Under Submission

E-STATISTICS, GROUP INVARIANCE AND ANYTIME VALID TESTING

By Muriel Felipe Pérez-Ortiz[1,a], Tyron Lardy[2,c] Rianne de Heide[3,d] and Peter D. Grünwald[1,b]

[1]Centrum Wiskunde & Informatica, Amsterdam, The Netherlands, [a]muriel.perez@cwi.nl; [b]pdg@cwi.nl

[2]Leiden University, Leiden, The Netherlands, [c]t.d.lardy@math.leidenuniv.nl

[3]Vrije Universiteit, Amsterdam, The Netherlands, [d]r.de.heide@vu.nl

We study worst-case growth-rate optimal (GROW) e-statistics for hypothesis testing between two group models. If the underlying group G acts freely on the observation space, there exists a maximally invariant statistic of the data. We show that among all e-statistics, invariant or not, the likelihood ratio of the maximally invariant is GROW and that an anytime valid test can be based on this likelihood ratio. By virtue of a representation theorem of Wijsman, it is equivalent to a Bayes factor with a right Haar prior on G. Such Bayes factors are known to have good frequentist and Bayesian properties. We show that reductions through sufficiency and invariance can be made in tandem without affecting optimality. A crucial assumption on the group G is its amenability, a well-known group-theoretical condition, which holds for general scale- and location families as well as finite-dimensional linear regression.

1. Introduction. Classically, hypothesis tests for group-invariant situations have been studied in great detail both for fixed-sample-size and sequential experiments (Cox, 1952; Hall, Wijsman and Ghosh, 1965; Eaton, 1989; Lehmann and Romano, 2005). Nevertheless, due to methodological concerns about combining evidence from multiple experiments using classical methods, a theory of testing based on e-statistics has been developed (Vovk and Wang, 2021; Grünwald, de Heide and Koolen, 2020; Ramdas et al., 2020). (e-statistics are more commonly known as e-variables or, in analogy to p-values, e-values; we call them e-statistics here to emphasize that they are, in fact, statistics of the data) The main concern that is successfully addressed by testing with e-statistics is that of error control in two common situations: when experiments are optionally stopped, and when aggregating the evidence of interdependent experiments that may themselves have been optionally stopped. The first of these situations is often referred to as anytime validity; the second, as optional continuation. As a contribution to this line of work, we characterize optimal e-statistics in group-invariant situations. As we will see, such situations include testing under linear-model and Gaussian assumptions. We focus on testing procedures that are simultaneously anytime valid and allow for optional continuation.

We concern ourselves with testing composite hypotheses where both null and alternative models remain unchanged under a group of transformations. In particular, we study the situation where the parameter of interest is a function $\delta = \delta(\theta)$ of the model parameter θ that is invariant under such transformations. For example, in the Gaussian case, the coefficient of variation is invariant under scale changes; the correlation coefficient, under affine transformations; and the variance of the principal components, under rotations around the origin. Roughly speaking, by replacing the data $X^n = (X_1, \ldots, X_n)$ by an invariant function $M_n = m_n(X^n)$, one discards all information that is not relevant to the parameter δ. Through the lens of the invariance-reduced data M_n, the hypotheses about the parameter of interest δ may simplify. This reduction principle has been used successfully to obtain sequential tests for composite hypotheses (Hall, Wijsman and Ghosh, 1965).

In this article we characterize e-statistics that are growth rate optimal in the worst case (GROW), as defined by Grünwald, de Heide and Koolen (2020) (see Section 1.2 for definitions), and use them for sequential testing. The main result of this article is the following: under regularity conditions, when the test about the invariance-reduced data M_n becomes a simple-vs.-simple test, the GROW e-statistic is the likelihood ratio statistic for M_n. Our main result covers the case in which the null expresses that the parameter of interest is equal to some fixed δ_0, and the alternative expresses that it is equal to some fixed δ_1; later on in the paper we extend this to settings with sets Δ_0 and Δ_1, including the case in which prior distributions on these sets are available. In proving the main result, we use Theorem 1 of (Grünwald, de Heide and Koolen, 2020), which shows an equivalence between finding a GROW e-statistic and performing a joint minimization of the Kullback-Leibler (KL) divergence between the convex hulls of both null and alternative sets of distributions. The main technical contribution of this article is showing that the value of this joint minimization problem is the KL divergence between the distributions of a maximally invariant function M_n, a function that, informally, looses as little information as possible about the invariant component of the data. One of the key assumptions for this result to hold is the amenability of the group G, a well known group-theoretical condition (Bondar and Milnes, 1981). This condition plays a fundamental role in the celebrated theorem of Hunt and Stein (Lehmann and Romano, 2005, Section 8.5), that relates tests that are max-min optimal for statistical power to group invariant tests. The concepts of power and GROW are to some mild extent related (Grünwald, de Heide and Koolen, 2020, Section 7): one may view GROW as the analogue of power in an optional continuation setting, in which a direct optimization of power leads to useless tests. Thus, there is an analogy between Hunt-Stein and our results, but the proof techniques that are required for our result, and that we develop, are quite different (see Section 1.3). We further investigate e-statistics that are relatively GROW (abbreviated to REGROW by Grünwald, de Heide and Koolen (2020)), a closely related optimality criterion (see Section 1.2). We show that, as opposed to the general case, any GROW e-statistic is also relatively GROW in the group-invariant situation. If data are gathered sequentially, we show that the GROW e-statistic, the likelihood ratio for a maximally invariant function of the data, can be used for anytime valid sequential testing; the sequence of optimal e-statistics becomes a test martingale, i.e. a nonnegative martingale with starting value 1, the mathematical object that forms the basis for anytime-valid testing (Shafer, 2021; Grünwald, de Heide and Koolen, 2020). With an eye towards aggregating evidence from potentially randomly stopped experiments, we describe when the optionally stopped optimal e-statistic remains an e-statistic, and show how data can be further reduced if a sufficient statistic for the invariant parameter is available. For the latter purpose, a result of C. Stein, reported by Hall, Wijsman and Ghosh (1965), is instrumental.

The rest of this introduction is organized in the following manner. In Section 1.1, we introduce formally our setup for hypothesis testing under group invariance. There, in Example 1.1, we show how the t-test fits in our setup. In Section 1.2, we define e-statistics, our main objects of study, and define our optimality criteria. In Section 1.3, we give an informal exposition of our main result, Corollary 4.3, about the characterization of optimal e-statistics for group-invariant situations, and Theorem 4.2, the main technical contribution of this article, on joint KL divergence minimization. There, in a continuation of Example 1.1, we explain the consequences for the t-test. In Section 1.4 we make a brief recount of the main motivations for the use of e-statistics for testing. In Section 1.5 we highlight previous work made in group-invariant testing and in Section 1.6 we introduce notation. Finally, in Section 1.7 we outline the rest of the article. Happy reading.

1.1. *Group invariance.* In this section we describe the group-invariant hypotheses that are of our current interest. In Example 1.1, as our guiding example, we show how the t-test fits in this framework. More precisely, assume that a group G acts freely on both the observation space $\mathcal{X}$ and the parameter space Θ. Denote the action of G on $\mathcal{X}$ by $(g, X) \mapsto gX$ for $g \in G$ and $X \in \mathcal{X}$. For samples of size n, we extend the action of G on $\mathcal{X}$ to $\mathcal{X}^n$ componentwise, that is, by $(g, X^n) \mapsto gX^n := (gX_1, \ldots, gX_n)$ for $g \in G$ and $X^n \in \mathcal{X}^n$. By invariance of a probabilistic model $\mathcal{P} = \{\mathbf{P}_\theta : \theta \in \Theta\}$ on $\mathcal{X}^n$ we understand that, for any $g \in G$ and measurable $B \subseteq \mathcal{X}^n$ and parameter $\theta \in \Theta$, the distribution $\mathbf{P}_\theta$ satisfies

$$\mathbf{P}_\theta\{X^n \in B\} = \mathbf{P}_{g\theta}\{X^n \in gB\}, \tag{1}$$

where $gB = \{gb : b \in B\}$ is the left translate of the set B by g. In particular, we study situations where the parameter of interest $\delta = \delta(\theta)$ indexes the orbits in the parameter space Θ under the action of G. More formally, we assume that δ is a maximally invariant function of the parameter θ, meaning that, for any pair $\theta, \theta' \in \Theta$, there exists g such that $g\theta = \theta'$ any time that $\delta(\theta) = \delta(\theta')$. In that case, we say that δ is a maximally invariant parameter. We are prepared to state the main statistical hypothesis testing problem of interest for this work. For two possible values δ_1, δ_0 of δ, we consider the composite vs. composite testing problem

$$\mathcal{H}_0 : \delta(\theta) = \delta_0 \ \text{ vs. } \ \mathcal{H}_1 : \delta(\theta) = \delta_1. \tag{2}$$

As is known, many classical parametric problems can be cast in this shape. Let us call maximally invariant any G-invariant function $M_n = m_n(X^n)$ that indexes the orbits of the action of G on $\mathcal{X}^n$. The distribution of M_n depends on θ only through the maximal invariant parameter δ, and, under this reduction, the problem (2) becomes simple. It is with the optimality of this reduction that we are concerned. In Section 8, we study cases in which, even after the invariance reduction, the problem under study remains composite. Before defining *e*-statistics, our main objects of study, we spell out how the t-test fits in this framework.

EXAMPLE 1.1 (t-test under Gaussian assumptions). Consider an iid sample $X^n = (X_1, \ldots, X_n)$ of size n from an unknown Gaussian distribution $N(\mu, \sigma)$, and testing whether $\mu/\sigma = \delta_0$ or $\mu/\sigma = \delta_1$. The parameter space Θ consists of all pairs $\mu \in \mathbb{R}$ and $\sigma \in \mathbb{R}^+$ and the Gaussian model is invariant under scale transformations. The group $G = (\mathbb{R}^+, \cdot\,)$ of positive real numbers with multiplication acts on Θ by $(c, (\mu, \sigma)) \mapsto (c\mu, c\sigma)$ for each $c \in \mathbb{R}^+$ and $(\mu, \sigma) \in \Theta$. The parameter of interest is the ratio $\delta = \mu/\sigma$ between the mean μ and the standard deviation σ. The parameter δ is scale-invariant and indexes the orbits of the action of G on Θ. The group G acts on the observation space $\mathcal{X} = \mathbb{R}^n$ by coordinatewise multiplication. A maximally invariant statistic is $M_n = m_n(X^n) = (X_1/|X_1|, \ldots, X_n/|X_1|)$, and its distribution only depends on the maximally invariant parameter $\delta = \mu/\sigma$.

1.2. *The family of e-statistics.* We now define *e*-statistics, our measure of evidence for the alternative over the null hypothesis. Given two subsets Θ_0, Θ_1 of the parameter space Θ, interpreted as the null and an alternative hypothesis, the family of *e*-statistics comprises all nonnegative functions of the data $X^n \in \mathcal{X}^n$ whose expected value is bounded by one under all elements of the null (Grünwald, de Heide and Koolen, 2020), that is, all statistics $T_n(X^n) \geq 0$ such that

$$\sup_{\theta_0 \in \Theta_0} \mathbf{E}_{\theta_0}[T_n(X^n)] \leq 1. \tag{3}$$

The main implication of this definition is that the test $\mathbb{1}\{T_n(X^n) \geq 1/\alpha\}$ has type-I error smaller than α and that the evidence, measured with *e*-statistics from multiple experiments, can be easily aggregated (see Section 1.4). When the null is simple (Θ_0 is a singleton), *e*-statistics coincide with likelihood ratios and Bayes factors; when the null is composite, some

Bayes factors are still e-statistics but usually they are not (Grünwald, de Heide and Koolen, 2020). Shafer (2021) explains and emphasizes their transparent betting interpretation, and simply calls them *bets*.

The traditional optimality criterion for hypothesis tests satisfying a Type-I error guarantee is their fixed-sample size or fixed-stopping rule worst-case power maximization. This criterion cannot be used in a context with optional stopping, since the definition of power requires that we know the stopping rule in advance, contradicting the idea of 'optional stopping' (Grünwald, de Heide and Koolen, 2020). Instead, we concern ourselves with e-statistics that are growth-rate optimal in the worst case (GROW), as defined by Grünwald, de Heide and Koolen (2020). Should it exist, an e-statistic T_n^* is GROW if it maximizes the worst-case expected logarithmic value under the alternative hypothesis, that is, if it maximizes

$$T_n \mapsto \inf_{\theta_1 \in \Theta_1} \mathbf{E}_{\theta_1}[\ln T_n(X^n)] \tag{4}$$

over all e-statistics. The idea is that, under the alternative, one wants to gather evidence as fast as possible, so it makes sense to maximize expectation of $f(T_n(X^n))$ under the alternative, for some increasing function f. Shafer (2019); Grünwald, de Heide and Koolen (2020) extensively argue why it makes sense to take f as the logarithm, an idea also known as *Kelly betting* (Kelly, 1956). Given the worst-case nature of this criterion, Grünwald, de Heide and Koolen (2020) explain that the GROW e-statistic is too conservative in some scenarios and cannot be used if the alternative can be arbitrarily close the null (in the t-test example, this would mean that the effect size under the alternative is unknown). As a response to this issue, they propose to instead maximize a relative form of (4) to obtain less conservative e-statistics outside the worst-case regime (see also Turner, Ly and Grünwald (2021) who, in their contingency table setting, achieve excellent results in practice with this relative criterion, but not with the absolute criterion). With this in mind, we say that an e-statistic T_n^* is relatively GROW if it maximizes the gain in expected logarithmic value relative to an oracle that is given the particular distribution in the alternative hypothesis from which data are generated, that is, if T_n^* maximizes, over all e-statistics,

$$T_n \mapsto \inf_{\theta_1 \in \Theta_1} \left\{ \mathbf{E}_{\theta_1}[\ln T_n(X^n)] - \sup_{T_n' \ e\text{-stat.}} \mathbf{E}_{\theta_1}[\ln T_n'(X^n)] \right\}. \tag{5}$$

As we we will see and contrary to the general case, in our group-invariant setting, any GROW e-statistic is also relatively GROW, so we can avoid a discussion which of the two is more appropriate. While acknowledging that there may be situations in which an e-statistic optimality property distinct from GROW is more relevant, in the remainder of this paper we will simply take the goal of achieving (relative) GROW for granted, without further motivation. With this in mind, we now turn to an informal discussion of the main results of this article.

1.3. *Main results.* We now informally outline the main results of this article. The main result of this article is Corollary 4.3, a characterization of the GROW e-statistic for the group-invariant problem defined in (2). This corollary is a consequence of Theorem 4.2, our main technical contribution, which will be described in Section 1.3.1. We now describe them informally. Recall that once data are reduced through a maximally invariant function $M_n = m_n(X^n)$ for the action of G on $\mathcal{X}^n$, the testing problem (2) becomes simple. We extend our results to situations when the invariance-reduced problem is still composite in Section 8. Sidestepping technicalities, the main result is the following theorem.

COROLLARY 1.2 (Informal statement of Corollary 4.3). *Under a number of technical conditions, among all possible e-statistics, G-invariant or not, the likelihood ratio $T_n^* = p_{\delta_1}^{M_n}/p_{\delta_0}^{M_n}$ for any maximally invariant function $M_n = m_n(X^n)$ is GROW for (2).*

We show further in Proposition 4.4 that, in our group-invariant situation, any GROW e-statistic is also relatively GROW, as defined in Section 1.2. With this theorem at hand, we characterize optimal e-statistics for group-invariant situations in fixed-sample experiments. We now turn our attention to sequential experiments, where data $X_1, X_2, \dots$ are gathered one by one. Here, a sequential test is a sequence of zero-one-valued statistics $\xi = (\xi_n)_{n\in\mathbb{N}}$ adapted to the natural filtration generated by $X_1, X_2, \dots$. We consider the test defined by $\xi_n = \mathbb{1}\{T_n^* \geq 1/\alpha\}$ for some value α, whose anytime validity we prove. Additionally, we show that, for certain stopping times $N \leq \infty$, the optionally stopped e-statistic T_N^* remains an e-statistic, which validates its use for for optional continuation.

PROPOSITION 1.3. *Let $T^* = (T_n^*)_{n\in\mathbb{N}}$, where, for each n, T_n^* is the likelihood ratio for the maximal invariants $M_n = m_n(X^n)$ for the action of G on $\mathcal{X}^n$. Let $\xi = (\xi_n)_{n\in\mathbb{N}}$ be the sequential test given by $\xi_n = \mathbb{1}\{T_n^* \geq 1/\alpha\}$. Then the following hold:*

1. *The sequential test ξ is anytime valid at level α, that is,*

$$\sup_{\theta_0\in\Theta_0} \mathbf{P}_{\theta_0}\{\xi_n = 1 \textit{ for some } n \in \mathbb{N}\} \leq \alpha.$$

2. *Suppose that $N \leq \infty$ is a stopping time with respect to $M = (M_n)_{n\in\mathbb{N}}$. Then the optionally stopped e-statistic T_N^* is also an e-statistic, that is,*

$$\sup_{\theta_0\in\Theta_0} \mathbf{E}_{\theta_0}[T_N^*(X^N)] \leq 1. \tag{6}$$

The proof of this result is, by now, standard; we perform it in Section 6. The main ingredient is showing, using the ideas of Hall, Wijsman and Ghosh (1965), that the process $T^* = (T_n^*)_{n\in\mathbb{N}}$ is a nonnegative martingale with respect to the sequence of maximally invariants and that is has expected value 1. An inequality of Ville (1939) and standard optional stopping theorems give the desired results (see the work of Ramdas et al. (2020) for more details). In regards to item 2 in Proposition 1.3, it is natural to ask whether (6) also holds for stopping times that are adapted to the full data $(X^n)_n$ but not to $(M_n)_n$ can be allowed (in our t-test example, this could be a stopping time N^* such as '$N^* := 1$ if $|X_1| \notin [a, b]$; $N^* = 2$ otherwise' for some $0 < a < b$). The answer is negative: in Appendix B, we show that, for appropriate choice of a, b, this N^* provides a counterexample. This means that such non-adapted N^* cannot be used in an optional continuation context, a subtle point explained by Grünwald, de Heide and Koolen (2020, Section 5).

Before turning to our main technical contributions, we anticipate further results and show their implication to our guiding example, the t-test. In Section 5, we utilize the invariance and sufficiency reductions of Hall, Wijsman and Ghosh (1965) to conclude that monitoring the likelihood ratio for $M_1, M_2, \dots$ is equivalent to monitoring the likelihood ratio of a sufficient statistic for the maximally invariant parameter δ (see Proposition 5.2). In Section 7 we show two applications to testing under multivariate Gaussian assumptions: testing whether the population mean zero, and testing whether a linear regression coefficient is zero. In Section 8 we show further extend Corollary 1.2 to cases where the null and alternative hypotheses are still composite even after an invariance reduction of the data (see Proposition 8.1).

1.3.1. *Technical contributions.* From a technical point of view, our main contribution is Theorem 4.2, a computation the infimum value of the Kullback-Leibler (KL) divergence between elements in the convex hulls of the null and alternative models in (2). In Section 2, we show in detail how our approach operates in the simpler case when G is finite or compact.

The main contribution in this article then, is the extension of this result to a large class of noncompact groups for which almost right-invariant probability measures exist. The existence of such measures on G is known as amenability (Bondar and Milnes, 1981), and it is the key assumption in our results. The amenability condition, as will be stated Definition 2.1, is the same that is used in the classical theorem of Hunt and Stein (Lehmann and Romano, 2005, Section 8.5), and, as we will see, it guarantees the existence of almost-right-invariant priors on the group G. The proof techniques that are needed for the results of this work are, however, distinct. Hunt-Stein's theorem shows that, when looking for a test that is max-min optimal in the sense of power, it is enough to look among group-invariant tests. At the core of the proof of the Hunt-Stein theorem lays the fact that the power is a linear function of the test under consideration. In its proof, an approximate symmetrization of the test is carried using almost-right-invariant priors without affecting power guarantees. This line of reasoning cannot be directly translated to our setting because of the nonlinearity of the objective function that characterizes GROW e-statistics.

Besides the main technical contribution Theorem 4.2, additional novel mathematical results are in Proposition 4.4, relating GROW to relative GROW, and the propositions in Section 8, extending Theorem 4.2 to settings in which $\mathcal{H}_0$ and $\mathcal{H}_1$ refer to composite sets of δ's and may be equpied with a prior ditribution for these δ's.

1.4. *Motivation.* Testing based on e-statistics addresses methodological concerns about testing procedures based on p-values (Royall, 1997; Wagenmakers, 2007; Benjamin et al., 2018; Grünwald, 2022). The main concerns that are addressed by using e-statistics for testing is that of sequentially aggregating the evidence of possibly nonindependent experiments while keeping type-I error guarantees (Wang and Ramdas, 2020; Vovk and Wang, 2020). This is referred to as optimal continuation, for the decision to perform a new study at all may depend on the outcomes of the previous one. For example, in medical research, the decision to perform a new trial usually depends on already existing data. The intricate dependencies within trials might be very hard or impossible to model and make it almost impossible to provide Type-I error guarantees for existing meta-analysis methods (that often implicitly presume their independence (Ter Schure and Grünwald, 2019)). Therefore, the ability to aggregate evidence with error control under such dependencies, as is achieved with e-statistics, becomes crucial. Relatedly, e-statistics also offer advantages over p-values in some multiple testing settings. Consequently, a wide interest in e-statistics has kindled in recent years — as a small sample we mention (Shafer, 2021; Henzi and Ziegel, 2021; Ramdas et al., 2021; Wang and Ramdas, 2022; Ren and Barber, 2022). This article is a contribution to this growing body of work. The technical properties that allow for type-I error control under optional continuation are the following two:

1. The type-I error of the test that rejects the null hypothesis anytime that $T(X) \geq 1/\alpha$ is smaller than α, a direct consequence of Markov's inequality and the definition of e-statistic.
2. If X_1 and X_2 are the outcomes of two experiments and $X_2 \mapsto T_2(X_2)$ is an e-statistic conditionally on the value of the e-statistic $X_1 \mapsto T_1(X_1)$, then $T(X_1, X_2) = T_1(X_1)T_2(X_2)$ is also an e-statistic.

Hence, the test $\mathbb{1}\{T(X_1, X_2) \geq 1/\alpha\}$ based on the aggregated e-statistic $T(X_1, X_2) = T_1(X_1)T_2(X_2)$ still has type-I error guarantees. The extension to more outcomes is straightforward. It is in this sense that testing based on e-statistics allows for optional continuation.

1.5. *Previous work.* Invariance, as data-reduction method, has a long tradition in statistics (Eaton, 1989). Perhaps the closest result to the ones we present is the classical theorem of Hunt and Stein (see Lehmann and Romano, 2005, Section 8.5). It establishes that, in group-invariant models like the ones we treat here, there is no loss in considering only group-invariant tests when searching for most powerful tests. The relation of data reductions based in invariance and sufficiency are well understood (Hall, Wijsman and Ghosh, 1965). In the Bayesian literature, group-invariant inference with right Haar priors has been thoroughly studied (Dawid, Stone and Zidek, 1973; Berger, Pericchi and Varshavsky, 1998). It has been shown that, in contrast to some other improper priors, inference based on right Haar priors yields admissible procedures in a decision-theoretical sense (Eaton and Sudderth, 2002, 1999) However, there have also been concerns in the Bayesian literature (Sun and Berger, 2007; Berger and Sun, 2008) that in some situations, the right Haar prior is not uniquely defined, and different choices lead to different conclusions. Interestingly, as we discuss in Section 9, in our setting this issue cannot arise. Finally, we mention (Liang and Barron, 2004) who provide exact minimax procedures for predictive density estimation for general location- and scale-families under Kullback-Leibler loss. Although there are clearly some similarities, the precise minimax result they prove is quite different; we provide a more detailed comparison in Section 9.

1.6. *Notation.* We used the notation that has been most convenient for our understanding. We use linear functional notation for integrals: instead of $\int f(x)\mathrm{d}\mathbf{P}(x)$, we write $\mathbf{P}[f(X)]$ or $\mathbf{P}[f]$. When it is important to specify the variable of integration we write $\mathbf{P}^x[f(x,y)]$ instead of $\int f(x,y)\mathrm{d}\mathbf{P}(x)$. We use lower case symbols in superscript for variables that are being integrated over. For the probability that $X^n \in B$ under $\mathbf{P}$ we write $\mathbf{P}\{X^n \in B\}$. We use $\mathbf{P}$ to refer to to the distribution of X^n according to $\mathbf{P}$. For a measurable function $T = T(X^n)$, we write $\mathbf{P}^T$ for the image measure of $\mathbf{P}$ under T, that is, $\mathbf{P}^T\{T \in B\} = \mathbf{P}\{T(X) \in B\}$. When writing conditional expectations, we write $\mathbf{P}[f(X)|Y]$ instead of $\mathbf{E}_{\mathbf{P}}[f(X)|Y]$, and write $\mathbf{P}_Y^X$ for the conditional distribution of X given Y. We only deal with situations where such conditional distributions exist. For a prior distribution $\mathbf{\Pi}^\theta$ on some parameter space Θ, we write $\mathbf{\Pi}^\theta\mathbf{P}_\theta^X$ for the marginal distribution that assigns probability $\mathbf{\Pi}^\theta\mathbf{P}_\theta^X\{X \in B\} = \int \mathbf{P}_\theta\{X \in B\}\mathrm{d}\mathbf{\Pi}(\theta)$ to any measurable set B. For the posterior distribution of θ given X we write $\mathbf{\Pi}_X^\theta$. Given two subsets H, K of a group G we write $HK = \{hk : h \in H, k \in K\}$ for the set of all possible products between an element of H and an element of K. Similarly, for an element $g \in G$ and a subset K of G, we define $gK = \{gk : k \in K\}$, the translation of K by g, and $K^{-1} = \{k^{-1} : k \in K\}$, the set of inverses of K. We say that K is symmetric if $K = K^{-1}$.

1.7. *Outline.* The rest of this article is structured as follows. We begin by describing our approach for finite and compact groups in Section 2. There, we also describe the challenges that are encountered when dealing with general groups and introduce the main group-theoretical condition, amenability. Next, in Section 3, we lay down formally the conditions necessary for our main results. In Section 4, we state the main results of this article in full. We continue in Section 5 by discussing our approach in the presence of a sufficient statistic for the models under consideration. We show, under regularity conditions, that there is no loss in further reducing the data through a sufficient statistic. With regards to anytime-valid testing, the subject of Section 6 is to show Proposition 1.3. In Section 7 we apply our results to two examples: one about testing whether a linear regression coefficient is zero, and other about testing whether a multivariate population mean is zero under Gaussian assumptions. In Section 8 we extend our results to cases in which, even after an invariance reduction of the data, the hypotheses at hand remain composite. We end this article with Section 9, where we discuss our results; and Section 10, where we give the proofs omitted from the rest of the text.

2. Technical outline. This section shows our techniques in the simple case when the group G in question is finite, and is intended to delineate our approach. Next, we describe how we generalize the result to noncompact amenable groups, and point at the difficulties that are found. Consider again the problem described in (2). Using that the action of the group on the parameter space is free, we can reparametrize each orbit in Θ/G with G. Indeed, we can pick an arbitrary but fixed element in the orbit $\theta_0 \in \delta_0$ and, for any other element $\theta \in \delta_0$, we can identify θ with the group element $g(\theta)$ that transports θ_0 to θ, that is, such that $g(\theta)\theta_0 = \theta$. Hence, with a slight abuse of notation, we can identify $\theta \in \delta_0$ with $g = g(\theta) \in G$ and identify $\mathbf{P}_\theta = \mathbf{P}_{g(\theta)\theta_0}$ with $\mathbf{P}_g$. With analogous definitions, for a fixed $\theta_1 \in \delta_1$, the same identification can carried in the alternative model by an analogous choice θ_1. In order to make notation more succinct, we use $\mathcal{Q} = \{\mathbf{Q}_g\}_{g\in G}$ to denote the alternative hypothesis to $\mathcal{P} = \{\mathbf{P}_g\}_{g\in G}$. We assume that each member of $\mathcal{Q}$ is absolutely continuous with respect to each member of $\mathcal{P}$. With these remarks at hand, the starting problem (2) can be rewritten in the form

$$\mathcal{H}_0 : X^n \sim \mathbf{P}_g, \text{ for some } g \in G, \text{ vs. } \mathcal{H}_1 : X^n \sim \mathbf{Q}_g, \text{ for some } g \in G. \tag{7}$$

As will follow from our discussion, our results are insensitive to the choices of θ_0 and θ_1. Using the invariance of the models, we show in Proposition 4.4 that, in our setting, an e-statistic is GROW if and only if it is relatively GROW (see Section 1.2 for definitions).

2.1. *Finite groups.* Start by assuming that G is a finite group. For instance, a group of permutations. Then, the representation theorem of Wijsman states that, if $M_n = m_n(X^n)$ is a maximally invariant function of X^n, the distribution of M_n can be computed by averaging over the group. Since M_n is G-invariant, then its distribution does not depend on g. We call $\mathbf{P}^n$ and $\mathbf{Q}^{M_n}$ the distributions of M_n under of the respective members of $\mathcal{P}$ and $\mathcal{Q}$, and p^{M_n} and q^{M_n} their respective densities. Then, the so far hypothesized GROW e-statistic T_n^*, the likelihood ratio for the maximal invariant $M_n = m_n(X^n)$, satisfies

$$T_n^*(X^n) = \frac{q^{M_n}(m_n(X^n))}{p^{M_n}(m_n(X^n))} = \frac{\frac{1}{|G|}\sum_{g\in G} q_g(X^n)}{\frac{1}{|G|}\sum_{g\in G} p_g(X^n)}. \tag{8}$$

For finite parameter spaces, as in our current case, Theorem 1 of Grünwald, de Heide and Koolen (2020) takes a simple form: the value of the max-min problem that defines a GROW e-statistic coincides with that of a KL minimization problem, that is,

$$\max_{T_n\ e\text{-stat.}} \min_{g\in G} \mathbf{Q}_g[\ln T_n(X^n)] = \min_{\mathbf{\Pi}_0,\mathbf{\Pi}_1} \mathrm{KL}(\mathbf{\Pi}_1^{g_1}\mathbf{Q}_{g_1}, \mathbf{\Pi}_0^{g_0}\mathbf{P}_{g_0}), \tag{9}$$

where $\mathrm{KL}(\mathbf{Q},\mathbf{P}) = \mathbf{Q}[\ln(q/p)]$ is the KL divergence, and the minimum on the right hand side is taken over all pairs of distributions on the group G. An application of the information processing inequality implies that, for any pair of probability distributions on G,

$$\mathrm{KL}(\mathbf{\Pi}_1^{g_1}\mathbf{Q}_{g_1}, \mathbf{\Pi}_0^{g_0}\mathbf{P}_{g_0}) \geq \mathrm{KL}(\mathbf{Q}^{M_n}, \mathbf{P}^{M_n}) = \min_{g\in G} \mathbf{Q}_g[\ln T_n^*(X^n)],$$

where the last equality follows from the fact that T_n^* from (8) only depends on X^n through the invariant $M_n = m_n(X^n)$ and consequently its distribution is independent of $g \in G$. Thus, (8) shows that the minimum KL of the right hand side of (9) is achieved for the particular choice of two uniform priors on G. Consequently, T_n^*, defined in (8), is a GROW e-statistic, that is,

$$\min_{g\in G} \mathbf{Q}_g[\ln T_n^*(X^n)] = \max_{T_n\ e\text{-stat.}} \min_{g\in G} \mathbf{Q}_g[\ln T_n(X^n)]. \tag{10}$$

We now turn to the challenges encountered when dealing with more complicated groups.

2.2. *Noncompact groups.* As we will see, a similar reasoning to that of the previous section can be carried out for compact groups. In this section we show the difficulties that we run into when considering noncompact groups, and how we circumvent them under the assumption that the group G is amenable. Anytime that G is a locally compact topological group, there exist left and right invariant measures λ and ρ, respectively, on G (see Bourbaki, 2004, VII, §1,n° 2). This means that, for any $g \in G$ and any $B \subseteq G$ measurable, $\lambda\{gB\} = \lambda\{B\}$ and $\rho\{Bg\} = \rho\{B\}$. We refer to them as left and right Haar measures; they will take the place that the uniform distribution took on finite groups. For simplicity of exposition, let us assume that both probabilistic models are dominated by a left invariant measure ν on $\mathcal{X}$. In that case, the invariance assumption (1) implies that the densities w.r.t. ν take the form $p_g(X^n) = p_1(g^{-1}X^n)$ and $q_g(X^n) = q_1(g^{-1}X^n)$, where 1 makes reference to the unit element of the group G. The theorem of Wijsman (Andersson, 1982) implies that, in analogy to (8), under regularity assumptions, the likelihood ratio for the maximal invariant $M_n = m_n(X^n)$ can be computed by integration over the group G, that is,

$$T_n^*(X^n) = \frac{q^{M_n}(m_n(X^n))}{p^{M_n}(m_n(X^n))} = \frac{\int_G q_g(X^n)\mathrm{d}\rho(g)}{\int_G p_g(X^n)\mathrm{d}\rho(g)}. \tag{11}$$

If the right Haar measure ρ can be chosen to be a probability measure, we can carry out the analogous computations to those made in the finite case of Section 2.1 to conclude that T_n^* is indeed GROW. However, it can be shown that the right Haar measure ρ is finite if and only if the group G at hand is compact (see Reiter and Stegeman, 2000, Proposition 3.3.5). This is a severe limitation; it would not even cover our guiding example, the t-test (see Example 1.1), because the group $(\mathbb{R}^+, \cdot\,)$ is not compact. The main technical contribution of this article is the extension of this result to noncompact *amenable* groups, defined next, for which there exist *almost right invariant* probability measures.

DEFINITION 2.1 (Amenability). A group G is amenable if there exists a sequence of *almost right-invariant* probability distributions, that is, a sequence $\mathbf{\Pi}_1, \mathbf{\Pi}_2, \ldots$ such that, for any measurable set $B \subseteq G$ and $g \in G$,

$$\lim_{k\to\infty} |\mathbf{\Pi}_k\{B\} - \mathbf{\Pi}_k\{Bg\}| = 0. \tag{12}$$

Amenable groups have been thoroughly studied (Paterson, 2000) and include, among others, all finite, compact, commutative, and solvable groups. An example of a nonamenable group is the free group in two elements and any group containing it. A prominent example of a nonamenable group is that of invertible $d \times d$ matrices with matrix multiplication. Under the amenability of G and our assumptions, we will show that, for the almost right-invariant sequence of probability distributions on G,

$$\lim_{k\to\infty} \mathrm{KL}(\mathbf{\Pi}_k^g \mathbf{Q}_g, \mathbf{\Pi}_k^g \mathbf{P}_g) = \mathrm{KL}(\mathbf{Q}^{M_n}, \mathbf{P}^{M_n}) = \min_{g\in G} \mathbf{Q}_g[\ln T_n^*(X^n)], \tag{13}$$

where the last equality follows from the fact that $T_n^*(X^n)$ depends on X^n only through the maximal invariant M_n and, consequently, its distribution does not depend on g. From this, via Theorem 1 of Grünwald, de Heide and Koolen (2020), the analogue of (10) holds and, consequently, as in the finite case of Section 2.1, T^* from (11) is GROW.

EXAMPLE 1.1 (continued). The group $G = (\mathbb{R}^+, \cdot)$ of the t-test setting is clearly amenable. The right Haar measure ρ is given by $\rho(\sigma) = 1/\sigma$, and the rightmost expression of (11) becomes, with $\bar{X} := n^{-1}\sum_{i=1}^n X_i$,

$$T_n^*(X^n) = \frac{\int_{\sigma>0} \frac{1}{\sigma} \exp\left(-\frac{n}{2}\left[\left(\bar{X}\sigma - \delta_1\right)^2 + \left[\frac{n^{-1}\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2}\right)\right]\right) d\sigma}{\int_{\sigma>0} \frac{1}{\sigma} \exp\left(-\frac{n}{2}\left[\left(\bar{X}\sigma - \delta_0\right)^2 + \left[\frac{n^{-1}\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2}\right)\right]\right) d\sigma}. \tag{14}$$

The expression (14) goes back to Cox (1952) who realized that it was equivalent to the likelihood ratio of the maximal invariant. Lai (1976) already used it in an anytime-valid context (essentially exploiting that it gives an e-statistic). Our results establish, for the first time, that (14) is also GROW and relatively GROW. Lai also considered putting a proper prior distribution on δ_1; the same is done in *Jeffreys' Bayesian t-test* (Jeffreys, 1961; Rouder et al., 2009). We return to this idea in Section 8.

Consider now the sufficient statistic $s_n(X^n) = (\hat{\mu}_n, \hat{\sigma}_n)$, where $\hat{\mu}_n$ is the maximum likelihood estimator for the mean μ; and $\hat{\sigma}_n$, for the standard deviation σ. The t-statistic $M_{\mathcal{S},n} = m_{\mathcal{S},n}(X^n) \propto \hat{\mu}_n/\hat{\sigma}_n$ is a maximally invariant function of the sufficient statistic. Our results imply that T_n^* also equals the likelihood ratio for $M_{\mathcal{S},n}$ is also relatively GROW, and that the test $\xi = (\xi_n)_{n\in\mathbb{N}}$ given by $\xi_n = \mathbb{1}\{T_n^* \geq 1/\alpha\}$ satisfies the conclusions of Proposition 1.3.

3. Assumptions. In this section we describe the assumptions made in our main results, their part in the proofs, and discuss their role for the purpose of parametric inference. We gather all assumptions below, in Assumption 1, for ease of reference. We start by laying out the assumptions on the spaces involved, followed by those on the probabilistic models under scrutiny. Our additional assumptions on the group G, the parameter space Θ and the observation space $\mathcal{X}$ are topological in nature. They have two purposes. The first, in relation to the discussion of Section 2, is to ensure that the representation theorem of Wijsman (Andersson, 1982) holds. The second purpose of our assumptions is to ensure that the observation space $\mathcal{X}^n$ can be put in bijective and bimeasurable[1] correspondence with a subset of $G \times \mathcal{X}^n/G$, where the group G acts naturally by multiplication on the first component. To this end, a theorem of Bondar (1976) is instrumental (see Remark 3.2). We assume that G is a topological group, that is, a group equipped with a topology whose operation is continuous. We assume that all topological spaces under consideration are equipped with their Borel σ-algebra, the one generated by their topology. As topological spaces, we assume that both G and $\mathcal{X}$ are separable, completely metrizable, and locally compact. We assume that the action of G on $\mathcal{X}^n$ is continuous and proper; the latter means that the map $G \times \mathcal{X}^n \to \mathcal{X}^n \times \mathcal{X}^n$ defined by

$$(g, x^n) \mapsto (gx^n, x^n)$$

is proper, that is, the inverse of compact sets is compact. Properness ensures that the induced topology on the orbits $\mathcal{X}^n/G$ is Hausdorff, locally compact, and σ-finite (see Andersson, 1982). We further assume that both probabilistic models are dominated by a common relatively left-invariant measure μ on $\mathcal{X}^n$ with some multiplier χ, that is, a measure μ such that, for any measurable set $B \subseteq \mathcal{X}^n$ and any group element $g \in G$, satisfies $\mu\{gB\} = \chi(g)\mu\{B\}$. We gather these assumptions for ease of reference.

ASSUMPTION 1. Let G be a topological group acting on $\mathcal{X}^n$, a topological space. The group G, the observation space $\mathcal{X}^n$, and the probabilistic models under consideration satisfy the following properties:

1. As topological spaces, G and $\mathcal{X}^n$ are separable, complete, metrizable and locally compact.
2. The action of G on $\mathcal{X}^n$ is free, continuous and proper.
3. The models $\{\mathbf{P}_g\}_{g\in G}$ and $\{\mathbf{Q}_g\}_{g\in G}$ are invariant and have densities with respect to a common measure μ on $\mathcal{X}^n$ that is relatively left-invariant with some multiplier χ.

[1] We call an invertible map bimeasurable if both the map and its inverse are measurable.

REMARK 3.1. Assumption 1 holds in most cases of interest for the purpose of parametric inference. We summarize some situations in which Assumption 1 holds which will be helpful in Section 7, where we apply our results in two examples. Let $\mathcal{X} = \mathbb{R}^d$ and identify $\mathcal{X}^n$ with set of $d \times n$ matrices. The properness of the following two actions on $\mathcal{X}^n$ are consequences of the more general results of Wijsman (1985).

1. The linear group in d dimensions $\mathrm{GL}(d)$, consisting of all $d \times d$ invertible real matrices with multiplication, acts continuously on $\mathcal{X}^n$ by left matrix multiplication. The continuous action of $\mathrm{GL}(d)$ on the restriction of $\mathcal{X}^n$ to matrices of rank d is free and proper any time that $n \geq d$. Seen as a subset of $\mathbb{R}^{d\times n}$, the restriction of other Lebesgue measure to $\mathcal{X}^n$ is relatively left invariant with multiplier $\chi(g) = |\det(g)|^n$, for $g \in \mathrm{GL}(d)$.
2. The affine linear group $\mathrm{AL}(d)$, all pairs (A, b) with $A \in \mathrm{GL}(d)$ and $b \in \mathbb{R}^d$ with group operation $(A, v)(B, u) = (AB, Au + v)$, also acts continuously on $\mathcal{X}^n$. An action is given by $((A, b), X^n) \mapsto [Ax_1 + b, \ldots, Ax_n + b]$, where $x_1, \ldots, x_n$ are the columns of $X^n \in \mathcal{X}^n$, and the square brackets make reference to the matrix with the given columns. This action is proper on the restriction of $\mathcal{X}^n$ to matrices of rank d any time that $n \geq d + 1$. Seen as a subset of $\mathbb{R}^{d\times n}$, the restriction of the Lebesgue measure to $\mathcal{X}^n$ is relatively left invariant with multiplier $\chi(g) = |\det(A)|^n$ for $g = (A, v) \in \mathrm{AL}(d)$.

REMARK 3.2. We use in the proof of the main theorem that, under these assumptions, the space $\mathcal{X}^n$ can be put in 1-to-1 bimeasurable correspondence with a subset of $G \times \mathcal{X}^n/G$, where G acts naturally by multiplication in the first component. More explicitly, under assumptions 1 and 2, Theorem 2 of Bondar (1976) guarantees the existence of a one-to-one map $r : \mathcal{X}^n \to G \times \mathcal{X}^n/G$ such that both r and its inverse are measurable, and, anytime that $x^n \mapsto (h(x^n), m(x^n))$, then, for any $g \in G$, the image of gx^n under r is $(gh(x^n), m(x^n))$.

REMARK 3.3. In our proofs, it will be useful to use, without loss of generality, the following modification to item 3 in Assumption 1:

3' The models $\{\mathbf{P}_g\}_{n\in\mathbb{N}}$ and $\{\mathbf{Q}_g\}_{n\in\mathbb{N}}$ are invariant and have densities with respect to a common measure ν on $\mathcal{X}^n$ that is left-invariant.

The reason that there is no loss in generality is that from any left-invariant measure μ with multiplier χ, a left-invariant measure ν can be constructed. Indeed, Bourbaki (2004, Chap. 7, §2 Proposition 7) shows that for any multiplier χ there exists a function $\varphi : \mathcal{X}^n \to \mathbb{R}$ with the property that $\varphi(gx) = \chi(g)\varphi(x)$ for any $x \in \mathcal{X}$ and $g \in G$. With this function at hand, one can define the measure $\mathrm{d}\nu(x) = \mathrm{d}\mu(x)/\varphi(x)$, which is left invariant. After multiplication by φ, probability densities with respect to μ are readily transformed into probability densities with respect to ν.

REMARK 3.4. On any locally compact group G there exists a left-invariant measure λ, called left Haar measure. It can be shown that λ is relatively right invariant with a multiplier Δ, that is, for any measurable $B \subseteq G$ and $g \in G$ it holds that $\lambda^h\{Bg\} = \Delta(g)\lambda^h\{B\}$ for any $g \in G$. This multiplier is called the (right) modulus of the group G. A computation shows that the measure ρ defined by $\rho^h\{B\} = \lambda^h\{B^{-1}\}$ for each measurable $B \subseteq G$, is right invariant, in other words, ρ is a right Haar measure. In the following, we always refer to right and left Haar measures that are related to each other by that identity. In our proofs we will use that for any integrable function f defined on G, the identities $\rho^h[f(h)] = \lambda^h[f(h)/\Delta(h)]$ and $\lambda^h[f(h^{-1})] = \rho^h[f(h)]$ hold (see Eaton, 1989, Section 1.3).

4. Main Result. In this section, we state in full the main result of this article, Corollary 4.3, a characterization of the GROW statistic for the statistical hypothesis testing problem (7). In Corollary 4.5 we will show that any GROW e-statistic is also relatively GROW in our group-invariant setting. Our main result stems from an application of the main technical contribution of this article, Theorem 4.2, which shows that the infimum Kullback-Leibler (KL) divergence between the elements of the convex hulls of the null and alternative hypotheses is exactly equal to the KL divergence between the distributions of the maximal invariant under both models. Theorem 4.2 will allow us to directly apply GHK's Theorem 1, which provides a general recipe constructing the GROW e-statistic in terms of the KL minimization problem (or joint information projection in information theoretic terminology). For simplicity and completeness we present here a special case of GHK's Theorem 1 that will be used in our group-invariant setting.

THEOREM 4.1 (Theorem 1 of Grünwald, de Heide and Koolen (2020), most general version given in their Section 4.3). Let $\mathcal{P} = \{\mathbf{P}_\theta\}_{\theta\in\Theta_0}$ and $\mathcal{Q} = \{\mathbf{Q}_\theta\}_{\theta\in\Theta_1}$ be two families of probability distributions on $\mathcal{X}^n$ that are dominated by a common measure. Suppose that there exists a random variable $V_n = v(X^n)$ such that

$$\inf_{\mathbf{\Pi}_0,\mathbf{\Pi}_1} \mathrm{KL}(\mathbf{\Pi}_1^{\theta_1}\mathbf{Q}_{\theta_1}, \mathbf{\Pi}_0^{\theta_0}\mathbf{P}_{\theta_0}) = \min_{\mathbf{\Pi}_0,\mathbf{\Pi}_1} \mathrm{KL}(\mathbf{\Pi}_1^{\theta_1}\mathbf{Q}_{\theta_1}^{V_n}, \mathbf{\Pi}_0^{\theta_0}\mathbf{P}_{\theta_0}^{V_n}) < \infty, \tag{15}$$

where the minimum and the infimum are over all pairs of proper probability distributions on Θ_0 and Θ_1. Assume further that there exists a pair of probability distributions $\mathbf{\Pi}_0^\star$ and $\mathbf{\Pi}_1^\star$ such that the previous minimum is achieved, that is,

$$\min_{\mathbf{\Pi}_0,\mathbf{\Pi}_1} \mathrm{KL}(\mathbf{\Pi}_1^{\theta_1}\mathbf{Q}_{\theta_1}^{V_n}, \mathbf{\Pi}_0^{\theta_0}\mathbf{P}_{\theta_0}^{V_n}) = \mathrm{KL}(\mathbf{\Pi}_1^{\star\theta_1}\mathbf{Q}_{\theta_1}^{V_n}, \mathbf{\Pi}_0^{\star\theta_0}\mathbf{P}_{\theta_0}^{V_n}). \tag{16}$$

Then

$$\max_{T_n\ e\text{-stat.}} \inf_{\theta_1\in\Theta_1} \mathbf{Q}_{\theta_1}[\ln T_n(X^n)] = \mathrm{KL}(\mathbf{\Pi}_1^{\star\theta_1}\mathbf{Q}_{\theta_1}^{V_n}, \mathbf{\Pi}_0^{\star\theta_0}\mathbf{P}_{\theta_0}^{V_n}). \tag{17}$$

In that case, the maximum on the left hand side of the previous display is achieved by the e-statistic T_n^* given by

$$T_n^*(X^n) := \frac{\mathbf{\Pi}_1^{\star\theta_1}[q_{\theta_1}^{V_n}(v_n(X^n))]}{\mathbf{\Pi}_0^{\star\theta_0}[p_{\theta_0}^{V_n}(v_n(X^n))]},$$

that is, T_n^* is GROW for testing $\mathcal{P}$ against $\mathcal{Q}$.

Once the connection between GROW e-statistics and KL divergence minimization is established, our next step is Theorem 4.2. In this section, we only treat the case in which, after the invariance-reduced data, both null and alternative hypothesis become simple so that the minimum in (15) trivializes. Theorem 4.2 establishes that, under our assumptions, (15) does indeed hold where V_n plays the role of the maximal invariant M_n and $\Theta_0 = \Theta_1 = G$ refer to the group. In Section 8 we investigate the case when this the hypotheses are still composite after the invariance reduction. Theorem 4.2 below immediately implies that the likelihood ratio for the maximal invariant is GROW; we delay its proof to Section 10.

THEOREM 4.2 (Main technical result). Let $M_n = m_n(X^n)$ be a maximally invariant function of the data X^n under the action of the group G on $\mathcal{X}^n$. Under Assumption 1, assume further that the group G is amenable as in Definition 2.1, and that there is $\varepsilon > 0$ such that

$$\mathbf{Q}_1\left[\left|\log\frac{q_1(X^n)}{p_1(X^n)}\right|^{1+\varepsilon}\right], \mathbf{Q}^{M_n}\left[\left|\log\frac{q^{M_n}(M_n)}{p^{M_n}(M_n)}\right|^{1+\varepsilon}\right] < \infty, \tag{18}$$

where the subindex in $\mathbf{Q}_1$ refers to the unit element of G, and $\mathbf{Q}^{M_n}$ and $\mathbf{P}^{M_n}$ are the distributions of M_n under any of the members of $\{\mathbf{Q}_g\}_{g\in G}$ and $\{\mathbf{P}_g\}_{g\in G}$, respectively. Then,

$$\inf_{\mathbf{\Pi}_0,\mathbf{\Pi}_1} \mathrm{KL}(\mathbf{\Pi}_1^g \mathbf{Q}_g, \mathbf{\Pi}_0^g \mathbf{P}_g) = \mathrm{KL}(\mathbf{Q}^{M_n}, \mathbf{P}^{M_n}), \tag{19}$$

where the infimum is taken over all pairs $(\mathbf{\Pi}_0, \mathbf{\Pi}_1)$ probability distributions on the group G.

From our previous discussion and with Theorem 4.2 at hand, the main result of this article follows.

COROLLARY 4.3 (Main result). *Under the assumptions of Theorem 4.2, a GROW e-statistic T^* for (7) is given by*

$$T_n^*(X^n) = \frac{q^{M_n}(m_n(X^n))}{p^{M_n}(m_n(X^n))},$$

the likelihood ratio for any maximally invariant statistic $M_n = m_n(X)$.

We end by showing that, in our group-invariant setting, any statistic that is GROW is also relatively GROW, meaning that any *e*-statistic that maximizes (5) also maximizes (4). This is not true in general (see Turner, Ly and Grünwald, 2021), and the result relies crucially on the invariant structure of the models under consideration. We give the proof of the following proposition at the end of the section.

PROPOSITION 4.4. *Suppose that the models $\{\mathbf{P}_g\}_{g\in G}$ and $\{\mathbf{Q}_g\}_{g\in G}$ satisfy item 3 of Assumption 1 and suppose that, for each $g \in G$, there exists $h \in G$ such that $\mathrm{KL}(\mathbf{Q}_g, \mathbf{P}_h)$ is finite. Then the map defined by*

$$g \mapsto \sup_{T_n \text{ e-stat.}} \mathbf{Q}_g[\ln T_n(X^n)]$$

is constant. Consequently, any maximizer of (5) also maximizes (4), that is, an e-statistic is relatively GROW if and only if it is also GROW for the hypothesis testing problem (7).

After inspecting that Proposition 4.4 indeed applies under the assumptions of Corollary 4.3, we can conclude the following corollary, the main objective of this section.

COROLLARY 4.5. *Not only is T^* from Corollary 4.3 GROW, it is also relatively GROW.*

PROOF. It is only left to check that, under the assumptions of Corollary 4.3, Proposition 4.4 applies. This is indeed the case because, by the invariance of the model and Hölder's inequality

$$\mathrm{KL}(\mathbf{Q}_g, \mathbf{P}_g) = \mathrm{KL}(\mathbf{Q}_1, \mathbf{P}_1) \le \left(\mathbf{Q}_1 \left[\left| \log \frac{q_1(X^n)}{p_1(X^n)} \right|^{1+\varepsilon} \right] \right)^{\frac{1}{1+\varepsilon}},$$

which was assumed to be finite. □

PROOF OF PROPOSITION 4.4. Let g be a fixed group element of G. Recall from Remark 3.3 that we may assume that both models are dominated by a left invariant measure ν on $\mathcal{X}$. By Theorem 1 of GHK (simplest instantiation in Section 2), any time that

$\inf_{h\in G} \mathrm{KL}(\mathbf{Q}_g, \mathbf{P}_h) < \infty$, there exists a subprobability density $\bar{p}$ on $\mathcal{X}^n$ relative to the left-invariant measure ν with two key properties: first, the function $T_n^\star(X^n) = q_g(X^n)/\bar{p}(X^n)$ is an e-statistic; second, $T_n^\star$ achieves the supremum in (5). Moreover the theorem implies that

$$\sup_{T \text{ } e\text{-stat.}} \mathbf{Q}_g[\ln T_n(X^n)] = \mathbf{Q}_g\left[\ln \frac{q_g(X^n)}{\bar{p}(X^n)}\right] = \inf_{\mathbf{\Pi}_0} \mathrm{KL}(\mathbf{Q}_g, \mathbf{\Pi}_0^{g'}\mathbf{P}_{g'}), \tag{20}$$

where the infimum is over all distributions $\mathbf{\Pi}_0$ on G. We will show that for any $g, h \in G$ and any prior $\mathbf{\Pi}$ on G, there exists a prior $\tilde{\mathbf{\Pi}}$ such that

$$\mathrm{KL}(\mathbf{Q}_g, \mathbf{\Pi}_0^{g'}\mathbf{P}_{g'}) = \mathrm{KL}(\mathbf{Q}_h, \tilde{\mathbf{\Pi}}^{g'}\mathbf{P}_{g'}). \tag{21}$$

From this our claim will follow: by symmetry, it implies that $g \mapsto \sup_{T_n \text{ } e\text{-stat.}} \mathbf{Q}_g[\ln T_n(X^n)]$ is constant over G because of its relation to the KL minimization in (20). Use the invariance of ν to compute

$$\begin{aligned}
\mathrm{KL}(\mathbf{Q}_g, \mathbf{\Pi}^{g'}\mathbf{P}_{g'}) &= \mathbf{Q}_g\left[\log \frac{q_g(X^n)}{\mathbf{\Pi}^{g'} p_{g'}(X^n)}\right] \\
&= \nu^{x^n}\left[q_g(x^n) \log \frac{q_g(x^n)}{\mathbf{\Pi}^{g'} p_{g'}(x^n)}\right] \\
&= \nu^{x^n}\left[q_h(hg^{-1}x^n) \log \frac{q_h(hg^{-1}x^n)}{\mathbf{\Pi}^{g'} p_{g'}(x^n)}\right].
\end{aligned}$$

Next, define $\tilde{\mathbf{\Pi}}$ as the probability distribution on G that assigns $\tilde{\mathbf{\Pi}}\{B\} = \mathbf{\Pi}\{gh^{-1}B\}$ for any measurable set $B \subseteq G$. Then $\mathbf{\Pi}^{g'} p_{g'} = \tilde{\mathbf{\Pi}}^{g'} p_{gh^{-1}g'}$. Plugging this back in the previous display, we see

$$\begin{aligned}
\mathrm{KL}(\mathbf{Q}_g, \mathbf{\Pi}^{g'}\mathbf{P}_{g'}) &= \nu^{x^n}\left[q_h(hg^{-1}x^n) \log \frac{q_h(hg^{-1}x^n)}{\tilde{\mathbf{\Pi}}^{g'} p_{gh^{-1}g'}(x^n)}\right] \\
&= \nu^{x^n}\left[q_h(hg^{-1}x^n) \log \frac{q_h(hg^{-1}x^n)}{\tilde{\mathbf{\Pi}}^{g'} p_{g'}(hg^{-1}x^n)}\right] \\
&= \nu^{x^n}\left[q_h(x^n) \log \frac{q_h(x^n)}{\tilde{\mathbf{\Pi}}^{g'} p_{g'}(x^n)}\right] \\
&= \mathrm{KL}(\mathbf{Q}_h, \tilde{\mathbf{\Pi}}^{g'}\mathbf{P}_{g'}),
\end{aligned}$$

where, in the second to last step, we use that ν is left-invariant. Hence, (21) follows and, by our previous discussion, so does our claim. ☐

5. Invariance and Sufficiency. The relationship between invariance and sufficiency has been thoroughly investigated (Hall, Wijsman and Ghosh, 1965; Hall, Wijsman and Ghosh, 1995; Berk, 1972; Nogales and Oyola, 1996). Consider a G-invariant hypothesis testing problem such that a sufficient statistic is available. If the action of G on the original data space induces a free action on the sufficient statistic, there must be a maximally invariant function of the sufficient statistic. With this structure in mind, the results presented thus far suggest two approaches for solving the hypothesis testing problem. The first is to reduce the data using the sufficient statistic, and to test the problem using the maximally invariant function of the sufficient statistic. The second approach is to use the maximal invariant for the original data. These two approaches yield two potentially different growth-optimal e-statistics, and one question arises naturally: are both approaches equivalent? In this section we show that this is indeed the case, under certain conditions.

We now introduce the set up formally. At the end of this section we revisit our guiding example, the t-test, and show how the results of this section apply to it. Let Θ be the parameter space, and let $\delta = \delta(\theta)$ be a maximal invariant function of θ for the action of G on Θ. Let $s_n : \mathcal{X}^n \to \mathcal{S}_n$ be a sufficient statistic for $\theta \in \Theta$. Consider again the hypothesis testing problem in the form presented in (2). Assume further that G acts freely and continuously on the image space $\mathcal{S}_n$ of the sufficient statistic $S_n = s_n(X^n)$, and assume that s_n is compatible with the action of G in the sense that, for any $X^n \in \mathcal{X}^n$ and any $g \in G$, the identity $gs_n(X^n) = s_n(gX^n)$ holds, where $(g, s) \mapsto gs$ makes reference to the action of G on $\mathcal{S}_n$. Let $M_{\mathcal{X},n} = m_{\mathcal{X},n}(X^n)$ and $M_{\mathcal{S},n} = m_{\mathcal{S},n}(S_n)$ be two maximally invariants for the actions of G on $\mathcal{X}^n$ and $\mathcal{S}_n$, respectively. Because of their invariance, the distributions of $M_{\mathcal{X},n}$ and $M_{\mathcal{S},n}$ depend only on the maximally invariant parameter δ. Hall, Wijsman and Ghosh (1965, Section II.3) proved that, under regularity conditions, if $S_{\mathcal{X},n} = s_{\mathcal{X},n}(X^n)$ is sufficient for $\theta \in \Theta$, then the statistic $M_{\mathcal{S},n} = m_{\mathcal{S},n}(s_n(X^n))$ is sufficient for δ. In that case, we call $M_{\mathcal{S},n}$ invariantly sufficient. Here we state the version of their result, attributed by Hall, Wijsman and Ghosh (1965) to C. Stein, that suits best our purposes[2].

THEOREM 5.1 (C. Stein). If there exists a Haar measure on the group G, the statistic $M_{\mathcal{S},n} = m_{\mathcal{S},n}(s_n(X^n))$ is invariantly sufficient, that is, it is sufficient for the maximally invariant parameter δ.

With this theorem at hand, and the fact that KL does not decrease by the application of sufficient transformations, we obtain the following theorem.

PROPOSITION 5.2. *Let $s_n : \mathcal{X}^n \to \mathcal{S}_n$ be sufficient statistic for $\theta \in \Theta$. Assume that G acts freely on $\mathcal{S}_n$ and that $s_n(gX^n) = gs_n(x^n)$ for all $X^n \in \mathcal{X}^n$ and $g \in G$. Let $m_{\mathcal{S},n}$ be a maximal invariant for the action of G on $\mathcal{S}_n$, and let $M_{\mathcal{S},n} = m_{\mathcal{S},n}(s_n(X^n))$. Then,*

$$\mathrm{KL}\left(\mathbf{P}_{\delta_1}^{M_{\mathcal{X},n}}, \mathbf{P}_{\delta_0}^{M_{\mathcal{X},n}}\right) = \mathrm{KL}\left(\mathbf{P}_{\delta_1}^{M_{\mathcal{S},n}}, \mathbf{P}_{\delta_0}^{M_{\mathcal{S},n}}\right).$$

PROOF. The function $M_{\mathcal{S},n} = m_{\mathcal{S},n}(s_n(X^n))$ is invariant, and consequently its distribution only depends on the maximal invariant parameter δ. Since $M_{\mathcal{X},n}$ is maximally invariant for the action of G on $\mathcal{X}^n$, there is a function f such that $M_{\mathcal{S},n} = f(M_{\mathcal{X},n})$. By Stein's theorem, Theorem 5.1, $M_{\mathcal{S},n}$ is sufficient for δ. Consequently, f is a sufficient transformation. Hence, from the invariance of the KL divergence under sufficient transformations, the result follows. □

Via the factorization theorem of Fisher and Neyman, the likelihood ratio for the maximal invariant $M_{\mathcal{X},n}$ coincides with that of the invariantly sufficient $M_{\mathcal{S},n}$. As a consequence, we obtain the answer to the motivating question of this section: performing an invariance reduction on the original data and on the sufficient statistic are equivalent.

COROLLARY 5.3. *Under the assumptions of Proposition 5.2,*

$$T_n^*(X^n) = \frac{p_{\delta_1}^{M_{\mathcal{X},n}}(m_{\mathcal{X},n}(X^n))}{p_{\delta_0}^{M_{\mathcal{X},n}}(m_{\mathcal{X},n}(X^n))} = \frac{p_{\delta_1}^{M_{\mathcal{S},n}}(m_{\mathcal{S},n}(S_n))}{p_{\delta_0}^{M_{\mathcal{S},n}}(m_{\mathcal{S},n}(S_n))}.$$

Hence, if assumptions of Corollary 4.3 also hold, the likelihood ratio for the invariantly sufficient statistic $M_{\mathcal{S},n}$ is relatively GROW.

[2]The assumption that there exists an invariant measure on G implies what Hall, Wijsman and Ghosh (1965) call Assumption A. (see Hall, Wijsman and Ghosh, 1965, discussion in p. 581)

EXAMPLE 1.1 (continued). We have seen that a maximally invariant function of the data is $M_{\mathcal{X},n} = m_{\mathcal{X},n}(X^n) = (X_1/|X_1|, \ldots, X_n/|X_1|)$ while the t-statistic $M_{\mathcal{S},n} = m_{\mathcal{S},n}(X^n) \propto \hat{\mu}_n/\hat{\sigma}_n$ is a maximally invariant function of the sufficient statistic $s_n(X^n) = (\hat{\mu}_n, \hat{\sigma}_n)$. Stein's theorem (Theorem 5.1) shows that the t-statistic $M_{\mathcal{S},n}$ is sufficient for the maximally invariant parameter $\delta = \mu/\sigma$. Corollary 5.3 shows that the likelihood ratio for the t-statistic is relatively GROW.

6. Anytime-valid testing under group-invariance. The main objective of this section is to prove Proposition 1.3, the main result of this article pertaining testing under optional stopping and continuation. We now assume that the observations are made sequentially. At the end of the section we describe the consequences to our main example, the t-test. We begin by defining our working model for this scenario. Let $X = (X_n)_{n\in\mathbb{N}}$ be a random process, where each X_n is an observation that takes values on a space $\mathcal{X}$. We consider the natural filtration associated to the observation process X, that is, the filtration $(\mathcal{F}_n)_{n\in\mathbb{N}}$ where and $\mathcal{F}_n$ is the sigma-algebra generated by $X^n = (X_1, \ldots, X_n)$. If $Z = (Z_n)_{n\in\mathbb{N}}$ is a random process and $(T_n)_{n\in\mathbb{N}}$ is a martingale adapted to the natural filtration associated to Z_n, we say that T is a Z-martingale. Similarly, we say that N is a Z-stopping time if N is a stopping time with respect to the natural filtration associated to Z. Let $(M_n)_{n\in\mathbb{N}}$ be a sequence where, for each n, $M_n = m_n(X^n)$ is a maximal invariant function for the action of G on $\mathcal{X}^n$. If, at each sample size n, the assumptions of Corollary 4.3 hold, we have shown that

$$T_n^*(X^n) = \frac{q^{M_n}(m_n(X^n))}{p^{M_n}(m_n(X^n))}, \tag{22}$$

the likelihood ratio for the maximal invariant $M_n = m_n(X^n)$, defines a sequence $T^* = (T_n^*)_{n\in\mathbb{N}}$ of relatively GROW e-statistics for (7). With an eye towards proving Proposition 1.3, in the next proposition we show, following the ideas of Hall, Wijsman and Ghosh (1965), that $T^* = (T_n^*)_{n\in\mathbb{N}}$ is a martingale.

PROPOSITION 6.1. *If $M = (M_n)_{n\in\mathbb{N}}$ is a sequence of maximal invariants $M_n = m_n(X^n)$ for the action of G on $\mathcal{X}^n$, the process $T^* = (T_n^*)_{n\in\mathbb{N}}$ given by (22) is a nonnegative M-martingale under any of the elements of the null hypothesis $\{\mathbf{P}_g\}_{g\in G}$.*

PROOF. Let $g \in G$ be arbitrary but fixed. We start by showing that T_n^* equals the likelihood ratio for $M^n = (M_1, \ldots, M_n)$ between $\mathbf{P}_g$ and $\mathbf{Q}_g$. For each $t > 1$, the maximal invariant at $n-1$, $M_{n-1} = m_{n-1}(X^{n-1})$ is invariant if seen as a function of X^n. Hence, by the maximality of m_n, M_{n-1} can be written as a function of M_n. Repeating this reasoning $n-1$ times yields that M_n contains all information about the value of $M^{n-1} = (M_1, \ldots, M_{n-1})$, all the maximal invariants at previous times. Two consequences fall from these observations. First, no additional information about T_n^* is gained by knowing the value of $M^{n-1} = (M_1, \ldots, M_{n-1})$ with respect to only knowing M_{n-1}, that is, $\mathbf{P}_g [T_n^* | M_{n-1}] = \mathbf{P}_g \left[T_n^* | M^{n-1}\right]$. Second, the likelihood ratio between $\mathbf{P}_g$ and $\mathbf{Q}_g$ for the sequence $M_1, \ldots, M_n$ equals the likelihood ratio for M_n alone, that is,

$$T_n^*(X^n) = \frac{q^{M_1,\ldots,M_n}(m_1(X^1), \ldots, m_n(X^n))}{p^{M_1,\ldots,M_n}(m_1(X^1), \ldots, m_n(X^n))}.$$

The previous two consequences, and a computation, together imply that T^* is an M-martingale under $\mathbf{P}_g$, that is, $\mathbf{P}_g \left[T_n^* | M^{n-1}\right] = T_{n-1}^*$. Since $g \in G$ was arbitrary, the result follows. $\square$

With this result at hand, we are in the position to prove Proposition 1.3 from Section 1.3, the main result in this work pertaining to sequential testing. We end this section with the implications to the t-test.

PROOF OF PROPOSITION 1.3. From Proposition 6.1, we know that $T^* = (T_n)_{n\in\mathbb{N}}$ is a nonnegative martingale with expected value equal to one. Let $\xi = (\xi_n)_n$ be the sequential test given by $\xi_n = \mathbb{1}\{T_n^* \geq 1/\alpha\}$. The anytime validity at level α of ξ, is a consequence of Ville's inequality, and the fact that the distribution of each T_n^* does not depend on g. Indeed, these two, together, imply that

$$\sup_{g\in G} \mathbf{P}_g\{T_n^* \geq 1/\alpha \text{ for some } n \in \mathbb{N}\} \leq \alpha.$$

Now, let $N \leq \infty$ be an M-stopping time. If the stopping time N is almost surely bounded, T_N^* is an e-statistic by virtue of the optional stopping theorem. However, since T^* is a nonnegative martingale, Doob's martingale convergence theorem implies the existence of an almost sure limit T_∞^*. Even when N might be infinite with positive probability, Theorem 4.8.4 of Durrett (2019) implies that T_N^* is still an e-statistic. □

EXAMPLE 1.1 (continued). In the previous section we saw that T_n^*, the likelihood ratio for the t-statistic is a GROW e-statistic. This, in conjunction with Proposition 1.3 implies that the test $\xi = (\xi_n)_{n\in\mathbb{N}}$ defined by $\xi_n = \mathbb{1}\{T_n^* \geq 1/\alpha\}$ is anytime valid at level α and that the randomly stopped e-statistic T_N^* remain one as long as the stopping time N is with respect to the sequence of maximally invariant statistics. In Appendix B we show a situation where the optionally stopped e-statistic is not an e-statistic if we take a stopping time that depends on the full data.

7. Testing multivariate normal distributions under group invariance. We show how the theory developed in the previous sections can be applied to hypothesis testing under normality assumptions. The family of d-dimensional normal distributions carries a natural invariance under scale-location transformations. The group of interest is the affine linear group $\mathrm{AL}(d)$, the group consisting of all pairs (v, A) with $v \in \mathbb{R}^d$, and A an invertible $d \times d$ matrix, and group operation $(v, A)(u, B) = (v + Au, AB)$. By considering amenable subgroups of $\mathrm{AL}(d)$, we obtain useful examples to which our results apply. We develop two in detail. The first is an alternative to Hotelling's T^2 for testing whether the mean of the distribution is identically zero, and results from the consideration $A \in \mathrm{LT}^+(d)$, the group of lower triangular matrices with positive entries on the diagonal, and $v = 0$. This test is in direct relation with the step-down procedure of Roy and Bargmann (1958)[3] (see also Subbaiah and Mudholkar, 1978). The second example that we consider is, in the setting of linear regression, a test for whether or not a specific regression coefficient is identically zero. It results from the restriction $A = cI$, a multiple of the $d \times d$ identity matrix.

7.1. *The lower triangular group.* Consider data $X^n = (X_1, \ldots, X_n)$ where $X_i \in \mathcal{X} = \mathbb{R}^d$. We assume each X_i to have a Gaussian distribution $N(\mu, \Sigma)$ with unknown mean $\mu \in \mathbb{R}^d$ and covariance matrix Σ. We consider a test for whether the mean μ of the distribution is zero. Before stating explicitly our hypothesis testing problem, we first reparametrize the Gaussian model using Cholesky's decomposition. Indeed, for a positive definite matrix Σ, its Cholesky

[3] Even though not explicitly in group-theoretic terms, the test of Roy and Bargmann (1958) test is based on a different maximally invariant function of the data. The fact that the test statistic of Roy and Bargmann (1958) is maximally invariant under the action of $\mathrm{LT}^+(d)$ is shown by Subbaiah and Mudholkar (1978)

decomposition is $\Sigma = \Lambda\Lambda'$ for a unique $\Lambda \in \mathrm{LT}^+(d)$. Consequently, $\mathrm{LT}^+(d)$ can be used to parametrize all covariance matrices. Hence, we may take the parameter space Θ to be $\Theta = \mathbb{R}^d \times \mathrm{LT}^+(d)$. In this parametrization, the likelihood of the original data $X^n = (X_1, \ldots, X_n)$ takes the form

$$p^{X^n}_{\Lambda,\delta}(X^n) = \frac{1}{(2\pi)^n(\det \Lambda)^n} \exp\left(-\frac{1}{2}\sum_{i=1}^n \left\| \Lambda^{-1}X_i - \delta \right\|^2 \right).$$

Consider the hypothesis testing problem,

$$\mathcal{H}_0 : \Lambda^{-1}\mu = \delta_0 \ \text{ vs. } \ \mathcal{H}_1 : \Lambda^{-1}\mu = \delta_1,$$

from which a test for whether μ is zero can be obtained by setting $\delta_0 = 0$. We now apply our results to this testing problem. Recall that the group $\mathrm{LT}^+(d)$ is amenable and acts on Θ by

$$(L, (\mu, \Lambda)) \mapsto (L\mu, L\Lambda) \tag{23}$$

for each $(\mu, \Lambda) \in \Theta$ and $L \in \mathrm{LT}^+(d)$, and a maximally invariant parameter is $\delta = \Lambda^{-1}\mu$. The group $\mathrm{LT}^+(d)$ acts on $\mathcal{X}^n$ by componentwise matrix multiplication, and the Gaussian model is invariant under this action. With the help of Remark 3.1, the assumptions of Corollary 4.3 are readily checked anytime that $n \geq d$, and we can conclude that, for any maximally invariant function $M_{\mathcal{X},n} = m_{\mathcal{X},n}(X^n)$ of the data, the likelihood ratio $T_n^* = p_{\delta_1}^{M_{\mathcal{X},n}} / p_{\delta_0}^{M_{\mathcal{X},n}}$ is GROW. However, from our discussion in Section 5, this likelihood ratio coincides with that of a invariantly sufficient statistic for δ. We now proceed to compute one such a statistic. Recall that the pair $S_n = s_n(X^n) = (\bar{X}_n, \bar{V}_n)$, consisting of the unbiased estimators $\bar{X}_n$ for the mean and the covariance matrix $\bar{V}_n$ is a sufficient statistic for (μ, Σ). We can apply to the sufficient statistic the same considerations that we applied to the parameter space. For $n \geq d$, we can perform the Cholesky decomposition of the empirical covariance matrix $\bar{V}_n = L_n L_n'$. The statistic $M_{\mathcal{S},n} = m_{\mathcal{S},n}(S_n) = \sqrt{\frac{n}{n-1}} L_n^{-1}\bar{Y}_n$ is maximally invariant under the action (23), and, by our discussion from Section 5, invariantly sufficient. In other words, $M_{\mathcal{S},n}$ is sufficient for δ. Hence, the GROW e-statistic can be written as $T_n^* = p_{\delta_1}^{M_{\mathcal{S},n}} / p_{\delta_0}^{M_{\mathcal{S},n}}$. For the purposes of sequential testing, Proposition 1.3 shows that the sequential test $(\xi^*_{n,\alpha} : n \in \mathbb{N})$ with $\xi^*_{n,\alpha} = \mathbb{1}\{T_n^* \geq 1/\alpha\}$ is anytime valid. For completeness, we give an explicit expression for the likelihood ratio $T^*_{\mathcal{S},n}$ when $\delta_0 = 0$. From this expression, the likelihood ratio for other values of δ_0 can be computed. We show the computations in Proposition A.1.

LEMMA 7.1. For the maximally invariant statistic $M_{\mathcal{S},n} = \sqrt{\frac{n}{n-1}} L_n^{-1}\bar{Y}_n$, we have

$$\frac{p_\delta^{M_{\mathcal{S},n}}(M_{\mathcal{S},n})}{p_0^{M_{\mathcal{S},n}}(M_{\mathcal{S},n})} = \mathrm{e}^{-\frac{n}{2}\|\delta\|^2} \mathbf{P}^T_{n,I}\left[\mathrm{e}^{n\langle \delta, TA_n^{-1}M_{\mathcal{S},n}\rangle} \right], \tag{24}$$

where A is the lower triangular matrix resulting from the Cholesky decomposition $I + M_{\mathcal{S},n}M_{\mathcal{S},n}' = A_nA_n'$, and $\mathbf{P}^T_{n,I}$ is the distribution according to which $nTT' \sim W(n, I)$.

PROOF. Use Proposition A.1 with $\gamma = \sqrt{n}\delta$, $X = \sqrt{n}\bar{X}$, $m = n - 1$, and $S = \bar{V}$. ☐

7.2. *A subset of the affine group* $\mathrm{AL}(d)$*: linear regression.* Consider the problem of testing whether one of the coefficients of a linear regression is zero under Gaussian error assumptions. Assume that the observations are of the form $(X_1, Y_1, Z_1), \ldots, (X_n, Y_n, Z_n)$, where, for each i, $X_i, Y_i \in \mathbb{R}$ and $Z_i \in \mathbb{R}^d$. We consider the the linear model given by

$$Y_i = \gamma X_i + \beta' Z_i + \sigma\varepsilon_i,$$

where $\gamma \in \mathbb{R}$, $\beta \in \mathbb{R}^d$ and $\sigma \in \mathbb{R}^+$ are the parameters, and $\varepsilon_1, \dots, \varepsilon_n$ are i.i.d. errors with standard Gaussian distribution $N(0,1)$. We are interested in testing

$$\mathcal{H}_0 : \gamma/\sigma = \delta_0 \;\text{ vs. }\; \mathcal{H}_1 : \gamma/\sigma = \delta_1. \tag{25}$$

A test for whether $\gamma = 0$ is readily obtained by taking $\delta_0 = 0$. This problem is invariant under the action of the subgroup G of $\mathrm{AL}(d)$ that results from the restriction to the pairs (A, v) where $A = cI$, a multiple of the $d \times d$ identity matrix, and $v \in \mathbb{R}^d$ (Kariya, 1980; Eaton, 1989). This group is amenable. On the observation space, G acts by $((cI, v), (X, Y, Z)) \mapsto (X, cY + v'Z, Z)$; on the parameter space, by $((cI, v), (\gamma, \beta, \sigma)) \mapsto (c\gamma, c\beta + v, c\sigma)$. A maximal invariant parameter is $\delta = \gamma/\sigma$. With this parametrization, the conditional density of Y becomes

$$p_{\delta,\beta,\sigma}(Y|X, Z) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left(-\frac{1}{2\sigma^2}(Y - \beta' Z - \sigma\delta X)^2\right).$$

Define the vectors $\boldsymbol{Y}_n = (Y_1, \dots, Y_n)'$ and $\boldsymbol{X}_n = (X_1, \dots, X_n)'$, and the $n \times d$ matrix $Z^n = [Z_1, \dots, Z_n]'$ whose rows are the vectors $Z_1, \dots, Z_n$. Assume that Z has full rank. A maximal invariant function of the data is given by $M_n = \left(\frac{A^{n\prime}\boldsymbol{Y}_n}{\|A^{n\prime}\boldsymbol{Y}_n\|}, \boldsymbol{X}_n, Z_n\right)$, where A^n is a $(n-d) \times n$ matrix whose columns form an orthonormal basis for the orthogonal complement of the column space of Z^n. It follows that $A^{n\prime}A^n = I^{n-d}$ and $A^n A^{n\prime} = I^n - Z^n(Z^{n\prime}Z^n)^{-1}Z^{n\prime}$, where and I^n is the $n \times n$ identity matrix (Kariya, 1980; Bhowmik, 2013). In order to compute the likelihood of the maximal invariant statistic M_n, we assume that the mechanism that generates $\boldsymbol{X}_n$ and Z^n is the same under both hypotheses. It only remains to compute the distribution of $\boldsymbol{U}_n = \frac{A^{n\prime}\boldsymbol{Y}_n}{\|A^{n\prime}\boldsymbol{Y}_n\|}$ conditionally on $\boldsymbol{X}_n$ and Z_n. Bhowmik (2013) shows that for arbitrary effect size δ, the density of this distribution is given by

$$\begin{aligned} p_\delta^{\boldsymbol{U}_n}(u|\boldsymbol{X}_n, Z^n) = \frac{1}{2}\Gamma\left(\frac{k}{2}\right)\pi^{-\frac{k}{2}}e^{c(\delta)}\Bigg[{}_1F_1\left(\frac{k}{2}, \frac{1}{2}, \frac{a^2(u,\delta)}{2}\right) \\ + \sqrt{2}a(u,\delta)\,\frac{\Gamma((1+k)/2)}{\Gamma(k/2)}\,{}_1F_1\left(\frac{1+k}{2}, \frac{3}{2}, \frac{a^2(u,\delta)}{2}\right)\Bigg], \end{aligned}$$

where $k = n - d$, u is a unit vector in $\mathbb{R}^k$, $a\,(u, \delta) = \delta \boldsymbol{X}_n' A^n u$, $c\,(\delta) = -\frac{1}{2}\delta^2 \boldsymbol{X}_n' A^n A^{n\prime} \boldsymbol{X}_n$, and ${}_1F_1$ is the confluent hypergeometric function. This can be used to compute the relatively GROW e-statistic in Theorem 4.2 for (25).

8. Composite invariant hypotheses. Until now we have considered null and alternative hypotheses that become simple when viewed through the lens of the maximally invariant statistic. As we saw, in the t-test this corresponds to testing simple hypotheses about the effect size δ. However, there are situations where it is desirable to contemplate hypotheses that are composite in the maximally invariant parameter. An example of such a situation is found in Hotelling's T^2 test (see Section 9). We also consider problems in which a fixed prior is placed on the maximally invariant parameter δ, in Corollary 8.3, thereby implementing the *method of mixtures*, a standard method to combine test martingales going back as far as Wald (1945) and Darling and Robbins (1968). It was already used in the context of our t-test example by Lai (1976).

Consider, as in the previous section, Θ to be the parameter space on which G acts freely and continuously. Let δ be a maximally invariant parameter. Suppose that the parameter space Θ can be decomposed as $\Theta \cong G \times \Theta/G$. Consider the testing problem

$$\mathcal{H}_0 : X^n \sim \mathbf{P}_{g,\delta},\;\; \delta \in \Delta_0,\; g \in G \;\text{ vs. }\; \mathcal{H}_1 : X^n \sim \mathbf{Q}_{g,\delta},\;\; \delta \in \Delta_1,\; g \in G, \tag{26}$$

where Δ_0, Δ_1 are two sets of possible values of the maximally invariant parameter $\delta = \delta(\theta)$. Recall that the distribution of a maximally invariant function of the data $M_n = m_n(X^n)$ depends on the parameter θ only through δ. Consequently, the alternatives in the previous hypothesis testing problem are not simple when data are reduced through invariance. The main objective of this section is to show that searching for a GROW e-statistic for (26) is equivalent to searching one for invariance-reduced problem

$$\mathcal{H}_0 : M_n \sim \mathbf{P}_\delta^{M_n}, \;\; \delta \in \Delta_0 \;\text{ vs. }\; \mathcal{H}_1 : M_n \sim \mathbf{Q}_\delta^{M_n}, \;\; \delta \in \Delta_1. \tag{27}$$

We follow the same steps that we followed in Section 4, and begin by showing that if there exists a minimizer for the KL minimization problem associated to (27), then it has the same value as that associated to (26).

PROPOSITION 8.1. *Assume that there exists a pair of probability distributions $\mathbf{\Pi}_0^\star, \mathbf{\Pi}_1^\star$ on Δ_0 and Δ_1 that satisfy*

$$\mathrm{KL}(\mathbf{\Pi}_1^{\star\delta}\mathbf{Q}_\delta^{M_n}, \mathbf{\Pi}_0^{\star\delta}\mathbf{P}_\delta^{M_n}) = \min_{\mathbf{\Pi}_0,\mathbf{\Pi}_1} \mathrm{KL}(\mathbf{\Pi}_1^\delta\mathbf{Q}_\delta^{M_n}, \mathbf{\Pi}_0^\delta\mathbf{P}_\delta^{M_n}). \tag{28}$$

For each $g \in G$, define the probability distributions $\mathbf{P}_g^\star = \mathbf{\Pi}_0^{\star\delta}\mathbf{P}_{g,\delta}$ and $\mathbf{Q}_g = \mathbf{\Pi}_1^{\star\delta}\mathbf{Q}_{g,\delta}$ on $\mathcal{X}^n$. If the models $\{\mathbf{P}_g^\star\}_{g\in G}$ and $\{\mathbf{Q}_g^\star\}_{g\in G}$ satisfy the assumptions of Theorem 4.2, then

$$\inf_{\mathbf{\Pi}_0,\mathbf{\Pi}_1} \mathrm{KL}(\mathbf{\Pi}_1^{g,\delta}\mathbf{Q}_{g,\delta}, \mathbf{\Pi}_0^{g,\delta}\mathbf{P}_{g,\delta}) = \min_{\mathbf{\Pi}_0,\mathbf{\Pi}_1} \mathrm{KL}(\mathbf{\Pi}_1^\delta\mathbf{Q}_\delta^{M_n}, \mathbf{\Pi}_1^\delta\mathbf{P}_\delta^{M_n}).$$

PROOF. Let $\mathbf{\Pi}_0^{g,\delta}, \mathbf{\Pi}_1^{g,\delta}$ be two probability distributions on $G \times \Delta_0$ and $G \times \Delta_1$, respectively. If we call $\mathbf{\Pi}_0^\delta$ and $\mathbf{\Pi}_1^\delta$ their respective marginals on Δ_0 and Δ_1, then, the information processing inequality implies that

$$\mathrm{KL}(\mathbf{\Pi}_1^{g,\delta}\mathbf{Q}_{g,\delta}, \mathbf{\Pi}_0^{g,\delta}\mathbf{P}_{g,\delta}) \geq \mathrm{KL}(\mathbf{\Pi}_1^\delta\mathbf{Q}_\delta^{M_n}, \mathbf{\Pi}_0^\delta\mathbf{P}_\delta^{M_n}) \geq \mathrm{KL}(\mathbf{\Pi}_1^{\star\delta}\mathbf{Q}_\delta^{M_n}, \mathbf{\Pi}_0^{\star\delta}\mathbf{P}_\delta^{M_n}).$$

This means that the right-most member of the previous display is a lower bound on our target infimum, that is,

$$\inf_{\mathbf{\Pi}_0,\mathbf{\Pi}_1} \mathrm{KL}(\mathbf{\Pi}_1^{g,\delta}\mathbf{Q}_{g,\delta}\mathbf{\Pi}_0^{g,\delta}\mathbf{P}_{g,\delta}) \geq \mathrm{KL}(\mathbf{\Pi}_1^{\star\delta}\mathbf{Q}_\delta^{M_n}, \mathbf{\Pi}_0^{\star\delta}\mathbf{P}_\delta^{M_n}). \tag{29}$$

To show that this is indeed an inequality, it suffices to prove that it is indeed the case if we limit ourselves to taking the infimum over a subset of all possible probability distributions $\mathbf{\Pi}_0, \mathbf{\Pi}_1$. We proceed to build such a subset. Let $\mathcal{P}(\mathbf{\Pi}_0^{\star\delta})$ be the set of probability distributions on $G \times \Delta_0$ with marginal distribution $\mathbf{\Pi}_0^{\star\delta}$. Define analogously the set of probability distributions $\mathcal{P}(\mathbf{\Pi}_1^{\star\delta})$ on $G \times \Delta_1$. By our assumptions, Theorem 4.2 can be readily used to conclude that

$$\inf_{(\mathbf{\Pi}_0,\mathbf{\Pi}_1)\in\mathcal{P}(\mathbf{\Pi}_0^{\star\delta})\times\mathcal{P}(\mathbf{\Pi}_1^{\star\delta})} \mathrm{KL}(\mathbf{\Pi}_1^{g,\delta}\mathbf{Q}_{g,\delta}, \mathbf{\Pi}_0^{g,\delta}\mathbf{P}_{g,\delta}) = \mathrm{KL}(\mathbf{\Pi}_1^{\star\delta}\mathbf{Q}_\delta^{M_n}, \mathbf{\Pi}_0^{\star\delta}\mathbf{P}_\delta^{M_n}) \tag{30}$$

holds; (29) and (30) together imply the result that we were after. □

From the previous proposition, using Theorem 4.1 and the steps used for Corollaries 4.3 and 4.5, we can conclude that the ratio of the Bayes marginals for the invariance-reduced data M_n using the optimal priors $\mathbf{\Pi}_0^\star$ and $\mathbf{\Pi}_1^\star$ is a relatively GROW e-statistic for (26).

COROLLARY 8.2. *Under the assumptions of Proposition 8.1, the statistic given by*

$$T^\star(X^n) = \frac{\mathbf{\Pi}_1^{\star\delta}[q_\delta(m_n(X^n))]}{\mathbf{\Pi}_0^{\star\delta}[p_\delta(m_n(X^n))]}$$

is a GROW and relatively GROW e-statistic for (26).

A standard approach to deal with unknown parameter values, both with Bayesian statistics and with e-statistics, is to employ proper prior distributions on the unknown parameters. In our setting, we may want to use specific priors $\tilde{\mathbf{\Pi}}_0$ and $\tilde{\mathbf{\Pi}}_1$ on Δ_0 and Δ_1. If we define for each g the probability distributions $\tilde{\mathbf{P}}_g = \tilde{\mathbf{\Pi}}_0^\delta \mathbf{P}_{g,\delta}$ and $\tilde{\mathbf{Q}}_g = \tilde{\mathbf{\Pi}}_1^\delta \mathbf{Q}_{g,\delta}$, and the resulting models $\{\tilde{\mathbf{P}}_g\}_{g\in G}$ and $\{\tilde{\mathbf{Q}}_g\}_{g\in G}$ also satisfy the conditions of Corollary 4.3, the proof of Proposition 8.1 also shows the following corollary.

COROLLARY 8.3. *Let $\tilde{\mathbf{\Pi}}_0$ and $\tilde{\mathbf{\Pi}}_1$ be two probability distributions on Δ_0 and Δ_1, respectively. Let $\{\tilde{\mathbf{P}}_g\}_{g\in G}$ and $\{\tilde{\mathbf{Q}}_g\}_{g\in G}$ be two probability models defined by $\tilde{\mathbf{P}}_g = \tilde{\mathbf{\Pi}}_0^\delta \mathbf{P}_{g,\delta}$ and $\tilde{\mathbf{Q}}_g = \tilde{\mathbf{\Pi}}_1^\delta \mathbf{Q}_{g,\delta}$. If $\{\tilde{\mathbf{P}}_g\}_{g\in G}$ and $\{\tilde{\mathbf{Q}}_g\}_{g\in G}$ satisfy the conditions of Corollary 4.3 (or more precisely, the conditions of Theorem 4.2 with $\tilde{\mathbf{P}}_g$ in the role of $\mathbf{P}_g$ and $\tilde{\mathbf{Q}}_g$ in the role of $\mathbf{Q}_g$), then*

$$\tilde{T}_n(X^n) = \frac{\tilde{\mathbf{\Pi}}_1^\delta[q_\delta(m_n(X^n))]}{\tilde{\mathbf{\Pi}}_0^\delta[p_\delta(m_n(X^n))]}$$

is a relatively GROW e-statistic for testing $\{\tilde{\mathbf{P}}_g\}_{g\in G}$ against $\{\tilde{\mathbf{Q}}_g\}_{g\in G}$.

9. Discussion, Related and Future Work. In this concluding section we bring up an issue that deserves further discussion and may inspire future work. It also highlights the differences between our work and related work in a Bayesian and information-theoretic context.

9.1. *Amenability is not always necessary.* We have shown that if a hypothesis testing problem is invariant under a group G and our assumptions are satisfied, then amenability of G is a sufficient condition for the likelihood ratio of the maximal invariant to be GROW. A natural question is whether amenability is also a necessary condition for the latter to hold. This is a relevant question, because there are some groups that are important for statistical practice, but are not amenable. For instance, $\mathrm{GL}(d)$ is the relevant group in Hotelling's test. The setup of this test is similar to that in Section 7.1, except that the hypotheses are given by

$$\mathcal{H}_0 : \|\Lambda^{-1}\mu\|^2 = 0 \text{ vs. } \mathcal{H}_1 : \|\Lambda^{-1}\mu\|^2 = \gamma. \tag{31}$$

A maximal invariant is the T^2-statistic $n\bar{Y}_n'\bar{V}_n^{-1}\bar{Y}_n$. Notice that this test is equivalent to the lower triangular test with the alternative expanded to $\Delta = \{\delta : \|\delta\|^2 = \gamma\}$, but that T^2 is not a maximal invariant under the lower triangular group. However, Giri, Kiefer and Stein (1963) have shown that for $d = 2$ and $n = 3$, the likelihood ratio of the T^2-statistic can be written as an integral over the likelihood ratio in (24) with a proper prior on $\delta \in \Delta$ as defined there. It follows as a result of Proposition 8.1 that the likelihood ratio of the T^2-statistics is also GROW in the case that $d = 2$ and $n = 3$. These results can be extended to the case that $d = 2$ with arbitrary n by the work of Shalaevskii (1971). As future work, it may be interesting to investigate whether amenity can be more generally replaced by a weaker condition, and/or whether a counterexample to Theorem 4.2 for non-amenable groups can be given.

9.2. *Comparison of our work to Sun and Berger (2007) and Liang and Barron (2004): two families vs. one.* As the above example illustrates, it is sometimes possible to represent the same $\mathcal{H}_0$ and $\mathcal{H}_1$ via (at least) two different groups, say G_a and G_b. Group G_a is combined with parameter of interest in some space Δ_a and priors $\mathbf{\Pi}_j^{*\delta_a}$ on Δ_a achieving (28) relative to group G_a, for $j = 0, 1$; group G_b has parameter of interest in Δ_b and priors $\mathbf{\Pi}_j^{*\delta_b}$ achieving (28) relative to group G_b; yet the tuples $\mathcal{T}_a = (G_a, \Delta_a, \{\mathbf{\Pi}_j^{*\delta_a}\}_{j=0,1})$ and $\mathcal{T}_b = (G_b, \Delta_b, \{\mathbf{\Pi}_j^{*\delta_b}\}_{j=0,1})$ define the same hypotheses $\mathcal{H}_0$ and $\mathcal{H}_1$. That is, the set of distributions $\{\mathbf{P}_g^*\}_{g\in G_a}$ obtained by applying Proposition 8.1) with group G_a (representing $\mathcal{H}_0$

defined relative to group G_a) coincides with the set of distributions $\{\mathbf{P}^*_g\}_{g \in G_b}$ obtained by applying Proposition 8.1) with group G_b (representing $\mathcal{H}_0$ defined relative to group G_a); and analogously for the set of distributions $\{\mathbf{P}^*_g\}_{g \in G_a}$ and the set of distributions $\{\mathbf{P}^*_g\}_{g \in G_b}$. (in the example above, G_a was $\mathrm{GL}(d)$ and the priors $\mathbf{\Pi}_0^{*\delta_a}, \mathbf{\Pi}_1^{*\delta_a}$ were degenerate priors on 0 and γ as in (31), respectively; G_b was the lower triangular group with a specific prior as indicated above). In such a case with multiple representations of the same $\mathcal{H}_0$ and $\mathcal{H}_1$, using the fact that the notion of 'GROW' does not refer to the underlying group, Corollary 8.2 can be used to identify the GROW e-statistic as soon as the assumptions of Proposition 8.1 hold for at least one of the tuples $\mathcal{T}_a$ or $\mathcal{T}_b$. Namely, if the assumptions hold for just one of the two tuples, we use Corollary 8.2 with that tuple; then $T^*(X^n)$ as defined in the corollary must be GROW, irrespective of whether $T^*(X^n)$ based on the other tuple is the same (as it was in the example above) or different. If the assumptions hold for *both* groups, then, using the fact that the GROW e-statistic is 'essentially' unique (see Theorem 1 of Grünwald, de Heide and Koolen (2020) for definition and proof), it follows that $T^*(X^n)$ as defined in Corollary 8.2 must coincide for both tuples.

Superficially, this may seem to contradict Sun and Berger (2007) who point out that in some settings, the right Haar prior is not uniquely defined, and different choices for right Haar prior give different posteriors, which are theire main object of interest. To resolve the paradox, note that, whereas we always formulate two models $\mathcal{H}_0$ and $\mathcal{H}_1$, Sun and Berger (2007) start with a single probabilistic model, say $\mathcal{P}$, that can be written as in (1) for two different groups G and G'. Their example shows that it is not always clear what group, and hence what Haar prior to pick, and their quantity of interest — the Bayesian posterior, i.e. a ratio between Bayes marginals for the same model $\mathcal{P}$ at different sample sizes n and $n-1$ — can depend on the choice. In contrast, our quantity of interest, the GROW e-statistic $T^*(X^n)$, a ratio between Bayes marginals for different models $\mathcal{H}_0$ and $\mathcal{H}_1$ at the same sample size, is uniquely defined as soon as there exists one group G with $\mathcal{H}_0$ and $\mathcal{H}_1$ as in (2) for which the assumptions of Theorem 4.2 hold; or more generally, as soon as there exists one tuple $\mathcal{T} = (G, \Delta, \{\mathbf{\Pi}_j^{*\delta}\}_{j=0,1})$ for which the assumptions of Proposition 8.1 hold, even if there exist other such tuples.

The consideration of two families $\mathcal{H}_0$ and $\mathcal{H}_1$ vs. a single $\mathcal{P}$ is also one of the main differences between our setting and the one of Liang and Barron (2004), who provide exact minimax procedures for predictive density estimation for general location- and scale-families under Kullback-Leibler loss. Their results apply to any invariant probabilistic model $\mathcal{P}$ as in (1) where the invariance is with respect to location or scale (and more generally, with respect to some other groups including the subset of the affine group that we consider in Section 7.2). Consider then such a $\mathcal{P}$ and let $p^{M_n}(m_n(X^n))$ be as in (11). As is well-known, provided that n' is larger than some minimum value, for all $n > n'$, $r(X_{n'+1}, \ldots, X_n \mid X_1, \ldots, X_{n'}) := p^{M_n}(m_n(X^n))/p^{M_{n'}}(m_{n'}(X^{n'}))$ defines a conditional probability density for $X_{n'+1}, \ldots, X_n$; this is a consequence of the formal-Bayes posterior corresponding to the right Haar prior becoming proper after n' observations, a.s. under all $\mathbf{P} \in \mathcal{P}$. For example, in the t-test setting, $n' = 1$. Liang and Barron (2004) show that the distribution corresponding to r minimizes the $\mathbf{P}^{n'}$- expected KL divergence to the conditional distribution $\mathbf{P}^n \mid X^{n'}$, in the worst case over all $\mathbf{P} \in \mathcal{P}$. Even though their optimal density r is defined in terms of the same quantities as our optimal statistic T_n^*, it is, just as in (Berger and Sun, 2008) considered above, a ratio between likelihoods for the same model at different sample sizes, rather than, as in our setting, between likelihoods for different models, both composite, at the same sample sizes. Our setting requires a joint KL minimization over two families, and therefore our proof techniques turn out quite different from their information- and decision-theoretic ones.

10. Proof of the main theorem, Theorem 4.2. For the proof of the main result, we use an equivalent definition of amenability to the one that was already anticipated in Section 2.2. We take the one that suits our purposes best (see Bondar and Milnes, 1981, p. 109, Condition A_1).

ASSUMPTION 2 (Amenability of G). There exists a increasing sequence of symmetric compact subsets $C_1 \subseteq C_2, \dots \subset G$ such that, for any compact set $K \subseteq G$,

$$\frac{\rho^h\{h \in C_i\}}{\rho^h\{h \in C_i K\}} \to 1$$

as $i \to \infty$.

In this formulation, amenability is the existence of *almost invariant* symmetric compact subsets of the group G. We use these sets to to build a sequence of *almost invariant* probability measures when G is noncompact.

PROOF OF THEOREM 4.2. Under our assumptions, Theorem 2 of Bondar (1976) implies the existence of a bimeasurable one-to-one map $\mathcal{X}^n \to G \times \mathcal{X}^n/G$ such that $r(x^n) = (h(x^n), m(x^n))$ and $r(gx^n) = (gh(x^n), m(x^n))$ for $h(x^n) \in G$ and $m(x^n) \in \mathcal{X}^n/G$ (see Remark 3.2). Hence, by a change of variables, we can assume that the densities are with respect to the image measure μ under r on $G \times \mathcal{X}^n/G$. Call the random variables $M = m(X^n)$ and $H = h(X^n)$. We can therefore assume, without loss of generality, that the data is of the form (H, M), that the group G acts canonically by multiplication on the first component, and that the measures are with respect to a G-invariant measure $\nu = \lambda \times \beta$ where λ is the Haar measure on G and β is some measure on $\mathcal{X}^n/G$ (see Remark 3.3). For each $g \in G$, write $\mathbf{P}^H_m$ and $\mathbf{Q}^H_{g,m}$ for the conditional probabilities $\mathbf{P}^H_g[\,\cdot\,|M = m]$ and $\mathbf{Q}^H_g[\,\cdot\,|M = m]$, which can be obtained through disintegration (see Chang and Pollard, 1997), and write $p(\,\cdot\,|m)$ and $q(\,\cdot\,|m)$ for their respective conditional densities with respect to the left Haar measure λ. For simplicity, simply write $\mathbf{P}$ and $\mathbf{Q}$ instead of $\mathbf{P}_1$ and $\mathbf{Q}_1$, any time that 1 is the unit element of the group.

We turn to our KL minimization objective. The chain rule for the KL divergence implies that, for any probability distribution $\mathbf{\Pi}$ on G,

$$\mathrm{KL}(\mathbf{\Pi}^g \mathbf{Q}_g, \mathbf{\Pi}^g \mathbf{P}_g) = \mathrm{KL}(\mathbf{Q}^M, \mathbf{P}^M) + \mathbf{Q}^m \left[\mathrm{KL}(\mathbf{\Pi}^g \mathbf{Q}^H_{g,m}, \mathbf{\Pi}^g \mathbf{P}^H_{g,m})\right], \tag{32}$$

where, recall, the superindex in $\mathbf{Q}^m$ signals integration over m. In order to prove our claim, we will build a sequence $\{\mathbf{\Pi}_i\}_{i \in \mathbb{N}}$ of probability distributions on G such that the term in (32) pertaining the conditional distributions given M goes to zero, that is, such that

$$\mathbf{Q}^m \left[\mathrm{KL}(\mathbf{\Pi}_i^g \mathbf{Q}^H_{g,m}, \mathbf{\Pi}_i^g \mathbf{P}^H_{g,m})\right] \to 0 \tag{33}$$

as $i \to \infty$. We define the distributions $\mathbf{\Pi}_i$ as the normalized restriction of the right Haar measure ρ to carefully chosen compact sets $C_i \subset G$, that we describe in brief. In other words, for $B \subseteq G$ measurable, we define $\mathbf{\Pi}_i$ by

$$\mathbf{\Pi}_i^g\{g \in B\} = \frac{\rho^g\{g \in B \cap C_i\}}{\rho^g\{g \in C_i\}}, \tag{34}$$

Next, the choice of sets C_i. Pick $C_i = J_i K_i L_i$ according to the following lemma.

LEMMA 10.1. Under the amenability of G there exist sequences $\{J_i\}_{i \in \mathbb{N}}$, $\{K_i\}_{i \in \mathbb{N}}$ and $\{L_i\}_{i \in \mathbb{N}}$ of compact symmetric neighborhoods of the unity of G, each increasing to cover G, such that

$$\frac{\rho^h\{h \in J_i\}}{\rho^h\{h \in J_i K_i L_i\}} \to 1$$

as $i \to \infty$.

There is no risk of dividing by ∞ in (34): by the continuity of the group operation each C_i is compact, hence $\rho\{C_i\} < \infty$. Proposition 10.1 ensures that $\mathbf{\Pi}_i^g\{g \in J_i\} \to 1$ as $i \to \infty$, a fact that will be useful later in the proof. Write $\mathbf{Q}_{i,m}^H := \mathbf{\Pi}_i^g \mathbf{Q}_{g,m}^H$, and $\mathbf{P}_{i,m}^H := \mathbf{\Pi}_i^g \mathbf{P}_{g,m}^H$, and $q_i(h|m)$ and $p_i(h|m)$ for their respective densities. Write our quantity of interest from (33) as

$$
\begin{aligned}
\mathbf{Q}^m\left[\mathrm{KL}(\mathbf{\Pi}_i^g \mathbf{Q}_{g,m}^H, \mathbf{\Pi}_i^g \mathbf{P}_{g,m}^H)\right] &= \int q_1(g^{-1}h, m) \log \frac{q_i(h|m)}{p_i(h|m)} \mathrm{d}\lambda(h) \mathrm{d}\mathbf{\Pi}_i(g) \mathrm{d}\beta(m) \\
&= \int q_1(h, m) \log \frac{q_i(gh|m)}{p_i(gh|m)} \mathrm{d}\lambda(h) \mathrm{d}\mathbf{\Pi}_i(g) \mathrm{d}\beta(m) \\
&= \mathbf{\Pi}_i^g \mathbf{Q}^{h,m}\left[\log \frac{q_i(gh|m)}{p_i(gh|m)}\right] \\
&= \underbrace{\mathbf{\Pi}_i^g \mathbf{Q}^{h,m}\left[\mathbb{1}\{gh \in J_i K_i\} \log \frac{q_i(gh|m)}{p_i(gh|m)}\right]}_{\mathrm{A}} + \\
&\qquad \underbrace{\mathbf{\Pi}_i^g \mathbf{Q}^{h,m}\left[\mathbb{1}\{gh \notin J_i K_i\} \log \frac{q_i(gh|m)}{p_i(gh|m)}\right]}_{\mathrm{B}}.
\end{aligned} \tag{35}
$$

We separate the rest of the proof in two steps, one for bounding each term in (35). These steps use two technical lemmas whose proof we give after showing how they help at achieving our goals.

Bound for A in (35): Notice that

$$\log \frac{q_i(gh|m)}{p_i(gh|m)} = \log \frac{\rho^{g'}[\mathbb{1}\{J_i K_i L_i\}\ q_{g'}(gh|m)]}{\rho^{g'}[\mathbb{1}\{J_i K_i L_i\}\ p_{g'}(gh|m)]}$$

Use $N = J_i K_i$ (which is not necessarily symmetric) and $L = L_i$ in the following

LEMMA 10.2. Let N and L be compact subsets of G. Assume that L is symmetric. Then, for each $m \in \mathcal{M}$ it holds that

$$\sup_{h \in N} \log \frac{\rho^g[\mathbb{1}\{NL\}\ q_g(h|m)]}{\rho^g[\mathbb{1}\{NL\}\ p_g(h|m)]} \leq -\log \mathbf{P}_m^H\{H \in L\}.$$

Conclude that for all $gh \in J_i K_i$ and $m \in \mathcal{M}$

$$\log \frac{q_i(gh|m)}{p_i(gh|m)} \leq -\log \mathbf{P}_m^h\{h \in L_i\}.$$

At the same time this implies that A in (35) is smaller than

$$-\mathbf{Q}^m\left[\log \mathbf{P}_m^h\{h \in L_i\}\right].$$

Since the sets L_i were chosen to satisfy $L_i \uparrow G$, the probability $\mathbf{P}_m^h\{h \in L_i\} \to 1$ goes to one for each value of m monotonically. Consequently the last display tends to 0 by the monotone convergence theorem, and so does A in (35).

Bound for B in (35): Our strategy at this point is to show that, as $i \to \infty$,

$$\mathbf{\Pi}_i^g \mathbf{Q}_1^h\{gh \notin J_i K_i\} \to 0, \tag{36}$$

and to use (18) to show our goal, that B in (35) tends to zero. To show (36), notice that if $g \in J_i$ and $h \in K_i$, then $gh \in J_i K_i$, which implies that

$$\mathbf{\Pi}_i^g \mathbf{Q}^h \{gh \in J_i K_i\} \geq \mathbf{\Pi}_i^g \{g \in J_i\} \, \mathbf{Q}^h \{h \in K_i\}.$$

Since K_i increases to cover G, we have $\mathbf{Q}^h \{h \in K_i\} \to 1$ as $i \to \infty$, and by our initial choice of sets J_i, K_i, L_i, the probability $\mathbf{\Pi}_i^g \{g \in J_i\} \to 1$, as $i \to \infty$. Hence (36) holds. To bound the second term, we use the following lemma.

LEMMA 10.3. Let $\mathbf{\Pi}^g$ be a distribution on G. Then for each $h \in G$ and $m \in \mathcal{M}$ it holds that

$$\log \frac{\int q_g(h|m) \mathrm{d}\mathbf{\Pi}(g)}{\int p_g(h|m) \mathrm{d}\mathbf{\Pi}(g)} \leq \mathbf{\Pi}_{h,m}^g \left[\log \frac{q_g(h|m)}{p_g(h|m)} \right].$$

where $\mathrm{d}\mathbf{\Pi}_{h,m}^g = \frac{q_g(h|m)\mathrm{d}\mathbf{\Pi}(g)}{\int_g q_g(h|m)\mathrm{d}\mathbf{\Pi}(g)}$

Apply Hölder's and Jensen's inequality consecutively to bound B in (35) by

$$\begin{aligned}
&\mathbf{\Pi}_i^g \mathbf{Q}^{h,m} \left[\mathbb{1}\{gh \notin J_i K_i\} \, \mathbf{\Pi}_{i,h,m}^{g'} \left[\log \frac{q_{g'}(gh|m)}{p_{g'}(gh|m)} \right] \right] \\
&\leq \left(\mathbf{\Pi}_i^g \mathbf{Q}^h \{gh \notin J_i K_i\} \right)^{1/q} \left(\mathbf{\Pi}_i^g \mathbf{Q}^{h,m} \left| \mathbf{\Pi}_{i,h,m}^{g'} \left[\log \frac{q_{g'}(gh|m)}{p_{g'}(gh|m)} \right] \right|^p \right)^{1/p} \\
&\leq \underbrace{\left(\mathbf{\Pi}_i^g \mathbf{Q}^h \{gh \notin J_i K_i\} \right)^{1/q}}_{\to 0 \text{ as } i \to \infty \text{ by (36)}} \left(\mathbf{\Pi}_i^g \mathbf{Q}^{h,m} \mathbf{\Pi}_{i,h,m}^{g'} \left| \log \frac{q_{g'}(gh|m)}{p_{g'}(gh|m)} \right|^p \right)^{1/p},
\end{aligned} \tag{37}$$

where $p = 1 + \varepsilon$ and q is p's Hölder conjugate, that is, $1/p + 1/q = 1$. Next, we show that the second factor in (37) remains bounded as $i \to \infty$. To this end, a change of variables shows that said factor can be rewritten as

$$\begin{aligned}
\mathbf{\Pi}_i^g \mathbf{Q}^{h,m} \mathbf{\Pi}_{i,h,m}^{g'} \left[\left| \log \frac{q_{g'}(gh|m)}{p_{g'}(gh|m)} \right|^p \right] &= \mathbf{\Pi}_i^g \mathbf{Q}_g^{h,m} \mathbf{\Pi}_{i,h,m}^{g'} \left[\left| \log \frac{q_{g'}(h|m)}{p_{g'}(h|m)} \right|^p \right] \\
&= \mathbf{Q}_i^{h,m} \mathbf{\Pi}_{i,h,m}^{g'} \left[\left| \log \frac{q_{g'}(h|m)}{p_{g'}(h|m)} \right|^p \right] \\
&= \mathbf{\Pi}_i^{g'} \mathbf{Q}_{g'}^{h,m} \left[\left| \log \frac{q_{g'}(h|m)}{p_{g'}(h|m)} \right|^p \right] \\
&= \mathbf{Q}^{h,m} \left[\left| \log \frac{q(h|m)}{p(h|m)} \right|^p \right].
\end{aligned}$$

Hence, as

$$\left(\mathbf{Q}^{h,m} \left[\left| \log \frac{q(h|m)}{p(h|m)} \right|^p \right] \right)^{1/p} \leq$$

$$\left(\mathbf{Q}^{h,m} \left[\left| \log \frac{q(h,m)}{p(h,m)} \right|^p \right] \right)^{1/p} + \left(\mathbf{Q}^m \left[\left| \log \frac{q(m)}{p(m)} \right| \right]^p \right)^{1/p} < \infty$$

by (18). We have shown that (37) tends to 0 as $i \to \infty$ and that consequently B in (35) tends to 0 in the same limit.

After completing these two steps, we have shown that both A and B in (35) tend to 0 as $i \to \infty$, and that consequently the claim of the theorem follows. All is left is to prove lemmas 10.1, 10.2, and 10.3. ☐

10.1. *Proof of technical lemmas 10.1, 10.2, and 10.3.*

PROOF OF LEMMA 10.1. Let $\{\varepsilon_i\}_i$ be a sequence of positive numbers decreasing to zero. Let $\{K_i\}_{i\in\mathbb{N}}$ and $\{L_i\}_{i\in\mathbb{N}}$ be two arbitrary sequences of compact symmetric subsets that increase to cover G. Fix $i \in \mathbb{N}$. The set K_iL_i is compact and by our assumption there exists a sequence $\{J_l\}_{l\in\mathbb{N}}$ and such that $\rho\{J_l\}/\rho\{J_lK_iL_i\} \to 0$ as $l \to \infty$. Pick $l(i)$ to be such that $\rho\{J_{l(i)}\}/\rho\{J_{l(i)}K_iL_i\} \geq 1 - \varepsilon_i$. The claim follows from a relabeling of the sequences. $\square$

PROOF OF LEMMA 10.2. Let $h \in N$. Then we can write

$$\begin{aligned}\rho^g[\mathbb{1}\{g \in NL\}\ q_g(h|m)] &= \rho^g[\mathbb{1}\{g \in NL\}\ q(g^{-1}h|m)]\\ &= \lambda^g[\mathbb{1}\{g \in (NL)^{-1}\}\ q(gh|m)]\\ &= \Delta(h^{-1})\lambda^g[\mathbb{1}\{g \in (NL)^{-1}h\}\ q(g|m)]\\ &= \Delta(h^{-1})\mathbf{Q}^g_m\{g \in (NL)^{-1}h\}\end{aligned}$$

The same computation can be carried out for p, and consequently

$$\begin{aligned}\log\frac{\rho^g[\mathbb{1}\{g \in NL\}\ q_g(h|m)]}{\rho^g[\mathbb{1}\{g \in NL\}\ p_g(h|m)]} &= \log\frac{\mathbf{Q}^g_m\{g \in (NL)^{-1}h\}}{\mathbf{P}^g_m\{g \in (NL)^{-1}h\}}\\ &\leq -\log\mathbf{P}^h_m\{h \in (NL)^{-1}h\}.\end{aligned}$$

By our assumption that $h \in N$, we have that $(NL)^{-1}h = L^{-1}N^{-1}h \supseteq L^{-1} = L$. This implies that the last quantity of the previous display is smaller than $-\log\mathbf{P}^H_m\{H \in L\}$. The result follows. $\square$

PROOF OF LEMMA 10.3. The result follows from a rewriting and an application of Jensen's inequality. Indeed,

$$\begin{aligned}-\log\frac{\int p_g(h|m)\mathrm{d}\mathbf{\Pi}(g)}{\int q_g(h|m)\mathrm{d}\mathbf{\Pi}(g)} &= -\log\frac{\int q_g(h|m)\frac{p_g(h|m)}{q_g(h|m)}\mathrm{d}\mathbf{\Pi}(g)}{\int q_g(h|m)\mathrm{d}\mathbf{\Pi}(g)}\\ &= -\log\mathbf{\Pi}^g_{h,m}\left[\frac{p_g(h|m)}{q_{g'}(h|m)}\right]\\ &\leq -\mathbf{\Pi}^g_{h,m}\left[\log\frac{p_g(h|m)}{q_g(h|m)}\right]\\ &= \mathbf{\Pi}^g_{h,m}\left[\log\frac{q_g(h|m)}{p_g(h|m)}\right].\end{aligned}$$

$\square$

11. Acknowledgements. Peter D. Grünwald is also affiliated with the Mathematical Institute of Leiden University. This work is part of the research program with project number 617.001.651, which is financed by the Dutch Research Council (NWO). We thank Wouter Koolen for useful conversations.

REFERENCES

ANDERSSON, S. (1982). Distributions of Maximal Invariants Using Quotient Measures. *Annals of Statistics* **10** 955–961. Publisher: Institute of Mathematical Statistics. https://doi.org/10.1214/aos/1176345885 MR663446

BENJAMIN, D. J., BERGER, J. O., JOHANNESSON, M., NOSEK, B. A., WAGENMAKERS, E.-J., BERK, R., BOLLEN, K. A., BREMBS, B., BROWN, L., CAMERER, C. et al. (2018). Redefine statistical significance. *Nature Human Behaviour* **2** 6.

BERGER, J. O., PERICCHI, L. R. and VARSHAVSKY, J. A. (1998). Bayes factors and marginal distributions in invariant situations. *Sankhyā: The Indian Journal of Statistics, Series A* 307–321.

BERGER, J. O. and SUN, D. (2008). Objective priors for the bivariate normal model. *Annals of Statistics* **36** 963–982. Publisher: Institute of Mathematical Statistics. https://doi.org/10.1214/07-AOS501 MR2396821

BERK, R. H. (1972). A Note on Sufficiency and Invariance. *The Annals of Mathematical Statistics* **43** 647–650. Publisher: Institute of Mathematical Statistics. https://doi.org/10.1214/aoms/1177692645

BHOWMIK, J. L. (2013). Constructing Locally Best Invariant Tests of the Linear Regression Model Using the Density Function of a Maximal Invariant. *American Journal of Mathematics and Statistics* **3** 45–52. https://doi.org/10.5923/j.ajms.20130301.07

BONDAR, J. V. (1976). Borel Cross-Sections and Maximal Invariants. *Annals of Statistics* **4** 866–877. Publisher: Institute of Mathematical Statistics. https://doi.org/10.1214/aos/1176343585 MR474589

BONDAR, J. V. and MILNES, P. (1981). Amenability: A survey for statistical applications of Hunt-Stein and related conditions on groups. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete* **57** 103–128. Publisher: Springer.

BOURBAKI, N. (2004). *Integration II: Chapters 7–9*. *Elements of Mathematics*. Springer-Verlag, Berlin Heidelberg. https://doi.org/10.1007/978-3-662-07931-7

CHANG, J. T. and POLLARD, D. (1997). Conditioning as disintegration. *Statistica Neerlandica* **51** 287–317. https://doi.org/10.1111/1467-9574.00056

COX, D. R. (1952). Sequential tests for composite hypotheses. *Mathematical Proceedings of the Cambridge Philosophical Society* **48** 290–299. Publisher: Cambridge University Press. https://doi.org/10.1017/S030500410002764X

DARLING, D. and ROBBINS, H. (1968). Some nonparametric sequential tests with power one. *Proceedings of the National Academy of Sciences* **61** 804–809.

DAWID, A. P., STONE, M. and ZIDEK, J. V. (1973). Marginalization Paradoxes in Bayesian and Structural Inference. *Journal of the Royal Statistical Society. Series B (Methodological)* **35** 189–233. Publisher: [Royal Statistical Society, Wiley].

DURRETT, R. (2019). *Probability: Theory and Examples*, 5th edition ed. *Cambridge Series in Statistical and Probabilistic Mathematics* **49**. Cambridge University Press Google-Books-ID: b22MDwAAQBAJ.

EATON, M. L. (1989). Group Invariance Applications in Statistics. *Regional Conference Series in Probability and Statistics* **1** i–133.

EATON, M. L. and SUDDERTH, W. D. (1999). Consistency and Strong Inconsistency of Group-Invariant Predictive Inferences. *Bernoulli* **5** 833–854. Publisher: International Statistical Institute (ISI) and Bernoulli Society for Mathematical Statistics and Probability. https://doi.org/10.2307/3318446

EATON, M. L. and SUDDERTH, W. D. (2002). Group invariant inference and right Haar measure. *Journal of Statistical Planning and Inference* **103** 87–99. https://doi.org/10.1016/S0378-3758(01)00199-9

GIRI, N., KIEFER, J. and STEIN, C. (1963). Minimax Character of Hotelling's T^2 Test in the Simplest Case. *The Annals of Mathematical Statistics* **34** 1524 – 1535. https://doi.org/10.1214/aoms/1177703884

GRÜNWALD, P. (2022). Beyond Neyman-Pearson. *arXiv preprint arXiv:2205.00901*.

GRÜNWALD, P., DE HEIDE, R. and KOOLEN, W. (2020). Safe Testing. *arXiv:1906.07801 [cs, math, stat]*. arXiv: 1906.07801.

HALL, W. J., WIJSMAN, R. A. and GHOSH, J. K. (1965). The Relationship Between Sufficiency and Invariance with Applications in Sequential Analysis. *The Annals of Mathematical Statistics* **36** 575–614. Publisher: Institute of Mathematical Statistics.

HALL, W. J., WIJSMAN, R. A. and GHOSH, J. K. (1995). Correction: The Relationship Between Sufficiency and Invariance with Applications in Sequential Analysis. *The Annals of Statistics* **23** 705–705. Publisher: Institute of Mathematical Statistics.

HENZI, A. and ZIEGEL, J. F. (2021). Valid sequential inference on probability forecast performance. *Biometrika* asab047. https://doi.org/10.1093/biomet/asab047

JEFFREYS, H. (1961). *Theory of Probability*, 3rd ed. Oxford University Press, London.

KARIYA, T. (1980). Locally Robust Tests for Serial Correlation in Least Squares Regression. *Annals of Statistics* **8** 1065–1070. https://doi.org/10.1214/aos/1176345143

KELLY, J. L. (1956). A New Interpretation of Information Rate. *Bell System Technical Journal* 917–926.

LAI, T. L. (1976). On confidence sequences. *The Annals of Statistics* **4** 265–280.

LEHMANN, E. L. and ROMANO, J. P. (2005). *Testing Statistical Hypotheses*, 3 ed. *Springer Texts in Statistics*. Springer-Verlag, New York. https://doi.org/10.1007/0-387-27605-X

LIANG, F. and BARRON, A. (2004). Exact minimax strategies for predictive density estimation, data compression, and model selection. *IEEE Transactions on Information Theory* **50** 2708–2726.

NOGALES, A. G. and OYOLA, J. A. (1996). Some Remarks on Sufficiency, Invariance and Conditional Independence. *The Annals of Statistics* **24** 906–909. Publisher: Institute of Mathematical Statistics.

PATERSON, A. L. (2000). *Amenability* **29**. American Mathematical Soc.

RAMDAS, A., RUF, J., LARSSON, M. and KOOLEN, W. (2020). Admissible anytime-valid sequential inference must rely on nonnegative martingales. *arXiv:2009.03167 [math, stat]*. arXiv: 2009.03167.

RAMDAS, A., RUF, J., LARSSON, M. and KOOLEN, W. M. (2021). Testing exchangeability: Fork-convexity, supermartingales and e-processes. *International Journal of Approximate Reasoning*. https://doi.org/10.1016/j.ijar.2021.06.017

REITER, H. and STEGEMAN, J. D. (2000). *Classical Harmonic Analysis and Locally Compact Groups*, second edition ed. *London Mathematical Society Monographs*. Oxford University Press, Oxford, New York.

REN, Z. and BARBER, R. F. (2022). Derandomized Knockoffs: Leveraging E-values for False Discovery Rate Control. *arXiv preprint arXiv:2205.15461*.

ROUDER, J. N., SPECKMAN, P. L., SUN, D., MOREY, R. D. and IVERSON, G. (2009). Bayesian t-tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review* **16** 225–237.

ROY, S. N. and BARGMANN, R. E. (1958). Tests of Multiple Independence and the Associated Confidence Bounds. *The Annals of Mathematical Statistics* **29** 491–503. Publisher: Institute of Mathematical Statistics. https://doi.org/10.1214/aoms/1177706624

ROYALL, R. (1997). *Statistical evidence: a likelihood paradigm*. Chapman and Hall.

SHAFER, G. (2019). The Language of Betting as a Strategy for Statistical and Scientific Communication.

SHAFER, G. (2021). Testing by betting: A strategy for statistical and scientific communication. *Journal of the Royal Statistical Society, Series A*. With Discussion.

SHALAEVSKII, O. V. (1971). Minimax Character of Hotelling's T2 Test. I. In *Investigations in Classical Problems of Probability Theory and Mathematical Statistics: Part I* (V. M. Kalinin and O. V. Shalaevskii, eds.) 74–101. Springer US, Boston, MA. https://doi.org/10.1007/978-1-4684-8211-9_2

SUBBAIAH, P. and MUDHOLKAR, G. S. (1978). A Comparison of Two Tests for the Significance of a Mean Vector. *Journal of the American Statistical Association* **73** 414–418. Publisher: Taylor & Francis _eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1978.10481592. https://doi.org/10.1080/01621459.1978.10481592

SUN, D. and BERGER, J. O. (2007). Objective Bayesian analysis for the multivariate normal model. *Bayesian Statistics* **8** 525–562.

TER SCHURE, J. and GRÜNWALD, P. (2019). Accumulation Bias in meta-analysis: the need to consider time in error control. *F1000Research* **8** 962. https://doi.org/10.12688/f1000research.19375.1

TURNER, R., LY, A. and GRÜNWALD, P. (2021). Two-Sample Tests that are Safe under Optional Stopping, with an Application to Contingency Tables. *arXiv:2106.02693 [cs, math, stat]*. arXiv: 2106.02693.

VILLE, J. (1939). *Étude Critique de la Notion de Collectif. Thèses de l'entre-deux-guerres* **218**.

VOVK, V. and WANG, R. (2020). E-values: Calibration, combination, and applications. *arXiv:1912.06116 [math, stat]*. arXiv: 1912.06116.

VOVK, V. and WANG, R. (2021). E-values: Calibration, combination and applications. *The Annals of Statistics* **49** 1736–1754. Publisher: Institute of Mathematical Statistics. https://doi.org/10.1214/20-AOS2020

WAGENMAKERS, E. J. (2007). A practical solution to the pervasive problems of p-values. *Psychonomic Bulletin & Review 14* 779-–804.

WALD, A. (1945). Sequential Tests of Statistical Hypotheses. *Ann. Math. Statist.* **16** 117-186.

WANG, R. and RAMDAS, A. (2020). False discovery rate control with e-values. *arXiv:2009.02824 [math, stat]*. arXiv: 2009.02824.

WANG, R. and RAMDAS, A. (2022). False discovery rate control with e-values. *Journal of the Royal Statistical Society Series B (Statistical Methodology)*.

WIJSMAN, R. A. (1985). Proper Action in Steps, with Application to Density Ratios of Maximal Invariants. *The Annals of Statistics* **13** 395–402. Publisher: Institute of Mathematical Statistics. https://doi.org/10.1214/aos/1176346600

APPENDIX A: COMPUTATIONS

PROPOSITION A.1. *Let $X \sim N(\gamma, I)$, and let $mS \sim W(m, I)$ be independent random variables. Let $LL' = S$ be the Cholesky decomposition of S, and let $M = \frac{1}{\sqrt{m}} L^{-1} X$. If $\mathbf{P}_{0,n}$ is the probability distribution under which $X \sim N(0, I)$, then, the likelihood $p^M_{\gamma,m}/p^M_{0,m}$ ratio is given by*

$$\frac{p^M_{\gamma,m}(M)}{p^M_{0,m}(M)} = \mathrm{e}^{-\frac{1}{2}\|\gamma\|^2} \mathbf{P}^T_{m+1,I}\left[\mathrm{e}^{\langle \gamma, TA^{-1}M \rangle}\right]$$

where $A \in \mathcal{L}^+$ is the Cholesky factor $AA' = I + MM'$, and $\mathbf{P}^T_{m+1,I}$ is the probability distribution on $\mathcal{L}^+$ such that $TT' \sim W(m+1, I)$.

PROOF. Let $\Sigma = \Lambda\Lambda'$ be the Cholesky decomposition of Σ. The density $p^X_{\gamma,\Lambda}$ of X with respect to the Lebesgue measure on $\mathbb{R}^d$ is

$$p^X_{\gamma,\Lambda}(X) = \frac{1}{(2\pi)^{d/2}\det(\Lambda)}\mathrm{etr}\left(-\frac{1}{2}(\Lambda^{-1}X-\gamma)(\Lambda^{-1}X-\gamma)'\right),$$

where, for a square matrix A, we define $\mathrm{etr}(A)$ to be the exponential of the trace of A. Let $W = mS$. Then, the density $p^W_{\gamma,\Lambda}$ of W with respect to the Lebesgue measure on $\mathbb{R}^{d(d-1)/2}$ is

$$p^W_{\gamma,\Lambda}(W) = \frac{1}{2^{md/2}\Gamma_d(n/2)\det(\Lambda)^m}\det(S)^{(m-d-1)/2}\mathrm{etr}\left(-\frac{1}{2}(\Lambda\Lambda')^{-1}W\right).$$

Now, let $W = TT'$ be the Cholesky decomposition of W. We seek to compute the distribution of the random lower lower triangular matrix T. To this end, the change of variables $W \mapsto T$ is one-to-one, and has Jacobian determinant equal to $2^d\prod_{i=1}^d t_{ii}^{d-i+1}$. Consequently, the density $p^T_{\gamma,\Lambda}(T)$ of T with respect to the Lebesgue measure is

$$p^T_{\gamma,\Lambda}(T) = \frac{2^d}{2^{md/2}\Gamma_d(m/2)}\det(\Lambda^{-1}T)^m\mathrm{etr}\left(-\frac{1}{2}(\Lambda^{-1}T)(\Lambda^{-1}T)'\right)\prod_{i=1}^d t_{ii}^{-i}. \tag{38}$$

We recognize $\mathrm{d}\nu(T) = \prod_{i=1}^d t_{ii}^{-i}\mathrm{d}T$ to be a left Haar measure on $\mathcal{L}_+$, and consequently

$$\tilde{p}^T_{\gamma,\Lambda}(T) = \frac{2^d}{2^{md/2}\Gamma_d(m/2)}\det(\Lambda^{-1}T)^m\mathrm{etr}\left(-\frac{1}{2}(\Lambda^{-1}T)(\Lambda^{-1}T)'\right) \tag{39}$$

is the density of T with respect to $\mathrm{d}\nu(T)$. After these rewritings, The density $\tilde{p}^{X,T}_{\gamma,\Lambda}(X,T)$ of the pair (X,T) with respect to $\mathrm{d}X \times \mathrm{d}\nu(T)$ is given by

$$\tilde{p}^{X,T}_{\gamma,\Lambda}(X,T) = \frac{2^d}{K}\frac{\det(\Lambda^{-1}T)^m}{\det(\Lambda)}\mathrm{etr}\left(-\frac{1}{2}(\Lambda^{-1}T)(\Lambda^{-1}T)' - \frac{1}{2}(\Lambda^{-1}X-\gamma)(\Lambda^{-1}X-\gamma)'\right)$$

with $K = (2\pi)^{d/2}2^{md/2}\Gamma_d(n/2)$. The change of variables $(X,T) \mapsto (T^{-1}X, T)$ has Jacobian determinant equal to $\det(T)$. If $M = T^{-1}X$, then, the density $\tilde{p}^{M,T}_{\gamma,\Lambda}$ of (M,T) with respect to $\mathrm{d}M \times \mathrm{d}\nu(T)$ is given by

$$\tilde{p}^{M,T}_{\gamma,\Lambda}(M,T) = \frac{\det(\Lambda^{-1}T)^{m+1}}{K''}\mathrm{etr}\left(-\frac{1}{2}(\Lambda^{-1}T)(\Lambda^{-1}T)' - \frac{1}{2}(\Lambda^{-1}TM-\gamma)(\Lambda^{-1}TM-\gamma)'\right).$$

We now marginalize T to obtain the distribution of the maximal invariant M. Since the integral is with respect to the left Haar measure $\mathrm{d}\nu(T)$, we have that

$$\int_{T\in\mathcal{L}^+}\tilde{p}^{M,T}_{\gamma,\Lambda}(M,T)\mathrm{d}\nu(T) = \int_{T\in\mathcal{L}^+}\tilde{p}^{M,T}_{\gamma,I}(M,\Lambda^{-1}T)\mathrm{d}\nu(T) = \int_{T\in\mathcal{L}^+}\tilde{p}^{M,T}_{\gamma,I}(M,T)\mathrm{d}\nu(T),$$

and consequently,

$$\begin{aligned}
p^M_{\gamma,\Lambda}(M) &= \frac{2^d}{K}\int_{T\in\mathcal{L}^+}\det(T)^{m+1}\mathrm{etr}\left(-\frac{1}{2}TT' - \frac{1}{2}(TM-\gamma)(TM-\gamma)'\right)\mathrm{d}\nu(T)\\
&= \frac{2^d}{K}\mathrm{e}^{-\frac{1}{2}\|\gamma\|^2}\int_{T\in\mathcal{L}^+}\det(T)^{m+1}\mathrm{etr}\left(-\frac{1}{2}T(I+MM')T' + \gamma(TM)'\right)\mathrm{d}\nu(T).
\end{aligned}$$

The matrix $I+MM'$ is positive definite and symmetric. It is then possible to perform its Cholesky decomposition $(I+MM')=AA'$. With this at hand, the previous display can be written as

$$p^M_{\gamma,\Lambda}(M)=\frac{\mathrm{e}^{-\frac{1}{2}\|\gamma\|^2}}{K}\int_{T\in\mathcal{L}^+}\det(T)^{m+1}\mathrm{etr}\left(-\frac{1}{2}(TA)(TA)'+\gamma(TM)'\right)\mathrm{d}\nu(T).$$

We now perform the change of variable $T\mapsto TA^{-1}$. To this end, notice that $\mathrm{d}\nu(A^{-1})=\mathrm{d}\nu(T)\prod_{i=1}^d a_{ii}^{-(d-2i+1)}$, and consequently

$$\begin{aligned}p^M_{\gamma,\Lambda}(M)&=\frac{2^d}{K}\frac{\mathrm{e}^{-\frac{1}{2}\|\gamma\|^2}\prod_{i=1}^d a_{ii}^{2i}}{\det(A)^{m+d+2}}\int_{T\in\mathcal{L}^+}\det(T)^{m+1}\mathrm{etr}\left(-\frac{1}{2}TT'+\gamma(TA^{-1}M)'\right)\mathrm{d}\nu(T)\\&=\frac{\Gamma_d\left(\frac{m+1}{2}\right)}{\pi^{d/2}\Gamma_d\left(\frac{m}{2}\right)}\frac{\prod_{i=1}^d a_{ii}^{2i}}{\det(A)^{m+d+2}}\mathrm{e}^{-\frac{1}{2}\|\gamma\|^2}\mathbf{P}^T_{m+1}\left[\mathrm{e}^{\langle\gamma,TA^{-1}M\rangle}\right],\end{aligned}$$

so that that at $\gamma=0$ the density $p^M_{0,\Lambda}(M)$ takes the form

$$p^M_{0,\Lambda}(M)=\frac{\Gamma_d\left(\frac{m+1}{2}\right)}{\pi^{d/2}\Gamma_d\left(\frac{m}{2}\right)}\frac{\prod_{i=1}^d a_{ii}^{2i}}{\det(A)^{m+d+2}},$$

and consequently the likelihood ratio is

$$\frac{p^M_{\gamma,\Lambda}(M)}{p^M_{0,\Lambda}(M)}=\mathrm{e}^{-\frac{1}{2}\|\gamma\|^2}\mathbf{P}^T_{m+1}\left[\mathrm{e}^{\langle\gamma,TA^{-1}M\rangle}\right].$$

□

REMARK A.2 (Numerical computation). Computing the optimal e-value is feasible numerically. We are interested in computing

$$\mathbf{P}^T_{m+1}\left[\mathrm{e}^{\langle x,Ty\rangle}\right],$$

where T is a $\mathcal{L}^+$-valued random lower triangular matrix such that $TT'\sim W(m+1,I)$, and $x,y\in\mathbb{R}^d$. Define, for $i\geq j$, the numbers $a_{ij}=x_iy_j$. Then $\langle x,Ty\rangle=\sum_{i\geq j}a_{ij}T_{ij}$. By Bartlett's decomposition, the entries of the matrix T are independent and $T_{ii}^2\sim\chi^2((m+1)-i+1)$, and $T_{ij}\sim N(0,1)$ for $i>j$. Hence, our target quantity satisfies

$$\mathbf{P}^T_{m+1}[\mathrm{e}^{\langle x,Ty\rangle}]=\mathbf{P}^T_{m+1}[\mathrm{e}^{\sum_{i\geq j}a_{ij}T_{ij}}]=\prod_{i\geq j}\mathbf{P}^T_{m+1}[\mathrm{e}^{a_{ij}T_{ij}}].$$

On the one hand, for the off-diagonal elements satisfy, using the expression for the moment generating function of a standard normal random variable,

$$\mathbf{P}^T_{m+1}[\mathrm{e}^{a_{ij}T_{ij}}]=\exp\left(\frac{1}{2}a_{ij}^2\right).$$

For the diagonal elements the situation is not as simple, but a numerical solution is possible. Indeed, for $a_{ii}\geq 0$, and $k_i=(m+1)-i+1$

$$\begin{aligned}\mathbf{P}^T_m[\mathrm{e}^{a_{ii}T_{ii}}]&=\frac{1}{2^{\frac{k_i}{2}}\Gamma\left(\frac{k_i}{2}\right)}\int_0^\infty x^{\frac{k_i}{2}-1}\exp\left(-\frac{1}{2}x+a_{ii}\sqrt{x}\right)\mathrm{d}x\\&={}_1F_1\left(\frac{k_i}{2},\frac{1}{2},\frac{a_{ii}^2}{2}\right)+\frac{\sqrt{2}a_{ii}\Gamma\left(\frac{k_i+1}{2}\right)}{\Gamma\left(\frac{k_i}{2}\right)}{}_1F_1\left(\frac{k_i+1}{2},\frac{3}{2},\frac{a_{ii}^2}{2}\right),\end{aligned}$$

where ${}_1F_1(a,b,z)$ is the Kummer confluent hypergeometric function. For $a_{ii} < 0$,

$$\frac{1}{2^{k_i/2}\Gamma\left(\frac{k_i}{2}\right)}\int_0^\infty x^{k_i/2-1}\exp\left(-\frac{1}{2}x + a_{ii}\sqrt{x}\right)\mathrm{d}x = \frac{\Gamma(k_i)}{2^{k_i-1}\Gamma\left(\frac{k_i}{2}\right)}U\left(\frac{k_i}{2},\frac{1}{2},\frac{a_{ii}^2}{2}\right),$$

and U is Kummer's U function.

APPENDIX B: E-STATISTICS T_N^* WITH VARIABLE STOPPING TIMES: THE FILTRATION MATTERS

Consider the the t-test as in Example 1.1. Fix some $0 < a < b$, and define the stopping time $N^* := 1$ if $|X_1| \notin [a,b]$. $N^* = 2$ otherwise. Then clearly N^* is not adapted to (hence not a stopping time relative to) $(M_n)_n$ as defined in that example, since $M_1 \in \{-1,1\}$ coarsens out all information in X_1 except its sign. Now let $\delta_0 := 0$ (so that $\mathcal{H}_0$ represents the normal distributions with mean $\mu = 0$ and arbitrary variance). Let $T_n^{*,\delta_1}(X^n)$ be equal to the GROW e-statistic $T_n^*(X^n)$ as in (14); here we make explicit its dependence on δ_1. For $\mathcal{H}_1$, to simplify computations, we put a prior $\tilde{\mathbf{\Pi}}_1^\delta$ on $\Delta_1 := \mathbb{R}$. We take $\tilde{\mathbf{\Pi}}_1^\delta$ to be a normal distribution with mean 0 and variance κ. We can now apply Corollary 8.3 (with prior $\tilde{\mathbf{\Pi}}_0^\delta$ putting mass 1 on $\delta = \delta_0 = 0$), which gives that $\tilde{T}_n(X^n)$ is an e-statistic, where

$$\tilde{T}_n(x^n) = \int \frac{1}{\sqrt{2\pi\kappa^2}}\exp\left(-\frac{\delta_1^2}{2\kappa^2}\right)\cdot T_n^{*,\delta_1}(x^n)d\delta_1$$

coincides with a standard type of Bayes factor used in Bayesian statistics. By exchanging the integrals in the numerator, this expression can be calculated analytically. The Bayes factor $\tilde{T}_1(x_1)$ for $x^1 = x_1$ is found to be equal to 1 for all $x_1 \neq 0$, and the Bayes factor for (x_1,x_2) is given by:

$$\tilde{T}_2(x_1,x_2) = \frac{\sqrt{2\kappa^2+1}\cdot(x_1^2+x_2^2)}{\kappa^2(x_1-x_2)^2+(x_1^2+x_2^2)}.$$

Now we consider the function

$$f(x) := \mathbf{E}_{X_2\sim N(0,1)}[\tilde{T}_2(x,X_2)].$$

$f(x)$ is continuous and even. We want to show that, with N^* as above, $\tilde{T}_{N^*}(X^{N^*})$ is not an E-variable for some specific choices of a, b and κ. Since, for any $\sigma > 0$, the null contains the distribution under which the X_i are i.i.d. $N(0,\sigma)$, the data may, under the null, in particular be sampled from $N(0,1)$. It thus suffices to show that

$$\mathbf{E}_{X_1,X_2\sim N(0,1)}[\tilde{T}_{N^*}(X^{N^*})] = P_{X_1\sim N(0,1)}(|X_1| \notin [a,b]) + \mathbf{E}_{X_1\sim N(0,1)}[\mathbf{1}_{|X_1|\in[a,b]}f(X_1)] > 1.$$

But from numerical integration we find that $f(x) > 1$ on $[a,b]$ and $[-b,-a]$ if we take $\kappa = 200$, $a \approx 0.44$ and $b \approx 1.70$. Using again numerical integration, we find that the above expectation is then approximately equal to 1.19, which shows that, even though $\tilde{T}_n$ is an e-statistic at each n by Corollary 8.3 (it is even a GROW one), $\tilde{T}_{N^*}$ is not an e-statistic (its expectation is 0.19 too large), providing the desired counterexample.