Electronic Journal of Statistics
Vol. 13 (2019) 2150–2193
ISSN: 1935-7524
https://doi.org/10.1214/19-EJS1576

Hypothesis testing near singularities and boundaries*

Jonathan D. Mitchell, Elizabeth S. Allman, and John A. Rhodes

Department of Mathematics & Statistics
University of Alaska Fairbanks
Fairbanks, Alaska 99775, USA
e-mail: jonathanmitchell88@gmail.com; esallman@alaska.edu; jarhodes2@alaska.edu

Abstract: The likelihood ratio statistic, with its asymptotic χ^2 distribution at regular model points, is often used for hypothesis testing. However, the asymptotic distribution can differ at model singularities and boundaries, suggesting the use of a χ^2 might be problematic nearby. Indeed, its poor behavior for testing near singularities and boundaries is apparent in simulations, and can lead to conservative or anti-conservative tests. Here we develop a new distribution designed for use in hypothesis testing near singularities and boundaries, which asymptotically agrees with that of the likelihood ratio statistic. For two example trinomial models, arising in the context of inference of evolutionary trees, we show the new distributions outperform a χ^2.

MSC 2010 subject classifications: Primary 62E17; secondary 92D15.
Keywords and phrases: Hypothesis testing, singularity, boundary, likelihood ratio statistic, chi-squared, phylogenomics, coalescent.

Received June 2018.

1. Introduction

The likelihood ratio statistic is commonly used to compare a null model to an alternative model. In many circumstances this statistic is asymptotically χ^2-distributed, which greatly facilitates testing of large data sets. As is well known, for smaller data sets, or when expectations are small for some outcomes, a χ^2 approximation may not be close enough to the true distribution for reliable testing. For example, minimum expected counts have been suggested to justify use of the approximation for contingency tables. But even for large data sets, a long thread of work has highlighted that problems can arise in using a χ^2 approximation at some points of the null model. Self and Liang [27] focused on non-standard asymptotics at boundary points of the null model, while more recently Drton [13] emphasized singularities. The asymptotic distribution at either of these points can be quite different from those at nearby regular points. Moreover, as shown by Andrews [2], the bootstrap may not consistently estimate the true distribution at such points.

[*] When this article was first made public, on June 28, 2019, its page numbering was incorrect (pp. 1250–1293). The article's page numbers were corrected to 2150–2193 on July 30, 2019.

Although [27] and [13] show how to understand and calculate asymptotic distributions at boundaries and singularities, they are not focused on how to use the distributions at such points in practice. Indeed, this is a difficult question, as the nature of these asymptotic distributions makes clear. For instance, one may find that an asymptotic distribution is χ^2_d with a fixed d degrees of freedom at almost all model points, but that at a boundary or singularity it discontinuously jumps to a different distribution — for instance, a mixture of several χ^2 distributions, or something more complicated. However, for the true non-asymptotic distribution, for any fixed sample size no matter how large, we do not expect such a jump to occur.

One might surmise that the asymptotics *at* the singularity or boundary could be relevant to testing even when the true parameter value is *near* that point, for fixed sample sizes. As the sample size is increased, the region on which the asymptotics give poor approximations shrinks, but no matter how large a sample is, the discontinuous behavior of the asymptotic distribution indicates there is some parameter region on which it is inappropriate for empirical use.

In this work we suggest a different approximation than the one obtained by standard asymptotics. While the usual arguments to derive the asymptotic distribution involve two approximations — the model is approximated by its tangent cone and the distribution of the random variables by a normal — we derive a different one by avoiding use of the tangent cone. This new approximate distribution is dependent on both sample size and parameter value, but has no discontinuous jump near boundaries or singularities. We explore it in detail using two particular models. These have both boundaries and singularities, yet are simple enough for a full exploration, using both theory and simulations.

For hypothesis testing with either the standard asymptotic distributions, or the new ones developed here, complications arise when the distributions depend on nuisance parameters. Methods of handling this include simply choosing the testing distribution giving the strictest test among those for all nuisance values, or among those in some confidence interval (Berger and Boos [8], Silvapulle and Sen [29]). More recent works of Andrews and Guggenberger [3] and McCloskey [20] adopt and extend these approaches, while bringing in non-standard asymptotic approximations along “drifting parameter sequences”. Through simulations we investigate how these methods apply to our new distributions for our example models.

Both of these last works [3, 20] (see also Andrews and Guggenberger [4, 5]) were motivated by issues with hypothesis testing near boundaries. They make use of limiting distributions obtained not in the standard way with sample size $n \to \infty$ while parameters are fixed, but with parameter values changing with n in a controlled way. One thus might consider the limits of distributions of finite sample likelihood ratio statistics along such drifting parameter sequences, and view them as approximations to the exact finite sample ones. In contrast, while the distributions of this paper are also approximations to the finite sample ones, they are found by using the asymptotics of the likelihood ratio process only, which is then scaled to approximate the process for finite sample size. Thus the relationship of these two approaches is not immediately clear. As we know of no general results on the form of limiting distributions along drifting parameter sequences, even in the case of likelihood ratio statistics, our approach is attractive both for computational tractability and for what might be considered a more intuitive basis rooted in the geometry of the model. Nonetheless, adoption of the approaches for dealing with nuisance parameters described by McCloskey [20] to the distributions of this paper are valuable for improved testing.

As amply demonstrated in the textbook of Silvapulle and Sen [29], models with boundaries arise commonly in empirical work, and continue to be of research interest [6, 7]. Although models with singularities have received considerably less attention, Drton [13] gives a number of natural examples. In this work, we focus on two simple models, one with a boundary, and one with a singularity, both of which arise in phylogenomics, but which we have not seen treated in depth elsewhere.

Phylogenomics is concerned with inferring evolutionary trees relating several different species from genomic-scale data. It builds on phylogenetics (the inference of trees based on sequences of a single gene), but brings in population-genetic effects that lead to many inferred gene trees differing from the species (or population) tree. Basics of the underlying multispecies coalescent model are explained below, though little familiarity with it is necessary for this work. It simply provides two motivating examples of nicely structured and accessible submodels of a trinomial (3-category multinomial) distribution, for which we can investigate behavior of tests near singularities and boundaries. While applications of the material developed here are highly relevant to phylogenomic practice, we defer discussion for empiricists to a later paper.

This paper is organized as follows. In Section 2 we lay out basic definitions, and illustrate with a simple example the problems that might arise when χ^2 distributions are used to approximate the distributions of likelihood ratio statistics near boundaries and singularities of null models. The specifics of the genomic models motivating our primary examples are then introduced.

The main theorem is given in Section 3, where an approximating distribution is defined for use in hypothesis testing. In Sections 4 and 5, we specialize to our examples, giving explicit forms of the finite sample approximating distributions. By simulation we show that using the standard χ^2_1 for hypothesis testing gives poor performance near a boundary or singularity; in contrast, the finite sample distributions we define perform very well for true parameters anywhere in the null model.

In Section 6 we use variation distances between the competing distributions (χ^2_1 and ours) to investigate the region of the null model where the standard χ^2_1 is good for testing, since this depends both on sample size and proximity to a singularity or boundary point. Section 7 investigates how various approaches to hypothesis testing with nuisance parameters behave in simulation. The final section is a discussion of our work and its potential for application beyond the examples developed here.

2. Definitions and examples

Let Θ, an open subset of $\mathbb{R}^k$, denote the parameter space for a family of probability distributions, and $\theta \in \Theta$ an unknown parameter vector. Submodels are specified by $\Theta_0 \subset \tilde{\Theta} \subseteq \Theta$, and we formulate the null hypothesis $H_0 : \theta \in \Theta_0$, with alternative $H_1 : \theta \in \Theta_1 = \tilde{\Theta} \setminus \Theta_0$. Given some data set, the *likelihood ratio statistic* is

$$\Lambda = 2 \left(\sup_{\theta \in \tilde{\Theta}} \ell(\theta) - \sup_{\theta \in \Theta_0} \ell(\theta) \right),$$

where $\ell(\theta) = \ell(\theta \mid \text{data})$ is the log-likelihood function. By determining the distribution of Λ under H_0, the decision as to how large Λ must be for rejection can be quantified.

While it is commonly assumed that the distribution of the likelihood ratio statistic under H_0 is well approximated by a χ^2 distribution, establishing this depends on a number of assumptions. Wilks [32] provided an early justification for sufficiently regular models defined by hyperplanes. Chernoff [9] extended the result to more general models, elucidating the role of the tangent space to the model, and making clear that asymptotic distributions other than χ^2 can arise. Other works emphasize that the statistic may not be asymptotically χ^2-distributed at boundaries of Θ_0 (e.g., [21], [27] and [28]).

Recent research of Drton [13] has emphasized that singularities pose problems as well. An asymptotic distribution of the statistic can be obtained at these problematic model points, as the distribution of the squared Euclidean distance between a standard normal sample and the appropriately linearly-transformed tangent cone of Θ_0 at the true parameter point θ_0 (Theorem 2.6 of [13]). Informally, the tangent cone is the set of all possible tangent vectors when approaching θ_0 along all possible paths in Θ_0. The tangent cone generalizes the tangent space which lead to the more familiar χ^2 distributions, but may lack the closure properties of a vector space that holds at smooth points of Θ_0.

To precisely define singularities and boundaries, we follow [13]. Assume Θ_0 is a *semialgebraic* subset of Θ. That is, Θ_0 is defined by a finite Boolean combination of polynomial equalities and inequalities, which ensures Chernoff regularity. The Zariski closure, $\overline{\Theta}_0$, of Θ_0 is the smallest algebraic variety (the zero set of a finite set of polynomials) containing Θ_0. This closure is the union of at most finitely many irreducible varieties, called components, which themselves cannot be expressed as a finite union of proper varieties.

A *singularity* of Θ_0 is then either *a*) a point in Θ_0 which lies on more than one irreducible component of $\overline{\Theta}_0$, or *b*) a point that lies on only one component, but at which the $n \times m$ Jacobian matrix of the defining equations of that component has lower rank than at generic points on the component. When a point lies on a single irreducible component Θ_0^i, the rank of the Jacobian is generically $m - \dim\left(\Theta_0^i\right)$. Lower rank indicates a problem with the notion of dimension at the point.

A subset of Θ_0 is said to be *open* if it is the intersection of $\overline{\Theta}_0$ with an open subset of $\mathbb{R}^k$. The *interior* of Θ_0 is the union of its open subsets, and the

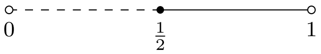

FIG 1. *The parameter space for a possibly biased coin. The solid segment is* $\Theta_0 = [1/2, 1)$*, while the dashed segment is* $\Theta_1 = (0, 1/2)$*.*

boundary of Θ_0 is the complement in Θ_0 of its interior. (In most applications, including all our example models, Θ_0 is closed in Θ under the standard topology, and this coincides with the usual definition of topological boundary in Θ.)

Note that the boundary and the set of singularities of a model need not be disjoint.

Example 2.1 (SIMPLE MODEL WITH BOUNDARY)**.** To test whether a coin, modeled by a Bernoulli random variable with probability of heads $\theta \in (0, 1)$, is biased towards tails, formulate hypotheses

$$H_0 : \ \theta \geq \frac{1}{2}, \qquad H_1 : \ \theta < \frac{1}{2}.$$

Here $\Theta = (0, 1) = \Delta^1$, the open simplex, and $\Theta_0 = [1/2, 1)$, as depicted in Figure 1. The Zariski closure of Θ_0 is the real line, and Θ_0 has no singularities but a single boundary point 1/2. At any θ_0 in the interior of Θ_0 the tangent cone is the full real line, $(-\infty, \infty)$. However, for $\theta_0 = 1/2$ the tangent cone is the half-line $[1/2, \infty)$.

From Theorem 2.6 of Drton [13], the asymptotic distribution of the likelihood ratio statistic is the distribution of the squared Euclidean distance between a normal random variable centered at θ_0 with variance 1 and the tangent cone at θ_0. For $\theta_0 > 1/2$, the squared Euclidean distance is 0 with probability 1 asymptotically, so the asymptotic distribution is a Dirac delta function δ_0. However, for $\theta_0 = 1/2$ the asymptotic distribution is a mixture $1/2\delta_0 + 1/2\chi^2_1$. Intuitively this is because samples from $\mathcal{N}(1/2, 1)$ lie on or off the tangent cone $[1/2, \infty)$ with probability 1/2, and the distributions of the squared distances are δ_0 and χ^2_1 respectively.

For this model, the maximum likelihood estimator (MLE) $\hat{\theta}_0$ of the parameter θ_0 is the maximum of 1/2 and the relative frequency of heads in a sample. If θ_0 lies in the interior of Θ_0, then for a sufficiently large sample $\hat{\theta}_0$ lies in the interior with probability arbitrarily close to 1. However, for a fixed sample size, no matter how large, there are points θ_0 close to 1/2 but still in the interior of Θ_0 for which this probability is much smaller (in fact, as close to 1/2 as desired). A better approximation to the distribution of the likelihood ratio statistic at such a point might be, for instance, a mixture of δ_0 and the square of a truncated normal centered at θ_0 with variance dependent on sample size. The mixing parameters depend on both θ_0 and the variance, while the truncation point of 1/2 is not generally the mean of the normal. When $\theta_0 = 1/2$, the normal distribution is truncated at the mean giving the asymptotic mixture distribution already described.

Of course for this model one can simply perform an exact binomial test, without any approximation. Nonetheless, this example highlights 1) that a likelihood ratio statistic's distribution can fail to converge uniformly to a χ^2 distribution even on the interior of Θ_0, 2) the role of the tangent cone in determining correct asymptotics, and 3) the inappropriateness of these asymptotic approximations for hypothesis testing for certain parameter values.

The next examples are the primary focus of our investigations. We briefly describe their motivation from phylogenomics, with more details supplied in Appendix A. The knowledge that these are submodels of a trinomial model is sufficient for the remainder of this work.

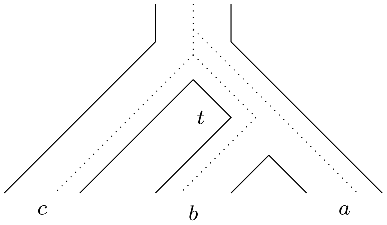

Fig 2. *An example of incomplete lineage sorting, where the dotted gene tree topology, $A|BC$, does not match the solid species tree topology, $c|ab$. This can occur because gene lineage coalescence events can predate species divergence events, when viewing time backward from the present (upwards).*

Example 2.2 (Model T1: Three species related by a specific species tree)**.** Suppose three species: a, b and c, are related by a rooted evolutionary species tree as shown in Figure 2, where the internal branch has length $t \geq 0$. Gene trees depicting evolutionary relationships for particular gene lineages (A, B, C) sampled from the three species may show differing topological relationships due to the population genetic effect of *incomplete lineage sorting*, illustrated in Figure 2. Under the *multispecies coalescent model* (see Appendix A), the three possible rooted gene tree topologies have probabilities

$$\left(p_{C|AB},\ p_{B|CA},\ p_{A|BC}\right) = \left(1 - \frac{2}{3}e^{-t},\ \frac{1}{3}e^{-t},\ \frac{1}{3}e^{-t}\right),$$

with $C|AB$ denoting the rooted topological gene tree matching the species tree topology with gene lineages A and B most closely related, and $B|AC$ and $A|BC$, interpreted analogously, gene tree topologies that do not match that of the species tree.

For a null hypothesis H_0 that the rooted topology of the species tree is $c|ab$, then

$$\Theta_0 = \left\{ (p_1,\ p_2,\ p_3) \mid p_1 \geq p_2 = p_3 > 0,\ \sum_i p_i = 1 \right\} \subset \Delta^2$$

is shown in Figure 3a. Here Δ^2 denotes the open 2-dimensional probability simplex. The alternative hypothesis, $\Theta_1 = \Delta^2 \smallsetminus \Theta_0$, can be interpreted as either that the species tree has a different tree structure $b|ca$ or $a|cb$, or that the model of a simple species tree under the multispecies coalescent is inadequate, perhaps due to introgression or hybridization of species populations, population structure within species, or other more complex biological issues.

Samples of n rooted gene trees drawn independently from the multispecies coalescent model on the species tree of Figure 2 are thus described by a submodel of the trinomial model with parameter space Θ_0.

Example 2.3 (Model T3: Three species related by any of the three possible species trees)**.** If the model T1 of Example 2.2 is modified, so that the specific species tree structure is not fixed, but any one of $a|bc$, $b|ac$, or $c|ab$ might be the species tree, then H_0 is that there is *some* species tree giving rise to the gene tree data. The alternative H_1 is that a simple species tree model does not fit the data. The null parameter space $\Theta_0 \subset \Delta^2$, shown in Figure 3b, is the union of three submodels of trinomial models.

As seen in Figure 3a, the model T1 has a boundary point at $(1/3, 1/3, 1/3) \in \Theta_0$, and no singularities. For model T3, the point $(1/3, 1/3, 1/3)$ is a singularity of Θ_0, since the Zariski closure of Θ_0 is three lines (irreducible components) crossing at that point. This point is also a boundary, though we will refer to it simply as the singularity.

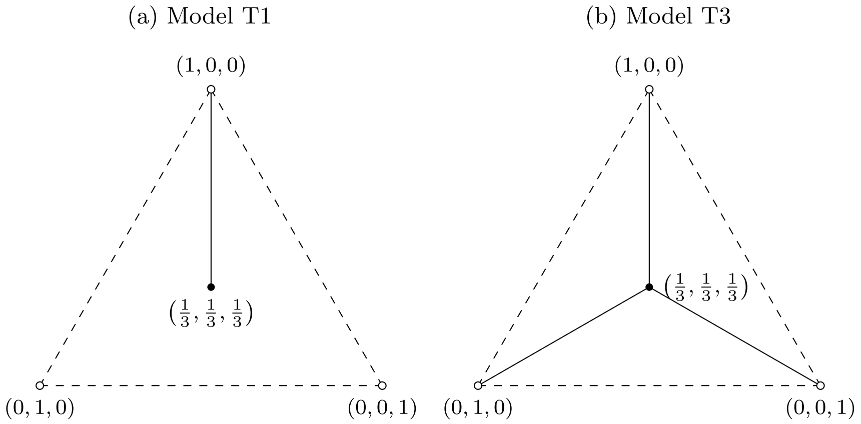
(a) Model T1
(1, 0, 0)
$(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$
(0, 1, 0)
(0, 0, 1)
(b) Model T3
(1, 0, 0)
$(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$
(0, 1, 0)
(0, 0, 1)

Fig 3. *Geometric view of the models: (a) T1 and (b) T3. The solid line segment(s) represent(s) Θ_0, while the region inside the dotted lines represents Θ, the open probability simplex Δ^2. The central point $(1/3, 1/3, 1/3)$ corresponding to $t = 0$ on any species tree is either a boundary (T1) or a singularity (T3).*

When a rooted species tree on three species has a short internal branch so that much incomplete lineage sorting occurs, the expected gene tree probabilities lie close to the boundary or singularity $(1/3, 1/3, 1/3)$ of the models. This is exactly the situation in which it is hardest to resolve species tree relationships, and therefore often one of pressing biological interest. Indeed, motivation for this paper is the recognition that the use of the standard asymptotic approximation is not reliable near boundaries and singularities, and a careful investigation of this problem is of practical as well as theoretical interest.

The models T_1 and T_3, and the more general multispecies coalescent model for larger trees, are increasingly used in inference of species trees from genomic-scale data, though typically little is done to test whether the model is appropriate for data. For relating three species, Degnan and Rosenberg [12] describe a hypothesis test using a χ^2 distribution, though our work here underscores that this test can be problematic near singularities and boundaries. Results of Allman, Degnan and Rhodes [1] show that this test extends to the unrooted 4-species trees this paper focuses on, though the same boundary and singularity issues arise in using the χ^2. Gaither and Kubatko [16] introduce a different hypothesis test for 4-species trees, but in a different framework, working from DNA sequence data under a combined model of coalescence with sequence evolution, and not on gene tree frequencies. Most empirical studies simply assume the coalescent model on a species tree is appropriate, even though several biological processes are known which could violate it.

3. Approximate distributions of likelihood ratio statistics

We now illustrate that, in principle, one can obtain an alternative, potentially more useful, approximation to the distribution of the likelihood ratio statistic than the standard asymptotic one.

For a statistical model with parameter spaces $\Theta_0 \subset \tilde{\Theta} \subseteq \Theta$, $\Theta_1 = \tilde{\Theta} \smallsetminus \Theta_0$, and parameter $\theta_0 \in \Theta_0$, let $X^{(1)}, \dots, X^{(n)}$ denote n independent and identically distributed random observations. The likelihood function for a sample realization $X^{(1)}, \dots, X^{(n)}$ is

$$\ell_n(\theta) = \sum_{i=1}^{n} \log p\left(x^{(i)} \mid \theta\right).$$

Maximizers of the likelihood over Θ_0 and $\tilde{\Theta}$ are the maximum likelihood estimators (MLEs) over the corresponding parameter spaces.

The likelihood ratio statistic for a sample then is

$$\Lambda_n = 2\left(\sup_{\theta \in \tilde{\Theta}} \ell_n(\theta) - \sup_{\theta \in \Theta_0} \ell_n(\theta)\right).$$

Under appropriate regularity conditions (see Theorem 16.7 of Van der Vaart [30]) the asymptotic distribution of this statistic, as $n \to \infty$, is that of

$$\left\| X - \mathcal{I}(\theta_0)^{\frac{1}{2}} T_0 \right\|^2 - \left\| X - \mathcal{I}(\theta_0)^{\frac{1}{2}} T \right\|^2,$$

for $X \sim \mathcal{N}(0, I)$, $\mathcal{I}(\theta_0)$ the Fisher information matrix at θ_0, T_0 and T the tangent cones to Θ_0 and $\tilde{\Theta}$ at θ_0, and where $\|x - B\|$ denotes the minimal

Euclidean distance between a point x and set B. In essence, establishing this theorem using local asymptotic normality depends on two approximations: the likelihood ratio process from sample realizations is approximately normal, and the model parameter space is approximated locally by its tangent cone.

Of these two approximations, it is that of the tangent cone which can lead to the discontinuous behavior of the asymptotic distribution, since the tangent cone's features can behave discontinuously as a function of the parameter. For example, if a model is parameterized by a closed ball in $\mathbb{R}^k$, at interior points the tangent space will be a k-dimensional Euclidean space, while at the boundary it becomes a half-space. For a model with parameter space a curve in the plane that crosses over itself, the tangent space will be a line at most points, but at the singularity it is two crossed lines.

Examining a derivation of the asymptotics of the likelihood ratio statistic more closely, local asymptotic normality allows for the approximation by a normal for large samples. For large samples the distribution's covariance approaches 0, and rescaling to a standard normal means the parameter space must be dilated around the true parameter. It is this dilation that allows the parameter space of the model to be approximated by a tangent cone. Thus these two approximations are interrelated, and are not made independently.

Nonetheless, we informally reason that while the normal approximation may be a good one even for a relatively small sample size, a much larger sample may be needed for the approximating normal to be sufficiently concentrated that the tangent approximation of the model is accurate. This motivates Theorem 3.1 below.

For parameter spaces $\Theta_0 \subset \tilde{\Theta} \subseteq \mathbb{R}^k$ and parameter value $\theta_0 \in \Theta_0$, define sequences of scaled translated parameter spaces $T_n = \sqrt{n}\left(\tilde{\Theta} - \theta_0\right)$ and $T_{n,0} = \sqrt{n}\left(\Theta_0 - \theta_0\right)$. Suppose $T_n \to T$ and $T_{n,0} \to T_0$ in the sense defined in [30]. As pointed out by [13], a condition such as Chernoff regularity ensures this convergence of spaces, with T and T_0 the tangent cones at θ_0 of $\tilde{\Theta}$ and Θ_0.

Theorem 3.1. *Consider n i.i.d. random observations from a model with parameter space Θ open in $\mathbb{R}^k$ and submodels determined by $\Theta_0 \subset \tilde{\Theta} \subseteq \Theta$, with $\Theta_1 = \tilde{\Theta} \setminus \Theta_0$. Let $\theta_0 \in \Theta_0$ be a true parameter point, with non-singular Fisher information matrix $\mathcal{I}\left(\theta_0\right)$ for a sample of size 1. Let $\mathcal{I}\left(\theta_0\right)^{\frac{1}{2}}$ be a matrix such that $\mathcal{I}\left(\theta_0\right) = \left(\mathcal{I}\left(\theta_0\right)^{\frac{1}{2}}\right)^T \mathcal{I}\left(\theta_0\right)^{\frac{1}{2}}$ and $Y \sim \mathcal{N}\left(\sqrt{n}\mathcal{I}\left(\theta_0\right)^{\frac{1}{2}}\theta_0, I\right)$.*

Then under the regularity assumptions of Proposition 16.7 of [30], for a sample of size n the likelihood ratio statistic Λ_n for H_0 vs. H_1 is approximately distributed as the random variable

$$W = \inf_{\tau \in \sqrt{n}\mathcal{I}(\theta_0)^{\frac{1}{2}}\Theta_0} \|Y - \tau\|^2 - \inf_{\tau \in \sqrt{n}\mathcal{I}(\theta_0)^{\frac{1}{2}}\tilde{\Theta}} \|Y - \tau\|^2,$$

in the sense that both the likelihood ratio statistic and this random variable converge in distribution to the same limit as $n \to \infty$.

Proof. By Theorem 16.7 of [30], the likelihood ratio statistic converges in distribution to

$$\left\|X-\mathcal{I}\left(\theta_0\right)^{\frac{1}{2}}T_0\right\|^2-\left\|X-\mathcal{I}\left(\theta_0\right)^{\frac{1}{2}}T\right\|^2,$$

for $X\sim\mathcal{N}\left(0,I\right)$.

However, with $Y=X+\sqrt{n}\mathcal{I}\left(\theta_0\right)^{\frac{1}{2}}\theta_0$,

$$\begin{aligned}W=&\inf_{\tau\in\sqrt{n}\mathcal{I}(\theta_0)^{\frac{1}{2}}\Theta_0}\left\|X+\sqrt{n}\mathcal{I}\left(\theta_0\right)^{\frac{1}{2}}\theta_0-\tau\right\|^2\\&-\inf_{\tau\in\sqrt{n}\mathcal{I}(\theta_0)^{\frac{1}{2}}\tilde{\Theta}}\left\|X+\sqrt{n}\mathcal{I}\left(\theta_0\right)^{\frac{1}{2}}\theta_0-\tau\right\|^2\\=&\inf_{\tau\in T_{n,0}}\left\|X-\mathcal{I}\left(\theta_0\right)^{\frac{1}{2}}\tau\right\|^2-\inf_{\tau\in T_n}\left\|X-\mathcal{I}\left(\theta_0\right)^{\frac{1}{2}}\tau\right\|^2\\=&\left\|X-\mathcal{I}\left(\theta_0\right)^{\frac{1}{2}}T_{n,0}\right\|^2-\left\|X-\mathcal{I}\left(\theta_0\right)^{\frac{1}{2}}T_n\right\|^2.\end{aligned}$$

Since $T_n\to T$ and $T_{n,0}\to T_0$, applying Lemma 7.13 of [30] yields the result. □

Note that the condition that the sample is i.i.d. is not necessary in the theorem; a more general result is possible if $\sqrt{n}\mathcal{I}\left(\theta_0\right)^{\frac{1}{2}}$ is replaced with the square root of the Fisher information matrix for a sample of size n.

Moreover, this theorem offers no measure of accuracy of the approximation for any finite sample size, and thus does not indicate whether it gives a better approximation than the standard asymptotic one in practice. This is typical of results on approximate distributions of test statistics. To highlight the theorem's potential for improved testing, in subsequent sections we present simulation results indicating that this distribution outperforms the standard asymptotic one in our example models T1 and T3.

Though the above theorem is stated for the likelihood ratio statistic, this is but one member of the *power-divergence family* of goodness-of-fit statistics of Cressie and Read [10]. For multinomially distributed data, with appropriate assumptions on the null model, all members of the family converge in distribution to the same asymptotic distribution. Thus the theorems and results in this paper are potentially useful for all members of the family. Although the Neyman-Pearson lemma (Neyman and Pearson [22]) states that the likelihood ratio test is the uniformly most powerful test for simple hypotheses, Cressie and Read [11] highlighted that in other scenarios other family members, such as Pearson's chi-squared statistic, may be better approximated by a χ^2 distribution than the likelihood ratio statistic is. It is of interest to investigate the use of the distribution of Theorem 3.1 for these other statistics.

For using the above distribution for practical testing, it is essential to note that while θ_0 and $\mathcal{I}\left(\theta_0\right)$ may be consistently estimated using the MLE $\hat{\theta}_0$ (Florescu [15]), the factors of $\sqrt{n}$ that appear with them in the specification of W in Theorem 3.1 produce quantities that are not consistently estimable. In Section 7 we return to this issue for our example models, discussing several approaches to handling it.

We emphasize that Theorem 3.1 can be expected to give a useful approximate distribution only when the normal approximation it depends upon is good. For instance for the models T1 and T3, with $\Theta = \Delta^2$, this is only when no counts of topologies are likely to be small, which occurs when the true parameter is away from the simplex's bounding triangle in a sense dependent upon sample size. If the true parameter is near a vertex of the triangle, then even for a large sample one may obtain very low frequencies of two of the three tree topologies, and must use other approaches.

4. Application to Model T3

We now apply Theorem 3.1 to determine an approximate distribution for the likelihood ratio statistic when testing the model T3 vs. an alternative of "no species tree". More formally, for $t^{(i)}$ the branch length in species tree $i \in \{1,2,3\}$ and taking $\phi_0^{(i)} = e^{-t^{(i)}} \in (0,1]$, the hypotheses are:

$$\begin{aligned} H_0\colon \quad \Theta_0 &= \left\{\left(1-\frac{2}{3}\phi_0^{(1)},\, \frac{1}{3}\phi_0^{(1)},\, \frac{1}{3}\phi_0^{(1)}\right)\right\} \cup \left\{\left(\frac{1}{3}\phi_0^{(2)},\, 1-\frac{2}{3}\phi_0^{(2)},\, \frac{1}{3}\phi_0^{(2)}\right)\right\} \\ &\qquad \cup \left\{\left(\tfrac{1}{3}\phi_0^{(3)},\, \tfrac{1}{3}\phi_0^{(3)},\, 1-\tfrac{2}{3}\phi_0^{(3)}\right)\right\}, \\ H_1\colon \quad \Theta_1 &= \Delta^2 \smallsetminus \Theta_0. \end{aligned}$$

We view the model $\tilde{\Theta} = \Theta_0 \cup \Theta_1 = \Delta^2$ as a subset of $\mathbb{R}^2$ through an appropriate affine transformation (see Appendix B for full details) which maps the singularity of Θ_0 to the origin and the true parameter point $\theta_0 = (1 - 2/3\phi_0,\, 1/3\phi_0,\, 1/3\phi_0)$, without loss of generality, to a point $(0, \mu_0)$ as in Figure 4. This affine transformation scales the simplex so that the normally distributed variable Y of Theorem 3.1 now has mean $(0, \mu_0)$ and identity covariance, where μ_0 is measured in standard deviations from the singularity and can be interpreted analogously for model T1. Unless $\theta_0 = (1/3, 1/3, 1/3)$ the affine transformation does not preserves angles. For other parameter values θ_0, the angle α_0 shown in Figure 4 is less than $\pi/6$.

We make one additional simplification, valid under the assumption that θ_0 is far from the triangle bounding the simplex $\tilde{\Theta}$, in a sense dependent on the sample size: the mass of the normal distribution of Y outside the image of $\tilde{\Theta}$ is negligible. This leads to the following proposition which is proved in Appendix B.

Proposition 4.1. For model T3, the likelihood ratio statistic for testing H_0 vs. H_1 at a true parameter point $\theta_0 = (1 - 2/3\phi_0,\, 1/3\phi_0,\, 1/3\phi_0)$ with sample size n is approximately distributed as the random variable

$$\tilde{\Lambda}_n = \min\left(Z^2 + \frac{1}{2}\left(1 - \operatorname{sgn}\left(\bar{Z}\right)\right)\bar{Z}^2,\ \left(\sin\alpha_0 Z + \cos\alpha_0 \operatorname{sgn}\left(Z\right)\bar{Z}\right)^2\right), \qquad (1)$$

where $Z \sim \mathcal{N}(0,1)$, $\quad \bar{Z} \sim \mathcal{N}(\mu_0, 1)$, $\quad \mu_0 = \sqrt{2n}\frac{1-\phi_0}{\sqrt{\phi_0(3-2\phi_0)}}$ and $\alpha_0 = \arctan\left(\frac{1}{\sqrt{3(3-2\phi_0)}}\right)$.

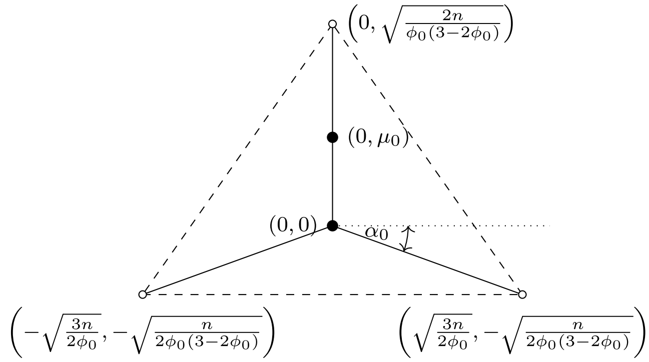
$\left(0, \sqrt{\frac{2n}{\phi_0(3-2\phi_0)}}\right)$
$(0, \mu_0)$
$(0,0)$
α_0
$\left(-\sqrt{\frac{3n}{2\phi_0}}, -\sqrt{\frac{n}{2\phi_0(3-2\phi_0)}}\right)$
$\left(\sqrt{\frac{3n}{2\phi_0}}, -\sqrt{\frac{n}{2\phi_0(3-2\phi_0)}}\right)$

Fig 4. *View of the image of model T3 after the affine transformation into* $\mathbb{R}^2$. *The singularity is mapped to the origin* $(0,0)$ *and the true parameter point* θ_0 *to* $(0,\mu_0)$. *The mapping is not conformal unless* θ_0 *is the singularity.*

Note that all the trigonometric functions in Equation (1) can be expressed as algebraic functions of ϕ_0.

To understand Equation (1), note that Z and $\bar{Z}$ are random variables corresponding to the x and y components of the sample point in the transformed space. The first argument then is simply the squared distance of $(Z, \bar{Z})$ to the vertical half-line in the null parameter space. The second argument is the squared distance to the other two half-lines, provided the closest point is not the origin. $(Z, \bar{Z})$ will be closest to the vertical half-line when the closest point on the other two half-lines is the origin. As shown in the proof, the distance predicted by the first argument of Equation (1) is then minimal. Thus, Equation (1) is the minimum squared Euclidean distance between the sample point and the transformed null parameter space.

By replacing $\operatorname{sgn}(Z)$ and $\operatorname{sgn}(\bar{Z})$ with ±1, the arguments are easily recognizable as χ^2 distributions. Moreover, suppose $\mu_0 > 0$ corresponds to any non-singular point in Θ_0, then as the sample size n goes to infinity, μ_0 also goes to infinity, causing the distribution of $\operatorname{sgn}(\bar{Z})$ to concentrate on 1, and the minimum in the formula tends toward selecting the first argument. It follows that $\tilde{\Lambda}_n$ is asymptotically χ^2_1-distributed as is the likelihood ratio statistic Λ_n, though for Λ_n the asymptotic behavior is typically determined more directly using the tangent cone approximation.

Now suppose $\mu_0 = 0$, so $\phi_0 = 1$; that is, the true parameter is the singularity. Then for any sample size n the approximate distribution in Equation (1) simplifies, with both Z and $\bar{Z}$ standard normal. Although this distribution is not a χ^2, it is exactly the standard asymptotic distribution, found using the tangent cone as in [13]. This is not surprising, as the tangent cone at this point locally agrees with the model itself.

Additional computations in Appendix B give the following.

Proposition 4.2. The probability density function for the random variable $\tilde{\Lambda}_n$ given for model T3 in Proposition 4.1 is, for $\lambda > 0$,

$$\begin{aligned} f_{\tilde{\Lambda}_n}(\lambda) = \frac{1}{2\sqrt{2\pi\lambda}} \Bigg[& \exp\left(-\frac{\lambda}{2}\right)\left(1 - \operatorname{erf}\left(\frac{1}{\sqrt{2}}\left(\sqrt{\lambda}\tan\beta_0 - \mu_0\right)\right)\right) \\ & + \exp\left(-\frac{1}{2}\left(\sqrt{\lambda} - \mu_0\cos\alpha_0\right)^2\right)\left(1 - \operatorname{erf}\left(\frac{1}{\sqrt{2}}\left(\sqrt{\lambda}\tan\beta_0 + \mu_0\sin\alpha_0\right)\right)\right) \\ & + \exp\left(-\frac{1}{2}\left(\sqrt{\lambda} + \mu_0\cos\alpha_0\right)^2\right)\left(1 - \operatorname{erf}\left(\frac{1}{\sqrt{2}}\left(\sqrt{\lambda}\tan\alpha_0 + \mu_0\sin\alpha_0\right)\right)\right)\Bigg], \end{aligned} \tag{2}$$

where $\mu_0 = \sqrt{2n}\frac{1-\phi_0}{\sqrt{\phi_0(3-2\phi_0)}}$, $\alpha_0 = \arctan\frac{1}{\sqrt{3(3-2\phi_0)}}$ and $\beta_0 = \frac{1}{2}\left(\frac{\pi}{2} - \alpha_0\right)$.

One can show that for $\phi_0 \in (0,1)$ as $n \to \infty$ Equation (2) gives the probability density function of χ^2_1.

Although Proposition 4.2 expresses the probability density function in terms of the error function, this density can quickly be integrated numerically to obtain a highly accurate approximation.

Figure 5 compares the density functions of Equation (2) at the singularity $\mu_0 = 0$ ($\phi_0 = 1$) and a regular point near the singularity $\mu_0 = 1$ ($\phi_0 \approx 0.9993$ when $n = 10^6$) to that of χ^2_1. At the singularity, the standard asymptotic density is given exactly by Equation (2), since there is no dependence on n. At all other points $\mu_0 > 0$, the standard asymptotic density is given by χ^2_1. The density plot for the parameter near the singularity, at $\mu_0 = 1$, lies between the other two plots, and can be considered a sort of interpolant that depends both on the sample size n and value of the parameter ϕ_0. Unlike the asymptotic densities, which have a jump discontinuity at the singularity, the density of Equation (2) is a continuous function of $\phi_0 \in [0,1)$ for any fixed n.

Simulations

We performed simulations to compare the use of the probability density function of Equation (2) to the χ^2_1 density for determining p-values of the likelihood ratio statistic when testing H_0 vs. H_1. We focused on true parameter values both at ($\mu_0 = 0$) and near the singularity ($\mu_0 = 1$, n varying). Near the singularity both distributions agree asymptotically, but at the singularity the χ^2_1 distribution is not the standard asymptotic distribution, while that of Equation (2) is. As the χ^2_1 distribution might naively be applied by an empiricist at the singularity, this last comparison is relevant. The value $\mu_0 = 1$ was chosen to be near enough, but not too near, to the singularity so that the χ^2_1 distribution and the asymptotic distribution at the singularity were both poor approximations. A range of

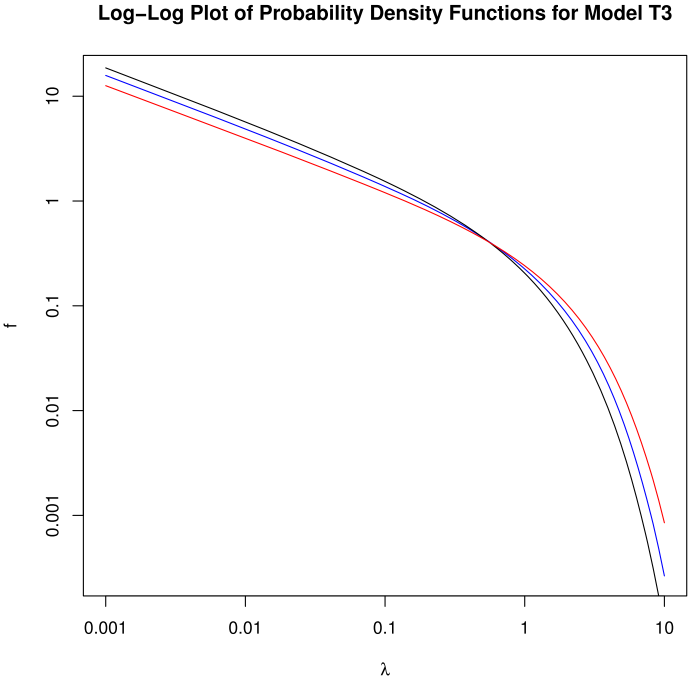

FIG 5. *Log-log plot of three approximating density functions over part of their support $\lambda \in (0, \infty)$. The asymptotic density of Equation (2) at the singularity $\mu_0 = 0$ ($\phi_0 = 1$) is in black; the approximating density at the nearby parameter value $\mu_0 = 1$ ($\phi_0 \approx 0.9993$ and $n = 10^6$) is in blue; and the asymptotic density at non-singular points, the χ_1^2 distribution, is in red. The blue approximating density can be viewed as an interpolant of the two asymptotic densities at and near the singularity.*

sample sizes was chosen, in part to demonstrate that near the singularity the χ_1^2 distribution can perform relatively poorly even for a large sample size, despite it being the asymptotic distribution.

Specifically, for the simulations presented in Figures 6, 7, (and later in Figures 9 and 10), $\theta_0 \in \Theta_0$ was chosen making $\mu_0 = 0$ or 1 for sample sizes $n = 30$, 10^3, 10^6. For each setting, μ_0, n, data was simulated from the multinomial distribution 10^6 times, and likelihood ratio statistics were calculated for each replicate. The probability density functions of Proposition 4.2 were used to determine p-values by numerical integration from the observed value of the statistic to infinity; p-values were also calculated using the χ_1^2 approximation by standard software. For each setting an empirical cumulative distribution function for 10^6 p-values was graphed.

In Figures 6 and 7, the discrete nature of the multinomial distribution is strongly apparent, particularly for $n = 30$. Since the possible likelihood ratio statistics form a discrete set and are unevenly spaced, jumps in the cumulative plots of p-values are unavoidable regardless of the simulation size.

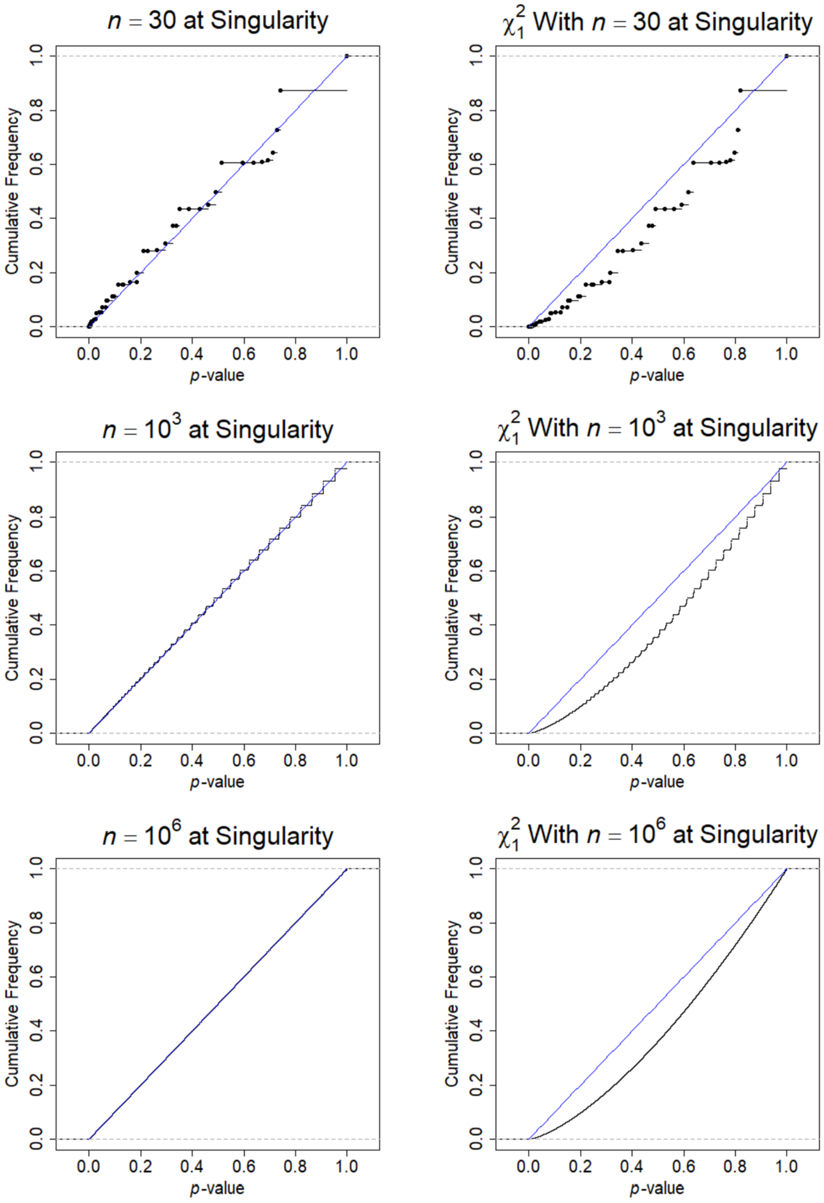

Fig 6. *Empirical cumulative distribution functions of p-values for the density function of Equation* (2) *(left column) and the* χ^2_1 *approximation (right column) for samples sizes* $n = 30, 10^3, 10^6$ *computed at the singularity,* $\mu_0 = 0$*, for model T3. The diagonal line, representing ideal behavior, is shown for comparison.*

Ideally, when Θ_0 has lower dimension than $\tilde{\Theta}$ (unlike Example 2.1) as for model T3, an approximate density function for the likelihood ratio statistic produces a simulated empirical cumulative distribution function of p-values close to $F_X(x) = x$ for $x \in (0, 1)$. The left column of Figure 6 shows that this holds for the density function of Equation (2) for the singularity, even for a relatively small sample size of $n = 30$. In contrast, this fails for the χ^2_1 distribution (which is not the asymptotic distribution), as seen in the right column.

In Figure 7, the results of these simulations are shown for the parameter near the singularity. Again, plots in the left column show that the density function of Equation (2) performs extremely well, even for a sample size of $n = 30$. The right column illustrates that the χ^2_1 distribution is a poor approximation for each of the three sample sizes, even though it is the standard asymptotic distribution. As an approximate density, the χ^2_1 performs better here than at the singularity where it is not the asymptotic distribution, but not as well as the approximating density of $\tilde{\Lambda}_n$. In summary, naively assuming the χ^2_1 distribution is an accurate approximation for the likelihood ratio statistic near (or at) a singularity can lead to inaccurate estimates of p-values.

Significantly, the right columns of Figures 6 and 7 suggest that the use of the χ^2_1 distribution gives a conservative test, as it produces larger p-values than desired, leading to rejecting H_0 less often than desired. Moreover, such a test is increasingly conservative closer to the singularity. This behavior has an intuitive geometric interpretation: When θ_0 is on the vertical line segment of Θ_0 and near, but not at, the singularity, then the observation can be substantially closer to an incorrect segment of Θ_0 than to the correct segment. The observation is then interpreted to be less extreme than it should be. Use of the χ^2_1 distribution then gives a larger p-value than desired.

5. Application to Model T1

We now examine our second example, model T1, in which the null hypothesis is that the species tree has a specific topology.

Our two hypotheses for this test are:

$$\begin{aligned} H_0\colon \quad & \Theta_0 = \left\{ \left(1 - \frac{2}{3}\phi_0, \frac{1}{3}\phi_0, \frac{1}{3}\phi_0 \right) \right\}, \text{ with } \phi_0 = e^{-t} \in (0, 1], \\ H_1\colon \quad & \Theta_1 = \Delta^2 \setminus \Theta_0. \end{aligned}$$

The model $\tilde{\Theta} = \Theta_0 \cup \Theta_1$ is again the open probability simplex Δ^2, which is viewed as a subset of $\mathbb{R}^2$ through the same affine transformation used for model T3. This is as depicted in Figure 4, but with the two non-vertical line segments erased.

Applying Theorem 3.1, an approximate distribution of the likelihood ratio statistic can be found. The proof of the following is given in Appendix C.

Proposition 5.1. For model T1, the likelihood ratio statistic for testing H_0 vs. H_1 at a true parameter point $\theta_0 = (1 - 2/3\phi_0, 1/3\phi_0, 1/3\phi_0)$ with sample size

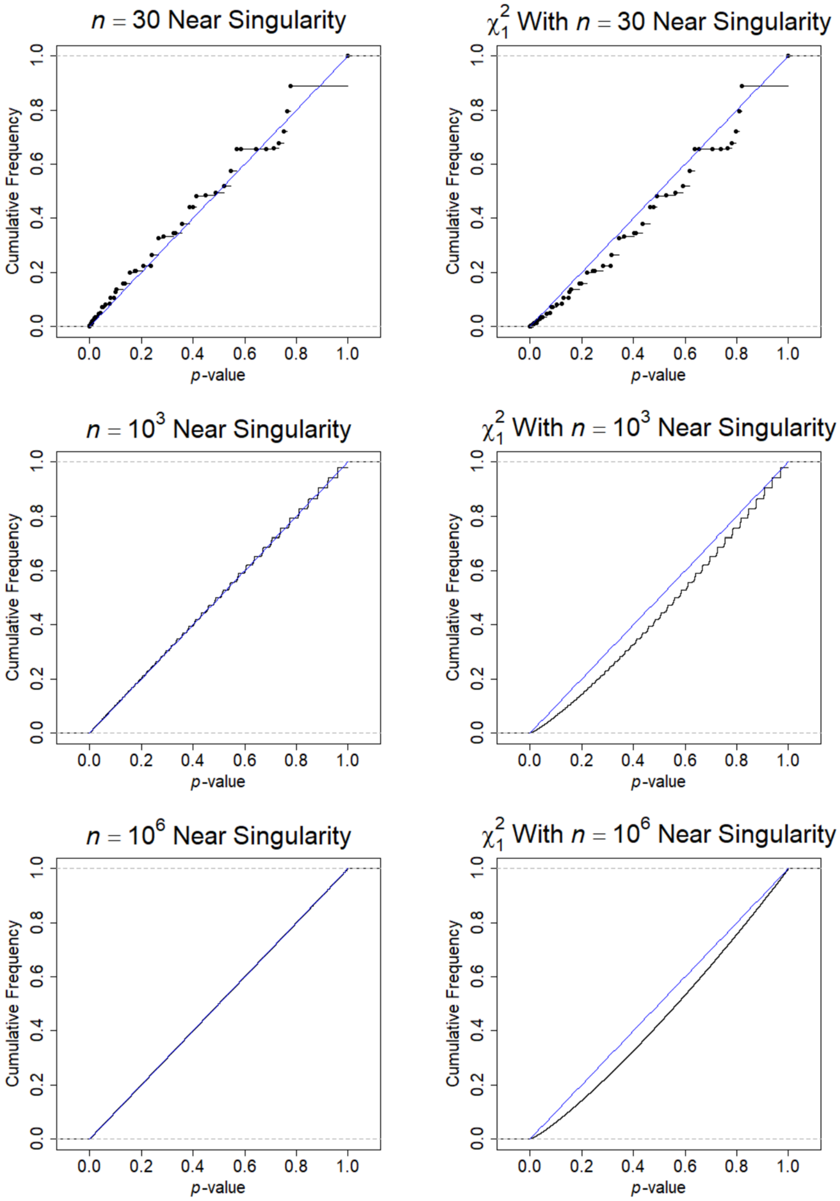

Fig 7. *Empirical cumulative distribution functions of p-values for the density function of Equation* (2) *(left column) and the* χ_1^2 *approximation (right column) for sample sizes* $n = 30, 10^3, 10^6$ *computed near the singularity,* $\mu_0 = 1$*, for model T3. The diagonal line, representing ideal behavior, is shown for comparison.*

n is approximately distributed as the random variable

$$\tilde{\Lambda}_n = Z^2 + \frac{1}{2}\left(1 - \operatorname{sgn}\left(\bar{Z}\right)\right)\bar{Z}^2,$$

where $Z \sim \mathcal{N}(0,1)$, $\bar{Z} \sim \mathcal{N}(\mu_0, 1)$ and $\mu_0 = \sqrt{2n}\frac{1-\phi_0}{\sqrt{\phi_0(3-2\phi_0)}}$.

Note that the distribution is the same as the first argument of the minimum in the distribution in Proposition 4.1 for model T3. This is expected as the first argument referred to the single line segment which is Θ_0 in this example.

Again, if $\operatorname{sgn}\left(\bar{Z}\right)$ was always positive then the distribution would be a χ^2_1 distribution, while if $\operatorname{sgn}\left(\bar{Z}\right)$ was always negative then it would be a χ^2_2 distribution. Further calculations in Appendix C yield the following.

Proposition 5.2. The probability density function of the random variable $\tilde{\Lambda}_n$ given for model T1 in Proposition 5.1 is, for $\lambda > 0$,

$$f_{\tilde{\Lambda}_n}(\lambda) = \frac{1}{4}\exp\left(-\frac{\lambda}{2}\right)\left[\sqrt{\frac{2}{\pi\lambda}}\left(1 + \operatorname{erf}\left(\frac{\mu_0}{\sqrt{2}}\right)\right) - \exp\left(-\frac{\mu_0^2}{2}\right)M_0\left(\mu_0\sqrt{\lambda}\right)\right], \tag{3}$$

where $M_0(x) = -\frac{2}{\pi}\int_0^{\frac{\pi}{2}}\exp(-x\cos\theta)\,d\theta$ is the modified Struve function 11.5.5 from Olver [23], and $\mu_0 = \sqrt{2n}\frac{1-\phi_0}{\sqrt{\phi_0(3-2\phi_0)}}$.

At the singularity, where $\mu_0 = 0$, Equation (3) gives the probability density function of $1/2\chi^2_1 + 1/2\chi^2_2$. This is as one expects from Example 1.2 of Drton [13]. One can also show that for $\phi_0 \in (0,1)$ as $n \to \infty$ Equation (3) gives the probability density function of χ^2_1, since $M_0(x) \to 0$ as $x \to \infty$.

Again the approximate probability density function can be integrated numerically quickly to obtain a highly accurate numerical approximation to the distribution.

Figure 8 gives a graphical comparison of the probability density functions of Equation (3) at $\mu_0 = 1$ ($\phi_0 \approx 0.9993$ and $n = 10^6$) and at $\mu_0 = 0$ (the probability density function $1/2\chi^2_1 + 1/2\chi^2_2$ at the boundary) to that of χ^2_1. The black and red densities are the standard asymptotic densities at and near the boundary, respectively. The graph for a parameter near the boundary ($\mu_0 = 1$) lies between those for the asymptotic distributions, interpolating them in a way dependent on both sample size n and parameter ϕ_0. Unlike the asymptotic distributions, which jump discontinuously at the singularity, the density of Equation (3) is a continuous function of ϕ_0.

Note that the χ^2_1 density (red curve) is closer to the approximate density (blue curve) in Figure 8 than in Figure 5, indicating it is closer to our distribution for T1 than for T3. This is not surprising, since the derivation of the asymptotic χ^2_1 is based on replacing the model with a single vertical line, which more closely matches the geometry of the model T1 than T3.

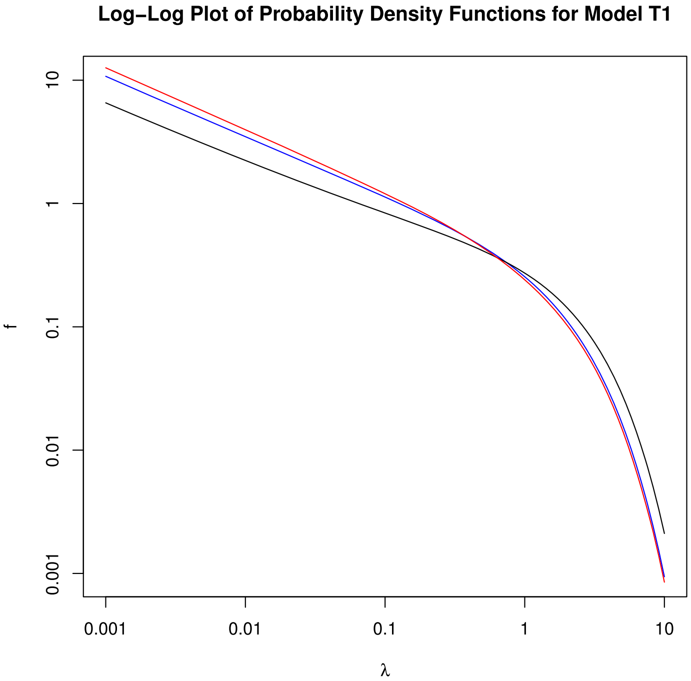

FIG 8. *Log-log plot of three probability density functions over part of their support, $\lambda \in (0, \infty)$. The density of Equation (3) at $\mu_0 = 1$ ($\phi_0 \approx 0.9993$ and $n = 10^6$) is in blue; the density of $1/2\chi_1^2 + 1/2\chi_2^2$ of the boundary is in black; and the density of the χ_1^2 distribution is in red. The black and red plots are the asymptotic distributions at and near the boundary, respectively.*

Simulations

The performance of the approximate density function of Proposition 5.2 was compared to the density function of the χ_1^2 distribution through simulations for model T1, similar to those previously described for T3.

In Figure 9, it can be seen that at the boundary our approximate density function outperforms the χ_1^2 approximation, which is biased towards smaller p-values. This is expected, since the distribution in Proposition 5.1 is the asymptotic distribution and χ_1^2 is not. We note that the χ_1^2 approximation rejects H_0 more often than it should and thus gives an anti-conservative test.

Near the boundary, as shown in Figure 10, our probability density function again fits the distribution of the likelihood ratio statistic better than the χ_1^2 does, though the improvement is small compared to that in Figure 9 for the boundary. This is expected as the χ_1^2 is now the asymptotic distribution. Moving away from the boundary (simulations not shown), the χ_1^2 distribution becomes a progressively better approximation, but remains biased towards smaller p-values. Thus the use of the χ_1^2 approximation leads to rejection of H_0 more often than it should, and is anti-conservative. Again, the χ_1^2 performs better for model T1 than for model T3 for some μ_0.

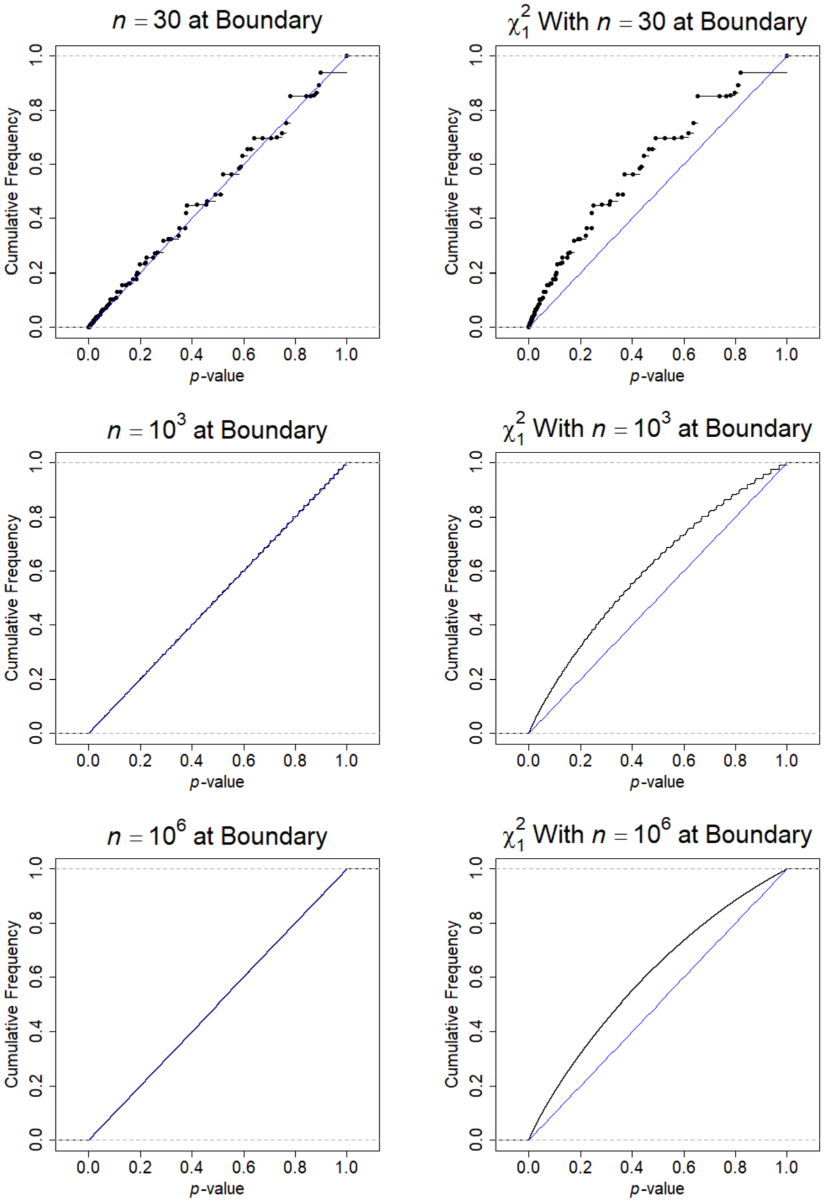

Fig 9. *Empirical cumulative distribution functions of p-values for the density function of Equation* (3) *(left column) and the* χ_1^2 *approximation (right column) for sample sizes* $n = 30, 10^3, 10^6$ *computed at the boundary,* $\mu_0 = 0$*, for model T1. The diagonal line, representing ideal behavior, is shown for comparison.*

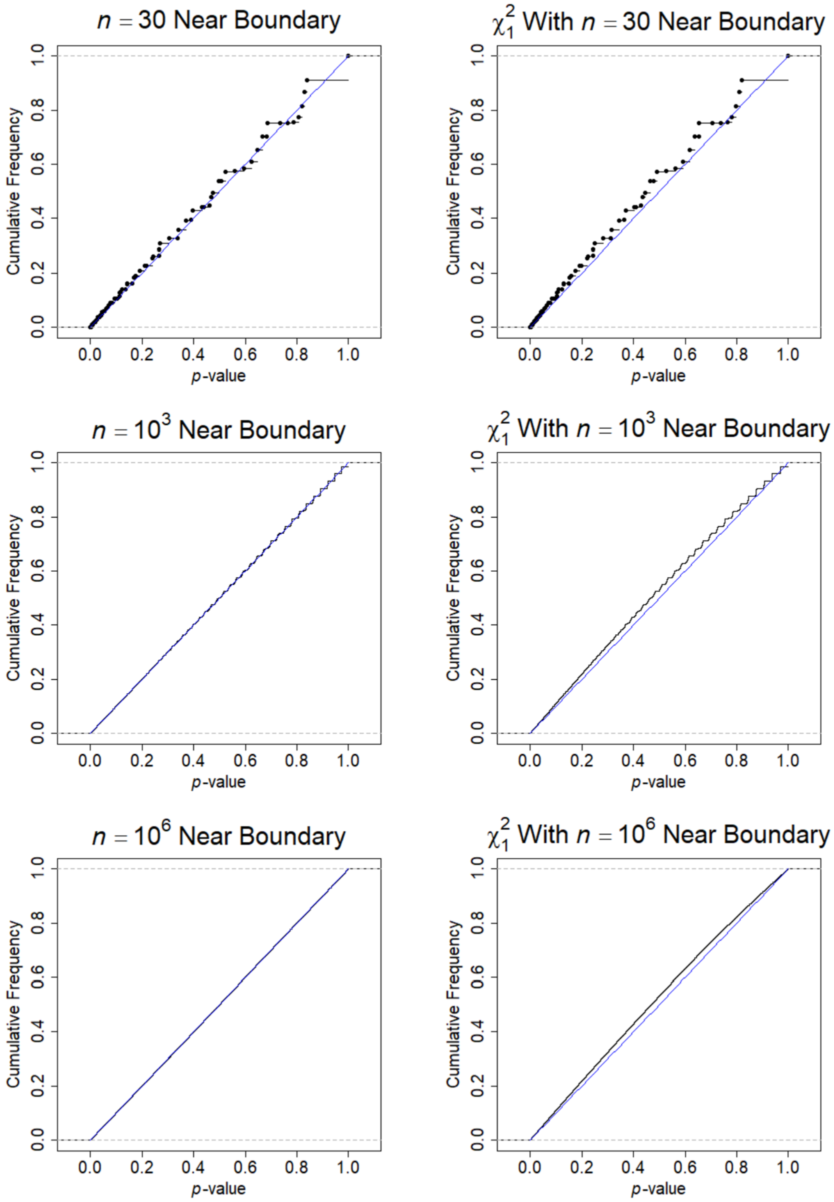

Fig 10. *Empirical cumulative distribution functions of p-values for the density function of Equation* (3) *and the* χ^2_1 *approximation (right column) for samples sizes* $n = 30, 10^3, 10^6$ *computed near the boundary,* $\mu_0 = 1$*, for model T1. The diagonal line, representing ideal behavior, is shown for comparison.*

The anti-conservative behavior of the χ_1^2 distribution is geometrically intuitive. For a true parameter θ_0 near the boundary point of Θ_0, some sample points will lie lower than the boundary, giving an MLE that is the boundary point. Such sample points are thus further from the MLE than they are from the vertical line extending Θ_0. However, the χ_1^2 distribution is appropriate for judging their squared distance from that line. This causes them to be viewed as more extreme than they should be, and their p-values to be calculated as smaller than desired.

6. Approximating likelihood ratio statistic distributions with χ^2

The distributions of Propositions 4.1 and 5.1 interpolate between the asymptotic distribution at the singularity or boundary, respectively, and the asymptotic χ_1^2 distribution far from the singularity or boundary. The further the true parameter point is from the singularity or boundary, the more accurate the χ_1^2 approximation is.

Indeed, while we have shown these approximate distributions for likelihood ratio statistics perform better than the asymptotic ones for finite sample sizes near the singularities and boundaries of our example models, it may still be desirable to use the asymptotic χ_1^2 distribution for testing sufficiently far from those points. The simpler form of these distributions and ready availability in standard software remains attractive. A natural problem, then, is how to decide when the simpler distribution is likely to lead to adequate performance in testing.

To approach this question quantitatively, we employ the *total variation distance* between our approximate distributions and the χ_1^2. The total variation distance between two continuous probability distributions F, G, with densities f, g, of support R, is

$$\delta\left(F, G\right) = \frac{1}{2}\int_R \mid f\left(\lambda\right) - g\left(\lambda\right) \mid d\lambda,$$

which can be interpreted as the maximum absolute difference of probabilities of events.

Using the distribution of Proposition 4.1 or Proposition 5.1, one can choose an acceptable upper bound ϵ on the total variation distance between this distribution and the χ_1^2. Then, using a numerical optimization routine, one can determine the values of ϕ_0, n for which this bound is not exceeded. The χ_1^2 approximation might be considered acceptable for such ϕ_0 and n.

Application to Model T1

For model T1, the dependence of the distribution from Proposition 5.1 on ϕ_0 and n is only through $\mu_0 = \mu_0\left(\phi_0, n\right)$, so let F_{μ_0} denote this distribution viewed as a function of μ_0. From the derivation of the density in Appendix C, it is clear that $\delta\left(F_{\mu_0}, \chi_1^2\right)$ is a decreasing function of μ_0. It is thus sufficient to determine numerically the value $\tilde{\mu}_0$ for which $\delta\left(F_{\tilde{\mu}_0}, \chi^2_1\right) = \epsilon$. Then $\mu_0 > \tilde{\mu}_0$ characterizes the parameters and sample sizes for which the χ^2_1 approximation might be considered acceptable.

Table 1 summarizes, for several choices of ϵ, the threshold value $\tilde{\mu}_0$. It also shows for several choices of sample size n, the corresponding thresholds $\phi_0 < \tilde{\phi}_0$ and $t > \tilde{t} = -\log\left(\tilde{\phi}_0\right)$, since μ_0 is a function of n and ϕ_0. For a given bound ϵ, larger sample sizes allow for shorter internal branches of the tree in Figure 2, while maintaining the χ^2_1 distribution as a reasonable approximation for the distribution of the likelihood ratio statistic.

Table 1

For model T1, the threshold values $\tilde{\mu}_0$ are given for which $\mu_0 > \tilde{\mu}_0$ ensures the total variation distance between the distribution of Proposition 5.1 and χ^2_1 is less than ϵ, for various ϵ. For a fixed sample size n, the thresholds are also given in terms of $\tilde{\phi}_0$ or $\tilde{t}$.

	$\epsilon = 5 \times 10^{-3}$, $\tilde{\mu}_0 = 1.84$		$\epsilon = 5 \times 10^{-4}$, $\tilde{\mu}_0 = 2.64$		$\epsilon = 5 \times 10^{-5}$, $\tilde{\mu}_0 = 3.28$	
n	$\tilde{\phi}_0$	$\tilde{t}$	$\tilde{\phi}_0$	$\tilde{t}$	$\tilde{\phi}_0$	$\tilde{t}$
30	0.748	0.291	0.642	0.443	0.565	0.572
10^2	0.863	0.147	0.802	0.220	0.754	0.283
10^3	0.958	0.0429	0.939	0.0626	0.924	0.0787
10^4	0.987	0.01320	0.981	0.0190	0.977	0.0237
10^5	0.996	0.00414	0.994	0.00594	0.993	0.00739
10^6	0.999	0.00130	0.998	0.00187	0.998	0.00233

Table 2 shows similar threshold values that ensure the null rejection probability for a test based on the χ^2_1 distribution exceeds a nominal level of $\alpha = 0.05$ by small amounts. As these calculations concern only the tail of the distribution, the thresholds are smaller than in Table 1.

Table 2

For model T1, the threshold values $\tilde{\mu}_0$ are given for which $\mu_0 > \tilde{\mu}_0$ ensures the exceedance in null rejection probability using the χ^2_1 is less than ϵ above $\alpha = 0.05$. For a fixed sample size n, the thresholds are also given in terms of $\tilde{\phi}_0$ or $\tilde{t}$.

	$\alpha = 0.05$, $\epsilon = 5 \times 10^{-3}$, $\tilde{\mu}_0 = 1.02$		$\alpha = 0.05$, $\epsilon = 5 \times 10^{-4}$, $\tilde{\mu}_0 = 1.83$		$\alpha = 0.05$, $\epsilon = 5 \times 10^{-5}$, $\tilde{\mu}_0 = 2.52$	
n	$\tilde{\phi}_0$	$\tilde{t}$	$\tilde{\phi}_0$	$\tilde{t}$	$\tilde{\phi}_0$	$\tilde{t}$
30	0.862	0.148	0.749	0.288	0.657	0.419
10^2	0.926	0.0770	0.864	0.146	0.812	0.209
10^3	0.977	0.0232	0.958	0.0426	0.942	0.0595
10^4	0.993	0.00724	0.987	0.0131	0.982	0.0181
10^5	0.998	0.00228	0.996	0.00411	0.994	0.00567
10^6	0.999	0.000719	0.999	0.00130	0.998	0.00179

Application to Model T3

For model T3, the dependence of the density of Proposition 4.2 on parameter ϕ_0 and sample size n is through both μ_0 and α_0. However, it is clear from the derivation in Appendix B that an upper bound on the variation distance is obtained by setting α_0 to its minimum value, $\alpha_0 = \arctan(1/3)$, for any value of μ_0. This simplifies the computations and leads to a conservative estimate of the threshold $\tilde{\mu}_0$. Table 3 summarizes thresholds found in this way.

TABLE 3

For model T3, conservative threshold values $\tilde{\mu}_0$ are given for which $\mu_0 > \tilde{\mu}_0$ ensures the total variation distance between the distribution of Proposition 4.1 and χ^2_1 is less than ϵ, for various ϵ. For a fixed sample size n, the thresholds are also given in terms of $\tilde{\phi}_0$ or $\tilde{t}$.

	$\epsilon = 5\times10^{-3}$, $\tilde{\mu}_0 = 2.74$		$\epsilon = 5\times10^{-4}$, $\tilde{\mu}_0 = 3.64$		$\epsilon = 5\times10^{-5}$, $\tilde{\mu}_0 = 4.40$	
n	$\tilde{\phi}_0$	$\tilde{t}$	$\tilde{\phi}_0$	$\tilde{t}$	$\tilde{\phi}_0$	$\tilde{t}$
30	0.629	0.463	0.525	0.645	0.449	0.802
10^2	0.795	0.230	0.727	0.319	0.672	0.398
10^3	0.937	0.0650	0.916	0.0879	0.898	0.108
10^4	0.980	0.0198	0.974	0.0264	0.968	0.0321
10^5	0.994	0.00617	0.992	0.00820	0.990	0.00993
10^6	0.998	0.00194	0.997	0.00258	0.997	0.00312

For model T3, a test based on the χ^2_1 distribution is conservative, as suggested by Figures 6 and 7 and as will be more formally established in the next section. Thus we give no analog of Table 2 for this model.

7. Hypothesis testing in practice

In using the distributions of Propositions 4.1 and 5.1 in a practical hypothesis test, one additional issue arises. Since the true parameter ϕ_0 is unknown, it is natural to use an estimate of it to determine the testing distribution. However, while the maximum likelihood estimator of ϕ_0 is consistent, it does not lead to a consistent estimator of the distribution parameter μ_0. Even though the variance in the estimate of ϕ_0 goes to zero as sample size $n \to \infty$, the factor of $\sqrt{n}$ in the formula for μ_0 results in the variance of its estimator not approaching zero.

In this section we explore methods of addressing this. These include using either the "least favorable" estimate over the full parameter space, or over a confidence interval for μ_0 [8, 29], or analogs of these approaches using drifting parameter sequences [3, 20].

7.1. Model T1

For model T1 we begin by examining drifting parameter sequences. As described by [20], near a boundary point of a parameter space one often finds that standard asymptotic distributions, obtained by holding parameters fixed and letting

$n \to \infty$, behave poorly for hypothesis testing, while "asymptotic distributions derived under appropriate drifting sequences of parameters often provide very good approximations to finite sample null distributions." A drifting sequence of parameters is, roughly, a sequence of parameter values γ_n that approaches a boundary point 0, in such a way that a limiting distribution exists when parameter γ_n is paired with sample size of n.

To be precise for the model T1, in the notation of Proposition 5.1 introduce a transformed parameter

$$\gamma = \frac{\sqrt{2}(1-\phi_0)}{\sqrt{\phi_0(3-2\phi_0)}} \in [0,\infty),$$

so that $\gamma = 0$ at the boundary point. For any fixed value of $\mu_0 \in [0,\infty)$, define a sequence $\gamma_n \to 0$ by

$$\mu_0 = \sqrt{n}\gamma_n.$$

Then $\{\gamma_n\}$ is a drifting parameter sequence, with *localization parameter* μ_0. By Proposition 5.1, the distribution along the drifting sequence is constant, so taking the limit is trivial. Thus for these distributions the concept of drifting sequences adds nothing new; the limit distributions along these drifting sequences with localization parameter μ_0 are exactly those of Proposition 5.1. This is not surprising, as the distributions were derived already accounting for the geometry of the model near the boundary, which is exactly what the limits along drifting sequences are intended to address.

As previously commented, though, we cannot consistently estimate μ_0 from data. A simple solution to this, the *least favorable* (LF) approach, is to adopt as a critical value for a test at level α the largest critical value at level α across all values of the localization parameter. This has also been called the size-corrected fixed critical value [3], and the resulting hypothesis test the p_{sup} test [29], but the idea goes back at least to [27]. The following corollary of Lemma C.1 of Appendix C is useful for determining the critical value.

Proposition 7.1. For localization parameter μ, let L_μ be the cdf of the distribution of Proposition 5.1 for model T1. For a given level $0 < \alpha < 1$ and localization parameter μ, let $CV_\mu(\alpha) = L_\mu^{-1}(1-\alpha)$ be the critical value with level α. Then $CV_\mu(\alpha)$ is a decreasing function of μ.

Thus if we consider, for some fixed α, the critical values $CV_\mu(\alpha)$ for all μ in some interval $[a,b]$, the largest will be from $\mu = a$, and thus using $CV_a(\alpha)$ as a cutoff for a test will have the smallest null rejection probability regardless of the value of μ in the interval.

In particular, the LF approach to testing for the model T1 is to always use the distribution for $\mu_0 = 0$. That is, one would use the distribution $1/2\chi^2_1 + 1/2\chi^2_2$, even though this gives a much more conservative test than the χ^2_1 which standard asymptotics suggests for all points except the boundary, and which in fact performs well when the true (unknown) parameter is far from the boundary point. This indicates the potential value of a confidence interval approach.

The Simple Bonferroni critical value of [20] is one such approach, and for T1 its use coincides with what is called the p^* test in [29]. It is based on first finding a confidence interval for μ_0 at an adjusted level, and then using the supremum of the critical values for μ_0 in this interval (or equivalently defining a p-value as the supremum over those values given by the μ_0 in the interval.)

Consider a sample of size n drawn from the model T1 with parameter $p_1 \in [1/3, 1]$ in the notation of Example 2.2. Let (n_1, n_2, n_3) be the counts of the three rooted topologies in the sample, with n_1 that matching the true tree. Then the maximum likelihood estimator of p_1 is

$$\hat{p}_1 = \max\left(\frac{n_1}{n}, \frac{1}{3}\right), \tag{4}$$

which is consistent, but biased upward. Transforming parameters, this gives consistent MLEs of ϕ_0 and γ as

$$\hat{\phi}_0 = \frac{3}{2}\left(1 - \hat{p}_1\right), \;\; \hat{\gamma} = \frac{\sqrt{2}(1-\hat{\phi}_0)}{\sqrt{\hat{\phi}_0(3-2\hat{\phi}_0)}},$$

which are biased downward and upward respectively. Since we will need to work with parameterizations in terms of both p_1 and γ, define the increasing function

$$\gamma(p) = \frac{3p-1}{3\sqrt{(1-p)p}} \tag{5}$$

so that $\gamma = \gamma(p_1)$ and $\hat{\gamma} = \gamma(\hat{p}_1)$. Then an estimator of μ_0 is $\hat{\mu}_0 = \sqrt{n}\hat{\gamma}$, though this is not consistent.

Nonetheless we can construct a confidence interval for μ_0 from $\hat{\mu}_0$. To ultimately obtain the least conservative test, we prefer a confidence interval that, when possible, excludes those values of μ_0 which produce the largest critical values. In light of Proposition 7.1 this means we seek a 1-sided confidence interval of the form $[a, \infty)$.

As the count n_1 is binomially distributed with parameters p_1 and n, if the estimator were simply n_1/n, and not as in Equation (4), a confidence interval $[b, \infty)$ for p_1 at level $1-\alpha$ can be obtained by well-known methods, with lower bound $b = b(\hat{p}_1, n; \alpha)$. If we require that the MLE $\hat{p}_1$ be in the confidence interval, then for all levels below 0.5, the lower bound is $b = \hat{p}_1$, effectively raising the level to 0.5. Since $p_1 \geq 1/3$, the interval is then modified to $[a, \infty)$ where $a = \max(b, 1/3)$. Since $\gamma(p)$ is increasing, a confidence interval for μ_0 is

$$I_\alpha(\hat{\gamma}, n) = [\sqrt{n}\gamma(a), \infty).$$

For our simulations, we took a similar approach, using a truncated normal approximation to the distribution of $\hat{\mu}_0$ directly.

The Simple Bonferroni critical value for level α and choice of $\delta \in [0, \alpha]$ is then defined as

$$CV(\hat{\gamma}, n, \alpha, \delta) = \sup_{\mu \in I_{\alpha-\delta}(\hat{\gamma}, n)} CV_\mu(1-\delta).$$

By Proposition 7.1, for the model T1 this becomes

$$CV(\hat{\gamma}, n, \alpha, \delta) = CV_{\sqrt{n}\gamma(a)}(1-\delta), \tag{6}$$

where a depends on $\hat{\gamma}, n, \alpha - \delta$. Note that for a fixed choice of δ/α, one can search for an α for which the critical value equals the observed statistic, and view this α as a p-value.

Rejecting the null hypothesis based on the Simple Bonferroni critical value ensures good behavior of the test in the following sense: Define the *asymptotic size* of a test with test statistic T under the null hypothesis and any critical value CV_n (possibly dependent on the sample and sample size n) as

$$AsySz(CV_n) = \limsup_{n\to\infty} \sup_{\gamma\in[0,\infty)} \mathbb{P}_\gamma(T > CV_n).$$

Then as proved in [20], the Simple Bonferroni critical value satisfies

$$\delta \leq AsySz(CV(\hat{\mu}_0, n, \alpha, \delta)) \leq \alpha.$$

More informally, asymptotically the test will result in rejection of the null hypothesis with probability at most α. Note that the LF critical value ensures an asymptotic size of α, so this result alone does not indicate the Simple Bonferroni critical value results in an improved test. Moreover, the asymptotic size focuses on avoiding type I error, and says nothing about power. The Bonferroni critical value is chosen in hopes of producing a more powerful test than the LF one, by using the data to guide one to a potentially smaller critical value.

Using Equation (6), if $\delta = \alpha$, the Simple Bonferroni critical value becomes $CV_0(1-\alpha)$ which is the LF choice. At the other extreme of $\delta = 0$, the critical value becomes infinite. Although [20] provides no theoretical guidance as to a good choice of δ when one has no information about the localization parameter, a value of $\delta = 0.9\alpha$, or the consideration of multiple values of δ, is suggested. Optimally, we would choose the critical value to be the infimum over all values of δ with an appropriate size-correction factor, or an approximation obtained by considering a fine grid of values (that is, the Minimum Bonferroni method of [20] without considering choices of β from the Adjusted Bonferroni method). To reduce computational time, however, through simulations we can determine optimal choices of δ for various μ_0 in advance, and then use the value of δ determined by $\hat{\mu}_0$. We call this compromise approach the Minimum Simple Bonferroni method.

The idea of using a confidence interval for μ_0 is pushed further in the definition of the Adjusted Bonferroni critical value, the precise definition of which can be found in [20] and depends upon a choice of $\beta \in [0,1]$. Through simulation of the joint behavior of the statistic and the parameter estimate over the confidence interval, it chooses a new level so that, under certain assumptions, the test has correct asymptotic size. Similar to the Minimum Simple Bonferroni method, we can determine optimal choices of β for various μ_0 in advance, and then find β determined by $\hat{\mu}_0$. We call this the Minimum Adjusted Bonferroni method. (This is similar to the Minimum Bonferroni Method of [20] but without its additional size correction. Simulations in Table 4 and Table 5 suggest this correction is not needed for the models considered here.)

We also consider a naive hypothesis test, which simply uses the estimate $\hat{\mu}_0$ from data as the value of μ_0 for determining the distribution of Proposition 5.1, despite the fact that this is not consistently estimated. Doing so may not give an asymptotic size below the chosen level α (and in fact does not for true $\mu_0 \approx 0$), so this test can be anti-conservative. However, when the true parameter μ_0 is far from the boundary, so with high probability $\hat{\mu}_0$ is also, this should behave like the χ^2_1. We refer to the test obtained in this way as the *ML-estimate* test. In simulations (not shown), a parametric bootstrap test, in which the MLE determines the parameter value, closely follows the behavior of this test for all values of μ_0 we investigated. This is as expected since the normal approximation used in deriving the results of Proposition 5.1 approximates the bootstrap process.

A final test, useful for comparison, uses the χ^2_1 distribution that a standard asymptotic argument suggests is appropriate at all non-boundary points.

Table 4 and Figure 11 show the results of simulations with these seven methods, giving the empirical null rejection rates for a range of values of the true parameter μ_0 when the nominal level is $\alpha = 0.05$, and plots of these rejection rates for several values of α. This table and figure indicate that use of the χ^2_1 gives a strongly anti-conservative test for a large interval of true parameters near the boundary point. The LF approach, except for values of μ_0 quite near the boundary, is strongly conservative. The Simple Bonferroni and Minimum Simple Bonferroni methods are also strongly conservative over a wide range, though for μ_0 very far from the boundary they approach the desired rejection rate. The Adjusted Bonferroni method with $\beta = 0.5$ attains a null rejection rate much closer to the nominal one near the boundary, but far from it is matched by the Simple and Minimum Simple Bonferroni methods. The Minimum Adjusted Bonferroni method comes closest to the nominal level of all methods with asymptotic size guarantees. Finally, the ML-estimate method is conservative for all values of μ_0 except those in a small interval near the boundary, and except on that interval comes closest to the desired rejection rate of all the methods.

7.2. Model T3

For the model T3 we also consider drifting parameter sequences $\gamma_n \to 0$ defined for a choice of $\mu_0 \in [0, \infty)$ by

$$\mu_0 = \sqrt{n}\gamma_n.$$

Here $\gamma = 0$ corresponds to the singularity of the model. For such a drifting parameter sequence $\{\gamma_n\}$, the distributions of Proposition 4.1 for samples of size n are *not* identical, but a limit distribution exists. Since for any fixed $\mu_0 > 0$ the γ_n approach 0, reparameterizing in terms of ϕ_0 one obtains a parameter sequence approaching 1. Using the formula in Proposition 4.1 for α_0, this gives a sequence of angles approaching $\pi/6$. The limit distribution, then, is given by Equation 1 of Proposition 4.1, where $Z, \bar{Z}$ depend on μ_0 as stated there, but

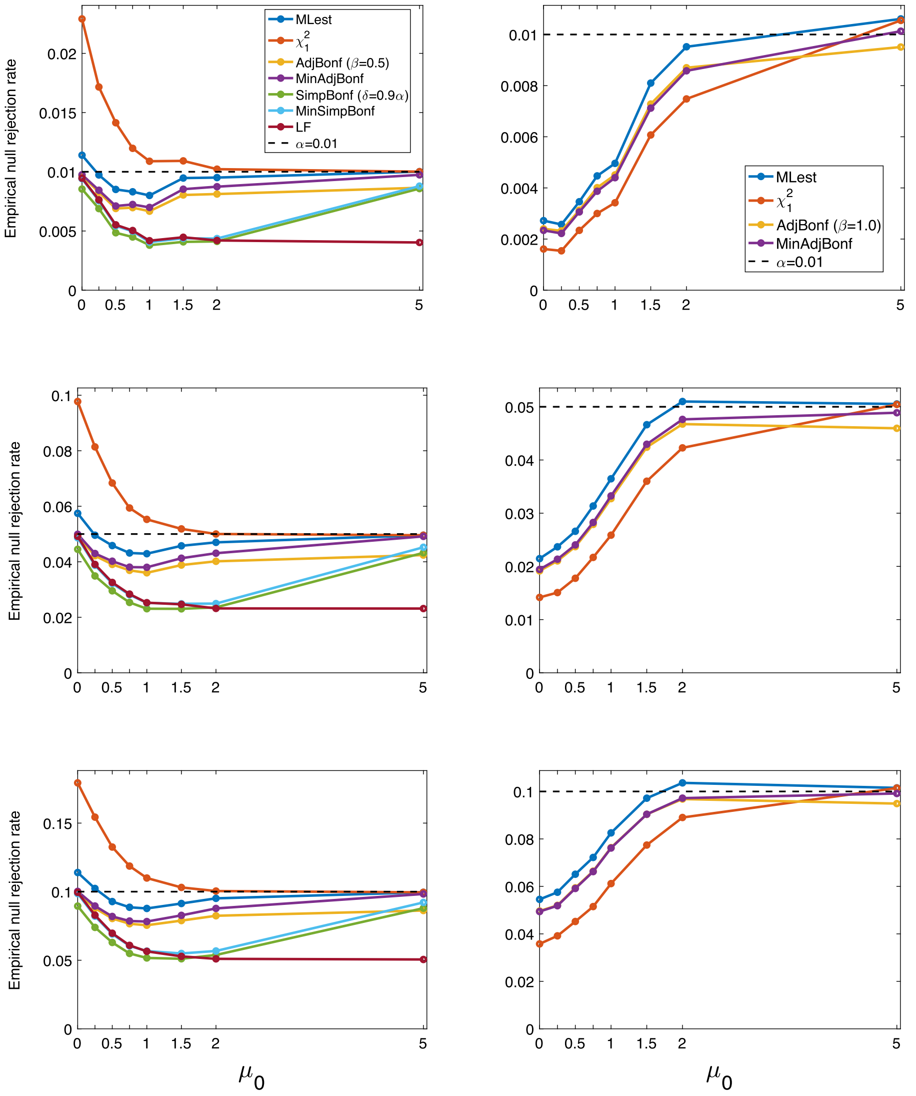

FIG 11. *Empirical null rejection rates for various hypothesis test methods for the model T1 (left) and T3 (right), at nominal levels* $\alpha = 0.01$ *(top),* 0.05 *(middle), and* 0.1 *(bottom). Note differences in vertical scales between columns. Plotted values for* $\alpha = 0.05$ *are as given in Tables 4 and 5. These were obtained from* 10^5 *repetitions of simulations with* $n = 10^6$.

Table 4

Empirical null rejection rates for various hypothesis test methods for the model T1, at nominal level $\alpha = 0.05$. Data was generated under T1 with $n = 10^6$ for the shown values of μ_0, with 10^5 repetitions to calculate rejection rates. For the Simple Bonferroni method, $\delta = 0.9\alpha$, and for the Adjusted Bonferroni method $\beta = 0.5$. Note that the ML estimate and χ^2_1 methods have no asymptotic guarantees of rejection at less than α. The Bonferroni methods are: Simple Bonferroni (SB), Minimum Simple Bonferroni (MSB), Adjusted Bonferroni (AB) and Minimum Adjusted Bonferroni (MAB).

	μ_0							
Method	0	0.25	0.50	0.75	1.0	1.5	2.0	5.0
MLest	0.0575	0.0485	0.0460	0.0435	0.0428	0.0457	0.0463	0.0501
LF	0.0492	0.0385	0.0331	0.0285	0.0257	0.0245	0.0224	0.0232
χ^2_1	0.0973	0.0780	0.0688	0.0601	0.0549	0.0520	0.0494	0.0501
SB	0.0448	0.0347	0.0297	0.0256	0.0234	0.0234	0.0235	0.0438
MSB	0.0489	0.0381	0.0327	0.0281	0.0256	0.0251	0.0247	0.0459
AB	0.0497	0.0419	0.0394	0.0375	0.0365	0.0391	0.0396	0.0431
MAB	0.0499	0.0422	0.0398	0.0382	0.0375	0.0408	0.0420	0.0489

where $\alpha_0 = \pi/6$. The density function is as in Equation 2 of Proposition 4.2, but with $\alpha_0 = \beta_0 = \pi/6$. Note that this limit distribution is not in the family of distributions given by these propositions, except in the case when $\mu_0 = 0$. Since the distributions of Proposition 4.1 have already accounted for the geometry of the model, the notion of a limit on a drifting sequence changes little.

As a corollary of Lemma B.1 of Appendix B we have the following.

Proposition 7.2. For localization parameter μ, let L_μ be the cdf of the limiting distribution along the drifting parameter sequence above, of the distributions of Proposition 4.1 for model T3. For a given level $0 < \alpha < 1$, let $CV_\mu(\alpha) = L_\mu^{-1}(1-\alpha)$ be the critical value with level α. Then $CV_\mu(\alpha)$ is an increasing function of μ.

For the distributions of Proposition 4.1, numerical computations show a similar relationship of critical values. That is, for fixed n as the parameter μ increases (or as ϕ decreases), the critical values at a fixed level increase. Because the angle $\alpha_0 = \alpha_0(\phi)$ in the formula for the distribution changes, this is considerably more difficult to formally prove than is Proposition 7.2.

In particular for the model T3 the LF approach to testing in either the framework of limit distributions of drifting sequences, or using the distributions of Proposition 4.1, is to always use the distribution obtained by letting $\mu_0 \to \infty$. In both cases, this is the χ^2_1 distribution appropriate for parameter values far from the singularity by standard arguments. However, this is quite conservative when the true parameter lies near the singularity.

To obtain a less conservative test through a confidence interval for the parameter, we seek to exclude values that would produce the largest critical values, so Proposition 7.2 and its following paragraph indicate we should find a 1-sided confidence interval for μ_0 of the form $[0, a]$. Because of the non-standard form of our model we sketch how this is done.

Consider a sample of size n drawn from the model T3 with parameters (p_1, p_2, p_3) in the notation of Example 2.3, on any of the three line segments of the model. Let (n_1, n_2, n_3) be the counts of the three rooted topologies in the sample. Then the maximum likelihood estimator of (p_1, p_2, p_3) might lie on any of the three line segments, so letting $p_{\max} = \max(p_1, p_2, p_3)$, one can show

$$\hat{p}_{\max} = \max\left(\frac{n_1}{n}, \frac{n_2}{n}, \frac{n_3}{n}\right),$$

which is consistent. Transforming parameters, this gives consistent ML estimators

$$\hat{\phi}_0 = \frac{3}{2}\left(1 - \hat{p}_{\max}\right), \; \hat{\gamma} = \frac{\sqrt{2}(1 - \hat{\phi}_0)}{\sqrt{\hat{\phi}_0(3 - 2\hat{\phi}_0)}}.$$

Using Equation (5), $\gamma = \gamma(p_1)$ and $\hat{\gamma} = \gamma(\hat{p}_{\max})$. We also consider the estimator of $\mu_0 = \sqrt{n}\gamma$ given by $\hat{\mu}_0 = \sqrt{n}\hat{\gamma}$.

The counts (n_1, n_2, n_3) are trinomially distributed with parameters (p_1, p_2, p_3) and n, and hence the $(n_1/n, n_2/n, n_3/n)$ are approximately normally distributed. To obtain a 1-sided confidence interval at level $1-\alpha$ of the form $[1/3, a]$ for $p_{\max}$, we seek the infimum of those a such that for all $p_{\max} > a$,

$$\mathbb{P}(X_{\max} < \hat{p}_{\max}) < \alpha, \tag{7}$$

where $X_{\max}$ is the maximum entry of a random draw (X_1, X_2, X_3) from the normal. The probability here can be calculated by integrating the appropriate normal density over a triangle bounded by the line segments orthogonal to the three model line segments at the points where $p_{\max} = \hat{p}_{\max}$ on them. A confidence interval for μ_0 is then

$$I_\alpha(\hat{\gamma}, n) = [0, \sqrt{n}\gamma(a)].$$

To ensure $\hat{\mu}_0$ lies in the interval, we increase its upper bound if necessary.

By Proposition 7.2, the Simple Bonferroni critical value in the drifting sequence setting is then

$$CV(\hat{\gamma}, n, \alpha, \delta) = CV_{\sqrt{n}\gamma(a)}(1 - \delta), \tag{8}$$

where a depends on $\hat{\gamma}, n, \alpha - \delta$ as above.

One can also consider the Adjusted Bonferroni and Minimum Adjusted Bonferroni tests while using the distributions of Proposition 4.1, and the naive approach of simply using the distribution whose parameter is the (inconsistent) ML estimate.

Table 5 and Figure 11 show the results of simulations with four methods, giving the empirical null rejection rates for a range of values of the true parameter μ_0. These were performed using the distributions of Proposition 4.1, and not their limits along drifting sequences, as there was little difference between the two. In our simulations we found no instances in which the Simple Bonferroni or Minimum Simple Bonferroni methods attained a null rejection probability higher than the χ_1^2, so those methods are omitted from the results.

Table 5

Empirical null rejection rates for various hypothesis test methods for the model T3, at nominal level $\alpha = 0.05$. Data was generated under T3 with $n = 10^6$ for the shown values of μ_0, with 10^5 repetitions to calculate rejection rates. For the Adjusted Bonferroni method $\beta = 1$. Note that the ML estimate and χ_1^2 methods have no asymptotic guarantees of rejection at less than α. The Bonferroni methods are: Adjusted Bonferroni (AB) and Minimum Adjusted Bonferroni (MAB).

	μ_0							
Method	0	0.25	0.50	0.75	1.0	1.5	2.0	5.0
MLest	0.0215	0.0237	0.0266	0.0314	0.0365	0.0466	0.0510	0.0505
χ_1^2	0.0142	0.0151	0.0178	0.0217	0.0259	0.0360	0.0423	0.0504
AB	0.0192	0.0210	0.0237	0.0279	0.0327	0.0424	0.0468	0.0460
MAB	0.0194	0.0214	0.0240	0.0283	0.0333	0.0430	0.0476	0.0489

As the plots show, all methods were quite conservative for μ_0 near the singularity, with the χ_1^2 (which is also the LF) being the most so. The Adjusted Bonferroni and Minimum Adjusted Bonferroni methods have better and very similar performance near the singularity, though far away from it the Minimum Adjusted Bonferroni method attains a null rejection rate closer to the desired one. The ML-estimate method, despite a lack of a theoretical guarantee, comes closest to the desired null rejection rate on a fairly large interval including the singularity, but appears to be slightly anti-conservative for μ_0 near 2, at least for larger α.

8. Discussion

The distributions commonly used in hypothesis testing are obtained through standard asymptotics with sample size $n \to \infty$, and may discontinuously jump between regular points of a model and its boundaries and singularities. As the examples of models T1 and T3 illustrate, even at regular points near boundaries and singularities such standard approximations as the χ_1^2 may behave poorly in testing. Although increasing sample size may lead to better performance at any specific point, the discontinuous behavior of such an asymptotic distribution means a region of poor performance can remain, though it shrinks in size. While Drton [13] commented that convergence to the asymptotics can be slow near a boundary or singularity, we further emphasize that the nonuniformity of the rate of convergence poses even more of a problem. Unless we have an *a priori* quantitative bound separating the true parameter from the singularities and boundaries, no finite sample size can be found which will lead to uniformly good performance of a standard asymptotic approximation.

Moreover, depending on the model, use of the χ^2 approximation may lead to either conservative or anti-conservative tests (or both, in different regions), depending on the geometry of the model beyond the singularity or boundary. Thus no simple rule can be adopted for adjusting one's test. Theorem 3.1 suggests an alternative approach of avoiding the approximation of the model by its tangent cone inherent in the derivation of the standard asymptotic distribution, and using a different approximate distribution dependent on both the true parameter and the sample size. For our example models this performed well, as illustrated by our simulations.

Even for our models, there are a number of hypothesis tests not presented here for which Theorem 3.1 will be useful. For instance, one may wish to test whether data fits a null hypothesis of a particular tree, model T1, vs. an alternative of the other trees, model T3$\smallsetminus$T1. Failure to reject the null hypothesis for each of the three choices of T1 would, in biological terminology, be interpreted as a soft polytomy, where an unresolved (star) tree represents ignorance of the true resolution. Similarly, one may wish to test whether data fits a simple hypothesis of an unresolved tree, $\theta_0 = (1/3, 1/3, /1/3)$, vs. an alternative of a resolved tree, model $T_3 \smallsetminus \{\theta_0\}$. For this test failure to reject the null hypothesis would, in biological terminology, be interpreted as a hard polytomy, where an unresolved tree represents what are believed to be true relationships.

Within phylogenetics, another possible use of Theorem 3.1 is for conducting hypothesis tests for distance data to fit a tree. An evolutionary distance $d\,(a, b)$ is typically a numerical measure of the amount of mutation between two species a and b, and under certain modeling assumptions should in expectation match the sum of lengths of branches between them on a tree. The 3-point condition states that for an ultrametric tree to exist relating species a, b, c, the expectations of $d\,(a, b)$, $d\,(a, c)$, and $d\,(b, c)$ must have the two largest equal, with the smallest pair indicating the correct tree topology. This is similar to models T1 and T3, with the inequality reversed, except that the distances may have any non-negative values. Again the model has a singularity or boundary.

Several works [18, 19] have proposed statistical tests involving distances. For instance, Gu and Li [18] tested the 3-point condition by focusing on the difference of the two distances that are assumed to be equal under H_0. Arguing that this difference is asymptotically normally distributed, a Z-test is performed. However, when all three distances are near equal, as they would be near the singularity or boundary point corresponding to a star tree, this test becomes inaccurate, as the smallest value may well not correspond to the true topology. Just as with models T1 and T3, the test could either be anti-conservative or conservative, depending on whether the null hypothesis was of a specific 3-species ultrametric tree or of any of the three possible trees, respectively.

Testing for genetic admixture between species and populations using the D statistic (also known as the ABBA-BABA test), as was done originally to understand Human-Neanderthal interactions [17, 14], is a means to reject a 4-species tree model of evolution. However this statistic depends only on counts of data *not* in accord with evolution on the presumed tree, which is similar to using only the counts n_2 and n_3 of tree topologies not in accord with the true tree for our model T1. In situations where the count n_1 of data concordant with the presumed true tree is of similar magnitude, the framework for testing we give here is likely more appropriate.

Our example models have rather special structure making them amenable to our approach. Since Θ_0 was locally linear, except at the singularities and boundaries, we were able to compute explicit density functions for the relevant distributions, so that using them was no more difficult than using a χ^2. For a model given by a k-dimensional half-space embedded in a $(k+1)$-dimensional space, the arguments for model T1 can be modified slightly to obtain an explicit density. More generally, it is likely the calculations generalize to give explicit densities for a larger class of models defined by linear equality and inequality constraints on parameters. As hypotheses of this form are common, this potentially gives a wide range of applications. Although models similar to T3, in which several linear half-spaces or spaces are joined at a singularity, are likely to be rarer, they should also be tractable in our framework. Although we do not believe explicit calculations such as those done here can be done for all models, within a restricted domain where they can be performed they may give improved tools.

With a broader perspective, Theorem 3.1 suggests that whenever the asymptotic distribution performs badly for hypothesis testing, one might do better by using a distribution taking the local geometry of the model into account in a more subtle way than just through the tangent cone. For instance, if a model were described by a curve in the plane, one should expect that even at regular points the asymptotic distribution may be less useful in regions of high curvature, where the tangent cone approximation of the model is poor. However, unlike in the case of singularities or boundaries one should be able to work out a sample size ensuring a reasonable fit by a χ^2, as long as the curvature is bounded. If obtaining a data set of that size is not possible, then even if the distribution of Theorem 3.1 cannot be computed, first approximating the model by a simpler curve with similar curvature, such as an appropriate polynomial, and then using the theorem might lead to a better distribution for use in hypothesis testing.

Acknowledgements

This work is supported by the US National Institutes of Health grant R01 GM117590, awarded under the Joint DMS/NIGMS Initiative to Support Research at the Interface of the Biological and Mathematical Sciences. The authors thank the associate editor and anonymous reviewers whose detailed suggestions improved the final form of this paper.

Appendix A: The multispecies coalescent model

We briefly introduce the multispecies coalescent model, which underlies models T1 and T3 of Examples 2.2 and 2.3. This model, introduced by Pamilo and Nei [24] (see also [25]), extends the Kingman coalescent model of population genetics, from applying to a single population, to a tree of populations, called a species tree. It is the fundamental model of the biological process of *incomplete lineage sorting*, by which gene trees of sampled lineages can fail to match the structure of the tree relating species overall. Incomplete lineage sorting is one of several processes that can make inference of species relationships from genetic data difficult. An example of a single such gene tree sampled for a particular species tree is shown in Figure 2.

The Kingman coalescent models a finite number of lineages, traced backward in time within a single population, as they merge, or coalesce, at common ancestors. The most convenient time scale is in coalescent units t, where $\Delta t = \Delta\tau / N(\tau)$, with time τ measured in number of generations and $N(\tau)$ the population size. In these units, if k lineages are sampled, the time to the first coalescence of the first pair of lineages is exponentially distributed with rate $\binom{k}{2}$. The pair that coalesces is then chosen uniformly at random. Then the coalescent process begins again with one less lineage, and hence rate $\binom{k-1}{2}$. Wakeley [31] provides a comprehensive introduction to this model.

While in population genetics, one often views the Kingman model as running until all lineages coalesce to a single one, in the multispecies coalescent that may not happen within a single population, which has a finite duration.

The parameters of the multispecies coalescent model are a rooted metric species tree, with branch lengths given in coalescent units. The branches of the species tree should be thought of as representing unstructured populations, which stretch back in time until they merge with another population. We also consider a population ancestral to the root of the species tree, which is considered to have infinite length, so that lineages in it coalesce into one with probability 1.

Specific finite numbers of lineages are to be sampled from each species' population at the leaves of the tree. Then the Kingman coalescent model applies for the duration of that population to its parental node in the tree. At that point, there are fewer lineages if any coalescent event occurred, but we gain more lineages from the other branch of the species tree which descends from that node. The combined collection of lineages then starts a new coalescent process on the branch leading towards the root. Continuing in this way, eventually a finite number of lineages reach the root, where a final Kingman coalescent process leads to a rooted metric gene tree. Finally, ignoring branch lengths yields a sampled rooted topological gene tree.

While for species trees with many species it is difficult to compute the probability of any gene tree (e.g., Rosenberg [26]), in the applications based on models T1 and T3, the species tree has only three species, and only one lineage is sampled from each. With only one lineage per species, coalescence can occur only in the internal branch of the tree or "above-the-root", and not in any terminal branch. Thus the only relevant branch length is the internal one.

Suppose that the true species tree is a rooted three species tree $((a, b) : t, c)$, as shown in Figure 2. There are three possible gene tree topologies,

$$AB|C,\ AC|B,\ BC|A.$$

In this case, the probability of gene trees discordant from the species tree are easiest to compute. For instance the gene trees $AC|B$ and $BC|A$ can only form if no coalescence occurs except above the root. From the exponential distribution of coalescent times, the probability of no coalescence of two lineages in a branch of length t is e^{-t}. Then, with three lineages present at the root, due to the exchangeability of lineages, the formation of each of the three rooted trees must have equal probability of 1/3. Thus $p_{AC|B} = 1/3e^{-t}$. The same argument gives $p_{BC|A} = 1/3e^{-t}$, which thus implies $p_{AB|C} = 1 - 2/3e^{-t}$.

Appendix B: Model T3

Here we prove Proposition 4.1 and Proposition 4.2 concerning the model T3.

For a fixed sample size, multinomial distributions form a regular exponential family if $\tilde{\Theta} = \Delta$ is the open simplex. The regularity conditions of Drton [13] are then satisfied, and thus Theorem 3.1 applies.

Since $\tilde{\Theta} = \Delta^2$ lies on a plane in $\mathbb{R}^3$, we first apply an affine transformation $M : \mathbb{R}^3 \to \mathbb{R}^2$,

$$M = \begin{pmatrix} 0 & -\frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \sqrt{\frac{2}{3}} & -\frac{1}{\sqrt{6}} & -\frac{1}{\sqrt{6}} \end{pmatrix},$$

to map $\tilde{\Theta}$ isometrically to the plane, sending the singularity to the origin. This maps a true parameter point, say $\theta_0 = \theta_0^{(1)} = (1 - 2/3\phi_0,\ 1/3\phi_0,\ 1/3\phi_0)$ without loss of generality, to $\left(0,\ \sqrt{\frac{2}{3}}\,(1-\phi_0)\right)$. Computing the Fisher information matrix $\mathcal{I}\,(\theta_0)$ for a sample of size $n = 1$ for θ_0 in planar coordinates, we obtain the transformation matrix

$$\sqrt{n}\mathcal{I}\,(\theta_0)^{\frac{1}{2}} = \begin{pmatrix} \sqrt{\frac{3n}{\phi_0}} & 0 \\ 0 & \sqrt{\frac{3n}{\phi_0(3-2\phi_0)}} \end{pmatrix},$$

which we apply to the planar image of Δ^2. The point θ_0 is mapped to $(0, \mu_0)$ with

$$\mu_0 = \sqrt{2n}\frac{1-\phi_0}{\sqrt{\phi_0\,(3-2\phi_0)}}.$$

Under these transformations the null parameter space Θ_0 is mapped non-conformally, provided $\phi_0 \in (0,1)$, to three line segments emanating from the origin, one to $\left(0, \sqrt{\frac{2n}{\phi_0(3-2\phi_0)}}\right)$ passing through the true parameter point $(0, \mu_0)$, and others to $\left(\pm\sqrt{\frac{3n}{2\phi_0}}, -\sqrt{\frac{n}{2\phi_0(3-2\phi_0)}}\right)$. (The parameter value $\phi_0 = 1$ corresponds to the singularity in Θ_0 and the transformation is conformal in this instance.) The full parameter space Δ^2 is mapped to the interior of the convex hull of the three points given above. See Figure 4.

The angle $\alpha_0 > 0$ formed between the positive x-axis and the line segment joining the origin to $\left(\sqrt{\frac{3n}{2\phi_0}}, -\sqrt{\frac{n}{2\phi_0(3-2\phi_0)}}\right)$, as shown is Figure 4, is $\alpha_0 = \arctan\left(\frac{1}{\sqrt{3(3-2\phi_0)}}\right)$, and varies from $\pi/6$ for $\phi_0 = 1$ down to $\arctan(1/3)$ as

$\phi_0 \to 0$. Letting $\gamma_0 = \tan(\alpha_0)$, in the transformed space the image of Θ_0 is contained in the union of the half-lines $\{(0, y) \mid y \geq 0\}$ and $y = -\gamma_0 \operatorname{sgn}(x)\, x$.

By Theorem 3.1, the approximate distribution of the likelihood ratio statistic is the distribution of the minimum squared Euclidean distance between a normal sample, $\mathcal{N}((0, \mu_0), I)$, and three line segments in the transformed space. Assuming that θ_0 is not too close to the boundary of $\tilde{\Theta}$ in a sense dependent on sample size, little of the mass of $\mathcal{N}((0, \mu_0), I)$ is outside the image of the simplex. Thus, for the remainder of the argument, we replace these line segments with half-lines emanating from the singularity $(0, 0)$.

Denote the marginal probability distributions of the bivariate normal sample by $Z \sim \mathcal{N}(0, 1)$ and $\bar{Z} \sim \mathcal{N}(\mu_0, 1)$. We next determine the minimum squared distance of a sample point $\left(Z, \bar{Z}\right)$ to the three half-lines.

Consider first the half-line $\{(0, y) \mid y \geq 0\}$. If $\bar{Z}$ is non-negative, then the squared Euclidean distance is Z^2, while if $\bar{Z}$ is negative, then the squared distance is $Z^2 + \bar{Z}^2$. Thus the squared Euclidean distance is

$$Z^2 + \frac{1}{2}\left(1 - \operatorname{sgn}\left(\bar{Z}\right)\right)\bar{Z}^2. \tag{9}$$

Now consider the half-lines $y = -\gamma_0 \operatorname{sgn}(x)\, x$ and denote the closest point to $\left(Z, \bar{Z}\right)$ by $(X, -\gamma_0 \operatorname{sgn}(X) X)$. Assuming $X \neq 0$, then $\operatorname{sgn}(X) = \operatorname{sgn}(Z)$, and minimizing

$$(Z - X)^2 + \left(\bar{Z} + \gamma_0 \operatorname{sgn}(Z)\, X\right)^2$$

yields

$$X = \frac{1}{1 + \gamma_0^2}\left(Z - \gamma_0 \operatorname{sgn}(Z)\, \bar{Z}\right).$$

Substituting into the previous expression gives the squared distance

$$\frac{\gamma_0^2}{1 + \gamma_0^2}\left(Z + \frac{1}{\gamma_0}\operatorname{sgn}(Z)\, \bar{Z}\right)^2 = \left(\sin\alpha_0 Z + \cos\alpha_0 \operatorname{sgn}(Z)\, \bar{Z}\right)^2. \tag{10}$$

In the case $X = 0$, the closest point to the two half lines is the origin. This can occur only when $\bar{Z} \geq 0$, so the squared distance to the two half-lines is at least Z^2, which is the squared distance to the vertical half-line given in Equation (9). Moreover, it can be shown that Z^2 is at most the value given in Equation (10) in this case.

It follows that the approximate distribution of the likelihood ratio statistic is that of the random variable

$$\tilde{\Lambda}_n = \min\left(Z^2 + \frac{1}{2}\left(1 - \operatorname{sgn}\left(\bar{Z}\right)\right)\bar{Z}^2,\ \left(\sin\alpha_0 Z + \cos\alpha_0 \operatorname{sgn}(Z)\, \bar{Z}\right)^2\right),$$

as given in Proposition 4.1.

To determine the probability density function for the approximate distribution of Proposition 4.1, we let $G_X(x)$ denote the cumulative distribution function of the (non-squared) Euclidean distance. This can be found by integrating the bivariate normal distribution $\mathcal{N}((0, \mu_0), I)$ over the tube of points within distance x from the transformed Θ_0, as shown in Figure 12. Calculations are simplified by the fact that the tubular region in Figure 12 has bilateral symmetry, as does the normal distribution.

Due to symmetry, we need only integrate over the shaded regions in the figure. Let $L_i = L_i(x)$ denote the half-lines forming the outer boundaries of these regions. Denote by β_0 the angle formed by the line segment joining the origin to the point of intersection of L_1 and L_2. This angle has measure $\beta_0 = 1/2\,(\pi/2 - \alpha_0)$.

With this setup, $G_X(x) = 2\,(G_1(x) + G_2(x) + G_3(x))$, where G_i is the integral over the shaded strip R_i, and the density of the Euclidean distance is

$$g_X(x) = 2\left(\frac{d}{dx}G_1(x) + \frac{d}{dx}G_2(x) + \frac{d}{dx}G_3(x)\right).$$

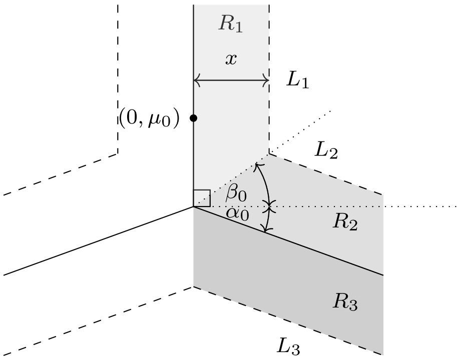

FIG 12. *The region of integration for $G_X(x)$ is between the dashed lines. The integral is evaluated as three integrals, over each of the shaded regions R_i.*

Considering $\frac{d}{dx}G_1(x)$ first, one sees that this derivative is the integral of the normal density over boundary L_1. We show this formally using polar coordinates:

$$\begin{aligned}\frac{d}{dx}G_1(x) &= \int_{\beta_0}^{\frac{\pi}{2}} \frac{d}{dx}\int_0^{\frac{x}{\cos\beta}} \frac{1}{2\pi}\exp\left(-\frac{1}{2}\left(r^2 - 2\mu_0 r\sin\beta + \mu_0^2\right)\right) r\,dr\,d\beta \\ &= \int_{\beta_0}^{\frac{\pi}{2}} \frac{1}{2\pi}\exp\left(-\frac{1}{2}\left(\frac{x^2}{\cos^2\beta} - 2\mu_0\frac{x}{\cos\beta}\sin\beta + \mu_0^2\right)\right)\frac{x}{\cos^2\beta}\,d\beta \\ &= \frac{1}{2\pi}\exp\left(-\frac{1}{2}x^2\right) \\ &\qquad \int_{\beta_0}^{\frac{\pi}{2}} \exp\left(-\frac{1}{2}\left(x^2\tan^2\beta - 2\mu_0 x\tan\beta + \mu_0^2\right)\right)\frac{x}{\cos^2\beta}\,d\beta.\end{aligned}$$

Substituting $y = x \tan \beta$ gives

$$\begin{aligned}\frac{d}{dx} G_1(x) &= \frac{1}{2\pi} \exp\left(-\frac{1}{2}x^2\right) \int_{x \tan \beta_0}^{\infty} \exp\left(-\frac{1}{2}(y-\mu_0)^2\right) dy \\ &= \frac{1}{2\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right)\left(1 - \operatorname{erf}\left(\frac{1}{\sqrt{2}}(x \tan \beta_0 - \mu_0)\right)\right).\end{aligned}$$

More briefly, over R_2 we have

$$\frac{d}{dx} G_2(x) = \int_{L_2} f(z, \bar{z})\, dx,$$

where f is the density of the bivariate normal. To evaluate this, we reflect about the line $y = \tan \beta_0 \, x$, mapping the mean of the Gaussian to $(\mu_0 \cos \alpha_0, -\mu_0 \sin \alpha_0)$, and sending L_2 to a vertical half-line (x, y), with $y \geq x \tan \beta_0$, so

$$\begin{aligned}\frac{d}{dx} G_2(x) =& \int_{x \tan \beta_0}^{\infty} \frac{1}{2\pi} \exp\left(-\frac{1}{2}\left((x - \mu_0 \cos \alpha_0)^2 + (y + \mu_0 \sin \alpha_0)^2\right)\right) dy \\ =& \frac{1}{2\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x - \mu_0 \cos \alpha_0)^2\right) \\ & \left(1 - \operatorname{erf}\left(\frac{1}{\sqrt{2}}(x \tan \beta_0 + \mu_0 \sin \alpha_0)\right)\right).\end{aligned}$$

Finally, since the same reflection maps L_3 to the vertical half-line $(-x, y)$ with $y \geq x \tan \alpha_0$,

$$\begin{aligned}\frac{d}{dx} G_3(x) =& \int_{x \tan \alpha_0}^{\infty} \frac{1}{2\pi} \exp\left(-\frac{1}{2}\left((-x - \mu_0 \cos \alpha_0)^2 + (y + \mu_0 \sin \alpha_0)^2\right)\right) dy \\ =& \frac{1}{2\sqrt{2\pi}} \exp\left(-\frac{1}{2}(x + \mu_0 \cos \alpha_0)^2\right) \\ & \left(1 - \operatorname{erf}\left(\frac{1}{\sqrt{2}}(x \tan \alpha_0 + \mu_0 \sin \alpha_0)\right)\right).\end{aligned}$$

Summing these three expressions and multiplying by 2, we obtain the density $g_X(x)$ for the distance. After a change of variable to convert to the squared Euclidean distance, the random variable $\tilde{\Lambda}_n$ has density function

$$\begin{aligned}f_{\tilde{\Lambda}_n}(\lambda) = \frac{1}{2\sqrt{2\pi\lambda}} \Bigg[& \exp\left(-\frac{1}{2}\lambda\right)\left(1 - \operatorname{erf}\left(\frac{1}{\sqrt{2}}\left(\sqrt{\lambda} \tan \beta_0 - \mu_0\right)\right)\right) \\ &+ \exp\left(-\frac{1}{2}\left(\sqrt{\lambda} - \mu_0 \cos \alpha_0\right)^2\right)\left(1 - \operatorname{erf}\left(\frac{1}{\sqrt{2}}\left(\sqrt{\lambda} \tan \beta_0 + \mu_0 \sin \alpha_0\right)\right)\right) \\ &+ \exp\left(-\frac{1}{2}\left(\sqrt{\lambda} + \mu_0 \cos \alpha_0\right)^2\right)\left(1 - \operatorname{erf}\left(\frac{1}{\sqrt{2}}\left(\sqrt{\lambda} \tan \alpha_0 + \mu_0 \sin \alpha_0\right)\right)\right)\Bigg],\end{aligned}$$

as given in Proposition 4.2.

For use in Section 7, we further investigate relationships between the distributions for different values of μ_0. Proposition 7.2 is a direct corollary of the following.

Lemma B.1. *Consider the limit of the distributions of Proposition 4.1 along a drifting parameter sequence defined by $\mu = \sqrt{n}\gamma_n$ for a fixed μ, and let $L_\mu(x)$ denote its cdf. Then for fixed x, the function $L_\mu(x)$ is strictly decreasing in μ.*

Proof. We follow a similar line of reasoning as used for proving Proposition 4.1, to show $\frac{\partial}{\partial \mu} L_\mu(x) < 0$.

With fixed $\alpha_0 = \pi/6$, let $G_{\mu_0}(x)$ denote the cdf for the non-squared Euclidean distance to the three lines depicted in Figure 12 for a standard normal with mean $(0, \mu_0)$. This is the integral of the normal density over the region depicted in Figure 12. Then $G_{\mu_0+\epsilon}(x)$ can be calculated as the integral of the same normal, with mean $(0, \mu_0)$, over a region obtained by shifting the region of Figure 12 downward by ϵ. Thus, for small ϵ, $G_{\mu_0+\epsilon}(x) - G_{\mu_0}(x)$ is the integral over two thin strips along the lower edges of the inverted V of the region minus two small strips along the upper edges of the inverted V. Since these strips have width $\frac{\sqrt{3}}{2}\epsilon$, with f denoting the density, we have

$$G_{\mu_0+\epsilon}(x) - G_{\mu_0}(x) \approx \sqrt{3}\epsilon \left(\int_{L_3} f\, ds - \int_{L_2} f\, ds \right),$$

so

$$\frac{\partial}{\partial \mu_0} G_{\mu_0}(x) = \sqrt{3} \left(\int_{L_3} f\, ds - \int_{L_2} f\, ds \right).$$

Since $\alpha_0 = \beta_0 = \frac{\pi}{6}$, by calculations in the proof of Proposition 4.1 above,

$$\frac{\partial}{\partial \mu_0} G_{\mu_0}(x) = \frac{\sqrt{3}}{2\sqrt{2\pi}} \left(1 - \operatorname{erf}\left(\frac{1}{\sqrt{2}} \left(\frac{1}{\sqrt{3}} x + \frac{1}{2}\mu_0 \right) \right) \right)$$
$$\left(\exp\left(-\frac{1}{2} \left(x + \frac{\sqrt{3}}{2}\mu_0 \right)^2 \right) - \exp\left(-\frac{1}{2} \left(x - \frac{\sqrt{3}}{2}\mu_0 \right)^2 \right) \right).$$

Since $|x - \frac{\sqrt{3}}{2}\mu_0| \leq |x + \frac{\sqrt{3}}{2}\mu_0|$ for all $x, \mu_0 \geq 0$, this shows $G_{\mu_0}(x)$ is decreasing in μ_0 for all x. Following a change of variables to the squared Euclidean distance, the claim is established. □

Appendix C: Model T1

We now prove Propositions 5.1 and 5.2, using the transformation and notation of Appendix B. Proposition 5.1 follows immediately by a simple modification to the argument in Appendix B. See Equation (9).

For Proposition 5.2, let $g_X(x)$ denote the probability density function for the (non-squared) distance x between a sample point $(Z, \bar{Z})$ and the mean $(0, \mu_0)$. Then $g_X(x) = \frac{d}{dx} G_X(x)$ is given by the integral of the Gaussian over the dashed curves shown in Figure 13. To compute this, we integrate over the dashed boundaries of R_1 and R_2 depicted in the figure. By symmetry,

$$g_X(x) = \frac{d}{dx} G_X(x) = 2\frac{d}{dx} G_1(x) + \frac{d}{dx} G_2(x),$$

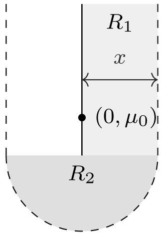

Fig 13. *The region of integration for model T1.*

where G_i is the contribution to the cdf over region R_i.

For R_1,

$$\begin{aligned}\frac{d}{dx}G_1(x) &= \int_0^\infty \frac{1}{2\pi}\exp\left(-\frac{1}{2}\left(x^2+(y-\mu_0)^2\right)\right)dy\\ &= \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{1}{2}x^2\right)\left(1+\operatorname{erf}\left(\frac{\mu_0}{\sqrt{2}}\right)\right).\end{aligned}$$

On R_2, using polar coordinates and C_2 for the dashed semi-circle, we find

$$\begin{aligned}\frac{d}{dx}G_2(x) &= \int_{C_2}\frac{1}{2\pi}\exp\left(-\frac{1}{2}\left(x^2+(y-\mu_0)^2\right)\right)d\sigma\\ &= \frac{1}{2\pi}x\exp\left(-\frac{1}{2}\left(x^2+\mu_0^2\right)\right)\int_\pi^{2\pi}\exp\left(\mu_0 x\sin\theta\right)d\theta\\ &= -\frac{1}{2}x\exp\left(-\frac{1}{2}\left(x^2+\mu_0^2\right)\right)M_0\left(\mu_0 x\right),\end{aligned}$$

where the last line is obtained after a change of variables, and $M_0(z)$ is the modified Struve function 11.5.5 from Olver [23].

Summing, and making a change of variable to the squared Euclidean distance, we find the probability density function for $\tilde{\Lambda}_n$ is

$$f_{\tilde{\Lambda}_n}(\lambda) = \frac{1}{4}\exp\left(-\frac{\lambda}{2}\right)\left[\sqrt{\frac{2}{\pi\lambda}}\left(1+\operatorname{erf}\left(\frac{\mu_0}{\sqrt{2}}\right)\right)-\exp\left(-\frac{\mu_0^2}{2}\right)M_0\left(\mu_0\sqrt{\lambda}\right)\right],$$

as given in Proposition 5.2.

For use in Section 7, we further investigate relationships between the distributions for different values of μ_0. Proposition 7.1 is a direct corollary of the following.

Lemma C.1. *Let L_{μ_0} denote the cdf of the distribution of Proposition 5.1. For fixed x, the function $L_{\mu_0}(x)$ is strictly increasing in μ_0.*

Proof. For fixed $x > 0$, if μ_0 is increased to $\mu_0 + \epsilon$ for positive ϵ, the region of integration for computing $L_{\mu_0+\epsilon}(x)$ is as shown in Figure 13, but with the center of the Gaussian moved upward to $(0, \mu_0 + \epsilon)$. Equivalently, we can compute the integral by instead considering a Gaussian centered at $(0, \mu_0)$, but extending the region R1 downward by ϵ, and moving the half disc R2 downward by ϵ. Since this enlarged region contains the region for computing $L_{\mu_0}(x)$ as well as an additional region of positive measure, the claim follows. □

References

[1] Allman, E. S., Degnan, J. H. and Rhodes, J. A. (2011). Identifying the Rooted Species Tree from the Distribution of Unrooted Gene Trees under the Coalescent. *J. Math Biol.* **6** 833–862. MR2795698

[2] Andrews, D. W. K. (2000). Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica* **68** 399–405. MR1748009

[3] Andrews, D. W. K. and Guggenberger, P. (2009a). Hybrid and size-corrected subsampling methods. *Econometrica* **77** 721–762. MR2531360

[4] Andrews, D. W. K. and Guggenberger, P. (2009b). Incorrect Asymptotic size of subsampling procedures based on post-consistent model selection estimators. *J. Econometrics* **152** 19–27. MR2562760

[5] Andrews, D. and Guggenberger, P. (2010). Asymptotic size and a problem with subsampling and with the m out of n bootstrap. *Econometric Theory* **26** 426–468. MR2600570

[6] Bartolucci, F. (2006). Likelihood inference for a class of latent Markov models under linear hypotheses on the transition probabilities. *J. R. Statist. Soc. B* **68** 155–178. MR2188980

[7] Bartolucci, F., Forcina, A. and Dardanoni, V. (2001). Positive quadrant dependence and marginal modeling in two-way tables with ordered margins. *J. Am. Stat. Assoc.* **96** 1497–1505. MR1946593

[8] Berger, R. L. and Boos, D. D. (1994). P values maximized over a confidence set for nuisence parameters. *J. Am. Stat. Assoc.* **89** 1012–1016. MR1294746

[9] Chernoff, H. (1954). On the distribution of the likelihood ratio. *The Annals of Mathematical Statistics* **25** 573–578. MR0065087

[10] Cressie, N. A. and Read, T. R. (1984). Multinomial Goodness-of-Fit Tests. *Journal of the Royal Statistical Society. Series B (Methodological)* 440–464. MR0790631

[11] Cressie, N. A. and Read, T. R. (1989). Pearson's X^2 and the Loglikelihood Ratio Statistic G^2: A Comparative Review. *International Statistical Review/Revue Internationale de Statistique* 19–43.

[12] Degnan, J. H. and Rosenberg, N. A. (2009). Gene tree discordance, phylogenetic inference and the multispecies coalescent. *Trends in Ecology & Evolution* **24** 332–340.

[13] Drton, M. (2009). Likelihood Ratio Tests and Singularities. *The Annals of Statistics* 979–1012. MR2502658

[14] Durand, E. Y., Patterson, N., Reich, D. and Slatkin, M. (2011). Testing for Ancient Admixture between Closely Related Populations. *Mol Biol Evol.* **28** 2239–2252.

[15] Florescu, I. (2014). *Probability and Stochastic Processes.* John Wiley & Sons. MR3241344

[16] Gaither, J. and Kubatko, L. (2016). Hypothesis tests for phylogenetic quartets, with applications to coalescent-based species tree inference. *Journal of Theoretical Biology* **408** 179–186. MR3548963

[17] Green, R. E., Krause, J., Briggs, A. W., Maricic, T., Stenzel, U. et al. (2010). A Draft Sequence of the Neandertal Genome. *Science* **328** 710–722.

[18] Gu, X. and Li, W.-H. (1996). Bias-corrected paralinear and LogDet distances and tests of molecular clocks and phylogenies under nonstationary nucleotide frequencies. *Molecular Biology and Evolution* **13** 1375–1383.

[19] Massingham, T. and Goldman, N. (2007). Statistics of the log-det estimator. *Molecular Biology and Evolution* **24** 2277–2285.

[20] McCloskey, A. (2017). Bonferroni-based size-correction for nonstandard testing problems. *J. Econometrics* **200** 17–35. MR3683659

[21] Miller, J. J. (1977). Asymptotic properties of maximum likelihood estimates in the mixed model of the analysis of variance. *The Annals of Statistics* 746–762. MR0448661

[22] Neyman, J. and Pearson, E. S. (1933). On the Problem of the Most Efficient Tests of Statistical Hypotheses. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character* **231** 289–337.

[23] Olver, F. W. (2010). *NIST handbook of mathematical functions hardback and CD-ROM.* Cambridge University Press.

[24] Pamilo, P. and Nei, M. (1988). Relationships between gene trees and species trees. *Mol Biol Evol.* **5** 568–583.

[25] Rannala, B. and Yang, Z. (2003). Bayes Estimation of Species Divergence Times and Ancestral Population Sizes Using DNA Sequences From Multiple Loci. *Genetics* **164** 1645–1656.

[26] Rosenberg, N. A. (2002). The probability of topological concordance of gene trees and species trees. *Theoretical Population Biology* **61** 225–247.

[27] Self, S. G. and Liang, K.-Y. (1987). Asymptotic Properties of Maximum Likelihood Estimators and Likelihood Ratio Tests Under Nonstandard Conditions. *Journal of the American Statistical Association* **82** 605–610. MR0898365

[28] Shapiro, A. (1985). Asymptotic distribution of test statistics in the analysis of moment structures under inequality constraints. *Biometrika* **72** 133–144. MR0790208

[29] Silvapulle, M. J. and Sen, P. K. (2001). *Constrained Statistical Inference: Inequality, Order, and Shape Restrictions.* Wiley. MR2099529

[30] Van der Vaart, A. W. (2000). *Asymptotic Statistics* **3**. Cambridge University Press. MR1652247

[31] WAKELEY, J. (2009). Coalescent theory: an introduction. *Roberts & Company.*

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