context of bandwidth-free inference and Kim and Ramdas (2024) use sample splitting to develop methods whose (asymptotic) validity does not depend on any assumption about the asymptotic relation between the dimensionality of the data d and the sample size n. We will not discuss these related inference methods in this thesis, and we will also not talk about the (very rich) literature on anytime valid inference based on test supermartingales and e-processes (see, for example, Ramdas et al., 2022, 2023; Waudby-Smith and Ramdas, 2024), which also has a strong connection to universal inference. Instead, this thesis is focused on the following topics: First, we will concentrate on a mathematically rigorous setup of universal inference in Chapter 2, which does not

yet exist in the literature. Specifically, we will take care of some details regarding the definitions of the statistics that the split likelihood ratio confidence sets and the split likelihood ratio tests are based on, which were not discussed in Wasserman et al. (2020). Chapter 2 also includes some simple, but illustrative examples of applying universal inference methods to specific models. Next, in Chapter 3, we will discuss the existing literature on universal inference. This includes the topics of extending universal inference from the split likelihood ratio method (which, in some sense, is the basis for any universal inference method), using universal inference in the presence of nuisance parameters, addressing the problem of choosing the optimal ratio according to which the data should be split and applications of universal inference with irregular models/hypotheses, which, as was mentioned before, are the only cases in which the usage of this method may be justified in practice. Finally, Chapter 4 will compare a specific version of the split likelihood ratio confidence set, namely the one that uses a maximum likelihood estimator, with the classical likelihood ratio set in various situations. Specifically, we will show that under very weak assumptions, the classical likelihood ratio confidence set will be a proper subset of the split likelihood ratio confidence set with probability 1 in statistical models of dimensions $d = 1$ and $d = 2$. Furthermore, we will prove in this chapter that the expected volume of the split set will be larger than the expected volume of the classical set in any multivariate Gaussian model with known covariance matrix and that in a fixed sample size n, increasing dimension d setting, the volume of the split set will grow exponentially faster compared to the classical one. This result is consistent with the curse of dimensionality of universal inference methods that was previously noticed by other authors. But first, we will introduce the reader to the notation and the general setting used throughout the remainder of this thesis.

1.1 Notation and general setting

In the following chapters, unless explicitly stated otherwise, we will always relate to the following (parametric) statistical model: Let $(\Omega, \mathcal{A}, \mathbb{P}_{\theta^*})$ be a probability space, where $\mathbb{P}_{\theta^*}$ belongs to $\{\mathbb{P}_\theta : \theta \in \Theta\}$, $\Theta \subseteq \mathbb{R}^d$ non-empty, and suppose that $X_1, \dots, X_n$ are i.i.d. $(\mathcal{X}, \mathcal{E})$-valued random variables on $(\Omega, \mathcal{A}, \mathbb{P}_{\theta^*})$. Furthermore, for any $\theta \in \Theta$, the distribution of X_i has a (known) density p_θ with respect to the (known) measure μ on $(\mathcal{X}, \mathcal{E})$, i.e.

$$\mathbb{P}_\theta(X_i \in E) = \int_E p_\theta \, d\mu$$

for any $E \in \mathcal{E}$, where $p_\theta : \mathcal{X} \mapsto [0, +\infty)$ is measurable. Note that the sample space $\mathcal{X}$ need not be necessarily real. Usually, θ^* is called the (unknown) *true parameter* and p_{θ^*} is called the (unknown) *true density*.

We also introduce the following notation regarding sample splitting: For any partition $\{\mathbf{D_0}, \mathbf{D_1}\}$ of the index set $\{1, \dots, n\}$ (where we explicitely require that both $\mathbf{D_0}$ and $\mathbf{D_1}$ are non-empty), we define $\hat{\theta}_1$ to be an estimator for θ^* that is based on $\mathbf{D_1} = \{j_1, \dots, j_m\}$

only, i.e.

$$\hat{\theta}_1 = h(X_{j_1}, \ldots, X_{j_m}),$$

where $h : \mathcal{X}^{|\mathbf{D_1}|} \mapsto \Theta$ is $\mathcal{E}^{\otimes \mathbf{D_1}} - \mathcal{B}(\Theta)-$measurable (here, $\mathcal{B}(\Theta)$ denotes the Borel $\sigma-$algebra on Θ). Moreover, we define $\mathcal{L}_0$ to be the likelihood function based on $\mathbf{D_0}$ *only*, i.e.

$$\mathcal{L}_0(\theta) = \prod_{i \in \mathbf{D_0}} p_\theta(X_i).$$

Having clarified the notation as well as the assumptions we make about the underlying statistical model from which the data is generated, we are now ready to start our discussion about universal inference.

2 A rigorous setup of Universal Inference

In the paper by Wasserman et al. (2020) that introduced the concept of universal inference (UI) to the public, the authors did not put a lot of effort into defining their method rigorously, overlooking some details and leaving some questions that naturally come up unanswered. The goal of this chapter now is to build a solid, mathematically rigorous basis for universal inference, as well as to improve in some respect the definitions given in the first section of the paper. If not mentioned otherwise, in the following we always refer to the setting and notation defined at the end of the introduction.

2.1 Split Likelihood Ratio confidence set

Suppose we are given independent and identically distributed (i.i.d.) data $X_1, \dots, X_n$, which we want to use to construct a confidence set for the true parameter θ^* of the underlying statistical model. To this end, we follow a certain sample splitting procedure: We partition the index set $\{1, \dots, n\}$ into two sets $\mathbf{D_0}$ and $\mathbf{D_1}$[1] and construct an estimator $\hat{\theta}_1$ for θ^*. Importantly, no assumptions such as consistency or unbiasedness are imposed on this estimator.

2.1.1 Implicit definition by Wasserman, Ramdas and Balakrishnan

First, the authors of the original paper on UI implicitly define the following $[0, +\infty]$-valued random variable for each $\theta \in \Theta$:

$$S_n(\theta) = \begin{cases} \frac{\mathcal{L}_0(\theta)}{\mathcal{L}_0(\theta^*)} & \text{if } \mathcal{L}_0(\theta^*) > 0, \\ 1 & \text{if } \mathcal{L}_0(\theta) = \mathcal{L}_0(\theta^*) = 0, \\ +\infty & \text{if } \mathcal{L}_0(\theta) > 0,\ \mathcal{L}_0(\theta^*) = 0. \end{cases}$$

Our way to building a confidence set with nominal coverage probability $1-\alpha$, $\alpha \in (0,1)$, now starts with an important (and quite intuitive) result that shows that the two latter cases of the definition in fact deal with null sets.

Lemma 1. *If θ^* is the true parameter of the model, then $\mathbb{P}_{\theta^*}(\mathcal{L}_0(\theta^*) = 0) = 0$.*

Proof. Recall that $\mathcal{L}_0(\theta) = \prod_{i \in \mathbf{D_0}} p_\theta(X_i)$, so $\{\mathcal{L}_0(\theta^*) = 0\} = \bigcup_{i \in \mathbf{D_0}} \{p_{\theta^*}(X_i) = 0\}$. Furthermore, due to the fact that p_{θ^*} is the density function of the distribution of the

[1] It does not matter whether this partition is performed in a randomized or a deterministic way because the data is i.i.d. However, we assume from now on that the sets $\mathbf{D_0}$ and $\mathbf{D_1}$ are fixed when performing the inference procedures.

X_i, it follows that for every $i = 1, \dots, n$

$$\mathbb{P}_{\theta^*}(p_{\theta^*}(X_i) = 0) = \mathbb{P}_{\theta^*}(X_i \in p_{\theta^*}^{-1}(\{0\})) = \int_{p_{\theta^*}^{-1}(\{0\})} p_{\theta^*}(x)\, d\mu(x) =$$

$$= \int_{p_{\theta^*}^{-1}(\{0\})} 0\, d\mu(x) = 0.$$

The claim then follows using the subadditivity of probability measures. □

With this result, we can now prove the following lemma that is the basis for all of universal inference:

Lemma 2. *With the definition above, it follows that* $\mathbb{E}_{\theta^*}[S_n(\theta)] \leq 1$ *for all* $\theta \in \Theta$.

Proof. Since the two latter cases of the definition of $S_n(\theta)$ deal with events that happen with probability 0 for every $\theta \in \Theta$, we can ignore them when computing the expectation. It follows that for every $\theta \in \Theta$, $\mathbb{E}_{\theta^*}[S_n(\theta)] = \mathbb{E}_{\theta^*}[S_n(\theta)\ \mathbf{1}\big(\mathcal{L}_0(\theta^*) > 0\big)]$.[2] Observe now that $S_n(\theta)\ \mathbf{1}\big(\mathcal{L}_0(\theta^*) > 0\big) = \prod_{i \in \mathbf{D_0}} Z_i(\theta)$, where

$$Z_i(\theta) = \begin{cases} \frac{p_\theta(X_i)}{p_\theta^*(X_i)} & \text{if } p_\theta^*(X_i) > 0, \\ 0, & \text{if } p_\theta^*(X_i) = 0. \end{cases}$$

Since the X_i are independent, the Z_i are independent, too. Consequently,

$$\mathbb{E}_{\theta^*}[S_n(\theta)\ \mathbf{1}\big(\mathcal{L}_0(\theta^*) > 0\big)] = \mathbb{E}_{\theta^*}\big[\prod_{i \in \mathbf{D_0}} Z_i(\theta)\big] = \prod_{i \in \mathbf{D_0}} \mathbb{E}_{\theta^*}[Z_i(\theta)].$$

The result now follows immediately if we can show that $\mathbb{E}_{\theta^*}[Z_i(\theta)] \leq 1$ for every $i \in \mathbf{D_0}$ and $\theta \in \Theta$. Using the fact that p_θ^* is the true density function of every X_i, we get

$$\mathbb{E}_{\theta^*}[Z_i(\theta)] = \int_{p_{\theta^*}^{-1}(\,(0,+\infty)\,)} \frac{p_\theta}{p_\theta^*}\, p_\theta^*\, d\mu = \int_{p_{\theta^*}^{-1}(\,(0,+\infty)\,)} p_\theta\, d\mu \leq \int_{\mathcal{X}} p_\theta\, d\mu = 1,$$

where the last step follows from the fact that p_θ is a density with respect to μ for every $\theta \in \Theta$. □

Remark. Note that the set $p_{\theta^*}^{-1}(\,(0,+\infty)\,)$ is nothing else than the support of the true distribution of the data. Therefore, if for every possible distribution in the model, its support does not depend on the corresponding parameter value (in other words $p_\theta^{-1}(\,(0,+\infty)\,) = \mathcal{X}$ for all $\theta \in \Theta$) it holds that $\mathbb{E}_{\theta^*}[Z_i(\theta)] = 1$, which in turn implies $\mathbb{E}_{\theta^*}[S_n(\theta)] = 1$ for every $\theta \in \Theta$.

Next, Wasserman et al. (2020) define the following $[0, +\infty]$- valued statistic (again, only in an implicit fashion).

[2] Regarding the latter random variable, we use the convention that $+\infty \cdot 0 = 0$.

Definition 3. *For every $\theta \in \Theta$, the $[0, +\infty]$-valued random variable $T_n(\theta)$ is defined as*

$$T_n(\theta) = \begin{cases} \frac{\mathcal{L}_0(\hat{\theta}_1)}{\mathcal{L}_0(\theta)} & \text{if } \mathcal{L}_0(\theta) > 0, \\ 1 & \text{if } \mathcal{L}_0(\hat{\theta}_1) = \mathcal{L}_0(\theta) = 0, \\ +\infty & \text{if } \mathcal{L}_0(\hat{\theta}_1) > 0,\ \mathcal{L}_0(\theta) = 0. \end{cases}$$

Note that to make sense of this definition, we need to ensure that the function $\omega \mapsto \mathcal{L}_0\Big(\omega, \hat{\theta}_1(\omega)\Big)$ is measurable. Since $\mathcal{L}_0\Big(\omega, \hat{\theta}_1(\omega)\Big) = \prod_{i \in \mathbf{D_0}} p\Big(X_i(\omega), \hat{\theta}_1(\omega)\Big)$, a necessary and sufficient condition (which is not mentioned by Wasserman et al. (2020)) to achieve this is to require the function $p : \mathcal{X} \times \Theta \mapsto [0, +\infty)$ with $p(x, \theta) = p_\theta(x)$ to be $\mathcal{E} \bigotimes \mathcal{B}(\Theta) - \mathcal{B}([0, +\infty))$-measurable. We will assume that this requirement is met for the remainder of this work.

Leaving the discussion of measurability, observe that $T_n(\theta^*) = S_n(\hat{\theta}_1)$. Not surprisingly, we now want to argue that $\mathbb{E}_{\theta^*}[T_n(\theta^*)] \leq 1$. In the original paper on universal inference, the authors try to achieve this by using the tower property and conditioning on $X_{j_1}, \ldots, X_{j_m}$, where $\{j_1, \ldots, j_m\} = \mathbf{D_1}$. However, it is not immediately clear that the random variable $\mathbb{E}_{\theta^*}[T_n(\theta^*)|X_{j_1}, \ldots, X_{j_m}]$ is well-defined in the first place because if we use the concept of conditional expectation in the usual sense, we first would need to show why $T_n(\theta^*)$ is integrable, which in turn seems impossible without conditioning. Luckily, the usual definition of conditional expectation can be extended to (not necessarily integrable) *non-negative* random variables, as the following lemma shows (see Billingsley, 1995, Exercise 34.8).

Lemma 4. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and let $\mathcal{G} \subseteq \mathcal{F}$ be a sub-σ-algebra of $\mathcal{F}$. Suppose that Z is a random variable on this probability space, taking values in $[0, +\infty]$ and with possibly infinite expectation. Then there exists a $[0, +\infty]$-valued random variable Y that, apart from integrability, satisfies the defining properties of a conditional expectation of Z given $\mathcal{G}$, namely that Y is $\mathcal{G}$-measurable and that*[3]

$$\mathbb{E}[\mathbf{1}(G)\, Z] = \mathbb{E}[\mathbf{1}(G)\, Y],\ G \in \mathcal{G}. \tag{2.1}$$

Moreover, the random variable Y is uniquely defined almost surely.

Proof. We first prove the existence of Y: Define a sequence of random variables $(Z_n)_{n \in \mathbb{N}}$ with $Z_n = \min(n, Z)$. Observe that this sequence is non-negative, monotonically non-decreasing and that $\lim_{n \to \infty} Z_n = \sup_{n \in \mathbb{N}} Z_n = Z$. Furthermore, every random variable in this sequence is bounded and therefore integrable. Consequently, the conditional expectation of Z_n given $\mathcal{G}$, as it is defined in the usual way, exists for any $n \in \mathbb{N}$, and we can define a sequence $(Y_n)_{n \in \mathbb{N}}$ with Y_n being any version of $\mathbb{E}[Z_n \,|\, \mathcal{G}]$. Next, for any $n \in \mathbb{N}$ define the set $G_n = \{Y_{n+1} < Y_n\}$. By one of the defining properties of conditional expectation, all of the Y_n are $\mathcal{G}$-measurable and therefore $G_n \in \mathcal{G}$ for every $n \in \mathbb{N}$.

[3] Here, we again use the convention $0 \cdot \infty = 0$.

Observe now that for all $n \in \mathbb{N}$, the random variable $\mathbf{1}(G_n)(Y_n - Y_{n+1})$ is well-defined (because $0 \leq Y_n < +\infty$) and non-negative. However, it also holds that

$$\mathbb{E}[\mathbf{1}(G_n)\,(Y_n - Y_{n+1})] = \mathbb{E}[\mathbf{1}(G_n)\,(Z_n - Z_{n+1})] \leq 0$$

because $Y_n - Y_{n+1}$ is a version of $\mathbb{E}[Z_n - Z_{n+1} \,|\, \mathcal{G}]$ and $Z_{n+1} \geq Z_n$. From this, we conclude that $\mathbf{1}(G_n)(Y_n - Y_{n+1}) = 0$ almost surely (a.s.). On the other hand, it is easy to see that

$$\{\mathbf{1}(G_n)\,(Y_n - Y_{n+1}) = 0\} = G_n^c$$

and therefore $Y_{n+1} \geq Y_n$ a.s. for all $n \in \mathbb{N}$. It can be easily verified that this is equivalent to the sequence $(Y_n)_{n\in\mathbb{N}}$ being monotonically non-decreasing a.s. and, obviously, it is also non-negative. Now, define A as the subset of Ω on which $(Y_n)_{n\in\mathbb{N}}$ is monotonically non-decreasing and the $[0, +\infty]$-valued random variable Y as

$$Y(\omega) = \begin{cases} \lim_{n\to\infty} Y_n(\omega) \text{ if } \omega \in A, \\ 0 \text{ if } \omega \in A^c. \end{cases}$$

Note that $A \in \mathcal{G}$ and thus, Y is $\mathcal{G}$-measurable. Moreover, for any $G \in \mathcal{G}$ both the sequences $(\mathbf{1}(G)\,Z_n)_{n\in\mathbb{N}}$ and $(\mathbf{1}(G)\,Y_n)_{n\in\mathbb{N}}$ are non-negative and monotonically non-decreasing a.s., and $\lim_{n\to\infty} \mathbf{1}(G)\,Z_n = \mathbf{1}(G)\,Z$, $\lim_{n\to\infty} \mathbf{1}(G)\,Y_n = \mathbf{1}(G)\,Y$ a.s. Therefore, we can use the monotone convergence theorem as well as the fact that we defined Y_n to be a conditional expectation of X_n given $\mathcal{G}$ to argue that for any $G \in \mathcal{G}$,

$$\mathbb{E}[\mathbf{1}(G)\,Z] = \lim_{n\to\infty} \mathbb{E}[\mathbf{1}(G)\,Z_n] = \lim_{n\to\infty} \mathbb{E}[\mathbf{1}(G)\,Y_n] = \mathbb{E}[\mathbf{1}(G)\,Y].$$

This proves the existence of a random variable Y which we can call a conditional expectation of Z given $\mathcal{G}$, even if Z is non-integrable.

It remains to show that the random variable Y is uniquely defined almost surely. Suppose that $\tilde{Y}$ is another $\mathcal{G}$-measurable random variable satisfying (2.1). For any $n \in \mathbb{N}$, define the set $G'_n = \{\tilde{Y} > Y,\, n > Y\}$. Obviously, $G'_n \in \mathcal{G}$ for all $n \in \mathbb{N}$. Furthermore, the random variables $\big(\mathbf{1}(G'_n)\,(\tilde{Y} - Y)\big)_{n\in\mathbb{N}}$ are well-defined because Y only takes values in $[0, +\infty)$ on the set G'_n and clearly, all of them are non-negative. Furthermore, for every $n \in \mathbb{N}$, the expectation $\mathbb{E}[\mathbf{1}(G'_n)\,Y]$ is finite because $\mathbf{1}(G'_n)\,Y$ is bounded. Consequently, we can use the linearity of expectation to get

$$\mathbb{E}[\mathbf{1}(G'_n)\,(\tilde{Y} - Y)] = \mathbb{E}[\mathbf{1}(G'_n)\,\tilde{Y}] - \mathbb{E}[\mathbf{1}(G'_n)\,Y] = \mathbb{E}[\mathbf{1}(G'_n)\,Z] - \mathbb{E}[\mathbf{1}(G'_n)\,Z] = 0,$$

where we used the fact that both Y and $\tilde{Y}$ are a conditional expectation of Z given $\mathcal{G}$. We conclude that $\mathbf{1}(G'_n)\,(\tilde{Y} - Y) = 0$ a.s. for all $n \in \mathbb{N}$. However, this immediately implies that all of the G'_n are null sets. Now, since $\{\tilde{Y} > Y\} \subseteq \{Y < \infty\}$, we may write the event $\{\tilde{Y} > Y\}$ as

$$\{\tilde{Y} > Y\} = \bigcup_{n\in\mathbb{N}} G'_n$$

and because $G_1' \subseteq G_2' \subseteq \cdots \subseteq \bigcup_{n\in\mathbb{N}} G_n' = \{\tilde{Y} > Y\}$, we can use the continuity from below of measures to arrive at the conclusion that $\{\tilde{Y} > Y\}$ is a null set, too. By repeating this argument with switched roles of Y and $\tilde{Y}$, we can verify that $\mathbb{P}(Y > \tilde{Y}) = 0$ as well. This proves that if $\tilde{Y}$ is another conditional expectation of Z given $\mathcal{G}$, then $\tilde{Y} = Y$ a.s. □

The result above is useful in our case because obviously, $T_n(\theta) > 0$ for all $\theta \in \Theta$. Furthermore, we can then use the following well-known property of conditional expectations (see, for instance, Durrett, 2019, Example 4.1.7):

Lemma 5. *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space with independent random variables $X : (\Omega, \mathcal{A}) \mapsto (\mathcal{X}, \mathcal{E})$ and $Y : (\Omega, \mathcal{A}) \mapsto (\mathcal{Y}, \mathcal{F})$. Furthermore, let $\varphi : (\mathcal{X} \times \mathcal{Y}, \mathcal{E} \bigotimes \mathcal{F}) \mapsto (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ be a measurable function with $\mathbb{E}[\,|\varphi(X,Y)|\,] < \infty$. Then*

$$\mathbb{E}[\varphi(X,Y)\,|\,Y] = g(Y) \text{ a.s.},$$

where $g : (\mathcal{Y}, \mathcal{F}) \mapsto ((\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is the measurable function defined by

$$g(y) = \mathbb{E}[\varphi(X, y)].$$

Here, we may replace the assumption that $\mathbb{E}[\,|\varphi(X,Y)|\,] < \infty$ with the requirement that $\varphi(X,Y) \geq 0$ because the proof of this result only relies on the evaluation of a double integral as an iterated integral, which by Tonelli's theorem is possible if the corresponding function is non-negative (but not necessarily integrable), too. With this property of conditional expectation in place, the following lemma is a direct result of Lemma 2:

Lemma 6. *With the definition of $T_n(\theta)$ given above, it follows that $\mathbb{E}_{\theta^*}[T_n(\theta^*)] = \mathbb{E}_{\theta^*}[S_n(\hat{\theta}_1)] \leq 1$, where θ^* is the true parameter of the model.*

Proof. Setting $\mathbf{D_0} = \{i_1, \ldots, i_l\}$, we first note that $\mathbb{E}_{\theta^*}[S_n(\hat{\theta}_1)|\,\hat{\theta}_1]$ is well-defined because $S_n(\hat{\theta}_1)$ is $\mathbb{P}_{\theta^*}$-almost surely non-negative. Furthermore, we can write $\mathbb{E}_{\theta^*}[S_n(\hat{\theta}_1)|\,\hat{\theta}_1)]$ as $\mathbb{E}_{\theta^*}[\varphi(Z_0, \hat{\theta}_1)|\,\hat{\theta}_1]$, where $Z_0 = (X_{i_1}, \ldots, X_{i_l})^T$ and $\varphi : \mathcal{X}^l \times \Theta \mapsto [0, +\infty)$ with

$$\varphi(z_0, \theta) = \begin{cases} \prod_{k=1}^{l} \frac{p_\theta\left(z_0^{(k)}\right)}{p_{\theta^*}\left(z_0^{(k)}\right)} & \text{if } \prod_{k=1}^{l} p_{\theta^*}\left(z_0^{(k)}\right) > 0, \\ 1 & \text{if } \prod_{j=1}^{l} p_\theta\left(z_0^{(j)}\right) = \prod_{k=1}^{l} p_{\theta^*}\left(z_0^{(k)}\right) = 0, \\ +\infty & \text{if } \prod_{j=1}^{l} p_\theta\left(z_0^{(j)}\right) > 0,\ \prod_{k=1}^{l} p_{\theta^*}\left(z_0^{(k)}\right) = 0. \end{cases}$$

Observe that Z_0 is independent of $\hat{\theta}_1$ because $\hat{\theta}_1$ is a measurable function of the data indexed by $\mathbf{D_1}$. By Lemma 5, it follows that $\mathbb{E}_{\theta^*}[\varphi(Z_0, \hat{\theta}_1)|\,\hat{\theta}_1] = g(\hat{\theta}_1)$ $\mathbb{P}_{\theta^*}$-a.s., where $g(\theta) = \mathbb{E}_{\theta^*}[\varphi(Z_0, \theta)] = \mathbb{E}_{\theta^*}[S_n(\theta)]$. Since we know from Lemma 2 that $\mathbb{E}_{\theta^*}[S_n(\theta)] \leq 1$ for all $\theta \in \Theta$, we get $\mathbb{E}_{\theta^*}[S_n(\hat{\theta}_1)|\,\hat{\theta}_1] = g(\hat{\theta}_1) \leq 1$ $\mathbb{P}_{\theta^*}$-a.s. and consequently by using the tower property (which holds for the extended definition of conditional expectation given above as well) it follows that $\mathbb{E}_{\theta^*}[T_n(\theta^*)] = \mathbb{E}_{\theta^*}[S_n(\hat{\theta}_1)] = \mathbb{E}_{\theta^*}[\,\mathbb{E}_{\theta^*}[S_n(\hat{\theta}_1)|\,\hat{\theta}_1]\,] \leq 1$. □

Remark. It follows immediately from the previous remark that in the case where the support of all the distributions in the model does not depend on the corresponding parameter value, we have $\mathbb{E}_{\theta^*}[T_n(\theta^*)] = 1$.

We now turn to the definition of the split likelihood ratio confidence set as it was stated by Wasserman et al. (2020).

Definition 7. *Suppose we observe data $X_1, \ldots, X_n$ from the statistical model previously described. Then for a given $\alpha \in (0,1)$ and $T_n(\theta)$ defined as above, we call*

$$C_n(\alpha) = \left\{\theta \in \Theta : T_n(\theta) \leq \frac{1}{\alpha}\right\}$$

the split likelihood ratio (split LR) confidence set for the parameter θ^* with nominal coverage probability $1-\alpha$.

The main theorem of their paper then explains why we can call $C_n(\alpha)$ a confidence set for θ^* with coverage probability *at least* $1-\alpha$. It is a direct consequence of Lemma 6.

Theorem 8. *With this definition of $C_n(\alpha)$, it follows that*

$$\mathbb{P}_{\theta^*}(\theta^* \in C_n(\alpha)) \geq 1 - \alpha.$$

Proof. We will show the equivalent claim that $\mathbb{P}_{\theta^*}(\theta^* \notin C_n(\alpha)) \leq \alpha$. Observe that $\{\omega \in \Omega : \theta^* \notin C_n(\alpha)\} = \{\omega \in \Omega : T_n(\theta^*) > \frac{1}{\alpha}\}$. Using Markov's inequality, which we can apply because $T_n(\theta) \geq 0$ for all $\theta \in \Theta$, as well as Lemma 6, we get the desired result:

$$\mathbb{P}_{\theta^*}(\theta^* \notin C_n(\alpha)) = \mathbb{P}_{\theta^*}\left(T_n(\theta^*) > \frac{1}{\alpha}\right) \leq \left(\frac{1}{\alpha}\right)^{-1} \mathbb{E}_{\theta^*}[(T_n(\theta^*)] \leq \alpha.$$

□

Remark. i) The confidence set defined in this way can also be expressed as

$$C_n(\alpha) = \left\{\theta \in \Theta : \mathcal{L}_0(\theta)\, \frac{1}{\alpha} \geq \mathcal{L}_0(\hat{\theta}_1)\right\}.$$

ii) The estimator for θ^* based on the subset $\mathbf{D_1}$ of the data, $\hat{\theta}_1$, is always included in this set because $\frac{1}{\alpha} > 1$ for $\alpha \in (0,1)$.

iii) If it exists, the maximum likelihood (ML) estimator based on the data indexed by $\mathbf{D_0}$, $\hat{\theta}_0{}^{(\mathrm{ML})}$, is always included in the split LR confidence set because $\mathcal{L}_0\left(\hat{\theta}_0{}^{(\mathrm{ML})}\right) \geq \mathcal{L}_0(\hat{\theta}_1)$.

iv) $C_n(\alpha)$ could also be defined as

$$C_n(\alpha) = \left\{\theta \in \Theta : T_n(\theta) < \frac{1}{\alpha}\right\},$$

because Markov's inequality holds for events of the type $\{Z \geq a\}$ (for non-negative random variables Z and $a > 0$) as well. Possibly, the authors did not want to define the split LR confidence set in this way because in many situations, $T_n(\theta)$ will be a continuous function of θ ($\mathbb{P}_{\theta^*}$-almost surely) and in this case, their definition of $C_n(\alpha)$ will give a closed subset of Θ, while the one stated here will give an open one. When $\Theta \subseteq \mathbb{R}$, confidence sets are usually written as closed intervals, while the variant of $C_n(\alpha)$ described here would lead to open intervals for continuous $T_n(\theta)$.

It was noted previously that the full definition of the split LR confidence set from above was stated by Wasserman et al. (2020) only implicitly. To be precise, the 'official' definition of $C_n(\alpha)$ given in their paper is not complete because the definition of the statistic $T_n(\theta)$, which the confidence set is based on, lacks a discussion of the case when for certain θ, we have both $\mathcal{L}_0(\hat{\theta}_1)$ and $\mathcal{L}_0(\theta)$ equal to 0, which yields $T_n(\theta) = \frac{0}{0}$. Only in an example included in section 3 of the paper, the authors clarify that they set $\frac{0}{0} = 1$ in the definition of $T_n(\theta)$. The example is presented here because it also points to the split LR confidence set being 'unnecessarily' large if it is defined in this way.

Example 9. Consider the following statistical model: $X_1, \ldots, X_n \overset{i.i.d}{\sim} U([0, \theta])$ with parameter space $\Theta = (0, +\infty)$ and true parameter $\theta^* > 0$. Assume that n is even and that $|\mathbf{D_0}| = |\mathbf{D_1}| = \frac{n}{2}$. Furthermore, set $\hat{\theta}_1$ to be the maximum likelihood estimator based on the subset of the data indexed by $\mathbf{D_1}$, so $\hat{\theta}_1 = \max_{i \in \mathbf{D_1}} X_i$. Obviously, $\mathbb{P}_{\theta^*}(\hat{\theta}_1 < \hat{\theta}_0) = \frac{1}{2}$, where $\hat{\theta}_0 = \max_{i \in \mathbf{D_0}} X_i$. However, if the event $\{\hat{\theta}_1 < \hat{\theta}_0\}$ happens, we have $\mathcal{L}_0(\hat{\theta}_1) = 0$ because then for $j \in \mathbf{D_0}$ with $X_j = \max_{i \in \mathbf{D_0}} X_i$, we get $p_{\hat{\theta}_1}(X_j) = 0$. As a consequence, conditional on this event we have $T_n(\theta) = \mathbf{1}(\mathcal{L}_0(\theta) = 0) < \frac{1}{\alpha}$ for all $\theta \in \Theta$ and $\alpha \in (0, 1)$, which leads to $C_n(\alpha) = \left\{\theta \in \Theta : T_n(\theta) \leq \frac{1}{\alpha}\right\} = \Theta$. In other words, in this statistical model, the split LR confidence set with the specification $\frac{0}{0} = 0$ for $T_n(\theta)$ gives the most trivial confidence set possible with probability $\frac{1}{2}$ when $\hat{\theta}_1$ is taken to be the ML estimator.

More generally, it is easy to see that if $\mathbb{P}_{\theta^*}\big(\mathcal{L}_0(\hat{\theta}_1) = 0\big) > 0$, then $\mathbb{P}_{\theta^*}\big(C_n(\alpha) = \Theta \,|\, \mathcal{L}_0(\hat{\theta}_1) = 0\big) = 1$. Wasserman et al. (2020) argue at this point that using a different approach based on the universal inference principle (which will be discussed later) leads to a more sensible confidence set for θ^*. However, there is an obvious possibility of improvement even without adapting the general approach when constructing the confidence set for this model: If we would set $\frac{0}{0} = +\infty$ in the context of $T_n(\theta)$, in Example 9, we would have $T_n(\theta) = 0$ for $\theta \geq \hat{\theta}_0$ and $T_n(\theta) = +\infty$ otherwise, given the event $\{\hat{\theta}_1 < \hat{\theta}_0\}$. This would lead to the confidence set $C_n(\alpha) = [\hat{\theta}_0, +\infty)$, which is still in some sense trivial given the observed data, but at least considers the observations for the construction of the set. Therefore, this example motivates a slight adaption regarding the definition of the split LR confidence set, which will be discussed in the following subsection.

2.1.2 Improved definition of the split LR confidence set

In which sense exactly is the split LR confidence set, as it was originally defined, 'unnecessarily' large? Note that we can write this set as $C_n(\alpha) = A_n(\alpha) \cup B_n$, where

$$A_n(\alpha) = \Big\{\theta \in \Theta : \mathcal{L}_0(\theta)\, \frac{1}{\alpha} \geq \mathcal{L}_0(\hat{\theta}_1),\ \mathcal{L}_0(\theta) > 0\Big\},$$

$$B_n = \Big\{\theta \in \Theta : \mathcal{L}_0(\theta)\ = \mathcal{L}_0(\hat{\theta}_1) = 0\Big\}.$$

Obviously, $\mathbb{P}_{\theta^*}(\theta^* \in B_n) = 0$ for all n. Therefore, $\mathbb{P}_{\theta^*}(\theta^* \in C_n) = \mathbb{P}_{\theta^*}(\theta^* \in A_n)$, so the set $A_n(\alpha)$ has a coverage probability of *at least* $1 - \alpha$, too. The set $A_n(\alpha)$ now can be easily defined, similar to $C_n(\alpha)$, via a statistic.

Definition 10. *For every $\theta \in \Theta$, we define the $[0, +\infty]$-valued random variable $T'_n(\theta)$ as*

$$T'_n(\theta) = \begin{cases} \frac{\mathcal{L}_0(\hat{\theta}_1)}{\mathcal{L}_0(\theta)} & \text{if } \mathcal{L}_0(\theta) > 0, \\ +\infty, & \text{if } \mathcal{L}_0(\theta) = 0. \end{cases}$$

Note that, just as in the definition of $T_n(\theta)$, the measurability of $T'_n(\theta)$ cannot be guaranteed without the assumption that the map $(x, \theta) \mapsto p_\theta(x)$ is joint measurable on $\mathcal{X} \times \Theta$. Using this new definition, we next define the split likelihood ratio confidence set based on the statistic $T'_n(\theta)$ analogously to the original definition by Wasserman et al. (2020).

Definition 11. *Suppose we observe data $X_1, \ldots, X_n$ from the statistical model previously described. Then for a given $\alpha \in (0, 1)$ and $T'_n(\theta)$ defined as above, we call*

$$C'_n(\alpha) = \Big\{\theta \in \Theta : T'_n(\theta) \leq \frac{1}{\alpha}\Big\}$$

*the **improved** split likelihood ratio (split LR) confidence set for the parameter θ^* with nominal coverage probability $1 - \alpha$.*

It is easy to see that this confidence set has coverage probability of *at least* $1 - \alpha$ for the true parameter θ^* as well because $C'_n(\alpha) = A_n(\alpha)$ for every $\alpha \in (0, 1)$, $n \in \mathbb{N}$. One could also verify this claim by arguing that $\mathbb{E}_{\theta^*}[\, T'_n(\theta^*)] = \mathbb{E}_{\theta^*}[\, T_n(\theta^*)\,] \leq 1$ because the values of the random variables $T'_n(\theta^*)$ and $T_n(\theta^*)$ agree almost everywhere (remember that $\mathbb{P}_{\theta^*}(\mathcal{L}_0(\theta^*) = 0) = 0$) and then applying the steps from the proof of Theorem 8 again.

Why can we call $C'_n(\alpha)$ an **improved** adaption of the original split LR confidence set $C_n(\alpha)$? Because obviously, $C'_n(\alpha) \subseteq C_n(\alpha)$ for all $\omega \in \Omega$, $\alpha \in (0, 1)$ and in the case where $\mathbb{P}_{\theta^*}(\mathcal{L}_0(\hat{\theta}_1) = 0) > 0$, the event $\{C'_n(\alpha) \subset C_n(\alpha)\}$ happens with positive probability for all $\alpha \in (0, 1)$. Therefore, we could say that in some sense, the confidence set $C'_n(\alpha)$ is never larger than $C_n(\alpha)$, but in some cases (see, for instance, Example 9) $C'_n(\alpha)$ is smaller than $C_n(\alpha)$ with positive probability.

Remark. The following statements hold for all $\alpha \in (0,1)$:

i) By an argument analogous to the one made in point iii) of the previous remark, $C'_n(\alpha)$ could also be defined as

$$C'_n(\alpha) = \Big\{\theta \in \Theta : T'_n(\theta) < \frac{1}{\alpha}\Big\}$$

and still have a coverage probability of *at least* $1-\alpha$ for θ^*. The confidence set can also be expressed as

$$C'_n(\alpha) = \Big\{\theta \in \Theta : \mathcal{L}_0(\theta)\ \frac{1}{\alpha} > \mathcal{L}_0(\hat{\theta}_1)\Big\}.$$

ii) Assume that $0 < \mathbb{P}_{\theta^*}\big(\mathcal{L}_0(\hat{\theta}_1) = 0\big) < 1$. Then $\mathbb{P}_{\theta^*}\big(\hat{\theta}_1 \in C'_n(\alpha) \,|\, \mathcal{L}_0(\hat{\theta}_1) = 0\big) = 0$ because $\{\mathcal{L}_0(\hat{\theta}_1) = 0\} = \{T'_n(\hat{\theta}_1) = +\infty\}$. On the other hand, $\mathbb{P}_{\theta^*}\big(\hat{\theta}_1 \in C'_n(\alpha) \,|\, \mathcal{L}_0(\hat{\theta}_1) > 0\big) = 1$ because $\{\mathcal{L}_0(\hat{\theta}_1) > 0\} = \{T'_n(\hat{\theta}_1) = 1\} \subseteq \{T'_n(\hat{\theta}_1) < \frac{1}{\alpha}\}$. If $\mathbb{P}_{\theta^*}\big(\mathcal{L}_0(\hat{\theta}_1) = 0\big) = 0$, then $\mathbb{P}^*_{\theta}\big(\hat{\theta}_1 \in C'_n(\alpha)\big) = 1$.

iii) Assume that $0 < \mathbb{P}_{\theta^*}\big(\mathcal{L}_0(\hat{\theta}_1) = 0\big) < 1$. Then $\mathbb{P}_{\theta^*}\big(C'_n(\alpha) \subset C_n(\alpha) \,|\, \mathcal{L}_0(\hat{\theta}_1) = 0\big) = 1$. On the other hand, $\mathbb{P}_{\theta^*}\big(C'_n(\alpha) = C_n(\alpha) \,|\, \mathcal{L}_0(\hat{\theta}_1) > 0\big) = 1$. If $\mathbb{P}_{\theta^*}\big(\mathcal{L}_0(\hat{\theta}_1) = 0\big) = 0$, then $\mathbb{P}_{\theta^*}\big(C'_n(\alpha) = C_n(\alpha)\big) = 1$.

iv) If the support of all the distributions in the model does not depend on the underlying parameter value, then we have $\mathbb{P}_{\theta^*}\big(\mathcal{L}_0(\hat{\theta}_1) = 0\big) = 0$.

v) If the support of all the distributions in the model does not depend on the underlying parameter value, we have $\mathbb{E}_{\theta^*}[T'_n(\theta^*)] = 1$.

vi) If it exists, the Maximum Likelihood estimator based on the data indexed by $\mathbf{D_0}$, $\hat{\theta}_0^{\,(\mathrm{ML})}$, is always included in the improved split LR confidence set.

2.1.3 What does minimal coverage probability mean in practice?

At this point, an important remark has to be made regarding the fact that the split LR confidence set with nominal coverage probability $1-\alpha$, regardless of which of the variants introduced above is chosen, always has an actual coverage probability of *at least* $1-\alpha$ for the true parameter. At first glance, this might not seem like a significant issue for statistical inference - for example, what would be the harm in getting a confidence interval with true coverage probability $\approx 96\%$ instead of the nominal value 95%? Two different aspects need to be taken into account here:

First, it has to be noted that the nominal coverage probability of the split LR confidence set in no way can be seen as an *approximation* to the actual one, as it is the case for so many other confidence sets constructed for parametric models. Usually, when we say that a confidence set has *nominal* coverage probability $1-\alpha$, this either means that the true coverage probability for the parameter asymptotically approaches this value, or, that in

some sense the actual coverage probability is as close to $1-\alpha$ as it can get - specifically, the latter interpretation often becomes relevant for confidence sets with nominal coverage probabilities in discrete models. Regarding the first type of approximation, confidence sets that attain their nominal coverage probability asymptotically are usually interpreted in the context of statistical applications as "*If the sample size is large, then the true coverage probability will be reasonably close to the nominal one*". However, a statement of this kind should be considered wrong in the context of the split LR principle. For this confidence set, we know for sure that the true coverage probability is *larger than or equal to* the nominal coverage probability for every sample size n, but without further investigations, we cannot be sure whether it ever will be close to the nominal value or not, even in an asymptotic sense.

This non-approximation feature of universal inference is due to the fact that the minimum coverage probability was derived via Markov's inequality, which in general should not be viewed as an approximation, but a bound that is only achieved in rare instances: It is easy to see that for any $a \in (0, +\infty)$, the equality $\mathbb{P}(X \geq a) = \frac{\mathbb{E}[X]}{a}$ only holds for those non-negative random variables X (living on the probability space $(\Omega, \mathcal{A}, \mathbb{P})$) which obey $\mathbb{P}(X \in \{0, a\}) = 1$. Furthermore, it is known that $\frac{\mathbb{E}[X]}{a}$ typically overestimates the probability $\mathbb{P}(X \geq a)$ significantly. For example, for X exponentially distributed with parameter $\lambda > 0$, Markov's inequality bounds $\mathbb{P}(X \geq a)$ from above with $\frac{1}{\lambda a}$, while the true value of this probability is $e^{-\lambda a}$. Here, the Markov bound is trivial for $a \leq \frac{1}{\lambda}$ and the difference between the bound and the actual value of $\mathbb{P}(X \geq a)$ monotonically decreases to 0 for $a \to \infty$, but it does so very slowly for small values of λ. As a consequence of this 'non-approximation' property of Markov's inequality, in most situations, the true coverage probability of a split LR confidence set will overshoot the nominal one significantly.

Whether or not this method leads to an in some sense reasonable confidence set depends on the statistical model considered, the proportion between the number of observations in the two data splits $\mathbf{D_0}$ and $\mathbf{D_1}$, as well as on the choice of the estimator $\hat{\theta}_1$: Since in the definition by Wasserman et al. (2020), there is virtually no restriction regarding this estimator, even some estimators that are usually considered useless, such as $\hat{\theta}_1 = \theta'$ for some $\theta' \in \Theta$, can be chosen. Naturally, such 'bad' choices for the estimator $\hat{\theta}_1$ can lead to actual coverage probabilities that overshoot the nominal ones by far for every sample size n. Due to the freedom regarding the estimator $\hat{\theta}_1$, it is even possible to derive confidence sets from the split LR approach that correspond to the whole parameter space Θ (with probability 1) for every sample size n, which of course means that the coverage probability of this set will be equal to 1 (see Example 12 below).

Secondly, this discrepancy between nominal and actual coverage probability, which is not guaranteed to disappear asymptotically, in many cases will have an impact on both the diameter[4] and the volume (Lebesgue measure) of the confidence set. Usually, a higher coverage probability is tied to a larger confidence set (with respect to its volume and/or

[4]Recall that for a metric space (M, d), the diameter of a set $A \subseteq M$ is defined as $\operatorname{diam}(A) = \sup_{x,y \in A} d(x, y)$.

diameter, on average or almost surely) and vice versa. Though no universal performance measure regarding confidence sets exists in the literature, one usually relies on one of these figures when comparing two confidence sets with the same (minimum) coverage probabilities and declares the one with the shorter diameter or volume (on average or almost surely) to be 'better' (see, for instance, Joshi, 1966; Casella and Hwang, 1991; Robins and Van der Vaart, 2006). Naturally, for many statistical models where confidence sets with exact coverage probabilities are known (see, for instance, Example 13), the set derived from the split likelihood ratio method will be inadmissible with respect to any useful definition of admissibility of confidence sets.

In light of these assessments regarding the split LR confidence set, Wasserman et al. (2020) state that the universal inference approach mostly is going to be useful in irregular models where no inference method with either finite sample or asymptotic guarantees is available (yet) - this is the case, for example, in Gaussian mixture models. In other words, the usefulness of the split LR approach in practice most likely is going to be limited to those models, in which it is not possible to compare it with other methods for constructing confidence sets because such methods are not known (yet).

Example 12. Consider the following statistical model: $X_i \overset{\text{i.i.d.}}{\sim} U(\{1,\dots,N\})$, where $N \in \{m,\dots,M\}$ for some $m, M \in \mathbb{N}$. Take $\hat{N}_1 = M$ as the estimator for the parameter N based on the subset of the data indexed with $\mathbf{D_1}$ and suppose that the true parameter N^* equals m. Then, we arrive for $T_n'(\cdot)$ from the improved definition of the split LR confidence set at

$$\begin{aligned} T_n'(N) &= \begin{cases} \left(\frac{\hat{N}_1{}^{-1}}{N^{-1}}\right)^{|\mathbf{D_0}|} & \text{if } \mathcal{L}_0(N) > 0, \\ +\infty & \text{if } \mathcal{L}_0(N) = 0 \end{cases} \\ &= \begin{cases} \left(\frac{M^{-1}}{N^{-1}}\right)^{|\mathbf{D_0}|} & \text{if } N \geq \max_{i\in\mathbf{D_0}} X_i, \\ +\infty & \text{if } N < \max_{i\in\mathbf{D_0}} X_i \end{cases} \\ &= \left(\frac{N}{M}\right)^{|\mathbf{D_0}|} \mathbb{P}_m\text{-a.s.} \end{aligned}$$

for any $N \in \{m,\dots,M\}$, since $\max_{i\in\mathbf{D_0}} X_i \leq m$ a.s. Consequently, for any $\alpha \in (0,1)$, the split LR confidence set (following the improved definition) boils down to

$$\begin{aligned} C_n'(\alpha) &= \left\{N \in \{m,\dots,M\} : T_n'(N) \leq \frac{1}{\alpha}\right\} \\ &= \left\{N \in \{m,\dots,M\} : \left(\frac{N}{M}\right)^{|\mathbf{D_0}|} \leq \frac{1}{\alpha}\right\} \\ &= \left\{N \in \{m,\dots,M\} : N \leq \left(\frac{1}{\alpha}\right)^{\frac{1}{|\mathbf{D_0}|}} M\right\} = \{m,\dots,M\}\ \mathbb{P}_m\text{-a.s.} \end{aligned}$$

where the last step follows from the fact that $\frac{1}{\alpha} > 1$ for any $\alpha \in (0,1)$. Therefore, we see that in this situation the split LR confidence set constructed with the specific estimator $\hat{\theta}_1 = M$ boils down to the whole parameter space for any sample size $n \in \mathbb{N}$ with probability 1, regardless of the values of $\mathbf{D_0}$ and $\mathbf{D_1}$.

Example 13. Consider the statistical model from Example 9 (uniform distribution). Just as in this example, take $\hat{\theta}_1 = \max_{i\in\mathbf{D_1}} X_i$ to be the maximum likelihood estimator based on the data indexed by $\mathbf{D_1}$, assume that the total number of observations n is even and that $|\mathbf{D_0}| = |\mathbf{D_1}| = \frac{n}{2}$. Then we have already seen that, even if we use the improved definition of the split LR confidence set, we have

$$\mathbb{P}_{\theta^*}\Big(C_n(\alpha) = [\hat{\theta}_0, +\infty)\Big) = \frac{1}{2}$$

because $\mathbb{P}_{\theta^*}\Big(C_n(\alpha) = [\hat{\theta}_0, +\infty) \,|\, \hat{\theta}_1 < \hat{\theta}_0\Big) = 1$. Moreover, it is easy to see that conditional on the event $\{\hat{\theta}_1 < \hat{\theta}_0\}$, the true coverage probability is $\mathbb{P}_{\theta^*}\Big(\theta^* \in C_n(\alpha) \,|\, \hat{\theta}_1 < \hat{\theta}_0\Big) = 1$. On the other hand, conditional on the event $\{\hat{\theta}_1 \geq \hat{\theta}_0\}$ (which happens with probability $\frac{1}{2}$, too) we get $T_n(\theta) = +\infty$ for $\theta < \hat{\theta}_0 = \max_{i\in\mathbf{D_0}} X_i$, and for $\theta \geq \hat{\theta}_0$

$$T_n(\theta) = \frac{\mathcal{L}_0(\hat{\theta}_1)}{\mathcal{L}_0(\theta)} = \frac{\hat{\theta}_1^{-|\mathbf{D_0}|}}{\theta^{-|\mathbf{D_0}|}} = \Big(\frac{\theta}{\hat{\theta}_1}\Big)^{\frac{n}{2}}.$$

Consequently, conditional on $\{\hat{\theta}_1 \geq \hat{\theta}_0\}$, the split LR confidence set turns out to be, for any $\alpha \in (0,1)$,

$$\begin{aligned} C_n(\alpha) &= \Big\{\theta \in (0,+\infty) : T_n(\theta) \leq \frac{1}{\alpha}\Big\} = \Big\{\theta \in (0,+\infty) : \theta \geq \hat{\theta}_0,\ \Big(\frac{\theta}{\hat{\theta}_1}\Big)^{\frac{n}{2}} \leq \frac{1}{\alpha}\Big\} \\ &= \Big[\hat{\theta}_0,\ \hat{\theta}_1\Big(\frac{1}{\alpha}\Big)^{\frac{2}{n}}\Big]. \end{aligned}$$

Furthermore, conditional on this event, the actual coverage probability for the true parameter θ^* is given by

$$\begin{aligned} &\mathbb{P}_{\theta^*}\Big(\theta^* \in C_n(\alpha) \,|\, \hat{\theta}_1 \geq \hat{\theta}_0\Big) = \mathbb{P}_{\theta^*}\Big(\theta^* \in \Big[\hat{\theta}_0,\ \hat{\theta}_1\Big(\frac{1}{\alpha}\Big)^{\frac{2}{n}}\Big] \,|\, \hat{\theta}_1 \geq \hat{\theta}_0\Big) \\ &= \mathbb{P}_{\theta^*}\Big(\hat{\theta}_1\Big(\frac{1}{\alpha}\Big)^{\frac{2}{n}} \geq \theta^* | \max_{i=1,\dots,n} X_i = \hat{\theta}_1\Big) = \mathbb{P}_{\theta^*}\Big(\max_{i=1,\dots,n} X_i \geq \alpha^{\frac{2}{n}}\,\theta^* | \max_{i=1,\dots,n} X_i = \hat{\theta}_1\Big) \\ &= \mathbb{P}_{\theta^*}\Big(\max_{i=1,\dots,n} X_i \geq \alpha^{\frac{2}{n}}\,\theta^*\Big) = 1-\alpha^2, \end{aligned}$$

where the last step follows from the independence of the random variables $\max_{i=1,\dots,n} X_i$ and $\mathbf{1}(\max_{i=1,\dots,n} X_i = \hat{\theta}_1)$.

All in all, we can describe the split LR confidence set in this model, with this specific choice of $\hat{\theta}_1$ and $|\mathbf{D_0}| = |\mathbf{D_1}| = \frac{n}{2}$, by

$$\mathbb{P}_{\theta^*}\Big(C_n(\alpha) = \Big[\hat{\theta}_0, +\infty\Big)\Big) = \mathbb{P}_{\theta^*}\Big(C_n(\alpha) = \Big[\hat{\theta}_0,\ \hat{\theta}_1\Big(\frac{1}{\alpha}\Big)^{\frac{2}{n}}\Big]\Big) = \frac{1}{2}.$$

The true coverage probability of this confidence set turns out to be

$$\begin{aligned} &\mathbb{P}_{\theta^*}\Big(\theta^* \in C_n(\alpha)\Big) = \frac{1}{2}\,\mathbb{P}_{\theta^*}\Big(\theta \in C_n(\alpha) \,|\, \hat{\theta}_1 < \hat{\theta}_0\Big) + \frac{1}{2}\,\mathbb{P}_{\theta^*}\Big(\theta \in C_n(\alpha) \,|\, \hat{\theta}_1 \geq \hat{\theta}_0\Big) \\ &= \frac{1}{2} + \frac{1-\alpha^2}{2} = 1 - \frac{\alpha^2}{2}. \end{aligned}$$

We see that in this case, the actual coverage probability overshoots the nominal one by $\alpha - \frac{\alpha^2}{2}$. Therefore, the larger the targeted coverage probability $1-\alpha$, the smaller the overshooting of this target by the split LR confidence set. As an example, for the common choice of confidence $1-\alpha = 0.95$, the actual coverage probability of this set is ≈ 0.9988. On the other hand, in this statistical model, the classical confidence set for the true parameter θ^* with (exact) coverage probability $1-\alpha$ is given by

$$K_n(\alpha) = \left[\hat{\theta}_{\mathrm{ML}},\, \hat{\theta}_{\mathrm{ML}} \left(\frac{1}{\alpha}\right)^{\frac{1}{n}}\right],$$

where $\hat{\theta}_{\mathrm{ML}} = \max_{i=1,\ldots,n} X_i$ is the maximum likelihood estimator for θ^* based on the whole sample.

It is easy to see that the length of $K_n(\alpha)$ is shorter than the one of $C_n(\alpha)$ with probability 1: If we denote the length of $C_n(\alpha)$ as $l_n^{(C)}(\alpha)$ and the length of $K_n(\alpha)$ as $l_n^{(K)}(\alpha)$, then conditional on $\{\hat{\theta}_1 < \hat{\theta}_0\}$, we have

$$l_n^{(C)}(\alpha) = +\infty > l_n^{(K)}(\alpha) = \hat{\theta}_{\mathrm{ML}} \frac{1-\alpha^{\frac{1}{n}}}{\alpha^{\frac{1}{n}}},$$

and conditional on $\{\hat{\theta}_1 > \hat{\theta}_0\}$, we get

$$\begin{aligned} l_n^{(C)}(\alpha) &= \hat{\theta}_1 \left(\frac{1}{\alpha}\right)^{\frac{2}{n}} - \hat{\theta}_0 = \hat{\theta}_{\mathrm{ML}} \left(\frac{1}{\alpha}\right)^{\frac{2}{n}} - \hat{\theta}_0 > \hat{\theta}_{\mathrm{ML}} \left(\frac{1}{\alpha}\right)^{\frac{2}{n}} - \hat{\theta}_{\mathrm{ML}} \\ &= \hat{\theta}_{\mathrm{ML}} \frac{1-\alpha^{\frac{2}{n}}}{\alpha^{\frac{2}{n}}} > \hat{\theta}_{\mathrm{ML}} \frac{1-\alpha^{\frac{1}{n}}}{\alpha^{\frac{1}{n}}} = l_n^{(K)}(\alpha). \end{aligned}$$

Consequently, following any reasonable notion of (in)admissibility of confidence sets, the split LR confidence will be viewed as inadmissible amongst the class of confidence sets with (minimum) coverage probability $1-\alpha$ for θ^* (see Casella and Hwang, 1991).

2.2 Split Likelihood Ratio test

It is well known that any method for constructing confidence sets also gives rise to a corresponding method for constructing statistical tests. In the case of the split likelihood ratio confidence set (regardless of which of the previously stated definition is used), if for a given statistical model with non-empty parameter space Θ and true parameter θ^*, we want to test the hypotheses

$$H_0 : \theta^* \in \Theta_0 \text{ vs. } H_1 : \theta^* \in \Theta_1$$

(where $\Theta_0, \Theta_1 \subseteq \Theta$, $\Theta_0, \Theta_1 \neq \emptyset$, $\Theta_0 \cap \Theta_1 = \emptyset$ and $\Theta_0 \cup \Theta_1 = \Theta$), with the type I error rate (size) $\sup_{\theta^* \in \Theta_0} \mathbb{P}_{\theta^*}(\text{Reject } H_0)$ controlled at level $\alpha \in (0,1)$, we can apply the following procedure:

$$\text{Reject } H_0 \text{ iff } C_n'(\alpha) \cap \Theta_0 = \emptyset. \tag{2.2}$$

Since $\mathbb{P}_{\theta^*}(C_n'(\alpha) \cap \Theta_0 = \emptyset) \leq \mathbb{P}_{\theta^*}(\theta^* \notin C_n'(\alpha)) \leq \alpha$ whenever the true parameter $\theta^* \in \Theta_0$, this test has level of significance α, by which we mean that (following Lehmann and Romano, 2005)

$$\sup_{\theta^* \in \Theta_0} \mathbb{P}_{\theta^*}(\text{Reject } H_0) \leq \alpha.$$

Not surprisingly, Wasserman et al. (2020) call this procedure the **split likelihood ratio test** (split LRT). They also provide an alternative, from a computational point of view more attractive formulation for it: For a given partition of the index set $\{1, \ldots, n\}$ into $\mathbf{D_0}$ and $\mathbf{D_1}$, define some estimator $\hat{\theta}_1$ for the true parameter θ^* based on the data indexed by $\mathbf{D_1}$ and compute the maximum likelihood estimator under H_0, based on the data indexed by $\mathbf{D_0}$, i.e.

$$\hat{\theta}_0^{H_0} = \operatorname*{argmax}_{\theta \in \Theta_0} \mathcal{L}_0(\theta).$$

Using these estimators, define the random variable

$$U_n = \frac{\mathcal{L}_0(\hat{\theta}_1)}{\mathcal{L}_0\left(\hat{\theta}_0^{H_0}\right)}.$$

According to Wasserman et al. (2020), the split LRT can then be reformulated as

$$\text{Reject } H_0 \text{ iff } U_n > \frac{1}{\alpha}.$$

However, just as in their definition of the split LR confidence set, this definition does not discuss all possible cases that may happen with positive probability. Specifically, the test statistic U_n fails to cover two cases, both of which can happen with positive probability for some statistical models: On the one hand, $\mathbb{P}_{\theta^*}(\mathcal{L}_0(\hat{\theta}_0^{H_0}) = 0) > 0$ is possible for models such as $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} U(0, \theta)$ with $\Theta = (0, +\infty)$, $\Theta_0 = (0, \theta']$ and $\Theta_1 = (\theta', +\infty)$ (where $\theta' > 0$) when the true parameter $\theta^* \in \Theta_1$. This relates to the blind spot in the definition of the split likelihood ratio confidence set - specifically, the case where $\mathcal{L}_0(\hat{\theta}_1) = \mathcal{L}_0(\hat{\theta}_0^{H_0}) = 0$ needs to be discussed. On the other hand, one also needs to consider the case where $\hat{\theta}_0^{H_0}$ does not exist. This happens, for example, when the data $X_1, \ldots, X_n$ is normally distributed with unknown parameters $\mu \in \mathbb{R}$ and $\sigma^2 \in (0, +\infty)$, $|\mathbf{D_0}| = 1$ and the null hypothesis does not restrict the value of σ^2 - in this case, the likelihood function based on $\mathbf{D_0}$ will be unbounded on Θ_0 with probability 1. With these caveats in mind, we propose the following, 'complete' version of the split likelihood ratio test statistic:

Definition 14. *Suppose that in a given statistical model with non-empty parameter space $\Theta \subseteq \mathbb{R}^d$, we want to test the hypotheses*

$$H_0 : \theta^* \in \Theta_0 \text{ vs. } H_1 : \theta^* \in \Theta_1,$$

where $\Theta_0, \Theta_1 \subseteq \Theta$, $\Theta_0, \Theta_1 \neq \emptyset$, $\Theta_0 \cap \Theta_1 = \emptyset$ and $\Theta_0 \cup \Theta_1 = \Theta$. We define some partition $\mathbf{D_0}$ and $\mathbf{D_1}$ of $\{1, \ldots, n\}$ and choose an estimator $\hat{\theta}_1$ based on the data subset indexed by

$\mathbf{D_1}$. *Then we call the following random variable the split likelihood ratio test statistic:*

$$U_n' = \begin{cases} \frac{\mathcal{L}_0(\hat{\theta}_1)}{\sup_{\theta\in\Theta_0}\mathcal{L}_0(\theta)} & \text{if } 0 < \sup_{\theta\in\Theta_0}\mathcal{L}_0(\theta) < +\infty, \\ 0 & \text{if } \sup_{\theta\in\Theta_0}\mathcal{L}_0(\theta) = +\infty, \\ +\infty & \text{if } \sup_{\theta\in\Theta_0}\mathcal{L}_0(\theta) = 0. \end{cases}$$

The following proposition then clarifies that we can use the test statistic defined above to reformulate the split likelihood ratio test presented at the beginning of this section:

Proposition 15. *Given hypotheses H_0 and H_1, a partition $\mathbf{D_0}$ and $\mathbf{D_1}$ of $\{1,\dots,n\}$, as well as a specific choice for the estimator $\hat{\theta}_1$ based on the data subset indexed by $\mathbf{D_1}$, the split likelihood ratio test with nominal level of significance α described in (2.2) and the one with the decision rule*

$$\text{Reject } H_0 \text{ iff } U_n' > \frac{1}{\alpha} \tag{2.3}$$

are equivalent, in the sense that for every possible set of observations, they lead to the same decision.

Proof. We will prove the statement by showing that the decision rules (2.2) and (2.3) agree in each of the following two possible cases:
Case 1: Suppose that $\sup_{\theta\in\Theta_0}\mathcal{L}_0(\theta) = 0$. By definition, in this case $U_n' = +\infty$ and therefore the test with decision rule (2.3) rejects H_0 for every possible value of α. On the other hand, $\sup_{\theta\in\Theta_0}\mathcal{L}_0(\theta) = 0$ implies that $\mathcal{L}_0(\theta) = 0$ and therefore $T_n'(\theta) = +\infty$ (see Definition 10) for all $\theta \in \Theta_0$. Consequently, $C_n'(\alpha) \cap \Theta_0 = \emptyset$ for every $\alpha \in (0,1)$, which means that for every possible significance level α, the test following (2.2) will reject as well.
Case 2: Now, assume that $\sup_{\theta\in\Theta_0}\mathcal{L}_0(\theta) > 0$. Observe that in this case, the following equivalence holds for all $\alpha \in (0,1)$:

$$\begin{aligned} U_n' > \frac{1}{\alpha} &\iff \mathcal{L}_0(\hat{\theta}_1) > \sup_{\theta\in\Theta_0}\mathcal{L}_0(\theta)\,\frac{1}{\alpha} \quad (\text{the direction } \Rightarrow \text{ holds because } \sup_{\theta\in\Theta_0}\mathcal{L}_0(\theta) > 0) \\ &\iff \forall\,\theta\in\Theta_0:\ \mathcal{L}_0(\hat{\theta}_1) > \mathcal{L}_0(\theta)\,\frac{1}{\alpha},\ \mathcal{L}_0(\hat{\theta}_1) > 0 \\ &\iff \forall\,\theta\in\Theta_0:\ T_n'(\theta) > \frac{1}{\alpha},\ \mathcal{L}_0(\hat{\theta}_1) > 0 \\ &\iff C_n'(\alpha)\cap\Theta_0 = \emptyset \quad (\text{the direction } \Leftarrow \text{ holds because } \sup_{\theta\in\Theta_0}\mathcal{L}_0(\theta) > 0). \end{aligned}$$

Therefore, the rules (2.2) and (2.3) will always agree in their decision in this case as well, no matter the level at which the size of the test is controlled. □

Remark. The following points hold for all $\alpha \in (0,1)$:

i) Reviewing Markov's inequality, any split LRT with significance level α could also be defined via the decision rule 'Reject H_0 iff $U_n' \geq \frac{1}{\alpha}$' and still control the type I error rate at this level.

ii) The decision rule stated in point i) could also be written as 'Reject H_0 iff $\mathcal{L}_0(\hat{\theta}_1) \geq \frac{1}{\alpha} \sup_{\theta \in \Theta_0} \mathcal{L}_0(\theta)$.'

iii) It can be shown along the lines of Lemma 6 that, under the null hypothesis, $\mathbb{E}_{\theta^*}[U_n] \leq 1$.

iv) Conditional on the event $\left\{\forall\ \theta \in \Theta_0 : \mathcal{L}_0(\theta) = 0\right\}$, any split LRT rejects H_0 with probability 1.

v) Conditional on the event $\left\{\hat{\theta}_1 \in \Theta_0\right\} \cap \left\{\exists\ \theta \in \Theta_0 : \mathcal{L}_0(\theta) > 0\right\}$, any split LRT rejects H_0 with probability 0.

vi) Conditional on the event $\left\{\sup_{\theta \in \Theta_0} \mathcal{L}_0(\theta) = +\infty\right\}$, any split LRT rejects H_0 with probability 0.

For the remainder of this paper, only the formulation presented in Proposition 15 will be used for the split LRT, as it is a lot easier to work with.

Similar points that were brought up in the discussion of split LR confidence sets in Subsection 2.1.3 also have to be taken into account when constructing tests with this method. By construction, it is impossible to get tests with *exact* or even *approximate* size α when using this approach. We only receive guarantees for the significance level, i.e. we can only bound the size of a split LRT from above. While, just like the minimum coverage probability guarantee of the split LR confidence set, at first glance, this feature might not seem like an issue to be concerned with, in many situations, it will entail a loss of *power* compared with other methods for constructing statistical tests (if they are available). For example, in most of the statistical models commonly studied, it will be the case that the power function of the split LRT is uniformly smaller on the parameter space than the power function of a different test (both with significance level α), either for any given sample size or asymptotically. In this case, the split LRT would certainly be said to be (asymptotically) outperformed by this test and, in the context of statistical decision theory, would be called *inadmissible* (see, for instance, Ferguson, 1967).

Even worse, in some models and for some specific choices of $|\mathbf{D_0}|$ and $\hat{\theta}_1$, it could occur that the critical region (the set of observations such that a given test rejects the H_0) of the split LRT that controls the size at level α is a proper subset of the critical region of a different test T with this significance level (see Example 17). This implies that one will never observe a rejection of H_0 by the split LRT whenever T does not reject the H_0. Even when other methods for constructing tests to compare the split LR approach with are not available, the power of the resulting tests may be so low that they may be deemed useless in practice (which of course is a subjective notion). Furthermore, as was discussed in the context of the split LR confidence set, the loss of power is determined by the choice of $|\mathbf{D_0}|$ and especially $\hat{\theta}_1$: As again, there is no restriction regarding the choice of $\hat{\theta}_1$, in theory, estimators such as $\hat{\theta}_1 = \theta'$ for some $\theta' \in \Theta$ are permitted. In this context, Example 16 shows that it is easy to construct tests that *never* reject the null hypothesis with probability 1 using the split likelihood ratio approach.

As a consequence of these considerations, Wasserman et al. (2020) again note that the usefulness of the split LRT method is restricted to statistical models and hypotheses for which other approaches for constructing tests with (finite sample or asymptotic) guarantees regarding the significance level are not available (yet).

Example 16. Suppose that for a given statistical model with non-empty parameter space Θ and true parameter θ^*, we want to test the hypothesis $H_0 : \theta^* \in \Theta_0$ against $H_1 : \theta^* \in \Theta_1$, where $\Theta_0, \Theta_1 \subseteq \Theta$, $\Theta_0, \Theta_1 \neq \emptyset$ and $\Theta_0 \cap \Theta_1 = \emptyset$. Furthermore, assume that all the distributions in the model have the same support. Then the following test that is derived from the split likelihood ratio approach rejects the null hypothesis with probability 0 for any nominal significance level $\alpha \in (0, 1)$: Define any partition $\mathbf{D_0}, \mathbf{D_1}$ of $\{1, \dots, n\}$ and set $\hat{\theta}_1$ to be an estimator based on the part of the data indexed by $\mathbf{D_1}$, which only takes values in Θ_0 (so $\hat{\theta}_1$ is a measurable function whose codomain is a subset of Θ_0). Note that because of our assumption that the support of the true distribution does not depend on the true parameter, $\mathbb{P}_{\theta^*}(U'_n = +\infty) = 0$, as then $\sup_{\theta \in \Theta_0} \mathcal{L}_0(\theta) > 0$ $\mathbb{P}_{\theta^*}$-a.s. Consequently, we get

$$U'_n \overset{\mathbb{P}_{\theta^*}\text{-a.s.}}{=} V'_n = \begin{cases} \frac{\mathcal{L}_0(\hat{\theta}_1)}{\sup_{\theta \in \Theta_0} \mathcal{L}_0(\theta)} & \text{if } 0 < \sup_{\theta \in \Theta_0} \mathcal{L}_0(\theta) < +\infty, \\ 0 & \text{if } \sup_{\theta \in \Theta_0} \mathcal{L}_0(\theta) \in \{0, +\infty\}. \end{cases}$$

Note that obviously, $V'_n \leq 1$ $\mathbb{P}_{\theta^*}$-a.s. because with our restriction of $\hat{\theta}_1$ only taking values in Θ_0, $\mathcal{L}_0(\hat{\theta}_1) \leq \sup_{\theta \in \Theta_0} \mathcal{L}_0(\theta)$ $\mathbb{P}_{\theta^*}$-a.s. However, this implies that $U'_n \leq 1$ $\mathbb{P}_{\theta^*}$-a.s., too, and consequently $\mathbb{P}_{\theta^*}(U_n > \frac{1}{\alpha}) = 0$ for every $\alpha \in (0, 1)$.

Example 17. Consider again the statistical model from Examples 9 and 13 (uniform distribution with true parameter θ^* and an even sample size n). We now want to test the hypothesis $H_0 : \theta^* \geq \theta'$ against $H_1 : \theta^* < \theta'$ for some $\theta' \in \Theta = (0, +\infty)$. Suppose we partition the index set into arbitrary $\mathbf{D_0}$ and $\mathbf{D_1}$ with $|\mathbf{D_0}| = |\mathbf{D_1}| = \frac{n}{2}$ and again take $\hat{\theta}_1 = \max_{i \in \mathbf{D_1}} X_i$ to be the maximum likelihood estimator based on the data subset indexed by $\mathbf{D_1}$. Writing $\Theta_0 = [\theta', +\infty)$, note that here $0 < \sup_{\theta \in \Theta_0} \mathcal{L}_0(\theta) < +\infty$ with probability 1, since $\sup_{\theta \in \Theta_0} \mathcal{L}_0(\theta) = \min\left(\theta'^{-\frac{n}{2}}, \hat{\theta}_0^{-\frac{n}{2}}\right)$, where $\hat{\theta}_0 = \max_{i \in \mathbf{D_0}} X_i$. Consequently, we get that conditional on the event $\{\hat{\theta}_1 < \hat{\theta}_0\}$

$$U'_n = \frac{\mathcal{L}_0(\hat{\theta}_1)}{\sup_{\theta \in \Theta_0} \mathcal{L}_0(\theta)} = 0 \ \mathbb{P}_{\theta^*}\text{-a.s.}$$

and conditional on the event $\{\hat{\theta}_1 \geq \hat{\theta}_0\}$

$$U'_n = \frac{\mathcal{L}_0(\hat{\theta}_1)}{\sup_{\theta \in \Theta_0} \mathcal{L}_0(\theta)} = \frac{\hat{\theta}_1^{-\frac{n}{2}}}{\min\left(\theta'^{-\frac{n}{2}}, \hat{\theta}_0^{-\frac{n}{2}}\right)} = \left(\frac{\max(\theta', \hat{\theta}_0)}{\hat{\theta}_1}\right)^{\frac{n}{2}} \ \mathbb{P}_{\theta^*}\text{-a.s.}$$

With this case by case definition of the test statistic U'_n, it follows that for any $\alpha \in (0, 1)$

$$\mathbb{P}_{\theta^*}\left(U'_n > \frac{1}{\alpha}, \hat{\theta}_1 < \hat{\theta}_0\right) = \mathbb{P}_{\theta^*}\left(0 > \frac{1}{\alpha}, \hat{\theta}_1 < \hat{\theta}_0\right) = 0,$$

as well as, following the same arguments as in Example 13,

$$\begin{aligned}
&\mathbb{P}_{\theta^*}\Big(U_n' > \frac{1}{\alpha},\, \hat{\theta}_1 \geq \hat{\theta}_0\Big) = \mathbb{P}_{\theta^*}\Big(\Big(\frac{\max(\theta', \hat{\theta}_0)}{\hat{\theta}_1}\Big)^{\frac{n}{2}} > \frac{1}{\alpha},\, \hat{\theta}_1 \geq \hat{\theta}_0\Big) \\
&= \mathbb{P}_{\theta^*}\Big(\max(\theta', \hat{\theta}_0) > \Big(\frac{1}{\alpha}\Big)^{\frac{2}{n}} \hat{\theta}_1,\, \hat{\theta}_1 \geq \hat{\theta}_0\Big) = \mathbb{P}_{\theta^*}\Big(\theta' > \Big(\frac{1}{\alpha}\Big)^{\frac{2}{n}} \hat{\theta}_1,\, \hat{\theta}_1 \geq \hat{\theta}_0\Big) \\
&= \mathbb{P}_{\theta^*}\Big(\hat{\theta}_1 < \alpha^{\frac{2}{n}}\, \theta',\, \max_{i=1,\ldots,n} X_i = \hat{\theta}_1\Big) = \mathbb{P}_{\theta^*}\Big(\max_{i=1,\ldots,n} X_i < \alpha^{\frac{2}{n}}\, \theta',\, \max_{i=1,\ldots,n} X_i = \hat{\theta}_1\Big) \\
&= \mathbb{P}_{\theta^*}\Big(\max_{i=1,\ldots,n} X_i < \alpha^{\frac{2}{n}}\, \theta'\Big)\, \mathbb{P}_{\theta^*}\Big(\max_{i=1,\ldots,n} X_i = \hat{\theta}_1\Big) = \min\Big(\alpha^2 \Big(\frac{\theta'}{\theta^*}\Big)^n, 1\Big)\, \frac{1}{2}.
\end{aligned}$$

Consequently, assuming that the null hypothesis is true, we get

$$\begin{aligned}
&\mathbb{P}_{\theta^*}\Big(U_n' > \frac{1}{\alpha}\Big) = \mathbb{P}_{\theta^*}\Big(U_n' > \frac{1}{\alpha},\, \hat{\theta}_1 < \hat{\theta}_0\Big) + \mathbb{P}_{\theta^*}\Big(U_n' > \frac{1}{\alpha},\, \hat{\theta}_1 \geq \hat{\theta}_0\Big) \\
&= 0 + \min\Big(\alpha^2 \Big(\frac{\theta'}{\theta^*}\Big)^n, 1\Big)\, \frac{1}{2} = \frac{\alpha^2}{2} \Big(\frac{\theta'}{\theta^*}\Big)^n,
\end{aligned}$$

and it immediately follows that the actual size of this test turns out to be $\sup_{\theta^* \in \Theta_0} \mathbb{P}_{\theta^*}\Big(U_n' > \frac{1}{\alpha}\Big) = \frac{\alpha^2}{2}$, which is not surprising given Example 13, where the actual coverage probability of the corresponding split LR confidence set was computed to be $1 - \frac{\alpha^2}{2}$.

The usual statistical test (with nominal significance level α) for these hypotheses (based on the likelihood ratio approach) would be to reject $H_0 : \theta^* \geq \theta'$ whenever $\hat{\theta}_{\mathrm{ML}} < \alpha^{\frac{1}{n}}\, \theta'$, where $\hat{\theta}_{\mathrm{ML}} = \max_{i=1,\ldots,n} X_i$ is the maximum likelihood estimator based on the whole sample. It is easy to see that under the null hypothesis, the rejection probability is

$$\mathbb{P}_{\theta^*}\Big(\hat{\theta}_{\mathrm{ML}} < \alpha^{\frac{1}{n}}\, \theta'\Big) = \alpha \Big(\frac{\theta'}{\theta^*}\Big)^n,$$

and therefore, the size of this test is exactly α.

We now show that for any $\alpha \in (0,1)$ and any given partition $\mathbf{D_0}, \mathbf{D_1}$ of the index set, the critical region of the split LRT described above is a proper subset of the critical region of the classical likelihood ratio test. Obviously, the critical region of the split LRT with significance level α can be written as

$$CR(sLRT) = \{x = (x_1, \ldots, x_n)^T \in (0, +\infty)^n : \max_{i \in \mathbf{D_0}} x_i < \max_{i \in \mathbf{D_1}} x_i < \alpha^{\frac{2}{n}}\, \theta'\},$$

while the rejection region of the likelihood ratio test is

$$CR(LRT) = \{x = (x_1, \ldots, x_n)^T \in (0, +\infty)^n : \max_{i=1,\ldots,n} x_i < \alpha^{\frac{1}{n}}\, \theta'\}.$$

It immediately follows that

$$CR(sLRT) \subset \{x = (x_1, \ldots, x_n)^T \in (0, +\infty)^n : \max_{i=1,\ldots,n} x_i < \alpha^{\frac{2}{n}}\, \theta'\} \subset CR(LRT),$$

where the second inclusion follows from $\alpha^{\frac{2}{n}} < \alpha^{\frac{1}{n}}$ for any $n \in \mathbb{N}$ because $\alpha \in (0, 1)$. Consequently, under any reasonable definition of admissibility of statistical tests, the split LRT with this choice of $\hat{\theta}_1$, as well as $|\mathbf{D_0}| = |\mathbf{D_1}| = \frac{n}{2}$, would be seen as inadmissible in this model and for these specific hypotheses, among the class of tests with significance level α.

3 Literature Review

In this chapter, we will give a brief overview of the existing literature on or related to universal inference. Since the idea of universal inference was introduced by Wasserman et al. (2020), many articles that mention this method have been published, however, most of them are concerned with the theories of e-variables and/or anytime valid inference (see, for example, Ramdas et al., 2022, 2023; Grünwald, 2024a), which we will not discuss in this thesis. In fact, so far, relatively few discussions of universal inference outside the context of e-values or anytime valid inference exist in the literature. We will divide the remaining literature on universal inference into three categories: Extensions of the split likelihood ratio method introduced by Wasserman et al. (2020), universal inference in the presence of nuisance parameters, determining a good choice for the splitting ratio and applying universal inference to irregular statistical models and hypotheses.

3.1 Extensions to the split LR method

We will begin the review by discussing tests and confidence sets derived from the universal inference principle that differ from the split likelihood ratio method discussed in Chapter 2. These can be viewed as natural extensions to the split LR approach, since they are all based on the fact that a split likelihood ratio statistic has expectation at most 1, too.

3.1.1 Using several sample splits

In the setup for the split LR setting, we required the specification of a partition of the index set $\{1, \dots, n\}$, $\mathbf{D_0}$ and $\mathbf{D_1}$ (independent of the data), according to which the sample split is performed. Consider now another split LR statistic (using the improved definition from Definition 10)), based on the same data and partition, but with reversed roles of $\mathbf{D_0}$ and $\mathbf{D_1}$:

$$T_n^{\text{swap}}(\theta) = \begin{cases} \frac{\mathcal{L}_1(\hat{\theta}_0)}{\mathcal{L}_1(\theta)} & \text{if } \mathcal{L}_1(\theta) > 0 \\ +\infty & \text{if } \mathcal{L}_1(\theta) = 0, \end{cases}$$

(where $\mathcal{L}_1(\theta)$ is the likelihood function based on the data subset indexed by $\mathbf{D_1}$ and $\hat{\theta}_0$ is any estimator based on the sample indexed by $\mathbf{D_0}$). Obviously, by the same argument as in Lemma 6, this statistic satisfies

$$\mathbb{E}_{\theta^*}\left(T_n^{\text{swap}}(\theta^*)\right) \leq 1$$

(where θ^* is the true parameter) as well and of course, it is non-negative, too. Consequently, the average of the two split LR statistics,

$$E_n(\theta) = \frac{T_n(\theta) + T_n^{\text{swap}}(\theta)}{2},$$

is also non-negative and satisfies $\mathbb{E}_{\theta^*}[E_n(\theta^*)] \leq 1$. By the same reasoning as in Theorem 8, the following confidence set, which is based on $E_n(\theta)$, has minimum coverage probability $1-\alpha$ (for any $\alpha \in (0,1)$) for the true parameter θ^* as well:

$$C_n^{swap}(\alpha) = \left\{\theta \in \Theta : E_n(\theta) \leq \frac{1}{\alpha}\right\} \tag{3.1}$$

Note that the two statistics $T_n(\theta)$ and $T_n^{\text{swap}}(\theta)$ need not be independent, but this does not affect the validity of the previous argument, since we are only dealing with their expectations. Wasserman et al. (2020) call $E_n(\alpha)$ the *cross-fit likelihood ratio statistic* and therefore, it makes sense to call $C_{E_n}(\alpha)$ the *cross-fit likelihood ratio confidence set*. They show that for some models such as the uniform distribution discussed in Examples 9, 13 and 17, this approach gives much more reasonable confidence sets than the split LR method, although in this specific case the average length of the resulting confidence interval will still be asymptotically larger by a constant factor greater than 1 (which in this case depends on the nominal significance level α) compared to the usual confidence interval $[\hat{\theta}, \hat{\theta}\left(\frac{1}{\alpha}\right)^{\frac{1}{n}}]$. Strieder and Drton (2022) note that it is of course not strictly necessary to combine the two split LR test statistics with equal weights, i.e. for any $a \in (0,1)$, not necessarily equal to $\frac{1}{2}$, the weighted average $a\,T_n(\theta) + (1-a)T_n^{\text{swap}}(\theta)$ can be used to construct a valid confidence set for θ^*.

Of course, we could continue with this strategy and repeat the calculation of a split LR statistic, based on another partition $\{\mathbf{D_2}, \mathbf{D_3}\}$ (which does not necessarily need to differ from $\{\mathbf{D_0}, \mathbf{D_1}\}$, but we again explicitely require that both $\mathbf{D_2}$ and $\mathbf{D_3}$ are non-empty) - actually, we could go on with creating partitions and determining split LR statistics as long as we wish: Using a set of $k \in \mathbb{N}$ (not necessarily different) partitions of $\{1, \ldots, n\}$ (where each partition has two blocks), B, and computing the corresponding split LR statistics (with the improved definition from Definition 10) $T_n^i(\theta)$; $i = 1, \ldots, k$; for each partition, we define

$$E_n^k(\theta) = \frac{1}{k} \sum_{i \in B} T_n^i(\theta). \tag{3.2}$$

Again, it is obvious that $\mathbb{E}_{\theta^*}[E_n^k(\theta^*)] \leq 1$, and so the confidence set built analogously to (3.1) has minimum coverage probability $1-\alpha$ as well. Wasserman et al. (2020) call a set of this kind the *subsampling variant* of the universal method. As a special case, one could consider removing all randomness that the sample splitting part of the universal approach usually adds to the resulting confidence set by computing split LR statistics for all $2^n - 2$[1] different choices for a two block-partition $\{\mathbf{D_0}, \mathbf{D_1}\}$ that the index set $\{1, \ldots, n\}$ permits.

[1] Note that the order of the partition matters in this context, as reversing the roles of $\mathbf{D_0}$ and $\mathbf{D_1}$ in general will result in different values of the statistic, and that $\emptyset$ and $\{1, \ldots, n\}$ of course are inadmissible choices for $\mathbf{D_0}$.

As another special case of the subsampling variant when the sample size n is even, Dunn et al. (2023) study the statistic $E_n^K(\theta) = \frac{1}{K_n}\sum_{i=1}^{K_n} T_n^i(\theta)$, where the $T_n^1(\theta), \ldots, T_n^{K_n}(\theta)$ are the split likelihood ratio statistics computed from all possible splits of the data having equal size, $B_1, \ldots, B_{K_n}$ (so, $B_i = \{\mathbf{D_0^i}, \mathbf{D_1^i}\}$ is a partition of $\{1, \ldots, n\}$, where $|\mathbf{D_0^i}| = |\mathbf{D_1^i}| = \frac{n}{2}$), with $K_n = \binom{n}{\frac{n}{2}}$. One can interpret $E_n^K(\theta)$ as $\mathbb{E}_{\theta^*}[T_n^{\text{half}}(\theta)|\mathcal{D}_n]$, where $T_n^{\text{half}}(\theta)$ is the split likelihood ratio statistic derived from a randomly chosen partition of the index set with two equally sized blocks and $\mathcal{D}_n = X_1, \ldots, X_n$ is the given data. Theorem 2 of Dunn et al. (2023) shows that in the setting where the (d-dimensional) data is i.i.d. distributed as $N(\theta^*, I_d)$, it holds that

$$E_n^K(\theta) \approx \exp\left(\frac{3n}{10}||\bar{Y}_n - \theta||^2\right)\left(\frac{2}{5}\right)^{\frac{d}{2}}$$

for large n. Although Dunn et al. (2023) explain that this approximation formally only holds in a very specific sense (which is explained in Theorem 2 of their paper), they verify its practical validity for large sample sizes by simulations. Note that the latter expressions equals $LR_n(\theta)^{\frac{3}{5}}\left(\frac{2}{5}\right)^{\frac{d}{2}}$, where $LR_n(\theta)$ is the classical likelihood ratio statistic for this model. Therefore, in large samples, a subsampling likelihood ratio confidence set that uses a high number of partitions $L \approx n$ may be approximated by

$$\begin{aligned} C_n^{\text{subsplit}}(\alpha) =& \left\{\theta \in \mathbb{R}^d : \frac{1}{L}\sum_{i=1}^{L} T_n^i(\theta) \leq \frac{1}{\alpha}\right\} \approx \\ & \left\{\theta \in \mathbb{R}^d : E_n^K(\theta) \leq \frac{1}{\alpha}\right\} \approx \\ & \left\{\theta \in \mathbb{R}^d : ||\bar{Y}_n - \theta||^2 \leq \frac{10}{3n}\left[\ln\left(\frac{1}{\alpha}\right) + \frac{d}{2}\ln\left(\frac{5}{2}\right)\right]\right\}. \end{aligned}$$

This approximation is useful because the computation of such a subsampling likelihood ratio confidence set will be computationally intensive even in the Gaussian model with known variance given here. Note that this approximative confidence set describes a ball with the same center $\bar{Y}_n$ as the classical likelihood ratio set. It is natural to propose that the latter will have a smaller radius (which is $\frac{Q_d(1-\alpha)}{n}$, where $Q_d(1-\alpha)$ is the upper α-quantile of a chi-square distribution with d degrees of freedom) and empirically, this seems to hold, too. We have not found an upper bound for $Q_d(1-\alpha)$ of exactly this form (which is both linear in d and $\ln\left(\frac{1}{\alpha}\right)$) in the literature.

Dunn et al. (2023) also present simulations that compare different variants of the universal method - namely, the split, cross-fit and subsampling likelihood ratio confidence sets, with the classical likelihood ratio set in the multivariate Gaussian case, where the LR set is known to achieve the nominal coverage probability exactly. With $n = 1000$ observations simulated from a $N(0, I_2)$-distribution, even data splits and $k = 100$ random partitions for the subsampling version, the results seem to indicate that the classical likelihood ratio set not only has the smallest volume and diameter, but even seems to be a proper subset of all the universal confidence sets almost surely. A corresponding result for the split LR confidence set in the Gaussian model, which only holds for dimensions

1 and 2, but also for possibly uneven choices of the sample splitting ratio, is presented in Theorem 30 of Chapter 4 of this thesis. Interestingly, the simulations indicate that this result holds for the other variants of the universal method as well, at least in the 2-dimensional normal case. Furthermore, they suggest that the average diameter/volume of the subsampling set is smaller than the one of the cross-fit set, which in turn is smaller than the simple split version. A theoretical proof of the first result is given in Theorem 4 of Dunn et al. (2023) for any multivariate Gaussian model, when an even sample split is used. Similarly, other simulations from Dunn et al. (2023) show that in this model, the subsampling variant of the universal test for the hypothesis $H_0 : \theta^* = 0$ has more power than the corresponding cross-fit version, whose power in turn is higher than the one of the split likelihood ratio test. Interestingly, the differences in power seem to increase with growing dimension of the model. Of course, in terms of power, the classical likelihood ratio test dominates all the universal tests in this model.

To compare the different variants of universal tests in an irregular setting (meaning a statistical model and a null hypothesis for which classical asymptotic theory of likelihood ratio statistics fails), Strieder and Drton (2022) obtain power simulations of these tests in a one-factor analysis setting with 12 observations per unit. Formally, assuming that the data is i.i.d. $N_{12}(0, \Sigma^*)$-distributed, where the unknown true covariance matrix $\Sigma^* \in \mathbb{R}^{12\times 12}$ is positive definite, they test the null hypothesis that $\Sigma^* \in F_{12,1} = \{\Psi + LL^T : \Psi \in \mathbb{R}^{12\times 12}$ diagonal, $L \in \mathbb{R}^{12}\}$ against the alternative that $\Sigma^* \notin F_{12,1}$. Theorem 6.1 of Drton (2009) shows that if $\Sigma^* = \Psi^* + L^* L^{*T}$, where $\Psi^* \in \mathbb{R}^{12\times 12}$ is diagonal and L^* has no more than 2 non-zero entries, then the likelihood ratio test statistic

$$LR = 2\Big(\sup_{\Sigma \in \mathbb{R}^{12\times 12}_{>0}} l_n(\Sigma) - \sup_{\Sigma \in F_{12,1}} l_n(\Sigma)\Big),$$

where $l_n(\cdot)$ is the log-likelihood function of the model, does not converge to the chi-square distribution with 54 degrees of freedom (the dimension of the whole parameter space is 78 and the dimension of the null hypothesis is 24 in this case); in fact, it does not converge to a chi-square distribution at all. Hence, for such L^*, the true size of the classical likelihood ratio test will not even be asymptotically equal to the nominal significance level. In contrast, tests based on universal inference will still control the type I error rate at the desired level because they do not require any assumptions regarding the underlying model or the null hypothesis. Simulating $n = 2000$ data points from $N_{12}\big(0, \Omega + L_1 L_1^T + L_2(h) L_2^T(h)\big)$, with $\Omega = \frac{1}{5} I$, $L_1 = (5, 5, 0, \ldots, 0) \in \mathbb{R}^{12}$ and $L_2(h) = \frac{h}{\sqrt{12}} (1, 1, \ldots, 1) \in \mathbb{R}^{12}$ (so the case $h = 0$ lies in the null hypothesis and corresponds to an irregular setting), Strieder and Drton (2022) look at the power of the split, cross-fit and subsampling likelihood ratio test as a function of $h \geq 0$. To this end, they use a certain splitting ratio they propose to be optimal for the first and the third method (see Section 3.3 of this thesis) and different weights and splitting ratios for the cross-fit method. They find that all the universal tests basically have zero power for $h \leq 0.3$, and that the simple split as well as the subsampling likelihood ratio test have slightly higher power than each of the cross-fit variants. Interestingly, in this irregular setting, the subsampling version does not seem to dominate the split LR test in terms of power.

3.1.2 Randomized improvements of universal inference

The finite sample validity of each of the universal inference methods relies on bounding the upper tail probabilities of split likelihood ratio statistics with Markov's inequality. Therefore, if we want to improve the power of the corresponding tests or decrease the volume of the corresponding confidence sets, the question naturally arises whether one can find a sharper bound on the upper α-quantile of a non-negative, integrable random variable, for which only an upper bound $c > 0$ on the first moment is known, than $\frac{c}{\alpha}$. In other words, does there exist a non-negative function g with $g(a, m) \leq \frac{m}{a}$ for all $a \in (0, 1]$, $m > 0$ and $g(a^*, m^*) < \frac{m^*}{a^*}$ for some $a^* \in (0, 1]$, $m^* > 0$ such that

$$\mathbb{P}\big(X > g(\alpha, \mathbb{E}[X])\big) \leq \alpha$$

for all $\alpha \in (0, 1]$, any probability space $(\Omega, \mathcal{A}, \mathbb{P})$, as well as any non-negative and integrable random variable X defined on it? The answer is negative for deterministic functions g (meaning that g is not allowed to depend on $\omega \in \Omega$): Of course $g(a, m) \geq m$ needs to hold for any $a \in (0, 1]$ because otherwise, the desired bound will not hold for random variables X with $\mathbb{P}(X = m) = 1$. If $g(\alpha^*, \mu^*) < \frac{\mu^*}{\alpha^*}$ for some $\alpha^* \in (0, 1]$, $\mu^* > 0$, choose some number $b \in \big(g(\alpha^*, \mu*), \frac{\mu^*}{\alpha^*}\big)$. Consider the random variable X^* with $\mathbb{P}\big(X^* = b\big) = \frac{\mu^*}{b} < 1$ (because $b > g(\alpha^*, \mu^*) \geq \mu^*$) and $\mathbb{P}(X^* = 0) = 1 - \frac{\mu^*}{b}$. Then $\mathbb{E}[X^*] = \mu^*$ and $\mathbb{P}\big(X^* > g(\alpha^*, \mathbb{E}[X^*])\big) = \mathbb{P}\big(X^* = b\big) = \frac{\mu^*}{b} > \alpha^*$ (because $0 < b < \frac{\mu^*}{\alpha^*}$), so the desired bound does not hold for the upper α^*-quantile of the specific random variable X^*.

However, it is possible to improve the Markov bound if one allows for 'random' quantiles, meaning that the function g can also depend on $\omega \in \Omega$. This is a consequence of the following result (Theorem 1.2 in Ramdas and Manole, 2023):

Theorem 18. *Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space and X be a non-negative, integrable random variable defined on it. Assume that there exists a random variable U on $(\Omega, \mathcal{A}, \mathbb{P})$ that is uniformly distributed on $[0, 1]$ and independent of X. Then, for any $\alpha \in (0, 1]$, it holds that:*

$$\mathbb{P}\big(X > \frac{E[X]}{\alpha} U\big) \leq \alpha.$$

Proof. The claim is trivial for $\mathbb{E}[X] = 0$, which of course implies $\mathbb{P}(X = 0) = 1$. Assume now that $\mathbb{E}[X] > 0$. Then

$$\mathbb{P}\big(X \geq \frac{E[X]}{\alpha} U\big) = \mathbb{P}\big(U \leq \frac{X}{E[X]}\, \alpha\big) = \mathbb{E}\Big[\mathbb{E}\Big[\mathbf{1}\big(U \leq \frac{X}{E[X]}\, \alpha\big) | X\Big]\Big] =$$
$$\mathbb{E}\Big[\min\big(\frac{X}{E[X]}\, \alpha,\, 1\big)\Big] \leq \mathbb{E}\Big[\frac{X}{E[X]}\, \alpha\Big] = \alpha,$$

where the last equality follows from the fact that U and X are independent. □

Ramdas and Manole (2023) call this result the *uniformly-randomized Markov inequality*. It is easy to see that the 'random' upper quantile bound $g^*(\alpha, \mu, \omega) = \frac{\mu}{\alpha} U(\omega)$ is

almost surely smaller than the deterministic Markov bound $g(\alpha, \mu) = \frac{\mu}{\alpha}$. As a consequence, Ramdas and Manole (2023) state that any of the universal inference methods can be improved by replacing the deterministic critical value $\frac{1}{\alpha}$ with the random critical value $\frac{U}{\alpha}$: Because of Theorem 18, the corresponding sets and tests will still control the coverage probability and the type I error rate at the nominal level, but because $\frac{U}{\alpha} \leq \frac{1}{\alpha}$ almost surely, these confidence sets almost surely will have smaller volume/diameter and the power of these tests will be larger compared to the ones using the deterministic critical value.

Moreover, Ramdas and Manole (2023) prove a generalized version of Markov's inequality for exchangeable random variables:

Theorem 19. *Assume that $X_1, X_2, \ldots$ is an exchangeable sequence of non-negative and integrable random variables with expectation μ. Then, for any $\alpha > 0$, it holds that:*

$$\mathbb{P}\big(\exists t \in \mathbb{N} : \frac{1}{t} \sum_{i=1}^{t} X_i > \frac{\mu}{\alpha}\big) \leq \alpha.$$

The result is a direct consequence of a time-reversed version of Ville's inequality, which Ramdas and Manole (2023) prove in Theorem 5.2. of their paper. Applying this to the subsampling variants of universal inference, the subsampling likelihood ratio statistic (3.2) may be replaced by

$$\sup_{j=1,\ldots,k} \frac{1}{j} \sum_{i=1}^{j} T_n^i, \tag{3.3}$$

where $T_n^1, \ldots, T_n^k$ are split LR statistics computed from $k \in \mathbb{N}$ random and independent splits of the data. The resulting tests and confidence sets will achieve the nominal significance/confidence level because with random and independent splitting, not only will the split LR statistics be non-negative and have their expectation bounded by 1, but they will be also be i.i.d. Therefore, Theorem 19 justifies the use of (3.3) in place of (3.2) with the same critical value $\frac{1}{\alpha}$. Observe that for given (random and independent) sample splits, (3.3) will be almost surely larger than or equal to (3.2).

Theorem 5.3 in Ramdas and Manole (2023) then combines the uniformly randomized and exchangeable version of Markov's inequality, from which it follows that the statistical test:

$$\text{Reject } H_0 \text{ if } T_n^1 > \frac{U}{\alpha} \text{ or if } \sup_{j=1,\ldots,k} \frac{1}{j} \sum_{i=1}^{j} T_n^i > \frac{1}{\alpha} \tag{3.4}$$

controls the probability of a Type I-error at level $\alpha \in (0, 1)$, too. Obviously, this test is more powerful than the one using (3.3) with critical value $\frac{1}{\alpha}$ or the one using only one split likelihood ratio statistic (derived from a random split according to the same rule) with the critical value $\frac{U}{\alpha}$. However, it is not clear whether (3.4) will be more powerful than the test

$$\text{Reject } H_0 \text{ if } \frac{1}{k}\sum_{i=1}^{k} T_n^i > \frac{U}{\alpha}, \tag{3.5}$$

whose type I error control follows from the uniformly randomized Markov inequality. In experiments conducted by Ramdas and Manole (2023) in the framework of a simple Gaussian mixture model, (3.4) and (3.5) seem to have similar power.

In practice, the use of all of these randomized improvements of universal inference methods will be difficult to justify because the external randomization (the result of a simulation of U for (3.4) and (3.5) or the specific order of $k \in \mathbb{N}$ independent partitions of the index set $\{1, \dots, n\}$ for (3.3) and (3.4)) will influence the results of the universal tests (if it is a 'close call' between the two hypotheses) and confidence sets greatly. As Ramdas and Manole (2023) themselves note, this potentially high dependence on the result of an external randomization implies problems with reproducibility and p-hacking in the context of science communication. They discuss several ideas how to solve this issue: For example, they suggest outsourcing the external randomization to a central repository from which uniform random numbers or random, independent partitions can be requested (and of course, these requests would need to be documented), or by generating the necessary random variables (which need to be independent of the data) from the data itself. For instance, if the data $X_1, \dots, X_n$ are i.i.d. and real-valued random variables following a continuous distribution, then the empirical cumulative distribution function evaluated at any data point X_i (where $i \in \{1, \dots, n\}$ is chosen independently of the data),

$$\hat{F}_n(X_i) = \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}\big(X_j \leq X_i\big),$$

is independent of $X_1, \dots, X_n$ and uniformly distributed on $\{\frac{1}{n}, \frac{2}{n}, \dots, 1\}$. Since this discrete uniform distribution stochastically dominates the continuous uniform distribution on the interval $(0, 1)$, one can use, for example, $\hat{F}_n(X_1)$ in place of an externally generated Unif$(0, 1)$-distributed random variable in (3.4) or (3.5) - this replacement is justified because Theorem 18 obviously also holds if the random variable U is stochastically larger than Unif$(0, 1)$. A similar strategy could be used to define random and independent splits of the index set $\{1, \dots, n\}$ that are generated from, but independent of the data.

3.1.3 RIPR split LRT

Recall the definition of the split likelihood ratio test given in (2.3): If for a given partition $\{\mathbf{D_0}, \mathbf{D_1}\}$ of the index set $\{1, \dots, n\}$ and a given null hypothesis $H_0 : \theta^* \in \Theta_0$, we have $0 < \sup_{\theta \in \Theta_0} \mathcal{L}_0(\theta) < +\infty$, with $\mathcal{L}_0(\theta) = \prod_{i \in \mathbf{D_0}} p_\theta(X_i)$, then the split LRT uses the test statistic

$$U_n = \frac{\mathcal{L}_0(\hat{\theta}_1)}{\sup_{\theta \in \Theta_0} \mathcal{L}_0(\theta)},$$

where $\hat{\theta}_1$ is an estimator for the true parameter θ^* based on the part of the data indexed by $\mathbf{D_1}$. To avoid trivial results, one might restrict this estimator to satisfy $\hat{\theta}_1 \notin \Theta_0$. The split LRT with nominal level $\alpha \in (0,1)$ then rejects the H_0 if $U_n > \frac{1}{\alpha}$. In Proposition 15, we argued why the type I error probability of this test indeed does not exceed α. Observe now that the estimator $\hat{\theta}_1$ does not influence the denominator of U_n at all and that we have maximized the likelihood (based on $\mathbf{D_0}$) under the null in the denominator, but not the likelihood under the alternative in the enumerator. Indeed, since this method relies on sample splitting, maximizing the likelihood (based on $\mathbf{D_0}$) under the alternative is in some sense prohibited. However, one might wonder if one can replace the denominator with a smaller expression and still achieve the desired significance level α when using the critical value $\frac{1}{\alpha}$. This is indeed the case if the statistical model considered is convex, as explained by Wasserman et al. (2020) in Section 6. Their reasoning relies heavily on the theory of the reversed information projection (RIPR) developed by Li (1999) and Grünwald (2024b).

Recall that a statistical model $\mathcal{P}$ on a measurable space $(\Omega, \mathcal{A})$ is called convex if, for any $\mathbb{P}_1$, $\mathbb{P}_2 \in \mathcal{P}$ and any $a \in [0,1]$, the probability measure $a\,\mathbb{P}_1 + (1-a)\,\mathbb{P}_2$ is also an element of $\mathcal{P}$. Examples of convex models are the set of all (finite or infinite) mixtures of a set of base distributions or the set of all distributions for which the expectation of a specific real-valued random variable $X(\omega)$ is equal to or bounded by $C \in \mathbb{R}$. If the distributions in the convex model $\mathbb{P}$ are all dominated by the same measure μ on $(\Omega, \mathcal{A})$, then the set of the corresponding densities $\mathcal{D} = \{p : p \text{ is the density of some } \mathbb{P} \in \mathcal{P} \text{ with respect to } \mu\}$ must be convex as well, in the sense that it satisfies $a\,p_1 + (1-a)\,p_2 \in \mathcal{D}$ for any p_1, $p_2 \in \mathcal{D}$ and $a \in [0,1]$. Furthermore, remember that the *Kullback-Leibler divergence* (sometimes also called *relative entropy*) between two probability measures $\mathbb{P}$ and $\mathbb{Q}$ on the same measurable space $(\Omega, \mathcal{A})$ is defined as

$$D_{\mathrm{KL}}(\mathbb{P} \,||\, \mathbb{Q}) = \begin{cases} \int_\Omega \ln\left(\frac{d\mathbb{P}}{d\mathbb{Q}}(\omega)\right) d\mathbb{P} & \text{if } \mathbb{P} \ll \mathbb{Q}, \\ +\infty & \text{otherwise.} \end{cases}$$

The first part of the definition may also be written as $\int_\Omega p(\omega)\, \ln\left(\frac{p(\omega)}{q(\omega)}\right) d\mu$, where μ is a σ-finite measure on $(\Omega, \mathcal{A})$ that dominates both $\mathbb{P}$ and $\mathbb{Q}$ and where we use the convention $0\,\ln(0) = 0\,\ln(+\infty) = 0$. The Kullback-Leibler divergence is an important concept in the fields of machine learning, Bayesian inference and information geometry. It is non-negative and equals 0 if and only if $\mathbb{P} \ll \mathbb{Q}$ and $p = q$ μ-almost everywhere. However, one should note that in general it does not define a metric on a given space of probability measures: It neither satisfies symmetry nor the triangle inequality in general.

For a given probability measure $\mathbb{P}$ and a set of probability measures $\mathcal{Q}$, we may define

$$D_{\mathrm{KL}}(\mathbb{P} \,||\, \mathcal{Q}) = \inf_{\mathbb{Q} \in \mathcal{Q}} D_{\mathrm{KL}}(\mathbb{P} \,||\, \mathbb{Q}).$$

An important conclusion from Chapter 4 of Li (1999), summarized in Theorem 1 of Lardy et al. (2024), is then the following:

Theorem 20. *Let $\mathcal{Q}$ be a convex statistical model and let $\mathbb{P}$ be a probability measure on the measurable space $(\Omega, \mathcal{A})$. Furthermore, let μ be a measure on $(\Omega, \mathcal{A})$ that dominates both $\mathbb{P}$ and any $\mathbb{Q} \in \mathcal{Q}$. Also, assume that $D_{KL}(\mathbb{P} \,||\, \mathcal{Q}) < +\infty$. Then:*

i) There exists a unique measure $\mathbb{Q}^*$ on $(\Omega, \mathcal{A})$ that is absolutely continuous with respect to μ, such that for any sequence of probability measures $(\mathbb{Q}_n)_{n\in\mathbb{N}}$ in $\mathcal{Q}$ with $D_{KL}(\mathbb{P} \,||\, \mathbb{Q}_n) \xrightarrow{n\to\infty} D_{KL}(\mathbb{P} \,||\, \mathcal{Q})$, we have $\ln(q_n) \xrightarrow{L_1(\mathbb{P})} \ln(q^*)$, where q_n, q^* are the densities of $\mathbb{Q}_n$, $\mathbb{Q}^*$ with respect to μ.

ii) $\mathbb{Q}^*(\Omega) \leq 1$.

iii) $D_{KL}(\mathbb{P} \,||\, \mathcal{Q}) = \int_\Omega p(\omega) \ln\left(\frac{p(\omega)}{q^*(\omega)}\right) d\mu$, where p is the density of $\mathbb{P}$ with respect to μ.

iv) If $q^* > 0$ μ-a.e, then for any $\mathbb{Q} \in \mathcal{Q}$, we have

$$\mathbb{E}_{\mathbb{Q}}\left[\frac{p(\omega)}{q^*(\omega)}\right] = \int_\Omega q(\omega)\frac{p(\omega)}{q^*(\omega)}\, d\mu \leq 1,$$

where q is the density of $\mathbb{Q}$ with respect to μ.

Lardy et al. (2024) call the sub-probability measure (meaning that $\mathbb{Q}^*(\Omega) \leq 1$) $\mathbb{Q}^*$ the *reverse information projection* (*RIPR*) of $\mathbb{P}$ on $\mathcal{Q}$. Note that $\mathbb{Q}^*$ need not be a member of $\mathcal{Q}$ - indeed, it need not be a probability measure at all, see Example 2 of Lardy et al. (2024) for an example. Theorem 20 is a very important result in the context of e-variables and the theory of anytime valid inference (see, for example, Grünwald, 2024b), but it also has an implication for universal inference if the null hypothesis is convex:

Corollary 21. *Let $X_1, \ldots, X_n$ be random variables taking values in $\mathcal{X}$ that are i.i.d. distributed according to an (unknown) density p^*, where p^* belongs to a set of density functions (with respect to some known measure μ), $\mathcal{D}$, that are strictly positive on $\mathcal{X}$ μ-almost everywhere. Suppose we want to test the null hypothesis $H_0 : p^* \in \mathcal{D}_0$ against the alternative $H_1 : p^* \in \mathcal{D}_1$, where $\mathcal{D}_0, \mathcal{D}_1$ are non-empty and disjoint subsets of $\mathcal{D}$. Furthermore, assume that $\mathcal{D}_0$ is a convex set of densities. Choose a partition $\{\mathbf{D_0}, \mathbf{D_1}\}$ of the index set $\{1, \ldots, n\}$ and an estimator $\hat{p}_1$ for the true density p^*, based on the data indexed by $\mathbf{D_1}$, that satisfies $D_{KL}(\hat{p}_1, \mathcal{D}_0) < +\infty$ almost surely. Define the test statistic*

$$U_{RIPR} = \frac{\mathcal{L}_0(\hat{p}_1)}{\mathcal{L}_0(\hat{p}_{RIPR})},$$

where $\mathcal{L}_0(p) = \prod_{i\in\mathbf{D_0}} p(X_i)$ and $\hat{p}_{RIPR}$ is the reverse information projection of $\hat{p}_1$ on $\mathcal{D}_0$ (which almost surely exists because $\mathcal{D}_0$ is convex and $D_{KL}(\hat{p}_1, \mathcal{D}_0) < +\infty$ almost surely). Then, for a given $\alpha \in (0, 1)$, the test with the decision rule

$$\textit{Reject } H_0 \textit{ iff. } U_{RIPR} \geq \frac{1}{\alpha}$$

controls the type I error probability at level α*. Also, if* $\hat{p}_{RIPR} \in \mathcal{D}_0$ *almost surely, then this test is uniformly more powerful than the split likelihood ratio test with the same choice of* $\{\mathbf{D_0}, \mathbf{D_1}\}$ *and* $\hat{p}_1$*, which uses the test statistic:*

$$U_{UI} = \begin{cases} \frac{\mathcal{L}_0(\hat{p}_1)}{\sup_{p \in \mathcal{D}_0} \mathcal{L}_0(p)} & \text{if } \sup_{p \in \mathcal{D}_0} \mathcal{L}_0(p) < +\infty, \\ 0 & \text{if } \sup_{p \in \mathcal{D}_0} \mathcal{L}_0(p) = +\infty \end{cases}$$

Wasserman et al. (2020) call the test using the statistic U_{RIPR} the *reverse information split likelihood ratio test* (RIPR split LRT). Note that all the likelihood ratios involved are well-defined because of our assumption that the statistical model only includes densities that have support on $\mathcal{X}$ almost everywhere. The result regarding the significance level then follows immediately from part iv) of Theorem 20 (because the densities in the model have full support, $\hat{p}_{\text{RIPR}}$ is strictly positive μ-a.e. on $\mathcal{X}$) and by using the same arguments as in Lemma 6 and Theorem 8 (conditioning on $\hat{p}_1$ and using Markov's inequality). For this result, we do not even need the assumption that $\hat{p}_{\text{RIPR}} \in \mathcal{D}_0$. However, if $\hat{p}_{\text{RIPR}} \in \mathcal{D}_0$ a.s., then it almost surely holds that

$$U_{\text{RIPR}} = \frac{\mathcal{L}_0(\hat{p}_1)}{\mathcal{L}_0(\hat{p}_{\text{RIPR}})} \geq U_{\text{UI}}.$$

This statement is trivial if $U_{\text{UI}} = 0$ and in the case of $U_{\text{UI}} > 0$, it follows from the fact that $\mathcal{L}_0(\hat{p}_{\text{RIPR}}) \leq \sup_{p \in \mathcal{D}_0} \mathcal{L}_0(p)$. Since both tests use the same critical value $\frac{1}{\alpha}$ to achieve the nominal significance level $\alpha \in (0, 1)$, the test using the reverse information projection of $\hat{p}_1$ on $\mathcal{D}_0$ in the denominator of the test statistic is more powerful than the one using the maximum likelihood estimator restricted to the null hypothesis. This assertion intuitively makes sense because as mentioned earlier, the usual split LRT only uses the maximum likelihood under the null hypothesis, while the RIPR split LRT neither uses maximum likelihood under the null, nor under the alternative - instead, it uses the density from the null hypothesis (or, in some sense, from the closure of the null hypothesis) that in some sense is the 'closest' one to $\hat{p}_1$. Using similar arguments, Grünwald (2024a) has shown that under certain conditions, one can use the RIPR approach without sample splitting to construct likelihood ratios whose expectations are bounded by 1 under the null hypothesis (such non-negative statistics are called *e-variables* or *e-statistics*) and which in some sense have optimal behavior under the alternative.

3.2 Dealing with nuisance parameters

Adapting universal inference methods to deal with nuisance parameters in statistical models is straightforward. Suppose that the parameter space Θ is a subset of R^d with $d \geq 2$ and we are only interested in inference on the first $p < d$ components of the true parameter θ_*, which we call $\theta_*^{(1)}$. Expressing any $\theta \in \Theta$ as $\theta = (\theta^{(1)}, \theta^{(2)})^T$ with $\theta^{(1)} \in \Theta^{(1)} \subseteq \mathbb{R}^p$ and $\theta^{(2)} \in \Theta^{(2)} \subseteq \mathbb{R}^{d-p}$ ($\theta^{(2)}$ is usually called a *nuisance component*), define the profile-likelihood function of $\theta^{(1)} \in \Theta^{(1)}$ (based on the data indexed by $\mathbf{D_0}$ for some partition $\{\mathbf{D_0}, \mathbf{D_1}\}$ of the index set $\{1, \ldots, n\}$) as:

$$\mathcal{L}_0^P(\theta^{(1)}) = \sup_{\vartheta^{(2)} \in \mathbb{R}^{d-p}} \mathcal{L}_0(\theta^{(1)}, \vartheta^{(2)}).$$

Now, to construct a confidence set for $\theta_*^{(1)}$ with minimum coverage probability $1-\alpha$, choose an estimator $\hat{\theta}_1$ of θ_* based only on the data indexed by $\mathbf{D_1}$, and define the statistic

$$T_n^P(\theta^{(1)}) = \begin{cases} \frac{\mathcal{L}_0(\hat{\theta}_1)}{\mathcal{L}_0^P(\theta^{(1)})} & \text{if } \mathcal{L}_0^P(\theta^{(1)}) > 0, \\ +\infty & \text{if } \mathcal{L}_0^P(\theta^{(1)}) = 0 \end{cases}$$

for any $\theta^{(1)} \in \Theta^{(1)}$. Then, by the same argument as in Lemma 6, it holds that for any true parameter $\theta_* = (\theta_*^{(1)}, \theta_*^{(2)})^T$:

$$\mathbb{E}_{\theta^*}\left[T_n^P(\theta_*^{(1)})\right] \leq \mathbb{E}_{\theta^*}\left[T_n'(\theta^*)\right] \leq 1,$$

where the statistic $T_n'(\theta)$ is defined as in Definition 10. The first inequality follows from the fact that $\mathcal{L}_0^P(\theta_*^{(1)}) \geq \mathcal{L}_0(\theta_*) > 0$ and therefore $T_n^P(\theta_*^{(1)}) \geq T_n'(\theta_*)$ almost surely. Consequently, for any $\alpha \in (0,1)$, the following set will achieve a confidence level of $1-\alpha$ for $\theta_*^{(1)}$:

$$C_n^P(\alpha) = \left\{\theta^{(1)} \in \Theta^{(1)} : T_n^P(\theta^{(1)}) < \frac{1}{\alpha}\right\}.$$

Again, this follows immediately from Markov's inequality.

Using a different argument, Wasserman et al. (2020) arrive at the same adaptation of the split likelihood ratio confidence set when nuisance parameters are present in the model. Tse and Davison (2022) then pointed out that in high-dimensional models, the presence of a lot of nuisance parameters implies that universal inference methods will become even more conservative. This is immediately clear from the fact that in most statistical models commonly considered, if the number of nuisance parameters $d-p$ is relatively large, then the profile likelihood $\mathcal{L}_0^P(\theta_*^{(1)}) = \sup_{\vartheta^{(2)} \in \mathbb{R}^{d-p}} \mathcal{L}_0(\theta_*^{(1)}, \vartheta^{(2)})$ will be significantly higher than the 'whole' likelihood $\mathcal{L}_0(\theta_*)$, and therefore, $T_n^P(\theta_*^{(1)})$ will be considerably smaller than $T_n'(\theta_*)$. To illustrate this, Tse and Davison (2022) consider the setting where the data $X_1, \ldots, X_n$ are i.i.d. $N_d(\mu_*, \Sigma_*)$-distributed, where $d \geq 2$, $\mu_* = \left(\mu_*^{(1)}, \mu_*^{(2)}\right)$ is unknown with $\mu_*^{(1)} \in \mathbb{R}^p$, $1 \leq p < d$, $\mu_*^{(2)} \in \mathbb{R}^q$ with $q = d-p$, and Σ_* is a known positive definite $d \times d$-matrix. Since all the distributions in this model have full support, it follows that

$$\mathbb{E}_{\mu_*, \Sigma_*}\left[T_n(\mu_*)\right] = \mathbb{E}_{\mu_*, \Sigma_*}\left[\frac{\mathcal{L}_0(\hat{\mu}_1, \Sigma_*)}{\mathcal{L}_0(\mu_*, \Sigma_*)}\right] = 1,$$

where $\mathcal{L}_0(\mu, \Sigma_*)$ is the likelihood function based on the data indexed by $\mathbf{D_0}$, $\hat{\mu}_1 = \frac{1}{|\mathbf{D_1}|}\sum_{i \in \mathbf{D_1}} X_i$ and $\{\mathbf{D_0}, \mathbf{D_1}\}$ is some partition of $\{1, \ldots, n\}$. However, if we are only interested in constructing a confidence set for $\mu_*^{(1)}$, then we would consider the statistic

$$T_n^P(\mu^{(1)}) = \frac{\mathcal{L}_0(\hat{\mu}_1, \Sigma_*)}{\mathcal{L}_0^P(\mu^{(1)}, \Sigma_*)},$$

where $\mathcal{L}_0^P(\mu^{(1)}, \Sigma_*) = \sup_{m^{(2)} \in \mathbb{R}^q} \mathcal{L}_0(\mu^{(1)}, m^{(2)}, \Sigma_*)$. Tse and Davison (2022) show that $T_n^P(\mu_*^{(1)})$ has the following distribution under the true parameter μ_*:

$$T_n^P(\mu_*^{(1)}) \overset{d}{=} \exp\Big(\frac{A(p,\rho) - B(q,\rho)}{2}\Big),$$

where

$$\begin{aligned} A(p,\rho) &= 2\sqrt{\rho}\, Y_p^T Z_p - \rho\, Z_p^T Z_p \\ B(q,\rho) &= (Y_q - \sqrt{\rho}\, Z_q)^T (Y_q - \sqrt{\rho}\, Z_q), \end{aligned}$$

with $\rho = \frac{|\mathbf{D_0}|}{|\mathbf{D_1}|}$; $Y_p, Z_p \sim N_p(0, I)$; $Y_q, Z_q \sim N_q(0, I)$ and Y_p, Z_p, Y_q, Z_q independent. One can then show that $\mathbb{E}\Big[\exp\big(\frac{A(p,\rho)}{2}\big)\Big] = 1$ and obviously, $B(q,\rho) \sim (1+\rho)\,\chi_q^2$, so

$$\mathbb{E}\Big[\exp\big(-\frac{B(q,\rho)}{2}\big)\Big] = MG_{\chi_q^2}\Big(-\frac{1+\rho}{2}\Big) = \Big(1 - 2\big(-\frac{1+\rho}{2}\big)\Big)^{-\frac{q}{2}} = (2+\rho)^{-\frac{q}{2}},$$

where $MG_{\chi_q^2}(\cdot)$ is the moment generating function of the chi-squared distribution with q degrees of freedom. Therefore, since $A(p,\rho)$ and $B(q,\rho)$ are obviously independent, we conclude that

$$\mathbb{E}_{\mu_*,\Sigma_*}\Big[T_n^P(\mu_*^{(1)})\Big] = \mathbb{E}_{\mu_*,\Sigma_*}\Big[\exp\Big(\frac{A(p,\rho)}{2}\Big)\Big]\, \mathbb{E}_{\mu_*,\Sigma_*}\Big[\exp\Big(-\frac{B(q,\rho)}{2}\Big)\Big] = (2+\rho)^{-\frac{q}{2}}.$$

Observe that for a large number of nuisance parameters q, this expectation will be very small, so Markov's inequality under the assumption that the expectation equals 1 will overestimate the upper tail probabilities of $T_n^P(\mu_*^{(1)})$ even more than it would do for a split likelihood ratio statistic in the normal model anyway. Of course, as Tse and Davison (2022) remark, this additional conservatism due to nuisance parameters can be eliminated in this situation because we know the exact value of the expectation of $T_n^P(\mu_*^{(1)})$: If, for a given $\alpha \in (0,1)$, we choose $\frac{(2+\rho)^{-\frac{q}{2}}}{\alpha}$ instead of $\frac{1}{\alpha}$ as the critical value for a confidence set using the statistic $T_n^P(\mu_*^{(1)})$, then the resulting set will almost surely have smaller diameter/volume, but due to Markov's inequality, it will still have minimum coverage probability $1-\alpha$ for $\mu_*^{(1)}$.

In the context of testing the null hypothesis $H_0 : \mu_*^{(1)} = \mu_0^{(1)}$ against the alternative $H_1 : \mu_*^{(1)} \neq \mu_0^{(1)}$, Tse and Davison (2022) perform some simulations to study the increase in power obtained by using the critical value that takes into account the number of nuisance parameters in the model. They test the null hypothesis on $p = 1$ parameter of interest and vary the number of nuisance parameters q between 5 and 640. Not surprisingly, they find that the power gained from using the adapted critical value increases with

the number of nuisance parameters. The adapted test still becomes catastrophically conservative for large q, which may be traced back to the general curse of dimensionality of the split likelihood ratio test (see Dunn et al., 2024; Strieder and Drton, 2022, as well as Theorem 42). Furthermore, the strategy of modifying the critical value of a universal test or confidence set according to the true expectation of $T_n^P(\theta_*^{(1)})$ will not be available in general, especially not in irregular models, which are the main source of motivation for universal inference. Tse and Davison (2022) show by simulations that in a specific mixture model of q-dimensional Gaussian distributions (where q ranges between 10 and 60), the split likelihood ratio test (with $\hat{\theta}_1$ being an EM-approximation of the maximum likelihood estimator under the alternative) of the null hypothesis that there is only one component against the alternative that there are two still has significance level α when the critical value $\frac{(2+\rho)^{-\frac{q}{2}}}{\alpha}$ instead of $\frac{1}{\alpha}$ is used. However, even with the larger critical value, the type I error probability of the split likelihood ratio test seems to be almost equal to 0 and the power gain is considerably smaller compared to the setting considered before.

3.3 Optimizing the splitting ratio

No matter which universal inference method is used to construct a confidence set or a statistical test, one needs to perform one or several sample splits in the process. Therefore, the question how the choice of the (non-random) ratio according to which the data is split (given the ratio, the actual split may be performed either in a deterministic or in a random way) affects the diameter/volume of the resulting confidence set or the power of the resulting test is of significant importance for applications. Specifically, given $\alpha \in (0,1)$, we should be interested in which value of $r_0 = \frac{|\mathbf{D_0}|}{n}$ (where n is the number of observations on which the confidence set/test is based) minimizes the expectation of a certain power of the diameter/volume of the split likelihood ratio confidence set with confidence level $1-\alpha$ or, given a null and alternative hypothesis, which value of r_0 maximizes the power of the corresponding split likelihood ratio test with significance level α. We will call r_0 the *splitting ratio* associated with a certain universal inference method. Regarding the split LR confidence set $C_n^{\mathrm{ML}}(\alpha, r_0)$ for the true mean parameter μ_* in the d-dimensional Gaussian model (with $\hat{\theta}_1$ being the maximum likelihood estimator for μ_* based on the data indexed by $\mathbf{D_1}$), the choice of r_0 that minimizes $\mathbb{E}_{\mu_*}\left[\lambda^{\frac{2}{d}}\left(C_n^{\mathrm{ML}}(\alpha, r_0)\right)\right]$, where λ is the Lebesgue measure of the set, is approximately given by

$$r_0^* = \frac{2\ln\left(\frac{1}{\alpha}\right) + d - \sqrt{d\left(2\ln\left(\frac{1}{\alpha}\right) + d\right)}}{2\ln\left(\frac{1}{\alpha}\right)}. \tag{3.6}$$

A proof of this result is given in Proposition 38 of the following chapter. The minimization of $\lambda^{\frac{2}{d}}\left(C_n^{\mathrm{ML}}(\alpha, r_0)\right)$ is only approximate in the sense that r_0^* may not be a feasible choice for the sample split proportion, which, given the sample size n, needs to be one of $\frac{1}{n}, \frac{2}{n}, \ldots, \frac{n-1}{n}$. However, it holds that one of the two 'closest' feasible choices to r_0^* is

indeed the minimizer amongst all possible values of r_0. Similarly, Theorem 3 of Dunn et al. (2023) shows that (3.6) also approximately minimizes the average squared radius of $C_n^{\mathrm{ML}}(\alpha, r_0)$ (it is not stated in the theorem that for given $n \in \mathbb{N}$, r_0^* can only be an approximate minimizer in the same sense as above, however, by the same argument as before, it is obviously true that r_0^* cannot be the exact minimizer of the expected squared radius amongst all feasible choices). Furthermore, Dunn et al. (2023) note that r_0^* converges to $\frac{1}{2}$ for high dimensions d. It is also easy to see that for $\alpha \to 0$, r_0^* goes to 1.

Given these results, one might wonder whether (3.6) also corresponds to an approximation of the choice of the splitting ratio which maximizes the power of statistical tests for any hypotheses in the multivariate normal model. The answer is no because, in general, the optimal choice will not only depend on the significance level α and the dimension of the model considered, but also on the dimension of the null hypothesis H_0 and the Euclidean distance of the true parameter θ^* to H_0. This follows from considerations made by Strieder and Drton (2022): They first determine the asymptotic distribution of the split likelihood ratio test statistic U_n for a given null hypothesis $H_0 : \theta \in \Theta_0$ (where Θ_0 is a non-empty subset of the parameter space Θ) in a specific setting, namely if $\hat{\theta}_1$ is taken to be the maximum likelihood estimator for the true parameter θ^* under the full model (based on the data indexed by $\mathbf{D_1}$) and if the statistical model considered satisfies the usual regularity assumptions which imply that the classical likelihood ratio statistic asymptotically follows a chi-square distribution (see Van der Vaart, 2007). Specifically, they assume for the statistical model, a parameter $\theta_0 \in \Theta_0$ and the null hypothesis:

i) θ_0 lies in the interior of the parameter space $\Theta \in \mathbb{R}^d$ and the model $\{\mathbb{P}_\theta : \theta \in \Theta\}$ is *differentiable in quadratic mean* at θ^*, meaning that there exists a measurable function $s_{\theta_0} : \mathcal{X} \mapsto \mathbb{R}^d$ such that, for $h \to 0$

$$\int_{\mathcal{X}} \left(\sqrt{p_{\theta_0+h}(x)} - \sqrt{p_{\theta_0}(x)} - \frac{h^T s_{\theta_0}(x) \sqrt{p_{\theta_0}(x)}}{2} \right)^2 d\mu = o(||h||^2).$$

$s_{\theta_0}(x)$ is usually called the *score function* of the model.

ii) The *Fisher information matrix* $I(\theta_0) = \mathbb{E}_{\theta_0}\left[s_{\theta_0}(X) s_{\theta_0}^T(X)\right]$ is invertible.

iii) There exists a neighborhood $U(\theta_0) \subseteq \Theta$ and a measurable function $\ell : \mathcal{X} \mapsto [0, +\infty)$ with $\mathbb{E}_{\theta_0}\left[\ell^2(X)\right] < \infty$ which satisfies

$$|\ln\left(p_{\theta_1}(x)\right) - \ln\left(p_{\theta_2}(x)\right)| \leq \ell(x)\, ||\theta_1 - \theta_2||$$

for all θ_1, θ_2 in $U(\theta_0)$.

iv) The maximum likelihood estimators $\hat{\theta}_{0,n}^{\mathrm{ML}}$ (based on $\mathbf{D_0}$) and $\hat{\theta}_{1,n}^{\mathrm{ML}}$ (based on $\mathbf{D_1}$) exist almost surely and are consistent for θ_0 under $\mathbb{P}_{\theta_0}$ for $n \to \infty$.

v) The sequence of sets $\mathcal{H}_n = \sqrt{n}(\Theta_0 - \theta_0)$ converges to a set $\mathcal{H}_0$, in the sense that we define

$$\mathcal{H}_0 = \{h : h = \lim_{n\to\infty} h_n \text{ for some converging sequence } (h_n)_{n\in\mathbb{N}} \text{ with } h_n \in \mathcal{H}_n \; \forall \; n \in \mathbb{N}\}$$

and now $\mathcal{H}_0$ should contain the limit $h' = \lim_{i\to\infty} h_{n_i}$ of any converging sequence $(h_{n_i})_{i\in\mathbb{N}}$ with $h_{n_i} \in \mathcal{H}_{n_i}$ for all $i \in \mathbb{N}$ (see Van der Vaart, 2007).

Under these regularity conditions, Strieder and Drton (2022) prove the following result in Theorem 1 of their paper:

Theorem 22. *Consider a statistical model $\{\mathbb{P}_\theta : \theta \in \Theta\}$ with $\Theta \in \mathbb{R}^d$ and a parameter $\theta_0 \in \Theta_0$ that satisfies the assumptions i) to iv) stated above and assume that the non-empty null hypothesis set $\Theta_0 \subseteq \Theta$ satisfies condition v). Fix the splitting ratio $r_0 \in (0,1)$ and assume that, for the split likelihood ratio test of the null hypothesis $H_0 : \theta_0 \in \Theta_0$ based on sample size $n \geq \frac{1}{r_0}$, we use a sample split $\{\mathbf{D_0}, \mathbf{D_1}\}$ with $|\mathbf{D_0}| = \lfloor r_0\, n \rfloor$ and the maximum likelihood estimator $\hat{\theta}_{1,n} = \arg\max_{\theta\in\Theta} \mathcal{L}_{1,n}(\theta)$ as the estimator based on the subset of the data indexed by $\mathbf{D_1}$. Then the log split likelihood ratio statistic*

$$LSR_n = 2\ln(U_n) = 2\left(l_0\big(\hat{\theta}_{1,n}\big) \;-\; l_0\big(\hat{\theta}_{0,n}^{H_0}\big)\right),$$

where $l_0(\theta) = \ln\big(\mathcal{L}_0(\theta)\big)$ and $\hat{\theta}_{0,n}^{H_0} = \arg\max_{\theta\in\Theta_0} \mathcal{L}_{0,n}(\theta)$, satisfies under $\mathbb{P}_{\theta_n}$ with $\theta_n = \theta_0 + \frac{h}{\sqrt{n}}$

$$LSR_n \xrightarrow{d} ||Y + \sqrt{r_0}\, I(\theta_0)^{\frac{1}{2}}\, h \;-\; I(\theta_0)^{\frac{1}{2}}\, \mathcal{H}_0||^2 \;-\; ||Y \;-\; \sqrt{\frac{r_0}{1-r_0}}\, Z||^2, \tag{3.7}$$

where $||x - A|| = \inf_{a\in A} ||x - a||$ for $x \in \mathbb{R}^d$, $A \subseteq \mathbb{R}^d$; $M\,\mathcal{H}_0 = \{M\,h_0 \;:\; h_0 \in \mathcal{H}_0\}$ for any matrix $M \in \mathbb{R}^{d\times d}$; $Y, Z \overset{i.i.d}{\sim} N_d(0, I_d)$ and $I(\theta_0)^{\frac{1}{2}}$ is the unique positive definite, symmetric $d \times d$-matrix that satisfies $I(\theta_0)^{\frac{1}{2}}\, I(\theta_0)^{\frac{1}{2}} = I(\theta_0)$.

Their proof mainly builds on ideas and results from Chapters 7 and 16 of Van der Vaart (2007). In some special cases, the asymptotic distribution of the log split likelihood ratio statistic in the setting of Theorem 22 reduces to a simpler expression, as the following result from Strieder and Drton (2022) shows.

Corollary 23. *Consider the same setting as in Theorem 22. If $I(\theta_0)^{\frac{1}{2}}\, \mathcal{H}_0 = \{0\}^p \times \mathbb{R}^{d-p}$ for some $p \in \{0, 1, \ldots, d\}$, then*

$$LSR_n \xrightarrow{d} ||Y_{[p]} + \sqrt{r_0}\, \tilde{h}_{[p]}||^2 \;-\; ||Y \;-\; \sqrt{\frac{r_0}{1-r_0}}\, Z||^2, \tag{3.8}$$

where $Y, Z \overset{i.i.d}{\sim} N_d(0, I_d)$, $x_{[p]}$ denotes the first p components of the vector $x \in \mathbb{R}^d$ (with the convention $x_{[0]} = 0$) and $\tilde{h} = I(\theta_0)^{\frac{1}{2}}\, h$.

Proof. If $I(\theta_0)^{\frac{1}{2}} H_0 = \{0\}^p \times \mathbb{R}^{d-p}$, then the first part of the asymptotic distribution in (3.7) reduces to, with $Y \sim N_d(0, I_d)$,

$$\begin{aligned}
&||Y + \sqrt{r_0}\, I(\theta_0)^{\frac{1}{2}} h \,-\, I(\theta_0)^{\frac{1}{2}} H_0||^2 = \\
\inf_{\vartheta \in \mathbb{R}^{d-p}} &\left(Y + \sqrt{r_0}\, I(\theta_0)^{\frac{1}{2}} h \,-\, \begin{pmatrix} 0_p \\ \vartheta \end{pmatrix} \right)^T \left(Y + \sqrt{r_0}\, I(\theta_0)^{\frac{1}{2}} h \,-\, \begin{pmatrix} 0_p \\ \vartheta \end{pmatrix} \right) = \\
&\left(Y_{[p]} + \left(\sqrt{r_0}\, I(\theta_0)^{\frac{1}{2}} h\right)_{[p]} \right)^T \left(Y_{[p]} + \left(\sqrt{r_0}\, I(\theta_0)^{\frac{1}{2}} h\right)_{[p]} \right) = \\
&||Y_{[p]} + \sqrt{r_0}\, \tilde{h}_{[p]}||^2.
\end{aligned}$$

The claim follows immediately. □

Corollary 23 implies that under the null hypothesis (remember that we assumed that $\theta_0 \in \Theta_0$), if the statistical model satisfies assumptions i) to v) at the true parameter θ^* and $I(\theta^*)\, H_0$ is a coordinate subspace, then

$$LSR_n \xrightarrow{d} ||Y_{[p]}||^2 \,-\, ||Y \,-\, \sqrt{\frac{r_0}{1 - r_0}}\, Z||^2$$

under $\mathbb{P}_{\theta^*}$, with $Y, Z \overset{\text{i.i.d}}{\sim} N_d(0, I_d)$. However, instead of adapting the critical value of split likelihood ratio tests according to this distribution (which for any given splitting ratio would be much smaller, as they show empirically) Strieder and Drton (2022) aim to understand how the splitting ratio r_0 influences the asymptotic distribution of the log split likelihood ratio statistic under the alternative, and therefore the power of the test if the Markov-critical value $\frac{1}{\alpha}$ is used. To this end, they define a new class of distributions:

Definition 24. *For $d \in \mathbb{N}$, $p \in \{0, 1, \ldots, d\}$, $\delta \in [0, +\infty)$ and $r_0 \in (0, 1)$, define the d-dimensional non-central split chi-square distribution with p degrees of freedom, non-centrality parameter δ and splitting ratio r_0 (abbreviated by $\text{split}_{r_0}\, \chi^2_{p,d}(\delta)$) as the distribution of*

$$||Y_{[p]} + \sqrt{r_0}\, h||^2 \,-\, ||Y \,-\, \sqrt{\frac{r_0}{1 - r_0}}\, Z||^2,$$

where $Y, Z \overset{i.i.d}{\sim} N_d(0, I_d)$ and h is any p-dimensional real vector such that $h^T h = \delta$.

Obviously, the asymptotic distribution in (3.8) corresponds to a $\text{split}_{r_0}\, \chi^2_{p,d}(\tilde{\delta}(p))$-distribution with $\tilde{\delta}(p) = \tilde{h}_{[p]}^T \tilde{h}_{[p]}$. It follows that for given significance level α, model dimension d, null hypothesis dimension $d - p$ and the non-centrality parameter (which depends on the true parameter as well as the null hypothesis dimension) $\tilde{\delta}(p)$, the optimal splitting ratio r_0^* that maximizes the asymptotic power of the split likelihood ratio test is exactly the splitting ratio that maximizes the cumulative distribution function of the $\text{split}_{r_0}\, \chi^2_{p,d}(\tilde{\delta}(p))$-distribution at input $2 \ln\left(\frac{1}{\alpha}\right)$ (note that the asymptotic result holds for the statistic $LSR_n = 2 \ln(U_n)$, not for U_n) among all $r_0 \in (0, 1)$. However, if we want

to find r_0^*, two problems arise: On the one hand, we do not know the true value of the non-centrality parameter $\tilde{\delta}(p)$ because it depends on the unknown true parameter. On the other hand, as the non-central split chi-square distribution is a new class of distributions, no numerical routines to evaluate the cumulative distribution function for given parameters exist yet. Regarding the first issue, Strieder and Drton (2022) propose to initialize a small non-centrality parameter δ_{init} (which implies that even with the optimal splitting ratio $r_0^*(\delta_{\text{init}})$, the approximate power of the split likelihood ratio test will be small), optimize the splitting ratio with respect to δ_{init}, then gradually increase the value of the non-centrality parameter until the power of the test with the calculated optimal splitting ratio reaches a pre-specified level (in the simulations, they use 0.8). The corresponding splitting ratio is then the output of this algorithm. Regarding the second issue, they approximate the value of the cdf of a given non-central split chi-square distribution at $2\ln(\frac{1}{\alpha})$ by repeatedly sampling from this distribution.

Approximating the optimal splitting ratio $r_0^*(d,p)$ for given model dimension d and null hypothesis dimension $d-p$ with the algorithm described above in combination with Monte Carlo approximation, they find that if $p << d$, then $r_0^*(d,p)$ is close to (3.6) (which does not depend on p!), the in some sense optimal choice if the goal is to construct a split likelihood ratio confidence set for θ^*. Notably, if $p \in \mathbb{N}$ is fixed, then $r_0^* > 0.5$ for all dimensions $d > p$. However, if $p \approx d$, then $r_0^*(d,p) < 0.5$ for large dimensions d. Surprisingly, it even seems to be the case that if $p(d) = d - k$ for some fixed $k \in \mathbb{N}$, then $r_0^*(d, p(d)) \to 0$ as $d \to \infty$. Furthermore, if $p(d) = c\,d$ for some fixed $c \in (0,1)$, then $r_0^*(d, p(d)) \to r(a) > 0$ as $d \to \infty$.

Via simulations, Strieder and Drton (2022) find that if the regularity conditions regarding the statistical model and the null hypothesis set (assumptions i) to v) plus the assumption that the rotated limiting hypothesis is a coordinate subspace) hold, then the asymptotic power of the split likelihood ratio test is indeed significantly improved for many values of p and d if an algorithmic approximation of $r_0^*(p,d)$ is used instead of the analytical value from (3.6). However, if the true parameter is fixed, then the power of the split LR test will still converge to 0 for growing dimension d even if the approximatively optimal splitting ratio is used. The extent of the power gain from using $r_0^*(p,d)$ therefore depends on the given p and d. If the distance of the true parameter to the null hypothesis set increases with growing dimension d, then splitting the data according the approximation of $r_0^*(p,d)$ via the algorithm described above manages to keep the power of the test stable with increasing dimension, while the power still goes to 0 if the data is split according to (3.6). Furthermore, Strieder and Drton (2022) use simulations to show that even in a specific irregular setting, namely, testing the null hypothesis of a one-factor model against the saturated alternative (which is irregular if the null hypothesis is true, but there are only 0, 1 or 2 positive loadings, see Subsection 3.1.1), the split likelihood ratio test has higher power when using the approximation of $r_0^*(p,d)$ instead of (3.6), although the power gain seems to be lower compared to the regular and asymptotic case. In general, the question whether $r_0^*(p,d)$ is the optimal splitting ratio with respect to maximizing power in (some) irregular settings as well remains open.

3.4 Applications of universal inference in irregular settings

Interestingly, the number of articles in the literature which specifically deal with the application of universal inference methods in irregular settings still seems to be very limited, despite this having been the main source of motivation behind developing this approach to inference in Wasserman et al. (2020), where it was originally introduced. Some authors have applied universal tests in irregular models although the corresponding articles mainly deal with other issues - to name three examples, Aguiar et al. (2024) use the split likelihood ratio test to test for a specific structure of a multilayer network, Ignatiadis and Wager (2022) suggest using the same test to evaluate whether the choice of a class of smooth priors is reasonable in the context of a Gaussian empirical Bayes problem and Lauritzen et al. (2021) apply it to test the total positivity assumption within an Ising model. Examples of use cases of universal methods in irregular models can also be found in Wasserman et al. (2020) and Tse and Davison (2022) (Gaussian mixture model with known mixture weights), as well as Strieder and Drton (2022) (one-factor model, see the end of Subsection 3.1.1 of this thesis), but these articles mainly focus on other aspects of universal inference. To our knowledge, there exist only three articles in the literature that specifically focus on applying universal inference methods in settings where classical inference approaches such as the likelihood ratio test or confidence set are unavailable. In the remainder of this section, we will discuss these briefly.

3.4.1 Empirical Bayesian inference

Consider the following (Bayesian) model setup that differs from the one we previously dealt with: Let $(\mathcal{X}, \mathcal{A})$ be a measurable space and let $\mathcal{T} \subseteq \mathbb{R}^d$. Furthermore, let $X_1, X_2, \ldots, X_n$ be a sequence of random variables living on $(\mathcal{X}, \mathcal{A})$, let $\Theta_1, \Theta_2, \ldots, \Theta_n$ be a sequence of $(\mathcal{T}, \mathcal{B}(\mathcal{T}))$-valued random variables and let $\mathcal{D}_\Psi = \{\pi_\psi : \psi \in \Psi\}$ be a set of densities on $(\mathcal{T}, \mathcal{B}(\mathcal{T}))$ with respect to some measure $\mu_\mathcal{T}$. Suppose that the Θ_i are identically distributed according to the density $\pi_{\psi^*}(\theta)$, where $\psi^* \in \Psi \subseteq \mathbb{R}^p$ is the unknown true parameter in the model and $\pi_{\psi^*}(\theta)$ is the corresponding density of each Θ_i. Moreover, suppose that the sequence $(X_1, \Theta_1), (X_2, \Theta_2), \ldots, (X_n, \Theta_n)$ is independent as well as that the conditional densities of X_i given $\Theta_i = \theta$ exist for any $\theta \in \mathcal{T}$ (with respect to some measure $\mu_\mathcal{X}$ on $(\mathcal{X}, \mathcal{A})$) and satisfy

$$f^{\psi^*}_{X_i|\Theta_i}(x|\theta) = f_{X|\Theta}(x|\theta)$$

for any true parameter $\psi^* \in \Psi$ (i.e. the distribution of X_i conditional on Θ_i does not depend on the unknown true parameter ψ^*). Consider now the situation where the realization of the sequence $\Theta_1, \Theta_2, \ldots, \Theta_n$ is unknown as well. Given such an unknown realization $\Theta_1 = \theta_1^*, \ldots, \Theta_n = \theta_n^*$, where $\theta_1^*, \ldots, \theta_n^* \in \mathcal{T}$, and observations of the data $X_1, \ldots, X_n$, Empirical Bayesian inference seeks to construct confidence sets for θ_i^* and statistical tests for null hypotheses of the form $H_0 : \theta_\mathbb{I}^* \in \mathcal{T}_{\mathbb{I},0}$, where $\mathbb{I} \subseteq \{1, \ldots, n\}$ is a non-empty index set, $\theta_\mathbb{I}^* = (\theta_i^*)_{i \in \mathbb{I}}$ and $\mathcal{T}_{\mathbb{I},0} \subseteq \mathcal{T}^{|\mathbb{I}|}$ is non-empty, too. Specifically, for any

$\alpha \in (0,1)$, one would like to have a random subset of $\mathcal{T}$, $EBC_n(\alpha, i)$, which satisfies

$$\mathbb{P}_{\psi^*}\left(\theta_i^* \in EBC_n(\alpha,i) \,|\, \Theta_1 = \theta_1, \Theta_1 = \theta_2, \ldots, \Theta_i = \theta_i^*, \ldots, \Theta_n = \theta_n\right) \geq 1-\alpha \tag{3.9}$$

for any $\psi^* \in \Psi$ and any $\theta_j \in \mathcal{T}$ with $j \neq i$, as well as a test with a rejection rule that satisfies (writing $\mathbb{I}^c$ for $\{1,\ldots,n\} \setminus \mathbb{I}$)

$$\mathbb{P}_{\psi^*}\left[\text{Reject } H_0 \mid \Theta_{\mathbb{I}} \in \mathcal{T}_{\mathbb{I},0}, \Theta_{\mathbb{I}^c} \in \mathcal{T}'\right] \leq \alpha \tag{3.10}$$

for any $\psi^* \in \Psi$ and any subset $\mathcal{T}'$ of $\mathcal{T}^{n-|\mathbb{I}|}$, with the convention that the second event in the conditioning part is ignored if $\mathbb{I} = \{1,\ldots,n\}$.

According to Nguyen and Gupta (2023), Empirical Bayes tests and confidence sets had been successfully constructed for certain classes of models, but a general approach to solving this inference problem had not existed yet prior to their publication. Their construction idea can be viewed as a direct transformation of split likelihood ratio confidence sets and tests to the empirical Bayesian setup: Suppose that for some $i = 1,\ldots,n$, we are interested in finding a confidence set for θ_i^* with conditional coverage probability $1-\alpha \in (0,1)$. First, choose an estimator $\hat{\psi}(X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots, X_n)$ for the unknown true parameter ψ^* which is based on all data points except for X_i. For any $\psi \in \Psi$, define the full likelihood of ψ as

$$\mathcal{L}_i^{\text{full}}(\psi) = \int_{\mathcal{T}} f_{X|\Theta}(X_i|\theta)\, \pi_\psi(\theta)\, d\mu_\theta$$

and for any $\theta \in \mathcal{T}$, define the conditional likelihood of θ as

$$\mathcal{L}_i^{\text{cond}}(\theta) = f_{X|\Theta}(X_i|\theta),$$

where both likelihood functions are only based on the observation of X_i. Defining the statistic

$$EBLR(i,\theta) = \begin{cases} \frac{\mathcal{L}_i^{\text{full}}(\hat{\psi})}{\mathcal{L}_i^{\text{cond}}(\theta)} & \text{if } \mathcal{L}_i^{\text{cond}}(\theta) > 0, \\ +\infty & \text{if } \mathcal{L}_i^{\text{cond}}(\theta) = 0. \end{cases}$$

for any $\theta \in \mathcal{T}$, it holds by arguments analogous to the ones made in the proofs of Lemmas 2 and 6 that

$$\mathbb{E}_{\psi^*}\left[EBLR(i,\theta_i^*) \,|\, \Theta_1 = \theta_1, \Theta_1 = \theta_2, \ldots, \Theta_i = \theta_i^*, \ldots, \Theta_n = \theta_n\right] \leq 1$$

for any $\psi^* \in \Psi$ and any $\theta_j \in \mathcal{T}$ with $j \neq i$. Consequently, the empirical Bayesian confidence set defined as

$$EBC_n(\alpha,i) = \left\{\theta \in \mathcal{T} : EBLR(i,\theta) > \frac{1}{\alpha}\right\} \tag{3.11}$$

satisfies the conditional minimum coverage probability stated in (3.9). Similarly, to construct a test for the null hypothesis $H_0 : \theta_{\mathbb{I}}^* \in \mathcal{T}_{\mathbb{I},0}$ (where $\mathbb{I}$ is a proper (!) and non-empty subset of $\{1,\ldots,n\}$, $\theta_{\mathbb{I}}^* = (\theta_i^*)_{i\in\mathbb{I}}$ and $\mathcal{T}_{\mathbb{I},0} \subseteq \mathcal{T}^{|\mathbb{I}|}$ is non-empty) against any

alternative, which controls the conditional type I error probability at level $\alpha \in (0,1)$, define an estimator $\hat{\psi}(\mathbb{I}^c)$ that depends only on the subset of the data indexed by $\mathbb{I}^c$ and define the full likelihood of ψ based on the data indexed by $\mathbb{I}$ as

$$\mathcal{L}_{\mathbb{I}}^{\text{full}}(\psi) = \prod_{i \in \mathbb{I}} \int_{\mathcal{T}} f_{X|\Theta}(X_i|\theta)\, \pi_{\psi}(\theta)\, d\mu_{\theta}$$

as well as the conditional likelihood of $\theta_{\mathbb{I}}$ derived from the same data as

$$\mathcal{L}_{\mathbb{I}}^{\text{cond}}(\theta_{\mathbb{I}}) = \prod_{i \in \mathbb{I}} f_{X|\Theta}(X_i|\theta_{\mathbb{I}}).$$

Next, we define the test statistic

$$EBT_n = \begin{cases} \dfrac{\mathcal{L}_{\mathbb{I}}^{\text{full}}\left(\hat{\psi}(\mathbb{I}^c)\right)}{\sup_{\vartheta_{\mathbb{I}} \in \mathcal{T}_{\mathbb{I},0}} \mathcal{L}_{\mathbb{I}}^{\text{cond}}(\vartheta_{\mathbb{I}})} & \text{if } 0 < \sup_{\vartheta_{\mathbb{I}} \in \mathcal{T}_{\mathbb{I},0}} \mathcal{L}_{\mathbb{I}}^{\text{cond}}(\vartheta_{\mathbb{I}}) < +\infty, \\ 0 & \text{if } \sup_{\vartheta_{\mathbb{I}} \in \mathcal{T}_{\mathbb{I},0}} \mathcal{L}_{\mathbb{I}}^{\text{cond}}(\vartheta_{\mathbb{I}}) = +\infty, \\ +\infty & \text{if } \sup_{\vartheta_{\mathbb{I}} \in \mathcal{T}_{\mathbb{I},0}} \mathcal{L}_{\mathbb{I}}^{\text{cond}}(\vartheta_{\mathbb{I}}) = 0. \end{cases}$$

For this statistic, it can be shown by a similar argument as in the proof of Theorem 3 of Wasserman et al. (2020) that

$$\mathbb{E}_{\psi^*}\left[EBT_n \,|\, \Theta_{\mathbb{I}} \in \mathcal{T}_{\mathbb{I},0}, \Theta_{\mathbb{I}^c} \in \mathcal{T}'\right] \leq 1$$

for any $\psi^* \in \Psi$ and any $\mathcal{T}' \subseteq \mathcal{T}^{n-|\mathbb{I}|}$. Therefore, the test which rejects the null hypothesis if and only if $EBT_n > \frac{1}{\alpha}$ indeed has level α in the sense of (3.10).

Nguyen and Gupta (2023) apply this strategy for Empirical Bayesian inference to three well-known Bayesian models (in all of them, it is assumed that $(X_1, \Theta_1), \ldots, (X_n, \Theta_n)$ is an i.i.d.-sequence): Stein's problem ($\Theta_i \sim N(0, \psi^2)$ with $\psi^2 > 0$ and $[X_i|\Theta_i = \theta_i] \sim N(\theta_i, 1)$), the Poisson-Gamma count model ($\Theta_i \sim \text{Gamma}(a, b)$ with $a, b > 0$ and $[X_i|\Theta_i = \theta_i] \sim \text{Poisson}(\theta_i\, w_i)$ with $w_i > 0$ known) and the Beta-Binomial model ($\Theta_i \sim \text{Beta}(\alpha, \beta)$ with $\alpha, \beta > 0$ and $[X_i|\Theta_i = \theta_i] \sim \text{Binom}(m_i, \theta_i)$ with $m_i > 0$ known). Using data simulated from these models, they found that the resulting confidence sets and tests on the sequence $\Theta_1, \ldots, \Theta_n$ to be very conservative and that for sample sizes up to 1000, no significant empirical decrease of average length (in the case of confidence intervals) or significant empirical increase of power (in the case of hypothesis tests) can be observed. For example, in Stein's problem, the conditional confidence interval for Θ_n constructed as in (3.11) has an empirical conditional coverage probability that can not be distinguished from 1 for $\alpha \leq 0.05$ and for almost all values of ψ^2 and the average length of this interval remains stable when increasing the sample size from 10 to 1000 observations. The same can be observed for confidence intervals resulting from data that is simulated from the Poisson-Gamma count model. However, while in Stein's model, the average length of the interval does not seem to depend on the true value of ψ^2, the empirical average lengths of the intervals resulting from simulations of the Poisson-Gamma model differ considerably for different true values of a and b. Testing the null hypothesis

$H_0 : \theta^*_{n-1} = \theta^*_n$ against the alternative $H_1 : \theta^*_{n-1} \neq \theta^*_n$ in the Poisson-Gamma model with simulated data, Nguyen and Gupta (2023) find that the empirical size of this test is more or less indistinguishable from 0 for $\alpha \leq 0.05$ as well as almost all true values of a and b and that the empirical power of this test, while depending on the true values of a and b, grows very slowly when increasing the sample size from 10 to 1000.

3.4.2 Testing for log-concavity

In non-parametric density estimation, it is a common procedure to impose shape constraints on the distributions considered to be possible for the data, as this can be viewed as a middle ground between a parametric and a fully non-parametric approach. A particularly popular constraint is *log-concavity*: A probability density function $f : \mathbb{R}^d \mapsto [0, +\infty)$ is called log-concave if

$$f(x) = e^{g(x)}$$

for some concave function $g : \mathbb{R}^d \mapsto \mathbb{R}$. The class of log-concave distributions includes many of the most-commonly encountered continuous families such as the normal, uniform (on a convex support), logistic and Laplace family (see Bagnoli and Bergstrom, 2005). Furthermore, d-dimensional log-concave densities have numerous properties one often considers to be desirable: They are uni-modal, their cumulative distribution function and the marginal densities of all subvectors are log-concave as well, their support is a convex subset of $\mathbb{R}^d$ and in the univariate case ($d = 1$), their right tail is at most exponential, which implies that all the moments of the corresponding distribution are guaranteed to exist if its support (l, u) satisfies $l > -\infty$ (see An, 1998, where several other properties are discussed, too). Regarding statistical inference, the class of log-concave densities has attractive properties as well: According to Theorem 1 in Cule et al. (2010), if one observes the d-dimensional data $X_1, X_2, \ldots, X_n$ (with $n > d$) which are i.i.d. according to an unknown probability density function f^*, then the maximum likelihood estimator for f^* among the class of d-dimensional log-concave densities $\mathcal{F}_d$,

$$\hat{f}_n = \arg\max_{f \in \mathcal{F}_d} \sum_{i=1}^{n} \ln\Big(f(X_i)\Big)$$

exists and is unique almost surely, regardless of whether $f^* \in \mathcal{F}_d$ or not. Moreover, while an analytic expression of $\hat{f}_n$ is typically unavailable, several algorithms to approximate this maximum likelihood estimator have been proposed and also implemented in statistical software (see, for example, Cule et al., 2009; Dümbgen and Rufibach, 2011). Importantly, these algorithms do not require the choice of a bandwidth, while for the commonly used Kernel density estimator, a $d \times d$ bandwidth matrix needs to be specified to estimate the true density from d-dimensional data.

Consequently, testing the assumption that the true probability density function is log-concave can be seen as a perfect opportunity to apply universal inference, specifically, the subsampling likelihood ratio test: The result of such a test can be used as a motivation to choose the method of log-concave density estimation and, although the underlying

statistical model is non-parametric, the split LR test can be implemented because the maximum likelihood estimator under the null hypothesis exists and can be approximated well. Furthermore, according to Dunn et al. (2024), while several ideas how to test this assumption exist in the literature, none of them have provable guarantees regarding their type I error probability, not even in an asymptotic sense. To define the hypotheses and the corresponding test statistic formally, suppose we observe $\mathbb{R}^d$-valued i.i.d. data $X_1, \ldots, X_n$ (with n > d), distributed according to the probability density function $f^*(x)$, and we want to test the null hypothesis $H_0 : f^* \in \mathcal{F}_d$ against the alternative $H_0 : f^* \notin \mathcal{F}_d$. To this end, we compute $B \in \mathbb{N}$ (possibly random) partitions of the index set $\{1, \ldots, n\}$, $\left(\{\mathbf{D_{0,b}}, \mathbf{D_{1,b}}\}\right)_{b=1,\ldots,B}$, and, for every $b = 1, \ldots, B$, define some estimator $\hat{f}_{1,b}$ for f^* based only on the data indexed by $\mathbf{D_{1,b}}$. Following (3.2), we define

$$U_{n,b} = \frac{\prod_{i \in \mathbf{D_{0,b}}} \hat{f}_{1,b}(X_i)}{\prod_{i \in \mathbf{D_{0,b}}} \hat{f}_{0,b}(X_i)}, \tag{3.12}$$

where $\hat{f}_{0,b}(x)$ is the maximum likelihood estimator for f^* among the class of log-concave densities, computed only with the subset of the data indexed by $\mathbf{D_{0,b}}$. Averaging these random variables, we define the test statistic

$$U_n^{\text{sub}} = \frac{1}{B} \sum_{b=1}^{B} U_{n,b} \tag{3.13}$$

and, following the discussion from Subsection 3.1.1, the test with the decision rule

$$\text{Reject } H_0 \text{ iff. } U_n^{\text{sub}} > \frac{1}{\alpha}, \tag{3.14}$$

has significance level α for any $\alpha \in (0, 1)$. Importantly, this result holds in finite samples and for any dimension d. Regarding the estimators $\hat{f}_{1,b}$, while no choice of them can compromise the significance level, it would be wise to choose ones that approximate the true density f^* well, as otherwise, the power of the resulting test might be a lot lower compared to a subsampling LR test that uses 'good' estimators for f^*. Dunn et al. (2024) suggest either to use a parametric approach, so to implicitly assume that $f^* \in \{f_\theta : \theta \in \Theta\}$ with $\Theta \subseteq \mathbb{R}^p$ and therefore use $\hat{f}_{1,b}(x) = f_{\hat{\theta}_{1,b}}(x)$, where $\hat{\theta}_{1,b} \in \arg\max_{\vartheta \in \Theta} \prod_{i \in \mathbf{D_{1,b}}} f_\vartheta(X_i)$ (assuming that such a maximizer for the likelihood function based on $\mathbf{D_1}$ exists), or to set $\hat{f}_{1,b}$ as the d-dimensional Kernel density estimator computed from the data subset indexed by $\mathbf{D_{1,b}}$, where the corresponding $d \times d$ bandwidth matrix may be pre-specified by the user or determined by procedures such as cross-validation (which, of course, may only be performed on the data indexed by $\mathbf{D_{1,b}}$). In theory, it is also possible to mix both approaches, i.e. to use a parametric maximum likelihood estimator for f^* when $b \in S_{\text{par}}$ and a Kernel density estimator for f^* when $b \in S_{\text{nonpar}}$, where S_{par} is a proper, non-empty subset of $\{1, \ldots, B\}$ and $S_{\text{nonpar}} = \{1, \ldots, B\} \setminus B_{\text{par}}$.

While the theory of universal inference guarantees that the test using the decision rule (3.14) controls the type I error probability at level $\alpha \in (0, 1)$, the power of this test for

a specific true density f^* needs to be evaluated empirically. As an interesting showcase, Dunn et al. (2024) consider a specific d-dimensional Gaussian mixture model, namely $X \sim f_{\mu^*}$ with

$$f_{\mu^*}(x) = \frac{1}{2}\,\phi_d(x) \,+\, \frac{1}{2}\,\phi_d(x-\mu^*),$$

where $\phi_d(x)$ is the probability density function of the $N_d(0, I_d)$-distribution. This is an interesting setting for our discussion because in this case, the true density f_{μ^*} is log-concave if and only if $||\mu^*|| \leq 2$. Therefore, they generate $n = 100$ observations from this model for different values of $\mu^* = (\mu_1^*, 0, \ldots, 0)^T \in \mathbb{R}^d$ and compute the result of the statistical test (3.14) (with $B = 100$) based on these observations. These simulations show that the power function (depending on μ_1^*) behaves reasonably well in the univariate case, but with growing dimension d, the subsampling LR test for log-concavity becomes catastrophically conservative, no matter which estimators $\hat{f}_{1,b}$ are used for f^*. As Dunn et al. (2024) note, the simulations indicate that $||\mu^*||$ would need to grow exponentially with respect to d to maintain a stable power level, even if for any subsampling split $b \in \{1, \ldots, B\}$, the estimator $\hat{f}_{1,b}$ is replaced by the true density f^*. This curse of dimensionality of universal inference was also noticed in the multivariate Gaussian model by Strieder and Drton (2022) and a heuristic explanation for it is given in Remark 4.2 of this work. Interestingly, the simulations by Dunn et al. (2024) from this mixture model indicate that empirically, the permutation test for log-concavity suggested by Cule et al. (2009) behaves exactly the opposite way: The power function of this test increases with the dimension d. However, it also seems to violate the nominal significance level $\alpha \in (0, 1)$ for $d \geq 3$.

Since the power of the subsampling likelihood ratio test becomes catastrophically low for large dimensions d, Dunn et al. (2024) suggest to combine the universal inference approach with a dimensionality reducing procedure. Their idea is to convert the d-dimensional density of the data into a univariate one via projections, as in this setting, the property to be tested (log-concavity) is preserved under projections: According to part (a) of Proposition 1 in Cule et al. (2010), if the random vector $Y \in \mathbb{R}^d$ is distributed according to a log-concave density, then for any given subspace V of $\mathbb{R}^d$, the marginal density of the orthogonal projection of Y onto V, $P_V(Y)$, is log-concave as well. Dunn et al. (2024) present two strategies for dimensionality reduction that build on this result. On the one hand, a trivial consequence of it, which was already mentioned above, is that under the null hypothesis, the d marginal densities of f^* (which correspond to the densities of the d axis-aligned projections of $X_i = (X_i^{(1)}, \ldots, X_i^{(d)})$ are log-concave, too. Therefore, it holds that under the null, for any collection of data splits $\left(\{\mathbf{D_{0,b}}, \mathbf{D_{1,b}}\}\right)_{b=1,\ldots,B}$, any estimator $\hat{f}_{1,b,k}$ for a one dimensional density function, which is based only on $\left(X_i^{(k)}\right)_{i \in \mathbf{D_{1,b}}}$ and any $k = 1, \ldots, d$, the expectation of the non-negative statistic

$$U_{n,b,k}^{\text{axis-proj}} = \frac{\prod_{i \in \mathbf{D_{0,b}}} \hat{f}_{1,b,k}(X_i^{(k)})}{\prod_{i \in \mathbf{D_{0,b}}} \hat{f}_{0,b,k}(X_i^{(k)})},$$

where $\hat{f}_{0,b,k}$ is the log-concave maximum likelihood estimator based only on the data $\left(X_i^{(k)}\right)_{i \in \mathbf{D_{0,b}}}$, is bounded by 1. Consequently, the same holds for the test statistic

$$U_n^{\text{sub,axis-proj}} = \frac{1}{d} \frac{1}{B} \sum_{k=1}^{d} \sum_{b=1}^{B} U_{n,b,k}$$

and thus, the test that rejects the null hypothesis of log-concavity iff. $U_n^{\text{sub,axis-proj}} > \frac{1}{\alpha}$ has level $\alpha \in (0,1)$. Similarly, instead of projecting the data onto the d axes, we may use n_{proj} random vectors (which of course should be independent of the data) $V_1, \ldots, V_{n_{proj}} \in \mathbb{R}^d$ and compute the projection $P_{V_k}(X_i) = V_k^T X_i$ for each $k = 1, \ldots, n_{proj}$ and each $i = 1, \ldots, n$. As an example, each of the random vectors V_k may be defined as $\frac{X_k}{||X_k||}$ with $X_1, \ldots, X_{n_{proj}} \overset{\text{i.i.d.}}{\sim} N(0, I_d)$, which would result in the V_k being independently and uniformly distributed on the d-dimensional unit sphere. Then, we may adapt the test defined via (3.12), (3.13) and (3.14) in the same way as with the axis-aligned projections above and the resulting decision rule will again control the type I error at level α.

To reveal that the subsampling likelihood ratio tests (with $B = 100$ data splits) in combination with axis-aligned or random projections perform much better in high-dimensional models than the subsampling LR tests without the help of dimension reduction, Dunn et al. (2024) simulate 100 observations from the multivariate Gaussian mixture model mentioned above for different values of $\mu^* = (\mu_1^*, 0, \ldots, 0)^T \in \mathbb{R}^d$ and compute the corresponding results of the three tests. The empirical power functions of the adapted subsampling LR tests decrease only very slowly between dimensions $d = 1$ and $d = 4$. The power functions of the two adaptions behave very similarly, with the power of the axis-aligned projection approach slightly beating the one of the random projection approach. In this specific parametrization of a multivariate Gaussian mixture model, the dominance of dimension reduction via axis-aligned projections is to be expected because under the alternative, all the empirical 'non-log-concavity' is due to the first component of X_i. Thus, Dunn et al. (2024) continue to show empirically that, if the data is simulated from the parametrization

$$\mu^* = \frac{1}{\sqrt{d}} \left(||\mu^*||, ||\mu^*||, \ldots, ||\mu^*||\right)^T \in \mathbb{R}^d$$

(so the empirical 'non-log-concavity' in the alternative case $||\mu^*|| > 2$ is split evenly among all the components of X_i), then, not surprisingly, the random projection approach beats the axis-aligned one. Consequently, we might conclude that if the 'non-log-concavity' of the true data generating distributions f^* is due to one or few of its components, then dimension reduction with axis-aligned projections will perform better and if the 'non-log-concavity' of f^* is distributed rather evenly among its components, then the random projection method will have higher power. All in all, a useful conclusion of this use case of universal inference is that in high-dimensional models, one should try to reduce the dimension of the inference problem before applying universal inference methods.

3.4.3 Testing for homogeneity in a two-component Gaussian mixture model

Gaussian mixture models are one of the most commonly used irregular statistical models. In this context, irregularity means that one of the conditions that imply that the likelihood ratio test statistic $LRT_n = 2\left(\sup_{\vartheta\in\Theta}\mathcal{L}(\vartheta) - \sup_{\vartheta\in\Theta_0}\mathcal{L}(\vartheta)\right)$ for testing a null hypothesis $H_0 : \theta \in \Theta_0$ against the alternative $H_1 : \theta \notin \Theta_0$ asymptotically follows a χ^2-distribution with $d-p$ degrees of freedom (where d is the dimensionality of the whole parameter space Θ and p is the dimensionality of the null hypothesis set Θ_0) is violated. Given this irregularity, it is not surprising that the application of universal inference in the context of two-component Gaussian mixture models was already suggested by Wasserman et al. (2020). In general, a d-dimensional Gaussian mixture model with two components may be defined as the set of probability measures with probability density functions:

$$f_{p,\mu^{(1)},\mu^{(2)},\Sigma_1,\Sigma_2}(x) = p\ \phi_d\Big(\Sigma_1^{-\frac{1}{2}}\left(x-\mu^{(1)}\right)\Big)\ +\ (1-p)\ \phi_d\Big(\Sigma_2^{-\frac{1}{2}}\left(x-\mu^{(2)}\right)\Big), \tag{3.15}$$

where $p \in [0,1]$; $\mu_1^{(1)}, \mu^{(2)} \in \mathbb{R}^d$; Σ_1, Σ_2 are symmetric, positive definite $d \times d$-matrices and $\phi_d(x)$ is the density function of the $N_d(0, I_d)$-distribution. From a Bayesian point of view, this model may be interpreted in the following way: First, a random variable G with $\mathbb{P}(G = 1) = p$ and $\mathbb{P}(G = 2) = 1 - p$ is realized. Depending on the value of G, the observation X is generated from one of two possible distributions: If $G = 1$, then X is drawn from a $N_d(\mu^{(1)}, \Sigma_1)$-distribution. If $G = 2$, then X follows a $N_d(\mu^{(2)}, \Sigma_2)$-distribution. In this context, $p \in [0,1]$ is usually called the mixture weight. Many different specifications can be derived from the general model formulation (3.15): The mixture weight p may be known, one or both of the group mean parameters $\mu^{(1)}, \mu^{(2)}$ may be known or one may assume that $\Sigma_1 = \Sigma_2$. The latter specification is called the *homoscedastic* case. It already gets rid of one of the problems regarding inference in the general multivariate Gaussian mixture model, namely that the likelihood function is unbounded with probability 1 (to see why, consider the univariate case $d = 1$: Fix some $p \in (0,1]$ and $\mu^{(2)} \in \mathbb{R}$, set $\mu^{(1)} = X_1$ and let $\Sigma_1 = \sigma_1^2 \to 0$). However, even in the homoscedastic case, the test for homogeneity, which is probably the most interesting hypothesis in the context of mixture models, fails to satisfy the usual regularity assumptions: Consider the univariate, homoscedastic version of (3.15) with known variance $\sigma^2 = \sigma_1^2 = \sigma_2^2 = 1$ and known group mean parameter $\mu^{(1)} = 0$, so the corresponding densities are

$$f_{p,\mu^{(2)}}(x) = p\ \phi\big(x\big)\ +\ (1-p)\ \phi\big(x-\mu^{(2)}\big), \tag{3.16}$$

where $\phi(x)$ denotes the density function of the standard normal distribution. Homogeneity in the context of mixture models means that the data is in fact generated by only one distribution instead of two different ones. In the model formulation (3.16), the null hypothesis of homogeneity can be written as

$$H_0 : p^* = 1 \text{ or } \mu_*^{(2)} = 0, \tag{3.17}$$

while the alternative hypothesis of heterogeneity may be specified as

$$H_1 : p^* \in (0,1) \text{ and } \mu_*^{(2)} \neq 0.$$

Note that we have further restricted the null hypothesis to the assumption that all the data is generated from a standard normal distribution, i.e. if the data is generated from only one normal distribution, we automatically assume it to be the standard one. It is easy to see that under the null hypothesis, the Fisher information matrix is singular (note that $\frac{\partial}{\partial \mu^{(2)}} \ln[f_{p,\mu^{(2)}}(x)]\big|_{p=1} = \frac{\partial}{\partial p} \ln[f_{p,\mu^{(2)}}(x)]\big|_{\mu^{(2)}=0} = 0$ for all $x \in \mathbb{R}$), thus, the classical regularity conditions from Wilks' theorem are violated. Indeed, Liu and Shao (2004) prove in Theorem 2 of their paper that under the null hypothesis (the data is distributed according to the standard normal distribution), the likelihood ratio test statistic LRT_n for the null hypothesis (3.17) does not converge in distribution to χ_1^2, but diverges to $+\infty$ at rate $O_p(\log\log(n))$. Specifically, they show that $LRT_n - \log\log n + \log(2\pi^2)$ converges in distribution to a Gumbel distribution with mean parameter $\mu = 0$ and scale parameter $\beta = 2$.

Theorem 25. *Suppose that* $X_1, \ldots, X_n \overset{i.i.d}{\sim} N(0,1)$. *Then, for the likelihood ratio test statistic for the hypothesis (3.17) in the model (3.16),*

$$LRT_n = 2\Big(\sup_{p\in[0,1],\, \mu^{(2)}\in\mathbb{R}} \sum_{i=1}^{n} \ln\big(p\,\phi(X_i) \;+\; (1-p)\,\phi(X_i - \mu^{(2)})\big) - \sum_{i=1}^{n} \ln\big(\phi(X_i)\big)\Big),$$

it holds that

$$\lim_{n\to\infty} \mathbb{P}(LRT_n - \log\log n + \log(2\pi^2) \leq x) = \exp\Big(-\exp\big(-\frac{x}{2}\big)\Big) \tag{3.18}$$

for any $x \in \mathbb{R}$.

On the other hand, as the universal inference methods do not require any regularity assumptions on the statistical model or the null hypothesis to be tested, the split likelihood ratio test for homogeneity with nominal significance level $\alpha \in (0,1)$ will be valid, even in finite samples. In this situation, a natural choice for the estimator $\hat{\theta}_1 = (\hat{p}_1, \hat{\mu}_1^{(2)})^T$ is the maximum likelihood estimator for $\theta^* = (p^*, \mu_*^{(2)})^T$ based on the data indexed by $\mathbf{D_1}$. This maximum likelihood estimator exists almost surely and, while not available analytically, may be approximated with the expectation maximization algorithm (see, for example, McLachlan and Krishnan, 2008).

Shi and Drton (2024) study the asymptotic properties of the log split likelihood ratio test statistic in this setting: To this end, they fix a splitting ratio $r_0 \in (0,1)$ and assume that, for n large enough, the index set $\{1,\ldots,n\}$ is partitioned into $\{\mathbf{D_0}, \mathbf{D_1}\}$, with $|\mathbf{D_0}| = \lfloor r_0 n \rfloor$. Then, they show that under the null hypothesis, the log split likelihood ratio test statistic, $SLR_n = 2\ln(U_n)$ (where U_n is the split LR test statistic as defined in Definition 14), tends to $-\infty$ at order $O_p(\log\log n)$ and, with an appropriate scaling, is asymptotically normal (Theorem 3.3 of Shi and Drton, 2024).

Theorem 26. *In the setting discussed above of a splitting ratio r_0 that is constant with respect to the sample size n, the log split likelihood ratio test statistic SLR_n for the null hypothesis (3.17) in the model (3.16) is defined as*

$$SLR_n = 2\Big(\ell_0\big(\hat{p}_{1,n}, \hat{\mu}_1^{(2)}\big) - \sum_{i \in \mathbf{D_0}} \ln\big(\phi(X_i)\big)\Big),$$

where $(\hat{p}_{1,n}, \hat{\mu}_{1,n}^{(2)})^T$ is the maximum likelihood estimator for $(p^*, \mu_*^{(2)})^T$ based only on the data subset indexed by $\mathbf{D_1}$ and $\ell_0\big(p, \mu^{(2)}\big) = \sum_{i \in \mathbf{D_0}} \ln\big(f_{p,\mu^{(2)}}(X_i)\big)$. Suppose that the null hypothesis is true, i.e. we observe $X_1, \ldots, X_n \overset{i.i.d}{\sim} N(0,1)$. Then we have

$$\frac{SLR_n + \frac{r_0}{1-r_0} \log\log n}{2\sqrt{\frac{r_0}{1-r_0} \log\log n}} \xrightarrow{d} N(0,1) \tag{3.19}$$

This result is quite surprising given the fact that the classical likelihood ratio test statistic for homogeneity in this model is not asymptotically normal, but converges (with an appropriate scaling) to an extreme value distribution (the Gumbel distribution). Shi and Drton (2024) use (3.19) to show that the split likelihood ratio test in this setting (which rejects H_0 if $SLR_n > 2\big(\ln\big(\frac{1}{\alpha}\big)\big)$) can distinguish the null hypothesis from a local alternative at the rate $\sqrt{\frac{1}{1-r_0}}\sqrt{\frac{\log\log n}{n}}$. This is an interesting observation: It means that when testing the hypothesis of homogeneity in a contaminated Gaussian mixture model, the rate of detection depends on the splitting ratio r_0. In other words, choosing a small r_0 has the benefit of a smaller detection rate. However, Shi and Drton (2024) find in their simulations that, as long as the local alternatives shrink at a rate $\gamma\sqrt{\frac{\log\log n}{n}}$ with $|\gamma| > \sqrt{\frac{1}{1-r_0}}$, choosing a higher value of r_0 leads to a higher empirical power. If we instead want to consider a test for homogeneity built on the classical likelihood ratio test statistic LRT_n with asymptotic size α, then, following (3.18), we should use the decision rule

$$\text{Reject } H_0 \text{ iff. } LRT_n > \log\log n - \log(2\pi^2) - 2\log\big(-\log(1-\alpha)\big). \tag{3.20}$$

Theorem 2.1 of Hall and Stewart (2005) shows that this (asymptotically valid) likelihood ratio test can detect alternative models at rate $\sqrt{\frac{\log\log n}{n}}$. Therefore, Shi and Drton (2024) conclude that in the contaminated Gaussian mixture model, the split likelihood ratio test and the likelihood ratio test (3.20) have the same detection rate for homogeneity up to a constant (which depends on the splitting ratio r_0).

4 Comparisons of the split Likelihood Ratio and the classical Likelihood Ratio confidence set

We now turn to a discussion about the relation between the split likelihood ratio confidence set defined in Chapter 2 and the classical likelihood ratio confidence set, which is used for inference in many statistical models of practical interest. Specifically, we will compare the split likelihood ratio confidence set that uses a maximum likelihood estimator based on $\mathbf{D_1}$ (if it exists) for $\hat{\theta}_1$ with the classical likelihood ratio set in terms of subset relations as well as their respective volumes. Only considering this specific type of split LR confidence sets is a sensible restriction because then, both sets will be very similar, in the sense that they both will be based on likelihood ratio statistics (the split set uses the likelihood function based on $\mathbf{D_0}$, while the classical set uses the likelihood function based on the whole data), where in the denominator, the corresponding likelihood function is evaluated at a maximum likelihood estimator (for the split set, the estimator is based on $\mathbf{D_1}$, while for the classical set, the estimator is based on the whole data). Moreover, in the context of universal inference, using maximum likelihood estimation (if it is available) in general seems like a natural choice and arriving at conclusions about the split LR confidence set in general is difficult if no restriction on the type of estimator used for $\hat{\theta}_1$ is imposed. In this context, it could be conjectured that (probably under weak conditions regarding the statistical model), among all possible choices for $\hat{\theta}_1$, the average diameter and/or volume (Lebesgue measure) of the resulting split likelihood ratio confidence set is minimized by the maximum likelihood estimator. However, to our knowledge, a result of this kind has not been published yet.

In this chapter, we will mainly present results which, to our knowledge, are not yet known in the literature. As was mentioned above, we will restrict the following discussion to a special case of the split likelihood ratio confidence set, namely the one which uses maximum likelihood estimation for $\hat{\theta}_1$. We will call the resulting confidence set a **split maximum likelihood ratio confidence set** (**split MLR** confidence set).

Definition 27. *Consider a statistical model $\{\mathbb{P}_\theta : \theta \in \Theta\}$, where $\Theta \subseteq \mathbb{R}^d$ is non-empty, a given sample size n and a partition $\{\mathbf{D_0}, \mathbf{D_1}\}$ of the index set $\{1, \dots, n\}$. Assume that for any true parameter $\theta_* \in \Theta$, the likelihood function based on the data indexed by $\mathbf{D_1}$ has at least one maximizer $\mathbb{P}_{\theta_*}$-a.s., i.e. $\mathbb{P}_{\theta_*}(\operatorname{argmax}_{\vartheta\in\Theta} \mathcal{L}_1(\vartheta) \neq \emptyset) = 1$. Define $\hat{\theta}_1^{ML}$ to be a maximum likelihood estimator based on $\mathbf{D_1}$, by which we mean some measurable selection (with respect to the data indexed by $\mathbf{D_1}$) of $\operatorname{argmax}_{\vartheta\in\Theta} \mathcal{L}_1(\vartheta)$. Then we call the split likelihood ratio confidence set with $\hat{\theta}_1 = \hat{\theta}_1^{ML}$ and nominal coverage probability $1-\alpha$*

*the **split maximum likelihood ratio confidence set** (**split MLR** confidence set) with nominal coverage probability $1-\alpha$ and denote it as $C_n^{ML}(\alpha)$. Furthermore, we denote the statistic on which this confidence set is based as $T_n^{ML}(\alpha)$.*

The next definition then deals with the well-known likelihood ratio confidence set. While the case of parameter values θ with $\mathcal{L}(\theta) = 0$ is usually not discussed in the literature, we explicitly include it here because it was taken into account in our definition of the split likelihood ratio confidence set as well. We also consider the possibility that the maximum likelihood estimator for θ_* based on the whole sample may not exist, which includes the case that the likelihood function is unbounded.

Definition 28. *For any statistical model $\{\mathbb{P}_\theta : \theta \in \Theta\}$, where $\Theta \subseteq \mathbb{R}^d$ is non-empty, and given sample size n, we define the $(0, +\infty]$-valued-**likelihood ratio statistic** as:*

$$LR_n(\theta) = \begin{cases} \frac{\sup_{\vartheta \in \Theta} \mathcal{L}(\vartheta)}{\mathcal{L}(\theta)} \text{ if } \mathcal{L}(\theta) > 0, \\ +\infty \text{ if } \mathcal{L}(\theta) = 0. \end{cases}$$

*Following this, we define the **classical likelihood ratio confidence set** (**LR** confidence set) with nominal coverage probability $1-\alpha$ to be[1]:*

$$LRC_n(\alpha) = \Big\{\theta \in \Theta : 2\ln\big(LR_n(\theta)\big) \leq Q_d(1-\alpha)\Big\},$$

*where $Q_d(\cdot)$ is the quantile function of the **chi-squared** distribution with d degrees of freedom.*

It is a well-known result (see, for example, Theorem 12.4.2 in Lehmann and Romano, 2005) that under certain conditions, $2\ln\big(LR_n(\theta_*)\big) \xrightarrow{w} \chi^2_d$, where θ_* is the true parameter. Consequently, if a statistical model satisfies these conditions, then $LRC_n(\alpha)$ is an asymptotically valid confidence set with nominal coverage probability $1-\alpha$ for θ_*, in the sense that $\mathbb{P}_{\theta_*}\Big(\theta_* \in LRC_n(\alpha)\Big) \xrightarrow{n\to\infty} 1-\alpha$, for any $\alpha \in (0,1)$.

4.1 Comparisons in dimensions 1 and 2

For the first part of our discussion about the relation between the split MLR and the classical LR confidence set, we will restrict ourselves to one- and two-dimensional models. The reason for this is the specific relationship between the upper quantile function $Q_d(1-\alpha)$ of a chi-square distribution with $d \in \mathbb{N}$ degrees of freedom (which is the critical value used for the classical LR confidence set with nominal coverage probability $1-\alpha$ in a d-dimensional model) and the function $\alpha \mapsto \frac{1}{\alpha}$ (which is the critical value used for the split LR confidence set with nominal coverage probability $1-\alpha$, in any dimension d). This is illustrated by Lemma 29. Using this result, the central theorem of this section, Theorem 30, shows that the classical LR set is always a subset of the split MLR set as soon as the latter exists. Proposition 31 then gives a minimum condition under which

[1] With the convention that $\ln(+\infty) = +\infty$.

the classical set is even a proper subset of the split set for some (data dependent) α. Following this, we show that slightly stronger conditions on the statistical model (which are still satisfied in most of the models commonly considered in applications) imply that the volume (Lebesgue measure) of the split MLR confidence set is almost surely larger than the one of the classical LR confidence set, as long as for the given $\alpha \in (0,1)$, there is a positive probability that the classical set with confidence level $1-\alpha$ does not already cover the whole parameter space: Theorem 32 expalins these conditions for dimension 1, while Theorem 35 describes the (slightly stronger) ones needed in dimension 2. We want to stress that even in Theorem 35, the required assumptions on the statistical model are very weak, such that for most models of interest in dimension 1 and 2, we can conclude that the classical LR confidence set will always turn out to be a subset of the split MLR one and that with positive probability, the latter will have a larger volume than the former.

Lemma 29. *Define $Q_d : (0,1) \mapsto (0,+\infty)$ as the quantile function of the chi-squared distribution with $d \in \mathbb{N}$ degrees of freedom. Then, for any $\alpha \in (0,1)$:*

$$\begin{aligned} &i)\ e^{\frac{Q_1(1-\alpha)}{2}} < \frac{1}{\alpha} \\ &ii)\ e^{\frac{Q_2(1-\alpha)}{2}} = \frac{1}{\alpha} \\ &iii)\ e^{\frac{Q_d(1-\alpha)}{2}} > \frac{1}{\alpha};\ d \geq 3. \end{aligned}$$

Proof. We start by proving the second identity. Recall that the probability density function (pdf) of the chi-squared distribution with d degrees of freedom is given by

$$f_d(x) = \frac{1}{2^{\frac{d}{2}}\,\Gamma\left(\frac{d}{2}\right)}\, x^{\frac{d}{2}-1}\, e^{\frac{-x}{2}},$$

where Γ denotes the Gamma function. For $d = 2$, this boils down to

$$f_2(x) = \frac{1}{2}\, e^{\frac{-x}{2}},$$

which is the pdf of the exponential distribution with parameter $\lambda = \frac{1}{2}$. Knowing the formula for the cumulative distribution function (cdf) of an exponential distribution, we see that the cdf of the chi-squared distribution with d degrees of freedom is $F_2(x) = 1 - e^{-\frac{1}{2}x}$. Inverting this gives the corresponding quantile function:

$$Q_2(\beta) = -2\,\ln(1-\beta),$$

and identity ii) follows directly from this by plugging in $\beta = 1-\alpha$. From here, it is easy to see that the inequalities i) and iii) hold because the sequence of quantile functions $(Q_d)_{d\in\mathbb{N}}$ obviously is strictly increasing, and therefore, the sequence $(G_d)_{d\in\mathbb{N}}$ defined by $G_d(\alpha) = e^{\frac{Q_d(1-\alpha)}{2}}$ is strictly increasing as well. □

We see that if both sets are written in terms of their respective likelihood ratios (without logarithms), the critical value of a split MLR confidence set with minimum coverage probability $1-\alpha$, $\frac{1}{\alpha}$, is guaranteed to be larger than or equal to the critical value of the corresponding likelihood ratio confidence set, $e^{\frac{Q_d(1-\alpha)}{2}}$, in dimensions 1 and 2. Consequently, if we can show that the likelihood ratio function of the classical set is always larger than or equal to the corresponding function of the split MLR confidence set, it follows that the classical set will always be a subset of the split MLR set. This is the content of the following important theorem.

Theorem 30. *Consider a statistical model $\{\mathbb{P}_\theta : \theta \in \Theta\}$, where the non-empty parameter space Θ is either a subset of $\mathbb{R}$ or $\mathbb{R}^2$, a given sample size n as well as any partition $\{\mathbf{D_0}, \mathbf{D_1}\}$ of the index set $\{1, \dots, n\}$. Suppose now that for this sample size and partition, $\mathbb{P}_{\theta_*}(\operatorname{argmax}_{\vartheta\in\Theta} \mathcal{L}_1(\vartheta) \neq \emptyset) = 1$ for any $\theta_* \in \Theta$. Then, for any $\alpha \in (0,1)$ and any true parameter $\theta_* \in \Theta$, the classical LR confidence set is a subset of the split MLR confidence set with probability 1:*

$$\mathbb{P}_{\theta_*}\big(LRC_n(\alpha) \subseteq C_n^{ML}(\alpha)\big) = 1.$$

Proof. First, rewrite the classical LR confidence set as:

$$LRC_n(\alpha) = \left\{\theta \in \Theta : LR_n(\theta) \leq e^{\frac{Q_d(1-\alpha)}{2}}\right\}$$

Note that the theorem is restricted to dimensions $d = 1$ and $d = 2$. In Lemma 29, we showed that $e^{\frac{Q_1(1-\alpha)}{2}} < \frac{1}{\alpha}$ and $e^{\frac{Q_2(1-\alpha)}{2}} = \frac{1}{\alpha}$ for all $\alpha \in (0,1)$, so it holds in the setting of the theorem that $e^{\frac{Q_d(1-\alpha)}{2}} \leq \frac{1}{\alpha}$. Thus, the result is proved if we can show that $LR_n(\theta) \geq T_n^{\mathrm{ML}}(\theta)$ for all $\theta \in \Theta$ (note that because of our assumption that the set of maximizers of the likelihood based on the data indexed by $\mathbf{D_1}$ is non-empty with probability 1, $T_n^{\mathrm{ML}}(\theta)$ as well as $C_n^{\mathrm{ML}}(\alpha)$ exist with probability 1). This statement is trivial if $\mathcal{L}(\theta) = 0$ because then $LR_n(\theta) = +\infty$. Consequently, we only need to compare the two statistics in the case where $\mathcal{L}(\theta) > 0$, which also implies that $\mathcal{L}_0(\theta), \mathcal{L}_1(\theta) > 0$ because $\mathcal{L}(\theta) = \mathcal{L}_0(\theta) \cdot \mathcal{L}_1(\theta)$. In this situation, we get:

$$\begin{aligned} LR_n(\theta) =& \frac{\sup_{\vartheta\in\Theta} \mathcal{L}(\vartheta)}{\mathcal{L}(\theta)} \geq \frac{\mathcal{L}\left(\hat{\theta}_1^{\mathrm{ML}}\right)}{\mathcal{L}(\theta)} = \frac{\mathcal{L}_0\left(\hat{\theta}_1^{\mathrm{ML}}\right)}{\mathcal{L}_0(\theta)} \frac{\mathcal{L}_1\left(\hat{\theta}_1^{\mathrm{ML}}\right)}{\mathcal{L}_1(\theta)} \\ &\geq \frac{\mathcal{L}_0\left(\hat{\theta}_1^{\mathrm{ML}}\right)}{\mathcal{L}_0(\theta)} = T_n^{\mathrm{ML}}(\theta), \end{aligned} \tag{4.1}$$

where the last inequality follows from $\frac{\mathcal{L}_1\left(\hat{\theta}_1^{\mathrm{ML}}\right)}{\mathcal{L}_1(\theta)} \geq 1$ for all $\theta \in \Theta$ and the last equality follows from the definition of $T_n^{\mathrm{ML}}(\theta)$ in the case where $\mathcal{L}_0(\theta) > 0$. □

We have shown that as soon as the split MLR confidence set exists in a statistical model of dimension 1 or 2, the classical likelihood ratio confidence set will always be a subset

of it. Naturally, the question now arises for which statistical models there is a positive probability that the likelihood ratio confidence set will be a *proper* subset of the split MLR confidence set (so, besides $LRC_n(\alpha) \subseteq C_n^{\mathrm{ML}}(\alpha)$, we have $C_n^{\mathrm{ML}}(\alpha) \setminus LRC_n(\alpha) \neq \emptyset$). The following proposition gives a very weak sufficient condition for this property with respect to the *families* of these confidence sets, $C_n^{\mathrm{ML}} = \{C_n^{\mathrm{ML}}(\alpha) : \alpha \in (0,1)\}$ and $LRC_n = \{LRC_n(\alpha) : \alpha \in (0,1)\}$, in the sense that the property only holds for some, but not necessarily every $\alpha \in (0,1)$.

Proposition 31. *Consider the same setting as in Theorem 30. Furthermore, assume that for the true parameter $\theta_* \in \Theta$, there exists a $\vartheta(\theta_*) \in \Theta$ such that*

$$\mu\big(p_{\theta_*}(x) > p_{\vartheta}(x) > 0\big) > 0, \tag{4.2}$$

where μ, p_{θ_*} and p_ϑ are defined as in Subsection 1.1. Then there exists an $\alpha'(\theta_*) \in (0,1)$ such that $\mathbb{P}_{\theta_*}\big(C_n^{ML}(\alpha') \setminus LRC_n(\alpha') \neq \emptyset\big) > 0$.

Proof. First, observe that condition (4.2) implies that there exists a $\vartheta(\theta_*)$ such that

$$\mathbb{P}_{\theta_*}\big(p_{\theta_*}(X_i) > p_{\vartheta}(X_i) > 0\big) > 0.$$

Next, defining the event $A_{\theta_*,\vartheta} = \bigcap_{i=1,\dots,n}\{p_{\theta_*}(X_i) > p_{\vartheta}(X_i) > 0\}$, we see that $\mathbb{P}_{\theta_*}(A_{\theta_*,\vartheta}) > 0$ as well as $A_{\theta_*,\vartheta} \subseteq \{\mathcal{L}(\theta_*) > \mathcal{L}(\vartheta)\}$ and $A_{\theta_*,\vartheta} \subseteq \{\mathcal{L}_1(\theta_*) > \mathcal{L}_1(\vartheta) > 0\}$, which implies $A_{\theta_*,\vartheta} \subseteq \{\mathcal{L}(\theta_*) > \mathcal{L}(\vartheta), \mathcal{L}_1(\theta_*) > \mathcal{L}_1(\vartheta) > 0\}$. Also, observe that from (4.1), it follows that $\{\mathcal{L}_1(\hat{\theta}_1^{\mathrm{ML}}) > \mathcal{L}_1(\vartheta) > 0\} \subseteq \{LR_n(\vartheta) > T_n^{\mathrm{ML}}(\vartheta)\}$. Consequently, it holds that

$$\begin{aligned}
0 < \mathbb{P}_{\theta_*}(A_{\theta_*,\vartheta}) &\leq \mathbb{P}_{\theta_*}\big(\mathcal{L}(\theta_*) > \mathcal{L}(\vartheta), \mathcal{L}_1(\theta_*) > \mathcal{L}_1(\vartheta) > 0\big) \\
&\leq \mathbb{P}_{\theta_*}\big(\sup_{\theta\in\Theta}\mathcal{L}(\theta) > \mathcal{L}(\vartheta), \mathcal{L}_1(\hat{\theta}_1^{\mathrm{ML}}) > \mathcal{L}_1(\vartheta) > 0\big) \\
&\leq \mathbb{P}_{\theta_*}\big(LR_n(\vartheta) > 1, LR_n(\vartheta) > T_n^{\mathrm{ML}}(\vartheta)\big).
\end{aligned}$$

Now, note that since $\mathbb{P}_{\theta_*}\big(LR_n(\vartheta) > 1, LR_n(\vartheta) > T_n^{\mathrm{ML}}(\vartheta)\big) > 0$, there exists a constant $C > 1$ such that

$$\mathbb{P}_{\theta_*}\big(LR_n(\vartheta) > C > T_n^{\mathrm{ML}}(\vartheta)\big) > 0.$$

To see why this is the case, observe that

$$\{LR_n(\vartheta) > 1, LR_n(\vartheta) > T_n^{\mathrm{ML}}(\vartheta)\} = \bigcup_{q\in\mathbb{Q},\, q>1}\{LR_n(\vartheta) > q > T_n^{\mathrm{ML}}(\vartheta)\}$$

and therefore

$$0 < \mathbb{P}_{\theta_*}\big(LR_n(\vartheta) > 1, LR_n(\vartheta) > T_n^{\mathrm{ML}}(\vartheta)\big) \leq \sum_{q\in\mathbb{Q},q>1}\mathbb{P}_{\theta_*}\big(LR_n(\vartheta) > q > T_n^{\mathrm{ML}}(\vartheta)\big).$$

This guarantees the existence of a rational number $q' > 1$ such that $\mathbb{P}_{\theta_*}\big(LR_n(\vartheta) > q' > T_n^{\mathrm{ML}}(\vartheta)\big) > 0$. Now, we just need to choose $\alpha' = \frac{1}{q'} \in (0,1)$ to get the desired result: Conditional on the event $\big\{LR_n(\vartheta) > q' > T_n^{\mathrm{ML}}(\vartheta)\big\}$, we have that

$$\begin{aligned}\vartheta \notin \Big\{\theta \in \Theta :\ LR_n(\theta) \leq q'\Big\} &= \Big\{\theta \in \Theta :\ LR_n(\theta) \leq \frac{1}{\alpha'}\Big\} \\ &\supseteq \Big\{\theta \in \Theta :\ LR_n(\theta) \leq e^{\frac{Q_d(1-\alpha')}{2}}\Big\} = LRC_n(\alpha'),\end{aligned}$$

because $e^{\frac{Q_d(1-\alpha)}{2}} \leq \frac{1}{\alpha}$ for $d = 1, 2$ and any $\alpha \in (0,1)$. However, since $T_n^{\mathrm{ML}}(\vartheta) \leq q'$ on this event,

$$\vartheta \in \Big\{\theta \in \Theta :\ T_n^{\mathrm{ML}}(\theta) \leq q'\Big\} = \Big\{\theta \in \Theta :\ T_n^{\mathrm{ML}}(\theta) \leq \frac{1}{\alpha'}\Big\} = C_n^{\mathrm{ML}}(\alpha').$$

We arrive at the conclusion that if we choose $\alpha' = \frac{1}{q'}$, then, conditional on the event $\big\{LR_n(\vartheta) > q' > T_n^{\mathrm{ML}}(\vartheta)\big\}$, we have $C_n^{\mathrm{ML}}(\alpha') \setminus LRC_n(\alpha') \supseteq \{\vartheta\}$. Since this event has positive probability under θ_*, we have completed the proof of *ii)*. □

Remark. The condition that there exists a $\vartheta(\theta_*)$ such that $\mu\big(p_{\theta_*}(x) > p_{\vartheta}(x) > 0\big) > 0$ for the true parameter θ_* essentially ensures that inference on θ_* is neither impossible nor trivial with positive probability. Specifically, the condition implies that with positive probability, the likelihood function is neither constant for all parameters in the model (in which case all sensible confidence sets will be equal to the whole parameter space Θ) nor only positive for the true parameter θ_* (in which case all sensible confidence sets will collapse to $\{\theta_*\}$). It is easy to see that if this condition is not satisfied by a model and $\sup_{\theta\in\Theta} \mathcal{L}(\theta) < \infty$ (which ensures that $LRC_n(\alpha)$ is non-empty), then with probability 1 and for all $\alpha \in (0,1)$, we would get either $LRC_n(\alpha) = \Theta$ or $C_n^{\mathrm{ML}}(\alpha) = \{\theta_*\}$. Obviously, the classical LR confidence set cannot be a proper subset of the split MLR confidence set for any $\alpha \in (0,1)$ in this case.

While Proposition 31 describes the minimal condition on a statistical model such that, with positive probability, $LRC_n(\alpha)$ may be a proper subset of $C_n^{\mathrm{ML}}(\alpha)$ for some (data driven) $\alpha \in (0,1)$, it does not tell us anything about the relationship between the two confidence sets for a fixed confidence level $1-\alpha$. Moreover, we cannot use Proposition 31 to conclude that $LRC_n(\alpha)$ has a smaller volume than $C_n^{\mathrm{ML}}(\alpha)$ (almost surely or on average), since even for α' with $C_n^{\mathrm{ML}}(\alpha') \setminus LRC_n(\alpha') \neq \emptyset$, the result does not tell us anything about the Lebesgue measure of this relative complement. The following two theorems give sufficient conditions to ensure that the volume of the classical LR confidence set is almost surely smaller than the one of the split MLR set for those α such that the first set does not already cover the whole parameter space with probability 1. First, Theorem 32 describes such a condition in the one-dimensional case.

Theorem 32. *Consider the same situation as in Theorem 30 with $\Theta \subseteq \mathbb{R}$. Furthermore, assume that for every fixed $x \in \mathcal{X}$, the function $\theta \mapsto p_\theta(x)$ is continuous. Then the*

following statements hold for the Lebesgue measure $\lambda[\cdot]$ on $\mathbb{R}$, any true parameter $\theta_* \in \Theta$ and any $\alpha \in (0,1)$:

i) $\mathbb{P}_{\theta_*}\Big(\lambda\big[\,C_n^{ML}(\alpha)\big] \geq \lambda\big[LRC_n(\alpha)\,\big]\Big) = 1.$

ii) $\mathbb{P}_{\theta_*}\Big(\lambda\big[\,C_n^{ML}(\alpha)\big] > \lambda\big[LRC_n(\alpha)\,\big]\Big) = \mathbb{P}_{\theta_*}\big(LRC_n(\alpha) \neq \Theta\big).$

Proof. Statement *i)* is a direct consequence of Theorem 30, in which we showed that $\mathbb{P}_{\theta_*}\big(LRC_n(\alpha) \subseteq C_n^{\mathrm{ML}}(\alpha)\big) = 1$ for any $\alpha \in (0,1)$ as soon as $\hat{\theta}_1^{\mathrm{ML}}$ (and therefore $C_n^{\mathrm{ML}}(\alpha)$) exists $\mathbb{P}_{\theta_*}$-a.s.

To prove Statement *ii)*, we will show that the events $A = \big\{\lambda\big[\,C_n^{\mathrm{ML}}(\alpha)\big] > \lambda\big[LRC_n(\alpha)\big]\big\}$ and $B = \big\{LRC_n(\alpha) \neq \Theta\big\}$ are equivalent: The fact that $A \subseteq B$ is trivial because if $LRC_n(\alpha) = \Theta$, then $\lambda\big[LRC_n(\alpha)\big] = \lambda\big[\Theta\big] \geq \lambda\big[\,C_n^{\mathrm{ML}}(\alpha)\big]$. To show that $B \subseteq A$, we consider two cases:

- **Case 1**: $\sup_{\vartheta\in\Theta} \mathcal{L}(\vartheta) = \infty$. On this event, it holds that $LRC_n(\alpha) = \emptyset$ for all $\alpha \in (0,1)$ because $LR_n(\theta) = \infty$ for all $\theta \in \Theta$. To prove $B \cap \big\{\sup_{\vartheta\in\Theta} \mathcal{L}(\vartheta) = \infty\big\} = \big\{\sup_{\vartheta\in\Theta} \mathcal{L}(\vartheta) = \infty\big\} \subseteq \big\{\lambda\big[\,C_n^{\mathrm{ML}}(\alpha)\big] > \lambda\big[LRC_n(\alpha)\big]\big\}$, it therefore suffices to show that $\lambda\big[\,C_n^{\mathrm{ML}}(\alpha)\big] > 0$ for any $\alpha \in (0,1)$. To see this, note that there always exists a (data dependent) $\theta' \in \Theta$ such that $T_n^{\mathrm{ML}}(\theta') \leq 1 < \frac{1}{\alpha}$ for all $\alpha \in (0,1)$: If $\mathcal{L}_0(\hat{\theta}_1^{\mathrm{ML}}) > 0$, then $T_n^{\mathrm{ML}}(\hat{\theta}_1^{\mathrm{ML}}) = \frac{\mathcal{L}_0(\hat{\theta}_1^{\mathrm{ML}})}{\mathcal{L}_0(\hat{\theta}_1^{\mathrm{ML}})} = 1$. If $\mathcal{L}_0(\hat{\theta}_1^{\mathrm{ML}}) = 0$, then $T_n^{\mathrm{ML}}(\theta_*) = \frac{\mathcal{L}_0(\hat{\theta}_1^{\mathrm{ML}})}{\mathcal{L}_0(\theta_*)} = 0$. Moreover, $T_n^{\mathrm{ML}}(\theta)$ is a continuous function with respect to $\theta \in \Theta$ ($\mathbb{P}_{\theta_*}$-a.s.) due to the assumption that $\theta \mapsto p_\theta(x)$ is continuous for every fixed $x \in \mathcal{X}$. Because of this and the fact that Θ is connected as well as the implication of $\sup_{\vartheta\in\Theta} \mathcal{L}(\vartheta) = \infty$ that the parameter space cannot be a single point $\{\theta\}$, we can find an interval $[\theta', \theta' + \epsilon(\alpha)]$ with $\epsilon(\alpha) > 0$, such that $T_n^{\mathrm{ML}}(\theta) < \frac{1}{\alpha}$ for any $\theta \in [\theta', \theta' + \epsilon(\alpha)]$. This implies that $\lambda\big[\,C_n^{\mathrm{ML}}(\alpha)\big] \geq \lambda\big[\,[\theta', \theta' + \epsilon(\alpha)]\,\big] = \epsilon(\alpha) > 0$ for any $\alpha \in (0,1)$.

- **Case 2**: $\sup_{\vartheta\in\Theta} \mathcal{L}(\vartheta) < \infty$. Note that conditional on the event $B \cap \{\sup_{\vartheta\in\Theta} \mathcal{L}(\vartheta) < \infty\}$, the two following (data dependent) points in the parameter space exist:

$$\begin{aligned}
&\theta_1 \in \Theta \text{ s.t. } LR_n(\theta_1) > e^{\frac{Q_1(1-\alpha)}{2}} \text{ (because } LRC_n(\alpha) \neq \Theta) \\
&\theta_2 \in \Theta \text{ s.t. } LR_n(\theta_2) \leq e^{\frac{Q_1(1-\alpha)}{2}} \text{ (because } \sup_{\vartheta\in\Theta} \mathcal{L}(\vartheta) < \infty \\
&\text{and therefore, } \forall\, \epsilon > 0\ \exists\, \theta'(\epsilon) \text{ s.t.} \frac{\sup_{\vartheta\in\Theta} \mathcal{L}(\vartheta)}{\mathcal{L}(\theta')} < 1 + \epsilon)
\end{aligned}$$

  In turn, the existence of θ_1 and θ_2 implies the existence of a θ_3 such that $e^{\frac{Q_1(1-\alpha)}{2}} < LR_n(\theta_3) < \frac{1}{\alpha}$ (note that $e^{\frac{Q_1(1-\alpha)}{2}} < \frac{1}{\alpha}$ for all $\alpha \in (0,1)$, as explained in Lemma 29): If $LR_n(\theta_1) < \frac{1}{\alpha}$, set $\theta_3 = \theta_1$; on the other hand, if $LR_n(\theta_1) \geq \frac{1}{\alpha}$, use the

intermediate value theorem (which can be applied because Θ is connected and $LR_n(\theta)$ is almost surely continuous with respect to $\theta \in \Theta$ due to the continuity assumption regarding the densities in the model) to get θ_3. The continuity of $LR_n(\theta)$ and connectedness of Θ then also implies that there exists a $\delta(\alpha) > 0$, such that

$$e^{\frac{Q_1(1-\alpha)}{2}} < LR_n(\theta) < \frac{1}{\alpha}$$

for all $\theta \in [\theta_3 - \delta(\alpha), \theta_3 + \delta(\alpha)]$. As a consequence of these considerations, we have that, conditional on $B \cap \{\sup_{\vartheta\in\Theta} \mathcal{L}(\vartheta) < \infty\}$,

$$\begin{aligned} C_n^{\mathrm{ML}}(\alpha) \setminus LRC_n(\alpha) &= \Big\{\theta \in \Theta :\ T_n^{\mathrm{ML}}(\theta) \leq \frac{1}{\alpha}\Big\} \setminus \Big\{\theta \in \Theta :\ LR_n(\theta) \leq e^{\frac{Q_1(1-\alpha)}{2}}\Big\} \\ &\supseteq \Big\{\theta \in \Theta :\ LR_n(\theta) \leq \frac{1}{\alpha}\Big\} \setminus \Big\{\theta \in \Theta :\ LR_n(\theta) \leq e^{\frac{Q_1(1-\alpha)}{2}}\Big\} \\ &\supseteq\ [\theta_3 - \delta(\alpha), \theta_3 + \delta(\alpha)], \end{aligned}$$

where the first superset relation follows from the fact that $LR_n(\theta) \geq T_n^{\mathrm{ML}}(\theta)$ as soon as $T_n^{\mathrm{ML}}(\theta)$ exists (see (4.1)). Consequently, we get

$$\begin{aligned} \lambda\big[C_n^{\mathrm{ML}}(\alpha)\big]\ -\ \lambda\big[LRC_n(\alpha)\big] &= \lambda\big[C_n^{\mathrm{ML}}(\alpha) \setminus LRC_n(\alpha)\big] \\ &\geq \lambda\big[[\theta_3 - \delta(\alpha), \theta_3 + \delta(\alpha)]\big] = 2\,\delta(\alpha) > 0. \end{aligned}$$

Putting the two cases together, we see that both events $B \cap \big\{ \sup_{\vartheta\in\Theta} \mathcal{L}(\vartheta) = \infty\big\} = \big\{ \sup_{\vartheta\in\Theta} \mathcal{L}(\vartheta) = \infty\big\}$ and $B \cap \big\{ \sup_{\vartheta\in\Theta} \mathcal{L}(\vartheta) < \infty)\big\}$ imply the event $A = \big\{\lambda\big[C_n^{\mathrm{ML}}(\alpha)\big] > \lambda\big[LRC_n(\alpha)\big]\big\}$. Therefore, the proof of Statement *ii)* is finished.

□

Remark. Theorem 32 implies that, if the described conditions on the statistical model hold, then for all $\alpha \in (0,1)$ and $\theta_* \in \Theta$ with $\mathbb{P}_{\theta_*}\big(LRC_n(\alpha) \neq \Theta\big) > 0$, the volume of the split MLR confidence set is almost surely larger than the one of the classical LR set, in the sense that

$$\begin{aligned} &\mathbb{P}_{\theta_*}\Big(\lambda\big[\,C_n^{\mathrm{ML}}(\alpha)\big]\ \geq \lambda\big[LRC_n(\alpha)\,\big]\Big) = 1 \text{ and} \\ &\mathbb{P}_{\theta_*}\Big(\lambda\big[\,C_n^{\mathrm{ML}}(\alpha)\big]\ > \lambda\big[LRC_n(\alpha)\,\big]\Big) > 0. \end{aligned}$$

Next, we will discuss an example of a popular one-dimensional statistical model in which Theorem 32 can be used to show that with probability 1, $\mathbb{P}_{\theta*}\Big(\lambda\big[\,C_n^{\mathrm{ML}}(\alpha)\big]\ > \lambda\big[LRC_n(\alpha)\,\big]\Big) = 1$ for any $\alpha \in (0,1)$.

Example 33. Consider the statistical model $\{\mathbb{P}_\theta : \theta \in (0,1)\}$, where the observations $X_1, \ldots, X_n$ are independent and Bernoulli-distributed with true parameter $\theta_* \in (0,1)$. Suppose that, using some partition $\{\mathbf{D_0}, \mathbf{D_1}\}$ of the index set $\{1, \ldots, n\}$, we construct the split MLR confidence set for θ_* with minimum coverage probability $1-\alpha$, , i.e. we set $\hat{\theta}_1^{(\mathrm{ML})} = \frac{1}{|\mathbf{D_1}|}\sum_{i\in\mathbf{D_1}} X_i$ for the estimator based on the data indexed by $\mathbf{D_1}$. Obviously, the conditions described in Theorem 32 are satisfied for this model, and, for any α, $\theta_* \in (0,1)$, we will have $\mathbb{P}_{\theta_*}\big(LRC_n(\alpha) \neq (0,1)\big) = 1$. Therefore, Theorem 32 tells us that $\mathbb{P}_{\theta_*}\big(\lambda\big[C_n^{\mathrm{ML}}(\alpha)\big] > \lambda\big[LRC_n(\alpha)\big]\big) = 1$ for all α, $\theta_* \in (0,1)$. In particular, the classical LR confidence in this case corresponds to

$$LRC_n(\alpha) = \Big\{\theta \in (0,1) : \theta^S (1-\theta)^{n-S} \geq e^{-\frac{Q_1(1-\alpha)}{2}} \big(\hat{\theta}^{(\mathrm{ML})}\big)^S \big(1-\hat{\theta}^{(\mathrm{ML})}\big)^{n-S}\Big\},$$

where $S = \sum_{i=1}^n X_i$ and $\hat{\theta}^{(\mathrm{ML})} = \frac{S}{n}$, while the split MLR confidence set can be written as

$$C_n^{\mathrm{ML}}(\alpha) = \Big\{\theta \in (0,1) : \theta^{S_0} \big(1-\theta\big)^{|\mathbf{D_0}|-S_0} \geq \alpha \big(\hat{\theta}_1^{(\mathrm{ML})}\big)^{S_0} \big(1-\hat{\theta}_1^{(\mathrm{ML})}\big)^{|\mathbf{D_0}|-S_0}\Big\},$$

where $S_0 = \sum_{i\in\mathbf{D_0}} X_i$. One can show that both $C_n^{\mathrm{ML}}(\alpha)$ and $LRC_n(\alpha)$ are intervals (with probability 1) and consequently, $LRC_n(\alpha)$ always will be a proper subinterval of $C_n^{\mathrm{ML}}(\alpha)$. Note that for both confidence intervals, it is impossible to get an explicit formula for their respective endpoints and that determining whether $LRC_n(\alpha)$ will always be a proper subinterval of $C_n^{\mathrm{ML}}(\alpha)$ in this model is a difficult task without using Theorem 32. Furthermore, because the assumptions of Wilks' theorem hold in this model, we know that, for any $\alpha \in (0,1)$, $LRC_n(\alpha)$ will have asymptotic coverage probability $1-\alpha$ for θ_*. However, in contrast to $C_n^{\mathrm{ML}}(\alpha)$, it will not achieve this coverage probability in finite samples. Therefore, one has to keep in mind that comparing the classical LR interval and the split MLR interval only in terms of their lengths cannot be justified in this situation. For a 'fair' competition, we would need to compare the split MLR confidence interval with another confidence interval that has minimum coverage probability $1-\alpha$ for θ_* in finite samples, such as the Clopper-Pearson interval (see Clopper and Pearson, 1934).

Naturally, we next turn to models in dimension $d = 2$. To get a similar result as in Theorem 32 regarding the relationship between the volumes of the split MLR and the classical LR confidence sets, we first need the following useful insight from real analysis.

Lemma 34. *Let X be a non-empty, open and connected subset of $\mathbb{R}^d$ and suppose that f, g are two continuous functions from X to $\mathbb{R}$ with $f(x) > g(x)$ for all $x \in X$. Assume that for some $c \in \mathbb{R}$, it holds that $\inf f < c < \sup f$, i.e. f is not constant. Then there exists a neighborhood $B_r(x') = \{x \in \mathbb{R}^d : ||x' - x|| < r\}$ (where $x' \in X$ and $r > 0$) such that $g(x) < c < f(x)$ for all $x \in B_r(x')$.*

Proof. Observe that the result is proved if we can show that there exists some point $x' \in X$ such that

$$f(x') > c, \; g(x') < c. \tag{4.3}$$

Since both f and g are continuous and X is open, we can then choose an $r_f > 0$ such that $B_{r_f}(x') \subseteq X$ and $f(x) > c$ for all $x \in B_{r_f}(x')$, as well as an $r_g > 0$ such that $B_{r_g}(x') \subseteq X$ and $g(x) < c$ for all $x \in B_{r_g}(x')$. The claim then follows by setting $r = \min(r_f, r_g)$.

To show that an $x' \in X$ which satisfies (4.3) exists, we first reduce the problem to the one-dimensional case by introducing the concept of a path: Because $\inf f < c < \sup f$, there exist points $x_1, x_2 \in X$ such that $f(x_1) < c < f(x_2)$. Since the set X is open and connected, it is also path-connected, so there exists a path in X from x_1 to x_2, i.e. a continuous function $P : [0,1] \mapsto X$ such that $P(0) = x_1$ and $P(1) = x_2$. We now want to show that there exists an $s \in (0,1)$ such that $f \circ P$ is strictly larger than c on $(s,1]$. To this end, we define the set $A = \{t \in (0,1) : f\big(P(t)\big) = c\}$ as well as $s = \sup A$. Since $f \circ P$ is continuous, the intermediate value theorem guarantees that there exists a point $u \in (0,1)$ such that $f\big(P(u)\big) = c$, which means that A is non-empty. Furthermore, because of continuity, we must have $f\big(P(s)\big) = c$. This implies that $s \in (0,1)$ because $f\big(P(1)\big) = f(x_2) > c$. Assume now that there exists a $t' \in (s,1)$ with $f\big(P(t')\big) \leq c$. Applying the intermediate value theorem again, we would then conclude that there exists a $t'' \in [t',1)$ such that $f\big(P(t'')\big) = c$, but obviously, this would contradict our definition of s being the supremum of the set A. Therefore, it must hold that $f\big(P(t)\big) > c$ for all $t \in (s,1)$.

Moreover, since $g(x) < f(x)$ for all $x \in X$, we know that $g\big(P(s)\big) < f\big(P(s)\big) = c$. Because $g \circ P$ is continuous, too, we know that $g\big(P(t)\big) < c$ for all $t \in [s, s+\delta]$ and $\delta \in (0, 1-s)$ small enough. Therefore, we conclude that for $x \in P\big((s, s+\delta]\big)$, it holds that $f(x) > c$ and $g(x) < c$, so any x' from $P\big((s, s+\delta]\big)$ satisfies (4.3). □

With this result in place, the proof of the following theorem is rather easy. Note that two details change in comparison to dimension $d = 1$ in order to arrive at a similiar statement regarding the volumes of the two confidence sets: The parameter space Θ is not only required to be connected, but also to be open, and the event that implies $\big\{\lambda\big[\,C_n^{\mathrm{ML}}(\alpha)\big] > \lambda\big[LRC_n(\alpha)\,\big]\big\}$ is restricted to $\Big\{\mathcal{L}\big(\hat{\theta}_1^{\mathrm{ML}}\big) < \sup_{\theta\in\Theta}\mathcal{L}(\theta),\, LRC_n(\alpha) \neq \Theta\Big\}$. This restriction is due to the fact that in the proof of Theorem 32, we relied on the result that $e^{\frac{Q_1(1-\alpha)}{2}} < \frac{1}{\alpha}$, but for dimension $d = 2$, we have $e^{\frac{Q_2(1-\alpha)}{2}} = \frac{1}{\alpha}$ for all $\alpha \in (0,1)$. It is easy to see from (4.1) that the event $\Big\{\mathcal{L}\big(\hat{\theta}_1^{\mathrm{ML}}\big) < \sup_{\theta\in\Theta}\mathcal{L}(\theta)\Big\}$ implies $\Big\{LR_n(\theta) > T_n^{\mathrm{ML}}(\theta)\ \forall\ \theta \in \Theta\Big\}$. Thus, we are able to use Lemma 34 with $d = 2$, $X = \Theta$, $f = LR_n$ and $g = T_n^{\mathrm{ML}}$ on this event.

Theorem 35. *Consider the same situation as in Theorem 32, except that we consider a two-dimensional parameter space $\Theta \subseteq \mathbb{R}^2$, which we also require to be open. Then the following statements hold for the Lebesgue measure $\lambda[\cdot]$ on $\mathbb{R}^2$, any true parameter θ_* and any $\alpha \in (0,1)$:*

i) $\mathbb{P}_{\theta_*}\Big(\lambda\big[\,C_n^{ML}(\alpha)\big] \geq \lambda\big[LRC_n(\alpha)\big]\Big) = 1.$

ii) $\mathbb{P}_{\theta_*}\Big(\lambda\big[\,C_n^{ML}(\alpha)\big] \;> \lambda\big[LRC_n(\alpha)\big]\Big) \geq \mathbb{P}_{\theta_*}\Big(\mathcal{L}\big(\hat{\theta}_1^{ML}\big) < \sup_{\theta\in\Theta}\mathcal{L}(\theta),\, LRC_n(\alpha) \neq \Theta\Big)$.

Proof. Statement *i)* again follows directly from Theorem 30.

Regarding Statement *ii)*, we will use a similar strategy as in the proof of Theorem 32 and show that

$$A = \Big\{\mathcal{L}\big(\hat{\theta}_1^{\mathrm{ML}}\big) < \sup_{\theta\in\Theta}\mathcal{L}(\theta),\, LRC_n(\alpha) \neq \Theta\Big\} \subseteq B = \Big\{\lambda\big[\,C_n^{\mathrm{ML}}(\alpha)\big] \;> \lambda\big[LRC_n(\alpha)\big]\Big\}.$$

Also, by the same arguments as in the proof of Theorem 32, we can restrict our attention to the case where $\sup_{\theta\in\Theta}\mathcal{L}(\theta) < \infty$. Again, the continuity assumption in the theorem implies that $LR_n(\theta)$ and $T_n^{\mathrm{ML}}(\theta)$ are continuous functions of θ ($\mathbb{P}_{\theta_*}$-a.s.). Furthermore, note that on the event A and in the case of bounded likelihood, we have

$$\begin{aligned} LR_n(\theta) > T_n^{\mathrm{ML}}(\theta)\ \forall\ \theta\in\Theta \text{ (because } \mathcal{L}\big(\hat{\theta}_1^{\mathrm{ML}}\big) < \sup_{\theta\in\Theta}\mathcal{L}(\theta)\text{)},\\ \sup_{\theta\in\Theta} LR_n(\theta) > e^{\frac{Q_2(1-\alpha)}{2}} = \frac{1}{\alpha} \text{ (because } LRC_n(\alpha)\neq\Theta\text{)},\\ \inf_{\theta\in\Theta} LR_n(\theta) < \frac{1}{\alpha} \text{ (because } \sup_{\theta\in\Theta}\mathcal{L}(\theta) < \infty\text{)}. \end{aligned}$$

To see why $\mathcal{L}\big(\hat{\theta}_1^{\mathrm{ML}}\big) < \sup_{\theta\in\Theta}\mathcal{L}(\theta)$ implies $LR_n(\theta) > T_n^{\mathrm{ML}}(\theta)\ \forall\ \theta\in\Theta$, review the calculations from (4.1): If $\mathcal{L}\big(\hat{\theta}_1^{\mathrm{ML}}\big) < \sup_{\theta\in\Theta}\mathcal{L}(\theta)$, then indeed,

$$LR_n(\theta) \;=\; \frac{\sup_{\vartheta\in\Theta}\mathcal{L}(\vartheta)}{\mathcal{L}(\theta)} \;<\; \frac{\mathcal{L}(\hat{\theta}_1^{\mathrm{ML}})}{\mathcal{L}(\theta)} \;\leq\; T_n^{\mathrm{ML}}(\theta)$$

for any $\theta\in\Theta$. Since all its conditions are fulfilled, we can now use Lemma 34 (with the special case $d=2$) to conclude that, conditional on $A\cap\{\sup_{\theta\in\Theta}\mathcal{L}(\theta)<\infty\}$, there exists a $\theta'\in\Theta$ and an $r(\theta')>0$ such that $LR_n(\theta) > \frac{1}{\alpha}$ and $T_n^{\mathrm{ML}}(\theta) < \frac{1}{\alpha}$ for all $\theta\in B_r(\theta')$, where $B_r(\theta') = \big\{\theta\in\mathbb{R}^2 : ||\theta'-\theta|| < r\big\}$. Consequently, we have that, conditional on the same event,

$$C_n^{\mathrm{ML}}(\alpha)\setminus LRC_n(\alpha) = \Big\{\theta\in\Theta:\ T_n^{\mathrm{ML}}(\theta)\leq\frac{1}{\alpha}\Big\}\cap\Big\{\theta\in\Theta:\ LR_n(\theta)>\frac{1}{\alpha}\Big\} \supseteq\ B_r(\theta'),$$

from which it follows that

$$\lambda\big[C_n^{\mathrm{ML}}(\alpha)\big] - \lambda\big[LRC_n(\alpha)\big] = \lambda\big[C_n^{\mathrm{ML}}(\alpha)\setminus LRC_n(\alpha)\big] \geq \lambda\big[B_r(\theta')\big] > 0.$$

Thus, we have shown that $A\subseteq B$, which concludes the proof of Statement *ii)*. □

Remark. i) Similar to the one-dimensional case, Theorem 35 implies that if for the given $\alpha\in(0,1)$ and $\theta_*\in\Theta$,

$$\mathbb{P}_{\theta_*}\Big(\mathcal{L}\big(\hat{\theta}_1^{\mathrm{ML}}\big) < \sup_{\theta\in\Theta}\mathcal{L}(\theta),\, LRC_n(\alpha)\neq\Theta\Big) > 0,$$

then the volume of the split LR confidence set will be almost surely larger than volume of the classical one.

ii) Note that in most statistical models of interest, $\mathbb{P}_{\theta_*}\big(\mathcal{L}\big(\hat{\theta}_1^{\mathrm{ML}}\big) < \sup_{\theta\in\Theta}\mathcal{L}(\theta)\big) = 1$. For example, if the maximum likelihood estimator based on the whole sample, $\hat{\theta}^{\mathrm{ML}}$, exists and is unique with probability 1, then the event $\big\{\mathcal{L}\big(\hat{\theta}_1^{\mathrm{ML}}\big) < \sup_{\theta\in\Theta}\mathcal{L}(\theta)\big\}$ reduces to $\big\{\hat{\theta}_1^{\mathrm{ML}} \neq \hat{\theta}^{\mathrm{ML}}\big\}$, which typically will have probability 1, too. If $\mathbb{P}_{\theta_*}\big(\mathcal{L}\big(\hat{\theta}_1^{\mathrm{ML}}\big) < \sup_{\theta\in\Theta}\mathcal{L}(\theta)\big) = 1$ holds, then Statement *ii)* from Theorem 35 simplifies to

$$\mathbb{P}_{\theta_*}\big(\lambda\big[C_n^{\mathrm{ML}}(\alpha)\big] > \lambda\big[LRC_n(\alpha)\big]\big) = \mathbb{P}_{\theta_*}\big(LRC_n(\alpha) \neq \Theta\big),$$

which is the same result as in Theorem 32.

Similar to Example 33 (which dealt with a one-dimensional model), we will now show that Theorem 35 can be applied to a widely used two-dimensional statistical model to conclude that, for any confidence level $1-\alpha \in (0,1)$, not only will the classical LR confidence set for the two parameters in this model always be a subset of the split MLR confidence set, but it also will have smaller volume (Lebesgue measure) with probability 1.

Example 36. Consider a Gaussian model with (unknown) true parameters $\mu_* \in \mathbb{R}$ and $\sigma_*^2 \in (0,+\infty)$ and i.i.d. observations $X_1,\ldots,X_n$ from this model, where $n \geq 3$. Suppose we are interested in a simultaneous confidence set for both μ_* and σ_*^2 (in other words, a confidence set for $\theta_* = (\mu_*,\sigma_*^2)^T \in \Theta = \mathbb{R}\times(0,+\infty)$) with nominal (minimum or asymptotically valid) coverage probability $1-\alpha$ for some $\alpha \in (0,1)$. Since this statistical model satisfies the conditions described in Theorem 35 (the densities are continuous with respect to the parameters θ and the two-dimensional parameter space is open and connected) and, for any $\alpha \in (0,1)$ and $\theta_* \in \Theta$

$$\mathbb{P}_{\theta_*}\big(\mathcal{L}\big(\hat{\theta}_1^{\mathrm{ML}}\big) < \sup_{\theta\in\Theta}\mathcal{L}(\theta)\big) = \mathbb{P}_{\theta_*}\big(LRC_n(\alpha) \neq \Theta\big) = 1,$$

where $LRC_n(\alpha)$ is the classical LR confidence set for θ_*, we know that, for any partition $\{\mathbf{D_0},\mathbf{D_1}\}$ of the index set $\{1,\ldots,n\}$ with $|\mathbf{D_1}| \geq 2$ (to ensure that the maximum likelihood estimator based on the data indexed by $\mathbf{D_1}$ exists with probability 1) used for the split MLR confidence set $C_n^{\mathrm{ML}}(\alpha)$,

$$\mathbb{P}_{\theta_*}\Big(\lambda\big[C_n^{\mathrm{ML}}(\alpha)\big] > \lambda\big[LRC_n(\alpha)\big]\Big) = 1$$

for any $\alpha \in (0,1)$ and any $\theta_* \in \Theta$. Because the assumptions of Wilks' theorem hold for this model, we also know that the coverage probability of $LRC_n(\alpha)$ converges to $1-\alpha$ for $n\to\infty$. However, just as in Example 33, the comparison between these two confidence sets is not entirely 'fair' because $LRC_n(\alpha)$ does not achieve minimum coverage probability $1-\alpha$ for θ_* in finite samples (see the results of the simulations in Arnold and Shavelle, 1998), while this of course is true for $C_n^{\mathrm{ML}}(\alpha)$. For a 'fair' competition of confidence sets, we could compare the split MLR set with, for example, the confidence set having *exact* coverage probability $1-\alpha$ for θ_* proposed in Section 3.3 of Chapter VIII from Mood (1950).

4.2 Comparisons in Gaussian models with dimension greater than 2

When we turn to models with dimension $d \geq 3$, it becomes difficult to arrive at conclusions regarding the relation between the classical LR and the split MLR confidence set which are as general as in the previous section, since, for every $\alpha \in (0,1)$, we then have $e^{\frac{Q_d(1-\alpha)}{2}} > \frac{1}{\alpha}$, while $e^{\frac{Q_1(1-\alpha)}{2}} < e^{\frac{Q_2(1-\alpha)}{2}} = \frac{1}{\alpha}$ (see Lemma 29). In other words, we can no longer base our considerations on the fact that the critical value of the split set is larger than the crictical value of the classical set for any confidence level $1-\alpha$. Specifically, this means that for $d \geq 3$, the probability that the classical LR confidence set will be a subset of the split MLR set will be less than 1 in general. Even more, the event that the volume of the split MLR set is smaller than the one of the classical set can have positive probability. Conclusions for a large class of models regarding the relation between the average volumes of the two confidence sets will also be difficult to obtain.

For this reason, we will restrict our following discussion of statistical models with dimension d possibly greater than 2 to the special case of d-dimensional Gaussian models with known covariance matrix. These models are also interesting to consider for comparison because in this case, the classical likelihood ratio set for the true mean parameter μ_* actually has *exact* coverage probability $1-\alpha$. It turns out that in these models, even in dimensions $d \geq 3$, the classical likelihood ratio confidence set will still have a smaller volume than the split MLR set with probability 1 for small α (remember that in dimensions 1 and 2, this statement holds for *all* $\alpha \in (0,1)$), see Proposition 41. The corresponding upper bound for α depends on the dimension d and goes to 0 as $d \to \infty$. On the other hand, the *average* volume (and average diameter, too, although this is not explicitly proved here) of the split MLR confidence set will be larger for *all* $\alpha \in (0,1)$ and $d \in \mathbb{N}$. Moreover, the relative performance of the split MLR set with respect to expected volume actually becomes worse for growing dimension d: Specifically, we will show that $\frac{\mathbb{E}_{\mu_*}\left[\lambda\left(C_n^{\mathrm{ML}}(\alpha)\right)\right]}{\lambda\left(LRC_n(\alpha)\right)} \xrightarrow{d\to+\infty} +\infty$ (note that in this model, $\lambda\left(C_n^{\mathrm{ML}}(\alpha)\right)$ is random, while $\lambda\left(LRC_n(\alpha)\right)$ is not). On a side note, we will find that in the d-dimensional Gaussian model, it is actually possible to minimize the expectation of $\lambda^{\frac{d}{2}}\left(C_n^{\mathrm{ML}}(\alpha)\right)$ with respect to the choice of $|\mathbf{D_0}|$.

As a starting point, we first need to determine a suitable expression of the (random) volume of the split MLR confidence set. This is achieved by the following lemma.

Lemma 37. *Consider the following statistical model: Let $X_1, \ldots, X_n$ (with $n \geq 2$) be i.i.d. d-dimensional random vectors that follow a multivariate normal distribution with unknown mean $\mu_* \in \mathbb{R}^d$ and known positive definite covariance matrix $\Sigma \in \mathbb{R}^{d\times d}$. For any given partition $\mathbf{D_0}$ and $\mathbf{D_1}$ of the index set $\{1,\ldots,n\}$, we write $n_i = |\mathbf{D_i}|$, $i = 0,1$; and define*

$$\hat{\mu}_1 = \frac{1}{n_1} \sum_{i \in \mathbf{D_1}} X_i$$

as the maximum likelihood estimator for μ_* based on the data indexed by $\mathbf{D_1}$ and compute the corresponding split likelihood ratio confidence set for μ_* with minimum coverage probability $1-\alpha$, $C_n^{ML}(\alpha)$, for a given $\alpha \in (0,1)$. Then the (random) Lebesgue measure of this set can be expressed as

$$\lambda\Big(C_n^{ML}(\alpha)\Big) = \frac{\pi^{\frac{d}{2}}}{\Gamma\big(\frac{d}{2}+1\big)} \left(\Big(\frac{2}{n_0}\Big) \ln\Big(\frac{1}{\alpha}\Big) + (\hat{\mu}_0 - \hat{\mu}_1)^T \Sigma^{-1} (\hat{\mu}_0 - \hat{\mu}_1) \right)^{\frac{d}{2}} \sqrt{\det(\Sigma)},$$

where Γ denotes the gamma function.

Proof. We first determine a suitable expression of the split maximum likelihood ratio confidence set. Note that this set can also be written as

$$C_n^{\mathrm{ML}}(\alpha) = \left\{ \mu \in \mathbb{R}^d : \ln\Big(T_n(\mu)\Big) \leq \ln\Big(\frac{1}{\alpha}\Big) \right\}.$$

Since $\mathcal{L}_0(\mu) > 0 \ \ \mathbb{P}_{\mu_*}-$almost surely for all $\mu \in \mathbb{R}^d$, we do not have to deal with the case $\mathcal{L}_0(\mu) = 0$ and the log split MLR statistic turns out to be

$$\begin{aligned}
&\ln\Big(T_n(\mu)\Big) = \ln\Big(\frac{\mathcal{L}_0(\hat{\mu}_1)}{\mathcal{L}_0(\mu)}\Big) = \ln\big(\mathcal{L}_0(\hat{\mu}_1)\big) - \ln\big(\mathcal{L}_0(\mu)\big) \\
&= \Big(-\frac{1}{2}\Big) \Big[\sum_{i \in \mathbf{D_0}} (X_i - \hat{\mu}_1)^T \Sigma^{-1} (X_i - \hat{\mu}_1) - \sum_{i \in \mathbf{D_0}} (X_i - \mu)^T \Sigma^{-1} (X_i - \mu) \Big] \\
&= \Big(-\frac{1}{2}\Big) \Big[\big(\sum_{i \in \mathbf{D_0}} (X_i - \hat{\mu}_1)^T \Sigma^{-1} (X_i - \hat{\mu}_1) \big) + n_0 \, \hat{\mu}_0^T \Sigma^{-1} \hat{\mu}_0 \\
&\qquad - \big(\sum_{i \in \mathbf{D_0}} (X_i - \mu)^T \Sigma^{-1} (X_i - \mu) \big) - n_0 \, \hat{\mu}_0^T \Sigma^{-1} \hat{\mu}_0 \Big] \\
&= \Big(-\frac{1}{2}\Big) \Big[\sum_{i \in \mathbf{D_0}} X_i^T \Sigma^{-1} X_i - 2\hat{\mu}_1^T \Sigma^{-1} \sum_{i \in \mathbf{D_0}} X_i + n_0 \, \hat{\mu}_1^T \Sigma^{-1} \hat{\mu}_1 + n_0 \, \hat{\mu}_0^T \Sigma^{-1} \hat{\mu}_0 \\
&\qquad - \sum_{i \in \mathbf{D_0}} X_i^T \Sigma^{-1} X_i + 2\mu^T \Sigma^{-1} \sum_{i \in \mathbf{D_0}} X_i - n_0 \, \mu^T \Sigma^{-1} \mu - n_0 \, \hat{\mu}_0^T \Sigma^{-1} \hat{\mu}_0 \Big] \\
&= \Big(-\frac{1}{2}\Big) \Big[n_0 \, (\hat{\mu}_1 - \hat{\mu}_0)^T \Sigma^{-1} (\hat{\mu}_1 - \hat{\mu}_0) - n_0 \, (\mu - \hat{\mu}_0)^T \Sigma^{-1} (\mu - \hat{\mu}_0) \Big] \\
&= \Big(\frac{n_0}{2}\Big) \Big[(\hat{\mu}_0 - \mu)^T \Sigma^{-1} (\hat{\mu}_0 - \mu) - (\hat{\mu}_0 - \hat{\mu}_1)^T \Sigma^{-1} (\hat{\mu}_0 - \hat{\mu}_1) \Big].
\end{aligned}$$

Consequently, the split MLR confidence set can be expressed as:

$$C_n^{\mathrm{ML}}(\alpha) = \left\{ \mu \in \mathbb{R}^d : \ \ln\left(T_n(\mu)\right) \leq \ln\left(\frac{1}{\alpha}\right) \right\}$$
$$= \left\{ \mu \in \mathbb{R}^d : \ \left(\frac{n_0}{2}\right) \left[(\hat{\mu}_0 - \mu)^T \Sigma^{-1} (\hat{\mu}_0 - \mu) \ - \ (\hat{\mu}_0 - \hat{\mu}_1)^T \Sigma^{-1} (\hat{\mu}_0 - \hat{\mu}_1) \right] \leq \ln\left(\frac{1}{\alpha}\right) \right\}$$
$$= \left\{ \mu \in \mathbb{R}^d : \ (\hat{\mu}_0 - \mu)^T \Sigma^{-1} (\hat{\mu}_0 - \mu) \ \leq \ \left(\frac{2}{n_0}\right) \ln\left(\frac{1}{\alpha}\right) \ + \ (\hat{\mu}_0 - \hat{\mu}_1)^T \Sigma^{-1} (\hat{\mu}_0 - \hat{\mu}_1) \right\}$$
$$= \left\{ \mu \in \mathbb{R}^d : \ (\hat{\mu}_0 - \mu)^T \left(A\big(n_0, \alpha, \Sigma, \hat{\mu}_0, \hat{\mu}_1\big) \right)^{-1} (\hat{\mu}_0 - \mu) \leq \ 1 \right\},$$

where $A\big(n_0, \alpha, \Sigma, \hat{\mu}_0, \hat{\mu}_1\big) = \left[\left(\frac{2}{n_0}\right) \ln\left(\frac{1}{\alpha}\right) \ + \ (\hat{\mu}_0 - \hat{\mu}_1)^T \Sigma^{-1} (\hat{\mu}_0 - \hat{\mu}_1) \right] \Sigma$ is a positive definite and symmetric $d \times d$-matrix. From the last expression, it is clear that the split MLR confidence set in fact is defined by the interior of a d-dimensional ellipsoid centered at $\hat{\mu}_0$, where the directions of the semi-axes are the eigenvectors of $A\big(n_0, \alpha, \Sigma, \hat{\mu}_0, \hat{\mu}_1\big)$ and the lengths of the semi-axes are given by the square roots of the corresponding eigenvalues (see Section 3.1 in Grötschel et al., 1988). Thus, the (random) Lebesgue measure of $C_n^{\mathrm{ML}}(\alpha, n_0)$ equals the (random) volume of this d-dimensional ellipsoid, which implies (see Grötschel et al., 1988):

$$\lambda\Big(C_n^{\mathrm{ML}}(\alpha, n_0)\Big) = \frac{\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}+1\right)} \sqrt{\det\left(A\big(n_0, \alpha, \Sigma, \hat{\mu}_0, \hat{\mu}_1\big)\right)}$$
$$= \frac{\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}+1\right)} \left(\left(\frac{2}{n_0}\right) \ln\left(\frac{1}{\alpha}\right) \ + \ (\hat{\mu}_0 - \hat{\mu}_1)^T \Sigma^{-1} (\hat{\mu}_0 - \hat{\mu}_1) \right)^{\frac{d}{2}} \sqrt{\det(\Sigma)}. \tag{4.4}$$

□

Following this result, we investigate the in some sense optimal choice for the number of observations to be included in $\mathbf{D_0}$ (which corresponds to the part of the data on which the likelihood function used for the split MLR statistic is based), $n_0 = |\mathbf{D_0}|$, which of course determines $n_1 = |\mathbf{D_1}|$ as well. We will see that the optimal choice for n_0 depends on the given $\alpha \in (0, 1)$ and the dimension d. In the context of this thesis, the result will be used in Theorem 42 to get a lower bound on the volume of the split MLR set for given dimension $d \in \mathbb{N}$ and $\alpha \in (0, 1)$. However, it is also of interest with respect to the question of how one should choose the proportion of observations to place in $\mathbf{D_0}$ in general (see Dunn et al., 2023).

Proposition 38. *Consider the same statistical model as in Lemma 37 and suppose now that we can choose the size of the blocks* $\mathbf{D_0}$ *and* $\mathbf{D_1}$ *that partition the index set* $\{1, \ldots, n\}$

(where $n \geq 2$), $n_0 = |\mathbf{D_0}|$ and $n_1 = |\mathbf{D_1}|$. Define the function $g_d : \{1, \dots, n-1\} \mapsto (0, +\infty)$, which takes the choice for n_0 (which of course determines n_1, too) as input, in the following way:

$$g_d(n_0) = \mathbb{E}_{\mu_*}\Big[\, \lambda^{\frac{2}{d}}\big(\, C_n^{ML}(\alpha, n_0)\,\big)\,\Big],$$

where μ_* denotes the true mean parameter. Then $g_d(n_0)$ is minimized either by $\lfloor r_{opt}\, n \rfloor$ or $\lceil r_{opt}\, n \rceil$, where

$$r_{opt}(\alpha, d) \;=\; \frac{2\,\ln\left(\frac{1}{\alpha}\right) + d \;-\; \sqrt{d\left(2\,\ln\left(\frac{1}{\alpha}\right) + d\right)}}{2\,\ln\left(\frac{1}{\alpha}\right)}.$$

Proof. Following the calculations from Lemma 37, we see that

$$\lambda^{\frac{2}{d}}\big(C_n^{\mathrm{ML}}(\alpha, n_0)\big) = \left(\frac{\pi^{\frac{d}{2}}}{\Gamma\big(\frac{d}{2}+1\big)}\right)^{\frac{2}{d}} \left(\left(\frac{2}{n_0}\right)\,\ln\left(\frac{1}{\alpha}\right) + (\hat{\mu}_0 - \hat{\mu}_1)^T\,\Sigma^{-1}\,(\hat{\mu}_0 - \hat{\mu}_1)\right) \left(\det(\Sigma)\right)^{\frac{1}{d}}$$

and consequently

$$g_d(n_0) = \mathbb{E}_{\mu_*}\Big[\, \lambda^{\frac{2}{d}}\big(\, C_n^{\mathrm{ML}}(\alpha, n_0)\,\big)\,\Big] \propto \left(\frac{2}{n_0}\right)\,\ln\left(\frac{1}{\alpha}\right) + \mathbb{E}_{\mu_*}\Big[\, (\hat{\mu}_0 - \hat{\mu}_1)^T\,\Sigma^{-1}\,(\hat{\mu}_0 - \hat{\mu}_1)\,\Big].$$

Next, note that since Σ^{-1} is symmetric and positive definite, there exists a unique symmetric and positive definite matrix $\Sigma^{-\frac{1}{2}}$ such that $\Sigma^{-1} = \Sigma^{-\frac{1}{2}}\,\Sigma^{-\frac{1}{2}}$. Furthermore, $\Sigma^{-\frac{1}{2}} = \left(\Sigma^{\frac{1}{2}}\right)^{-1}$, where $\Sigma^{\frac{1}{2}}$ is the unique symmetric and positive definite matrix that satisfies $\Sigma = \Sigma^{\frac{1}{2}}\,\Sigma^{\frac{1}{2}}$. Moreover, because $\hat{\mu}_0$ and $\hat{\mu}_1$ are independent, we have $\hat{\mu}_0 - \hat{\mu}_1 \sim N_d\Big(0, \frac{\Sigma}{n_0} + \frac{\Sigma}{n_1}\Big)$ and therefore, by the reproductive property of the multivariate normal distribution, $\big(\frac{1}{n_0} + \frac{1}{n_1}\big)^{-\frac{1}{2}}\,\Sigma^{-\frac{1}{2}}\,(\hat{\mu}_0 - \hat{\mu}_1) \sim N_d\Big(0, I_d\Big)$, where I_d denotes the $d \times d$-identity matrix. It follows that

$$\begin{aligned}
&(\hat{\mu}_0 - \hat{\mu}_1)^T\,\Sigma^{-1}\,(\hat{\mu}_0 - \hat{\mu}_1) \;\sim\; (\hat{\mu}_0 - \hat{\mu}_1)^T\,\Sigma^{-\frac{1}{2}}\,\Sigma^{-\frac{1}{2}}\,(\hat{\mu}_0 - \hat{\mu}_1) \\
\sim &\left(\frac{1}{n_0} + \frac{1}{n_1}\right)\Big(\big(\frac{1}{n_0} + \frac{1}{n_1}\big)^{-\frac{1}{2}}\,\Sigma^{-\frac{1}{2}}\,(\hat{\mu}_0 - \hat{\mu}_1)\Big)^T\Big(\big(\frac{1}{n_0} + \frac{1}{n_1}\big)^{-\frac{1}{2}}\,\Sigma^{-\frac{1}{2}}\,(\hat{\mu}_0 - \hat{\mu}_1)\Big) \\
\sim &\left(\frac{1}{n_0} + \frac{1}{n_1}\right)\chi^2_d
\end{aligned}$$

and consequently

$$\mathbb{E}_{\mu_*}\Big[\, (\hat{\mu}_0 - \hat{\mu}_1)^T\,\Sigma^{-1}\,(\hat{\mu}_0 - \hat{\mu}_1)\,\Big] = \left(\frac{1}{n_0} + \frac{1}{n_1}\right) d,$$

which implies that

$$g_d(n_0) \;\propto\; \left(\frac{2}{n_0}\right)\,\ln\left(\frac{1}{\alpha}\right) + \left(\frac{1}{n_0} + \frac{1}{n_1}\right) d.$$

Instead of minimizing $g_d(n_0)$ directly, we now turn to the related, but simpler problem of finding the minimizer of the function

$$h_d(r) = \left(\frac{2}{r\,n}\right) \ln\left(\frac{1}{\alpha}\right) + \left(\frac{1}{r\,n} + \frac{1}{(1-r)\,n}\right) d \tag{4.5}$$

for $r \in (0,1)$. Here, r can be interpreted as the proportion of the data that is assigned to $\mathbf{D_0}$, which we allow to range freely between 0 and 1, even if it may not lead to an actually feasible split of the dataset (for feasibility, we would need to restrict r to the set $\{\frac{1}{n}, \frac{2}{n}, \ldots, \frac{n-1}{n}\}$). The advantage of dealing with $h_d(r)$ is that it can easily be minimized on the interval $(0,1)$: Checking the first derivative reveals that it has exactly one critical point within $(0,1)$,

$$r_{\text{opt}}(\alpha, d) = \frac{2\,\ln\left(\frac{1}{\alpha}\right) + d - \sqrt{d\left(2\,\ln\left(\frac{1}{\alpha}\right) + d\right)}}{2\,\ln\left(\frac{1}{\alpha}\right)}.$$

Furthermore, checking the second derivative, we see that $h_d(r)$ is strictly convex on $(0,1)$, so r_{opt} is the unique minimizer of this function on this interval. Turning now to feasible options for the split proportion, the strict convexity as well as the existence of a unique minimizer r_{opt} on the interval $(0,1)$ implies that $h_d(r)$ is minimized over $r \in \{\frac{1}{n}, \frac{2}{n}, \ldots, \frac{n-1}{n}\}$ either by $r'_{\text{opt,feas}} = \frac{k}{n}$ or by $r''_{\text{opt,feas}} = \frac{k+1}{n}$, where $k \in \{1, \ldots, n-1\}$ is chosen such that

$$k \leq r_{\text{opt}}\, n < k+1.$$

Observe that $r_{\text{opt}} > 0.5$ for all $\alpha \in (0,1)$, $d \in \mathbb{N}$, and therefore, such a k always exists, but $r''_{\text{opt,feas}}$ is an infeasible option if $k = n-1$. Returning then to the original function $g_d(n_0)$, we see that this function is minimized (over $n_0 \in \{1, \ldots, n-1\}$) either by k or by $k+1$. If $r_{\text{opt}}\, n \notin \mathbb{N}$, then $k = \lfloor r_{\text{opt}}\, n \rfloor$ and $k+1 = \lceil r_{\text{opt}}\, n \rceil$. On the other hand, in the case that $r_{\text{opt}}\, n \in \mathbb{N}$, it is easy to see that the function $h_d(r)$ is minimized by $r'_{\text{opt,feas}} = r_{\text{opt}}$ and therefore, $g_d(n_0)$ in this case is minimized by $k = r_{\text{opt}}\, n = \lfloor r_{\text{opt}}\, n \rfloor = \lceil r_{\text{opt}}\, n \rceil$. This finishes the proof of the claim. □

Remark. As was mentioned in the proof of the theorem, it is easy to see that for the optimal splitting proportion $r_{\text{opt}}(\alpha, d)$, it holds that $r_{\text{opt}}(\alpha, d) \in (\frac{1}{2}, 1)$ for all $d \in \mathbb{N}$ and $\alpha \in (0,1)$. Furthermore, we have $\lim_{d\to\infty} r_{\text{opt}}(\alpha, d) = \frac{1}{2}$ for any given $\alpha \in (0,1)$ and $\lim_{\alpha\to 0} r_{\text{opt}}(\alpha, d) = 1$ as well as $\lim_{\alpha\to 1} r_{\text{opt}}(\alpha, d) = \frac{1}{2}$ for fixed dimension $d \in \mathbb{N}$.

To be able to relate the split MLR confidence set and the classical likelihood ratio set in models of dimension greater than two, we need a more accurate understanding of the relationship between the two critical values of these sets (if they are expressed in the form where the underlying statistics are log-likelihood ratios), $\ln\left(\frac{1}{\alpha}\right)$ and $\frac{Q_d(1-\alpha)}{2}$. Accordingly, the following two lemmas lead to an upper bound for $Q_d(1-\alpha)$ in terms of a strictly increasing function of $\ln\left(\frac{1}{\alpha}\right)$ and the degrees of freedom d. The first lemma is a variation of Lemma 8 from Birgé and Massart (1998).

Lemma 39. *Let X be a real-valued random variable on some probability space $(\Omega, \mathcal{A}, \mathbb{P})$. Suppose that the cumulant generating function $K(t)$ of X exists for $t \in [0, \frac{1}{b})$ and assume that*

$$K_X(t) = \ln\left(\mathbb{E}[e^{tX}]\right) \leq \frac{a\,t^2}{2\,(1-bt)};\ t \in \left[0, \frac{1}{b}\right), \tag{4.6}$$

where a and b are some positive constants. Then, for any $x \geq 0$, the probability distribution of X satisfies

$$\mathbb{P}(X > bx + \sqrt{2ax}) \leq e^{-x}.$$

Proof. We start by applying a Chernoff bound on the distribution of X. For any $\epsilon > 0$, the expression $\mathbb{P}(X > \epsilon)$ is bounded by

$$\begin{aligned}\mathbb{P}(X > \epsilon) &\leq \inf_{t \in [0, \frac{1}{b})} \frac{\mathbb{E}[e^{t\,X}]}{e^{t\,\epsilon}} = \inf_{t \in [0, \frac{1}{b})} \exp\left(K_X(t) - t\,\epsilon\right) \\ &= \exp\Big(\inf_{t \in [0, \frac{1}{b})} K_X(t) - t\,\epsilon\Big) \leq \exp\Big(\inf_{t \in [0, \frac{1}{b})} g(t)\Big),\end{aligned}$$

where $g(t) = \frac{a\,t^2}{2\,(1-bt)} - t\,\epsilon$. Obviously, g is twice-differentiable on the interval $\left[0, \frac{1}{b}\right)$, and checking the first derivative, we find that g has a local extremum at $t^* = \frac{1}{b}\left(1 - \sqrt{\frac{a}{a+2\epsilon\,b}}\right)$. Furthermore, checking the second derivative reveals that g is strictly convex on $\left[0, \frac{1}{b}\right)$, so the infimum (minimum) of g on the interval $\left[0, \frac{1}{b}\right)$ is achieved at t^* and

$$\inf_{t \in [0, \frac{1}{b})} g(t) = g(t^*) = \frac{\sqrt{(a + b\,\epsilon)^2 - (b\,\epsilon)^2} - (a + b\,\epsilon)}{b^2}.$$

Consequently, for any $\epsilon > 0$, we may write

$$\mathbb{P}(X > \epsilon) \leq \exp\big(-h(\epsilon)\big),$$

with $h(\epsilon) = \frac{(a+b\,\epsilon) - \sqrt{(a+b\,\epsilon)^2 - (b\,\epsilon)^2}}{b^2}$. Computing the inverse of this function for positive arguments then gives

$$h^{-1}(x) = bx + \sqrt{2ax},$$

which completes the proof. □

The following bound on the upper α-quantile of a chi-square distribution with d degrees of freedom then follows directly from Lemma 39. This result can also be found in Laurent and Massart (2000) and Inglot (2010).

Lemma 40. *Let $Q_d : (0,1) \mapsto (0, +\infty)$ be the quantile function of the chi-squared distribution with $d \in \mathbb{N}$ degrees of freedom. Then the following inequality holds for any $\alpha \in (0,1)$:*

$$Q_d(1-\alpha) \leq d + 2\,\ln\left(\frac{1}{\alpha}\right) + 2\sqrt{d\,\ln\left(\frac{1}{\alpha}\right)}. \tag{4.7}$$

Proof. We will prove bound (4.7) by relying on Lemma 39. For any $d \in \mathbb{N}$, let $Y_d \sim \chi^2_d - d$. To check that the assumption regarding the cumulant generating function mentioned in this lemma holds for Y_d, first recall the formula for the moment generating function of a chi-squared distribution with d degrees of freedom, which leads to

$$K_{Y_d}(t) = \ln\big(\mathbb{E}[e^{t\,Y_d}]\big) = \ln\big((1-2t)^{-\frac{d}{2}}\,e^{-d\,t}\big) = \Big(-\frac{d}{2}\Big)\,\ln(1-2t) - d\,t$$

for $t \in [0, \frac{1}{2})$. Next, we use the fact that $\ln(1+x) \geq \frac{x}{2}\,\frac{2+x}{1+x}$ for $x \in (-1, 0]$. This can be shown, for example, by noting that the differentiable function $g(x) = \ln(1+x) - \frac{x}{2}\,\frac{2+x}{1+x}$ is 0 at $x = 0$ and has negative derivative on the interval $(-1, 0)$, so $g(x) > 0$ for $x \in (-1, 0)$. The cumulant generating function of Y_d then can be bounded by

$$\begin{aligned} K_{Y_d}(t) &= \Big(-\frac{d}{2}\Big)\,\ln(1-2t) - d\,t \leq \Big(-\frac{d}{2}\Big)\,\frac{-(2t)}{2}\,\frac{2+(-2t)}{1+(-2t)} - d\,t \\ &= d\,t\,\frac{1-t}{1-2t} - d\,t = \frac{d\,t^2}{1-2t} = \frac{2d\,t^2}{2\,(1-2t)} \end{aligned}$$

for $t \in \big[0, \frac{1}{2}\big)$. This bound is of the desired form as in (4.6) with $a = 2d$ and $b = 2$. Thus, Lemma 39 gives

$$\mathbb{P}\big(Y_d > 2x + \sqrt{2\,(2d)\,x}\big) = \mathbb{P}\big(Y_d > 2x + 2\,\sqrt{dx}\big) \leq e^{-x}$$

for any $x > 0$. For $Z_d \sim \chi^2_d$ and any $\alpha \in (0, 1)$, this implies

$$\mathbb{P}\left(Z_d > d + 2\,\ln\Big(\frac{1}{\alpha}\Big) + 2\,\sqrt{d\ln\Big(\frac{1}{\alpha}\Big)}\right) \leq e^{-\ln\left(\frac{1}{\alpha}\right)} = \alpha.$$

This proves the desired upper bound for the quantile function of a chi-squared distribution with d degrees of freedom. □

We will now use this bound on $Q_d(1-\alpha)$ for two results which clarify the relationship between the volumes of the split MLR confidence set and the classical likelihood ratio set in multidimensional Gaussian models. These are especially interesting for dimensions $d \geq 3$, since in dimensions 1 and 2 we already know by Theorems 32 and 35 that

$$\mathbb{P}_{\mu_*}\Big(\lambda\big(C_n^{\mathrm{ML}}(\alpha)\big) > \lambda\big(LRC_n(\alpha)\big)\Big) = 1$$

for any $\alpha \in (0, 1)$ and $\mu_* \in \mathbb{R}^d$, $d = 1, 2$. The following result then reveals that the statement that $LRC_n(\alpha)$ has smaller volume than $C_n^{\mathrm{ML}}(\alpha)$ with probability 1 still holds for small α in dimensions $d \geq 3$. However, the corresponding upper bound on α decreases exponentially with growing dimension d.

Proposition 41. *Consider the statistical model from Lemma 37, as well as a given partition* $\{\mathbf{D_0}, \mathbf{D_1}\}$ *of the index set* $\{1, \ldots, n\}$. *For any* $\alpha \in (0, 1)$, *denote the split*

maximum likelihood ratio confidence set with nominal (minimum) coverage probability $1-\alpha$ *for* $\mu_* \in \mathbb{R}^d$ *as* $C_n^{ML}(\alpha)$ *and the corresponding classical likelihood ratio confidence set with (in this case exact) coverage probability* $1-\alpha$ *as* $LRC_n(\alpha)$*. Then, for any true parameter* μ_**, it holds that*

$$\mathbb{P}_{\mu_*}\Big(\lambda\big(LRC_n(\alpha)\big) < \lambda\big(C_n^{ML}(\alpha)\big)\Big) = 1 \;\forall\; \alpha \in (0, e^{-c^2}),$$

with $c = \frac{\sqrt{d}\left(n_0+\sqrt{n_0\,(2n-n_0)}\right)}{2\,(n-n_0)}$ *and* $n_0 = |\mathbf{D_0}|$.

Proof. It is well-known that in this model, the classical likelihood ratio confidence set with (exact) coverage probability $1-\alpha$ can be written as

$$LRC_n(\alpha) = \Big\{\mu \in \mathbb{R}^d :\; n\,(\hat{\mu}-\mu)^T\,\Sigma^{-1}\,(\hat{\mu}-\mu) \;\leq\; Q_d(1-\alpha)\Big\},$$

where $\hat{\mu} = \frac{1}{n}\sum_{i=1}^n X_i$ is the maximum likelihood estimator for μ_* (based on the whole sample) and $Q_d(\cdot)$ is the quantile function of the chi-squared distribution with $d \in \mathbb{N}$ degrees of freedom. Following the arguments from the proof of Lemma 37, it is easy to see that this confidence set is a d-dimensional ellipsoid, too, however, in this case its randomness is limited to its center, $\hat{\mu}$. Therefore, its (non-random) Lebesgue measure is given by the (non-random) volume of this ellipsoid, which is

$$\lambda\Big(LRC_n(\alpha)\Big) = \frac{\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}+1\right)}\left(\frac{Q_d(1-\alpha)}{n}\right)^{\frac{d}{2}}\sqrt{\det(\Sigma)}.$$

Using the expression for the (random) volume of the split MLR confidence set for μ_* obtained in Lemma 37, it follows that

$$\Big\{\lambda\big(LRC_n(\alpha)\big) < \lambda\big(C_n^{\mathrm{ML}}(\alpha)\big)\Big\} = \Big\{\frac{Q_d(1-\alpha)}{n} \;<\; \left(\frac{2}{n_0}\right)\ln\left(\frac{1}{\alpha}\right) + (\hat{\mu}_0-\hat{\mu}_1)^T\,\Sigma^{-1}\,(\hat{\mu}_0-\hat{\mu}_1)\Big\},$$

where $\hat{\mu}_i$ is the maximum likelihood estimator for μ_* based on the subset of the data indexed by $\mathbf{D_i}$; $i = 0, 1$. Observe that the right-hand side of the inequality in the second event is greater than $\left(\frac{2}{n_0}\right)\ln\left(\frac{1}{\alpha}\right)$ with probability 1 because Σ^{-1} is positive definite. The upper bound for chi-square quantiles (4.7) derived from Lemma 40 together with the fact that $n_0 < n$ then implies that

$$\frac{Q_d(1-\alpha)}{n} \leq \frac{d+2\ln\left(\frac{1}{\alpha}\right)+2\sqrt{d\ln\left(\frac{1}{\alpha}\right)}}{n} < \left(\frac{2}{n_0}\right)\ln\left(\frac{1}{\alpha}\right) \tag{4.8}$$

for $\ln\left(\frac{1}{\alpha}\right)$ large enough. Specifically, solving the second inequality for $x = \sqrt{\ln\left(\frac{1}{\alpha}\right)}$ yields

$$\frac{d+2x^2+2\sqrt{d}x}{n} < \left(\frac{2}{n_0}\right)x^2,\; x>0 \iff x > \frac{\sqrt{d}\left(n_0+\sqrt{n_0\,(2n-n_0)}\right)}{2\,(n-n_0)},$$

and therefore, the second inequality from (4.8) holds iff. $\alpha < e^{-c^2}$ with $c = \frac{\sqrt{d}\left(n_0+\sqrt{n_0\,(2n-n_0)}\right)}{2\,(n-n_0)}$. Consequently, for $\alpha \in (0, e^{-c^2})$, we have

$$\begin{aligned}
&\mathbb{P}_{\mu_*}\Big(\lambda\big(LRC_n(\alpha)\big) < \lambda\big(C_n^{\mathrm{ML}}(\alpha)\big)\Big) \\
&= \mathbb{P}_{\mu_*}\Big(\frac{Q_d(1-\alpha)}{n} < \big(\frac{2}{n_0}\big)\,\ln\big(\frac{1}{\alpha}\big) + (\hat{\mu}_0-\hat{\mu}_1)^T\,\Sigma^{-1}\,(\hat{\mu}_0-\hat{\mu}_1)\Big) \\
&\geq \mathbb{P}_{\mu_*}\Big(\frac{Q_d(1-\alpha)}{n} < \big(\frac{2}{n_0}\big)\,\ln\big(\frac{1}{\alpha}\big)\Big) = 1,
\end{aligned}$$

which proves the desired claim. □

The next theorem then completes our discussion about the relation between the split MLR and the classical LR confidence set in multivariate Gaussian models by showing that the average volume of the classical set will be smaller than the one of the split set for any $\alpha \in (0,1)$. Furthermore, for $d \to \infty$ and fixed sample size $n \in \mathbb{N}$, the ratio between the expectation of the (random) volume of $C_n^{\mathrm{ML}}(\alpha)$ and the (non-random) volume of $LRC_n(\alpha)$ diverges to infinity - in other words the expected volume of $C_n^{\mathrm{ML}}(\alpha)$ is of larger order (with respect to d and for fixed n) than the volume of $LRC_n(\alpha)$.

Theorem 42. *Consider the same situation as in Proposition 41 with $d \geq 2$. It then holds that:*

$$\frac{\mathbb{E}_{\mu_*}\Big[\lambda\big(C_n^{ML}(\alpha)\big)\Big]}{\lambda\big(LRC_n(\alpha)\big)} \geq \left(\frac{2\,\ln\left(\frac{1}{\alpha}\right) + 2\,d + 2\,\sqrt{\left(2\ln\left(\frac{1}{\alpha}\right)+d\right)\,d}}{2\,\ln\left(\frac{1}{\alpha}\right) + d + 2\,\sqrt{d\,\ln\left(\frac{1}{\alpha}\right)}}\right)^{\frac{d}{2}} \xrightarrow{d\to+\infty} +\infty. \tag{4.9}$$

In particular, this implies that $\mathbb{E}_{\mu_*}\Big[\lambda\big(C_n^{ML}(\alpha)\big)\Big] > \lambda\big(LRC_n(\alpha)\big)$ for any $\alpha \in (0,1)$ and $d \in \mathbb{N}$.

Proof. We begin by bounding $\mathbb{E}_{\mu_*}\Big[\lambda\big(C_n^{\mathrm{ML}}(\alpha)\big)\Big]$ from below (note that $\lambda\big(C_n^{\mathrm{ML}}(\alpha)\big)$, in contrast to $\lambda\big(LRC_n(\alpha)\big)$, is random). Following (4.4), we can use Jensen's inequality (since $d \geq 2$) to get:

$$\begin{aligned}
&\mathbb{E}_{\mu_*}\Big[\lambda\big(C_n^{\mathrm{ML}}(\alpha)\big)\Big] \\
&= \mathbb{E}_{\mu_*}\left[\frac{\pi^{\frac{d}{2}}}{\Gamma\big(\frac{d}{2}+1\big)}\left(\big(\frac{2}{n_0}\big)\,\ln\big(\frac{1}{\alpha}\big) + (\hat{\mu}_0-\hat{\mu}_1)^T\,\Sigma^{-1}\,(\hat{\mu}_0-\hat{\mu}_1)\right)^{\frac{d}{2}}\sqrt{\det(\Sigma)}\right] \\
&\geq \frac{\pi^{\frac{d}{2}}}{\Gamma\big(\frac{d}{2}+1\big)}\left(\mathbb{E}_{\mu_*}\left[\big(\frac{2}{n_0}\big)\,\ln\big(\frac{1}{\alpha}\big) + (\hat{\mu}_0-\hat{\mu}_1)^T\,\Sigma^{-1}\,(\hat{\mu}_0-\hat{\mu}_1)\right]\right)^{\frac{d}{2}}\sqrt{\det(\Sigma)} \\
&= \frac{\pi^{\frac{d}{2}}}{\Gamma\big(\frac{d}{2}+1\big)}\Big(h_d\Big(\frac{n_0}{n}\Big)\Big)^{\frac{d}{2}}\sqrt{\det(\Sigma)},
\end{aligned}$$

where $n_0 = |\mathbf{D_0}|$ and $h_d : (0,1) \mapsto (0,+\infty)$ is defined as in (4.5). From Proposition 38, we also know that $h_d(r)$ is minimized on $(0,1)$ by

$$r_{\text{opt}}(\alpha, d) = \frac{2\,\ln\left(\frac{1}{\alpha}\right) + d - \sqrt{d\left(2\,\ln\left(\frac{1}{\alpha}\right) + d\right)}}{2\,\ln\left(\frac{1}{\alpha}\right)}.$$

Plugging this minimizer into h_d yields

$$h_d\Big(\frac{n_0}{n}\Big) \geq h_d\Big(r_{\text{opt}}(\alpha, d)\Big) = \frac{2\,\ln\left(\frac{1}{\alpha}\right) + 2\,d + 2\sqrt{\left(2\ln\left(\frac{1}{\alpha}\right) + d\right)d}}{n}.$$

Therefore, we have the lower bound

$$\mathbb{E}_{\mu_*}\Big[\lambda\big(C_n^{\text{ML}}(\alpha)\big)\Big] \geq \frac{\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}+1\right)} \left(\frac{2\,\ln\left(\frac{1}{\alpha}\right) + 2\,d + 2\sqrt{\left(2\ln\left(\frac{1}{\alpha}\right) + d\right)d}}{n}\right)^{\frac{d}{2}} \sqrt{\det(\Sigma)}. \tag{4.10}$$

On the other hand, we can use the chi-square quantiles bound (4.7) to get an upper bound on $\lambda\Big(LRC_n(\alpha)\Big)$. Indeed, we have

$$\begin{aligned}
\lambda\big(LRC_n(\alpha)\big) &= \frac{\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}+1\right)} \left(\frac{Q_d(1-\alpha)}{n}\right)^{\frac{d}{2}} \sqrt{\det(\Sigma)} \\
&\leq \frac{\pi^{\frac{d}{2}}}{\Gamma\left(\frac{d}{2}+1\right)} \left(\frac{d + 2\,\ln\left(\frac{1}{\alpha}\right) + 2\sqrt{d\,\ln\left(\frac{1}{\alpha}\right)}}{n}\right)^{\frac{d}{2}} \sqrt{\det(\Sigma)}.
\end{aligned} \tag{4.11}$$

Putting the two bounds (4.10) and (4.11) together, we get the desired lower bound on the ratio between the expected volume (Lebesgue measure) of the split MLR confidence set and the (non-random) volume of the classical likelihood ratio set, both with nominal coverage probability $1-\alpha$:

$$\begin{aligned}
\frac{\mathbb{E}\Big[\lambda\big(C_n^{\text{ML}}(\alpha)\big)\Big]}{\lambda\big(LRC_n(\alpha)\big)} &\geq \frac{\left(\frac{2\,\ln\left(\frac{1}{\alpha}\right)+2\,d+2\sqrt{\left(2\ln\left(\frac{1}{\alpha}\right)+d\right)d}}{n}\right)^{\frac{d}{2}}}{\left(\frac{d+2\,\ln\left(\frac{1}{\alpha}\right)+2\sqrt{d\,\ln\left(\frac{1}{\alpha}\right)}}{n}\right)^{\frac{d}{2}}} \\
&= \left(\frac{2\,\ln\left(\frac{1}{\alpha}\right) + 2\,d + 2\sqrt{\left(2\ln\left(\frac{1}{\alpha}\right) + d\right)d}}{2\,\ln\left(\frac{1}{\alpha}\right) + d + 2\sqrt{d\,\ln\left(\frac{1}{\alpha}\right)}}\right)^{\frac{d}{2}}
\end{aligned} \tag{4.12}$$

for $d \geq 2$ and any $\alpha \in (0,1)$. It is easy to see that (4.12) diverges for $d \to +\infty$ and fixed $n \in \mathbb{N}$ because the fraction inside the brackets converges to 4. Moreover, note that (4.12) is greater than 1 for any $\alpha \in (0,1)$. Finally, due to Theorem 32, $\mathbb{E}_{\mu_*}\left[\lambda\big(C_n^{\mathrm{ML}}(\alpha)\big)\right] > \lambda\big(LRC_n(\alpha)\big)$ in dimension $d = 1$ as well, which completes the proof. □

Remark. By the exact same arguments as in the preceeding proof, one can show that, for $r^2\big(A\big)$ being the squared radius (which is the length of the longest semi-axis) of an ellipsoid $A \subseteq \mathbb{R}^d$, it holds that

$$\frac{\mathbb{E}_{\mu_*}\left[r^2\big(C_n^{\mathrm{ML}}(\alpha)\big)\right]}{r^2\big(LRC_n(\alpha)\big)} \geq \frac{2\,\ln\left(\frac{1}{\alpha}\right) + 2\,d + 2\sqrt{\left(2\ln\left(\frac{1}{\alpha}\right) + d\right)d}}{2\,\ln\left(\frac{1}{\alpha}\right) + d + 2\sqrt{d\,\ln\left(\frac{1}{\alpha}\right)}} \xrightarrow{d\to\infty} 4.$$

This is consistent with a result by Dunn et al. (2023), who showed by similar reasoning to ours that, indeed, for $|\mathbf{D_0}| = |\mathbf{D_1}| = \frac{n}{2}$,

$$\frac{\mathbb{E}_{\mu_*}\left[r^2\big(C_n^{\mathrm{ML}}(\alpha)\big)\right]}{r^2\big(LRC_n(\alpha)\big)} \xrightarrow{d\to\infty} 4.$$

At first glance, this might seem like a surprising result, given that we confirmed in Theorem 42 that the expected volume of the split MLR confidence set is actually of larger order with respect to d (for fixed $n \in \mathbb{N}$) than the volume of the classical likelihood ratio set. However, the connection between these two seemingly contradictory results becomes clear once we view the asymptotic relation between the two confidence sets in the following way: In high-dimensional models, the split MLR confidence set is approximately four times larger than the classical likelihood ratio set *in every dimension*. Specifically, every principal axis of the split MLR elliposid will be about 4 times longer than the corresponding principal axis of the classical LR ellipsoid. This implies that, for every extra dimension, the expected volume of $C_n^{\mathrm{ML}}(\alpha)$ is multiplied with a factor of approximately 2 (since the volume of an ellipsoid is proportional to the product of the lengths of its principal semi-axes) compared to the volume of $LRC_n(\alpha)$. Thus, $\mathbb{E}_{\mu_*}\left[\lambda\big(C_n^{\mathrm{ML}}(\alpha)\big)\right]$ will in some sense grow exponentially faster than $\lambda\big(LRC_n(\alpha)\big)$ for $d \to \infty$ and fixed $n \in \mathbb{N}$, though note that both

$$\mathbb{E}_{\mu_*}\left[\lambda\big(C_n^{\mathrm{ML}}(\alpha)\big)\right],\ \lambda\big(LRC_n(\alpha)\big) \xrightarrow{d\to\infty} \infty$$

for fixed n. The implicit conclusion present in both Wasserman et al. (2020) and Dunn et al. (2023), that in high dimensional Gaussian models, the split MLR approach does not lead to significantly larger confidence sets in comparison to the classical likelihood ratio sets, therefore is only valid if 'larger' refers to the (expected) diameter or the (squared) radius of a set, but not if we compare them by their respective volumes. Following the arguments from above, we would expect a similar relationship between the split MLR confidence set and the classical LR set in many high-dimensional statistical models of

interest: The (expected) radii/diameters of the sets will grow at the same rate, but the volume of the split MLR set will grow exponentially faster for increasing dimension d. Furthermore, we can expect a statistical test that uses the split MLR approach (by which we mean that $\hat{\theta}_1$ in Definition 14 is set to be the (unrestricted) maximum likelihood estimator based on the subset of the data indexed by $\mathbf{D_1}$) to become more and more conservative (compared to the classical likelihood ratio test) for high-dimensional null hypothesis sets Θ_0. This is consistent with results for specific testing situations discussed in Strieder and Drton (2022) (testing the hypothesis that the first $d-k$ entries of the unknown mean parameter μ_* in a Gaussian model with known covariance matrix are equal to 0, where $k < d$ is large, see Subsections 3.3 and 3.2 of this thesis), as well as Dunn et al. (2024) (testing the hypothesis that the true density of d-dimensional i.i.d. observations belongs to the class of log-concave densities, see Subsection 3.4 of this thesis). Consequently, from a practical point of view, one should always consider to perform some dimension reduction on the data when applying universal inference methods in high-dimensional statistical models (as described in, for example, Dunn et al., 2024).

5 Conclusion

In this thesis, we have tried to provide a compact overview of *universal inference*, a method for constructing confidence sets and statistical tests that was introduced by Wasserman et al. (2020). We have shown in Chapter 2 that, given $\alpha \in (0,1)$, the resulting confidence sets have minimum coverage probability $1-\alpha$ and the resulting tests have significance level α in any statistical model for which a likelihood function can be defined, irrespective of any regularity requirements on the model or the null hypothesis. In the course of this, we also filled in some mathematical details that, to our knowledge, have not been discussed in the literature so far; for example, we considered the $\frac{0}{0}$-case in the definition of the split likelihood ratio statistic and argued why the conditional expectation of this random variable given the estimator $\hat{\theta}_1$ exists. In the literature review (Chapter 3), we presented numerous extensions of the basic split likelihood ratio method, such as considering several sample splits, introducing randomization to improve the power of the resulting tests or using a reverse information projection approach for the split likelihood ratio test. We also discussed the literature on universal inference in the context of nuisance parameters and on the optimal choice for the splitting ratio. The discussion of the existing literature was completed by three examples which illustrated the application of universal inference in irregular models and hypotheses. The most important result of Chapter 4 (Theorem 30) was that, if the maximum likelihood estimator based on the corresponding part of the data exists and is used for the split likelihood ratio statistic, then the classical likelihood ratio confidence set will always be a subset of the split likelihood ratio set in 1 or 2-dimensional statistical models. Under very weak additional assumptions, we further concluded that the volume of the split likelihood ratio confidence set will be larger than the one of the classical likelihood ratio set with probability 1 (Theorems 32 and 35). If the model dimension is greater than 2, then such almost sure-relations regarding the volumes of the two sets do not hold anymore in general. However, we were able to show for the multivariate Gaussian model that the average volume of the split set is larger than the average volume of the classical set for every dimension d and sample size n, and that in a 'fixed n, increasing d'-setting, the average volume of the split set will grow exponentially faster compared to the one of the classical set (Theorem 42), adding to the ongoing discussion about the curse of dimensionality of universal inference (see Strieder and Drton, 2022; Dunn et al., 2024).

As universal inference is a relatively new method, there are still a lot of interesting open questions about it that have not been answered yet. For example, it is often (at least implicitly) assumed in the literature that the best choice for $\hat{\theta}_1$ (which in theory can be any estimator) is the maximum likelihood estimator, however, to our knowledge, a proof that this choice indeed maximizes the power of the resulting split likelihood ratio test

(possibly under some conditions on the model) is not yet known. Other interesting open questions include whether the average volume of the classical likelihood ratio confidence set will be smaller than the one of the split set in all regular (in the sense that the conditions of Wilks' theorem are satisfied) models and whether in dimensions $d = 1$ and $d = 2$, the classical likelihood ratio set will always be a subset of the *subsampling* likelihood ratio set as well. However, probably the most important task for researchers in universal inference right now is to study its performance in models/hypotheses where the usual general methods for constructing tests and confidence sets (most importantly, the classical likelihood ratio and the Bootstrap approach) fail. This was the main motivation for the proposal of the universal inference approach by Wasserman et al. (2020) in the first place, however, as far as we know, there currently exist only three publications which specifically deal with the application of universal inference in irregular settings. As in general, this method has great potential for inference in irregular models, we hope that the literature on this topic will grow significantly in the coming years.

Bibliography

Aguiar, I., Taylor, D., and Ugander, J. (2024). A tensor factorization model of multilayer network interdependence. *Journal of Machine Learning Research*, 25(282):1–54.

An, M. Y. (1998). Logconcavity versus Logconvexity: A Complete Characterization. *Journal of Economic Theory*, 80(2):350–369.

Arnold, B. C. and Shavelle, R. M. (1998). Joint Confidence Sets for the Mean and Variance of a Normal Distribution. *The American Statistician*, 52(2):133–140.

Bagnoli, M. and Bergstrom, T. (2005). Log-concave probability and its applications. *Economic Theory*, 26(2):445–469.

Billingsley, P. (1995). *Probability and Measure*. John Wiley & Sons, New York, 3rd edition.

Birgé, L. and Massart, P. (1998). Minimum contrast estimators on sieves: exponential bounds and rates of convergence. *Bernoulli*, 4(3):329–375.

Cai, Z., Lei, J., and Roeder, K. (2024). Asymptotic Distribution-Free Independence Test for High-Dimension Data. *Journal of the American Statistical Association*, 119(547):1794–1804.

Casella, G. and Hwang, J. T. (1991). Evaluating Confidence Sets Using Loss Functions. *Statistica Sinica*, 1(1):159–173.

Chen, J. and Li, P. (2009). Hypothesis test for normal mixture models: The EM approach. *The Annals of Statistics*, 37(5A):2523–2542.

Clopper, C. J. and Pearson, E. S. (1934). The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial. *Biometrika*, 26(4):404–413.

Cule, M., Gramacy, R. B., and Samworth, R. (2009). LogConcDEAD: An R Package for Maximum Likelihood Estimation of a Multivariate Log-Concave Density. *Journal of Statistical Software*, 29(2):1–20.

Cule, M., Samworth, R., and Stewart, M. (2010). Maximum Likelihood Estimation of a Multi-Dimensional Log-Concave Density. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 72(5):545–607.

Dalmasso, N., Masserano, L., Zhao, D., Izbicki, R., and Lee, A. B. (2024). Likelihood-free frequentist inference: bridging classical statistics and machine learning for reliable simulator-based inference. *Electronic Journal of Statistics*, 18(2):5045–5090.

Drton, M. (2009). Likelihood ratio tests and singularities. *The Annals of Statistics*, 37(2):979–1012.

Dunn, R., Gangrade, A., Wasserman, L., and Ramdas, A. (2024). Universal Inference Meets Random Projections: A Scalable Test for Log-Concavity. *Journal of Computational and Graphical Statistics*, 34(1):1–13.

Dunn, R., Ramdas, A., Balakrishnan, S., and Wasserman, L. (2023). Gaussian universal likelihood ratio testing. *Biometrika*, 110(2):319–337.

Durrett, R. (2019). *Probability: Theory and Examples*. Cambridge University Press, Cambridge, 5th edition.

Dümbgen, L. and Rufibach, K. (2011). logcondens: Computations Related to Univariate Log-Concave Density Estimation. *Journal of Statistical Software*, 39(6):1–28.

Ferguson, T. S. (1967). *Mathematical Statistics: A Decision Theoretic Approach*. Academic Press, New York, 1st edition.

Grötschel, M., Lovasz, L., and Schrijver, A. (1988). *Geometric Algorithms and Combinatorial Optimization*. Springer, Berlin, Heidelberg, 1st edition.

Grünwald, P. (2024a). Beyond Neyman–Pearson: E-values enable hypothesis testing with a data-driven alpha. *Proceedings of the National Academy of Sciences*, 121(39):e2302098121.

Grünwald, P. (2024b). Safe testing. *Journal of the Royal Statistical Society Series B: Statistical methodology*, 86(5):1091–1128.

Hall, P. and Stewart, M. (2005). Theoretical analysis of power in a two-component normal mixture model. *Journal of Statistical Planning and Inference*, 134(1):158–179.

Hao, Y. and Grünwald, P. (2024). E-Values for Exponential Families: the General Case. *arXiv preprint arXiv:2404.19465*.

Ignatiadis, N. and Wager, S. (2022). Confidence Intervals for Nonparametric Empirical Bayes Analysis. *Journal of the American Statistical Association*, 117(539):1149–1166.

Inglot, T. (2010). Inequalities for quantiles of the chi-square distribution. *Probability and Mathematical Statistics*, 30(2):339–351.

Joshi, V. M. (1966). Admissibility of Confidence Intervals. *The Annals of Mathematical Statistics*, 37(3):629–638.

Kim, I. and Ramdas, A. (2024). Dimension-agnostic inference using cross U-statistics. *Bernoulli*, 30(1):683–711.

Lardy, T., Grünwald, P., and Harremoës, P. (2024). Reverse Information Projections and Optimal E-Statistics. *IEEE Transactions on Information Theory*, 70(11):7616–7631.

Laurent, B. and Massart, P. (2000). Adaptive Estimation of a Quadratic Functional by Model Selection. *The Annals of Statistics*, 28(5):1302–1338.

Lauritzen, S., Uhler, C., and Zwiernik, P. (2021). Total positivity in exponential families with application to binary variables. *The Annals of Statistics*, 49(3):1436–1459.

Lehmann, E. and Romano, J. P. (2005). *Testing Statistical Hypotheses*. Springer Science+Business Media, New York, 3rd edition.

Li, Q. J. (1999). *Estimation of mixture models*. phd, Yale University.

Liu, X. and Shao, Y. (2004). Asymptotics for the likelihood ratio test in a two-component normal mixture model. *Journal of Statistical Planning and Inference*, 123(1):61–81.

Lundborg, A. R., Kim, I., Shah, R. D., and Samworth, R. J. (2024). The projected covariance measure for assumption-lean variable significance testing. *The Annals of Statistics*, 52(6):2851–2878.

McLachlan, G. J. (1987). On Bootstrapping the Likelihood Ratio Test Stastistic for the Number of Components in a Normal Mixture. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 36(3):318–324.

McLachlan, G. J. and Krishnan, T. (2008). *The EM Algorithm and Extensions*. John Wiley & Sons, Hoboken, 2nd edition.

McLachlan, G. J. and Rathnayake, S. (2014). On the number of components in a Gaussian mixture model. *WIREs Data Mining and Knowledge Discovery*, 4(5):341–355.

Mood, A. M. (1950). *Introduction to the theory of statistics*. McGraw-Hill, New York, 1st edition.

Nguyen, H. D. and Gupta, M. (2023). Finite sample inference for empirical Bayesian methods. *Scandinavian Journal of Statistics*, 50(4):1616–1640.

Ramdas, A., Grünwald, P., Vovk, V., and Shafer, G. (2023). Game-Theoretic Statistics and Safe Anytime-Valid Inference. *Statistical Science*, 38(4):576–601.

Ramdas, A. and Manole, T. (2023). Randomized and Exchangeable Improvements of Markov's, Chebyshev's and Chernoff's Inequalities. *arXiv preprint arXiv:2304.02611*.

Ramdas, A., Ruf, J., Larsson, M., and Koolen, W. M. (2022). Testing exchangeability: Fork-convexity, supermartingales and e-processes. *International Journal of Approximate Reasoning*, 141:83–109.

Robins, J. and Van der Vaart, A. (2006). Adaptive nonparametric confidence sets. *The Annals of Statistics*, 34(1):229–253.

Shi, H. and Drton, M. (2024). On universal inference in Gaussian mixture models. *arXiv preprint arXiv:2407.19361*.

Strieder, D. and Drton, M. (2022). On the choice of the splitting ratio for the split likelihood ratio test. *Electronic Journal of Statistics*, 16(2):6631–6650.

Tse, T. and Davison, A. C. (2022). A note on universal inference. *Stat*, 11(1):e501.

Van der Vaart, A. (2007). *Asymptotic statistics*. Cambridge University Press, Cambridge, 8th edition.

Vovk, V. and Wang, R. (2021). E-values: Calibration, combination and applications. *The Annals of Statistics*, 49(3):1736–1754.

Wasserman, L., Ramdas, A., and Balakrishnan, S. (2020). Universal inference. *Proceedings of the National Academy of Sciences*, 117(29):16880–16890.

Waudby-Smith, I. and Ramdas, A. (2024). Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1):1–27.

Zhang, Y. and Shao, X. (2024). Another look at bandwidth-free inference: a sample splitting approach. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1):246–272.