On the choice of the splitting ratio for the split likelihood ratio test[*]

David Strieder and Mathias Drton

Technical University of Munich; TUM School of Computation, Information and Technology, Munich Center for Machine Learning (MCML), Munich Data Science Institute (MDSI)
e-mail: david.strieder@tum.de*;* mathias.drton@tum.de

Abstract: The recently introduced framework of universal inference provides a new approach to constructing hypothesis tests and confidence regions that are valid in finite samples and do not rely on any specific regularity assumptions on the underlying statistical model. At the core of the methodology is a split likelihood ratio statistic, which is formed under data splitting and compared to a cleverly selected universal critical value. As this critical value can be very conservative, it is interesting to mitigate the potential loss of power by careful choice of the ratio according to which data are split. Motivated by this problem, we study the split likelihood ratio test under local alternatives and introduce the resulting class of non-central split chi-square distributions. We investigate the properties of this new class of distributions and use it to numerically examine and propose an optimal choice of the data splitting ratio for tests of composite hypotheses of different dimensions.

Contents

1. Introduction

Likelihood ratio tests provide powerful solutions to a broad range of hypothesis testing problems. However, their implementation generally relies on asymptotic approximations whose validity requires the underlying statistical models to satisfy a number of regularity conditions. When these conditions are not met, the needed distribution-theoretic insights may be difficult to obtain; see, e.g., [1, 3]. Recent work of Wasserman, Ramdas and Balakrishnan [8] addresses this challenge by providing a *split likelihood ratio test* that is universally applicable to problems with i.i.d. samples. In this *universal inference* methodology, the data are split into two parts: one part is used to form a maximum likelihood estimate of a distribution under the full model, and the remaining data are used to compare the likelihood under the estimate versus the null hypothesis. Crucially, the independence of the split data allows one to apply a universal critical value, which merely depends on the chosen significance level and is guaranteed to be conservative even for finite samples. The resulting methodology makes it possible to conduct rather simple analyses of complicated composite hypotheses. For example, Strieder et al. [5] recently used the approach to construct hypothesis tests for causal effects in a setting with unknown causal structure.

The initial work in [8] and the follow-up paper by Dunn et al. [2] investigate the performance of the universal inference framework in the Gaussian case and under consideration of point hypotheses/construction of confidence regions. Unsurprisingly, the universal framework is rather conservative. To cite the authors: "our methods may not be optimal, though we do not yet fully understand how close to optimal they are beyond special cases (uniform, Gaussian)." During the review period, the independent work of Tse and Davison [6] was announced to appear, which covers different aspects of related problems.

The goal of the present paper is to expand our insights about the behavior of the split likelihood ratio test, as introduced in more detail in Section 2. In particular, we seek to shed light on the impact of the dimensionality of the tested null and alternative hypotheses. To this end, we study the case of smooth hypotheses in regular parametric models that are differentiable in quadratic mean. Under similar conditions, Wasserman, Ramdas and Balakrishnan [8] studied in their initial work the diameter of confidence sets corresponding to the inverted split likelihood ratio tests. We extend our insights about the limit behaviour of the split likelihood ratio test by deriving precise limit theory and calculating the large-sample asymptotic distribution allowing for local alternatives. This distribution belongs to a "split-version" of noncentral chi-square distributions, for which moments may be derived explicitly. We then use this new class of *noncentral split chi-square distributions* to propose a new routine for calculating the optimal splitting ratio for the split likelihood ratio test based on the dimensionality of the tested null and alternative hypotheses (Section 3). Furthermore, we use this new class of distributions to conduct numerical experiments that analyze the power and the optimal choice of the splitting ratio for the split likelihood ratio (Section 4). The simulations suggest, in particular, that while in lower dimensional settings an even split performs well, in higher dimensions a lower

splitting ratio is advantageous and our proposed new splitting ratio significantly improves power. In an experiment with factor analysis models, we demonstrate that our proposal may also lead to a power gain in irregular settings. This and other findings are further discussed in the concluding Section 6.

Notation In the remainder, the symbols $\xrightarrow{P}$, $\xrightarrow{a.s.}$ and $\xrightarrow{\mathcal{D}}$ stand for convergence in probability, almost sure convergence and convergence in distribution, respectively. The stochastic Landau symbol $o_P(1)$ indicates convergence to zero in probability. If not stated otherwise, the limits refer to $n \to \infty$. With Id we denote the $d \times d$ identity matrix, and $\mathcal{N}_d(0, \mathrm{Id})$ is the standard normal distribution in $\mathbb{R}^d$. For a given vector $X \in \mathbb{R}^d$ we denote the vector of its first p components by $X_{[p]}$.

2. Background on the split likelihood ratio test

Let $\{P_\theta : \theta \in \Theta\}$ be a given (parametric) statistical model, with parameter space $\Theta \subset \mathbb{R}^d$. The distributions P_θ are assumed to be dominated by a measure μ, and we write p_θ for the μ-density of P_θ. Given an i.i.d. sample $X_1, \ldots, X_n$ from an unknown distribution P_θ in the model, we are interested in testing

$$H_0 : \theta \in \Theta_0 \quad \text{versus} \quad H_1 : \theta \in \Theta \setminus \Theta_0 \tag{1}$$

for a subset $\Theta_0 \subset \Theta$. Universal inference solves this problem by appealing to a likelihood ratio, however, one that is built using data splitting.

To split the data, one chooses a fraction $m_0 \in (0, 1)$ and partitions the n data points into two disjoint subsets $D_0 = \{X_{1,0}, \ldots, X_{\lfloor m_0 n \rfloor, 0}\}$ and $D_1 = \{X_{1,1}, \ldots, X_{\lceil m_1 n \rceil, 1}\}$, where $m_1 \equiv 1 - m_0$. In order to lighten notation in subsequent derivations, we simply write $m_0 n$ for $\lfloor m_0 n \rfloor$ and $m_1 n$ for $\lceil m_1 n \rceil$. Let

$$\ell_k(\theta) = \sum_{i=1}^{m_k n} \log p_\theta(X_{i,k}), \quad k = 0, 1,$$

be the log-likelihood functions based on D_0 and D_1, respectively. Let $\widehat{\theta}_{n,0} := \mathrm{argmax}_{\theta \in \Theta_0} \ell_0(\theta)$ be the maximum likelihood estimator (MLE) of θ under H_0 and based on D_0. Furthermore, let $\widehat{\theta}_{n,1} := \mathrm{argmax}_{\theta \in \Theta} \ell_1(\theta)$ be the MLE of θ under the full model and based on D_1. Now the split likelihood ratio statistic is defined as

$$\Lambda_n := 2 \left(\ell_0(\widehat{\theta}_{n,1}) - \ell_0(\widehat{\theta}_{n,0}) \right). \tag{2}$$

As shown in [8], an application of Markov's inequality yields for any $\alpha \in (0, 1)$ that under $H_0 : \theta \in \Theta_0$ we have

$$P_\theta\left(\Lambda_n > -2\log(\alpha)\right) \leq \alpha \mathbb{E}_\theta \left[\frac{\prod_{i=1}^{m_0 n} p_{\widehat{\theta}_{n,1}}(X_{i,0})}{\prod_{i=1}^{m_0 n} p_{\widehat{\theta}_{n,0}}(X_{i,0})} \right]$$

$$\leq \alpha \mathbb{E}_\theta \left[\mathbb{E}_\theta \left[\frac{\prod_{i=1}^{m_0 n} p_{\widehat{\theta}_{n,1}}(X_{i,0})}{\prod_{i=1}^{m_0 n} p_\theta(X_{i,0})} \middle| D_1 \right] \right] \leq \alpha.$$

Here, we used the fact that $\widehat{\theta}_{n,1}$ is fixed when we condition on D_1, and for any fixed $\theta^* \in \Theta$ it holds that

$$\mathbb{E}_\theta \left[\frac{\prod_{i=1}^{m_0 n} p_{\theta^*}(X_{i,0})}{\prod_{i=1}^{m_0 n} p_\theta(X_{i,0})} \right] \leq \int \prod_{i=1}^{m_0 n} p_{\theta^*}(X_{i,0}) = 1.$$

Therefore, the decision rule

$$\text{reject } H_0 \text{ if } \Lambda_n > -2\log(\alpha) \tag{3}$$

constitutes a valid level α test. Notably, this *split likelihood ratio test* (SLRT) holds level α in finite samples and without any regularity conditions.

Remark 2.1. The MLE $\widehat{\theta}_{n,1}$ could be replaced by any other estimator and the test would continue to be valid. While this may be of interest for computational reasons, we focus in the following on the asymptotically efficient maximum likelihood estimator. However, the analysis can be extended in a similar fashion for any asymptotic linear estimator, see Remark 3.2.

In the following, we derive the asymptotic distribution of the split likelihood ratio Λ_n and use it to study the power of the SLRT and the impact of the splitting ratio m_0. Our calculation of the limiting distribution of Λ_n is couched in the classical framework of local alternatives in models that are differentiable in quadratic mean.

3. Asymptotic theory for the SLRT

Let θ_0 be a point in the interior of Θ. Assume that $\theta_0 \in \Theta_0$ and define the sequence of parameters $\theta_n = \theta_0 + h/\sqrt{n}$ for a choice of $h \in \mathbb{R}^d$. Suppose then that for each (large) n we are given an i.i.d. sample of size n from the local alternative P_{θ_n}. Suppose further that the considered model possesses the usual smoothness properties that lead to chi-square limits for the ordinary likelihood ratio, see, e.g., [7]. Specifically, we assume that:

(A1) The model $\{P_\theta : \theta \in \Theta\}$ is differentiable in quadratic mean at θ_0, with derivative (i.e., score function) $\dot{\ell}_{\theta_0}$. Its Fisher information $\mathbb{E}_{\theta_0}[\dot{\ell}_{\theta_0}\dot{\ell}_{\theta_0}^T] = I(\theta_0)$ is nonsingular, and there exists a measurable function $\dot{\ell}$ with $\mathbb{E}_{\theta_0}[\dot{\ell}^2] < \infty$ such that

$$|\log p_{\theta_1}(x) - \log p_{\theta_2}(x)| \leq \dot{\ell}(x)\|\theta_1 - \theta_2\|$$

for all θ_1, θ_2 in a neighborhood of θ_0.

(A2) The maximum likelihood estimators $\widehat{\theta}_{n,0}$ and $\widehat{\theta}_{n,1}$ are consistent estimators for θ_0 under P_{θ_0}.

(A3) The local parameter spaces $H_n := \sqrt{n}(\Theta_0 - \theta_0)$ converge to a set H_0.

Assumption (A3) uses the notion of convergence of sets in the sense of [7], that is, the local parameter spaces H_n converge to the limit hypothesis H_0, if the set H_0

a) is the set of all limits $\lim h_n$ of converging sequences with $h_n \in H_n$ and

b) contains all limits $\lim_{i \to \infty} h_{n_i}$ of converging sequences with $h_{n_i} \in H_{n_i}$.

Further, we note that the set convergence is guaranteed to hold when Θ_0 is defined by polynomial equations and inequalities, in which case the limit H_0 is the tangent cone of Θ_0 at θ_0 [1].

3.1. Asymptotic distribution

Theorem 3.1 (Asymptotic distribution of the split likelihood ratio statistic)**.** *Suppose the considered statistical model $\{P_\theta : \theta \in \Theta\}$ satisfies assumptions (A1)-(A3). Then under P_{θ_n} with $\theta_n = \theta_0 + h/\sqrt{n}$, the split likelihood ratio statistic from* (2) *satisfies*

$$\Lambda_n \xrightarrow{\mathcal{D}} \|X + \sqrt{m_0} I(\theta_0)^{1/2} h - I(\theta_0)^{1/2} H_0\|^2 - \|X - \sqrt{\tfrac{m_0}{m_1}} Y\|^2,$$

where $X, Y \sim \mathcal{N}_d(0, \mathrm{Id})$ independent and $\|x - H_0\| = \inf_{h \in H_0} \|x - h\|$.

Proof. The proof is based on classical local asymptotic normality results. As shown in Theorem 7.12 of [7], our assumption (A1) implies the existence of the Fisher information and the uniform approximation

$$\sup_{\|h\| \le M_n} \left| \log \prod_{i=1}^n \frac{p_{\theta_0 + h/\sqrt{n}}}{p_{\theta_0}}(X_i) - \frac{1}{\sqrt{n}} \sum_{i=1}^n h^T \dot{\ell}_{\theta_0}(X_i) + \frac{1}{2} h^T I(\theta_0) h \right| = o_{P_{\theta_0}}(1)$$
(4)

for M_n a slowly diverging sequence in $\mathbb{R}$. Via the results collected in [7], assumption (A2) implies that consistent MLEs are $\sqrt{n}$-consistent, which entails that both our split sample MLEs $\hat{\theta}_{n,0}$ and $\hat{\theta}_{n,1}$ are $\sqrt{n}$-consistent under P_{θ_0}.

Define $\widehat{\psi}_{n,1} := \sqrt{m_1 n}(\widehat{\theta}_{n,1} - \theta_0)$, $G_{n,0} := \frac{1}{\sqrt{m_0 n}} \sum_{i=1}^{m_0 n} \dot{\ell}_{\theta_0}(X_{i,0})$ and $G_{n,1} := \frac{1}{\sqrt{m_1 n}} \sum_{i=1}^{m_1 n} \dot{\ell}_{\theta_0}(X_{i,1})$. Let $B(M_n) = \{h \in \mathbb{R}^d : \|h\| \le M_n\}$ be the ball of radius M_n. Similarly to the proof of Theorem 16.7 in [7] but accounting for the split sample, we obtain from (4) and the $\sqrt{n}$-consistency of $\hat{\theta}_{n,0}$ and $\hat{\theta}_{n,1}$ that

$$
\begin{aligned}
\Lambda_n &= 2\left(\ell_0(\widehat{\theta}_{n,1}) - \ell_0(\widehat{\theta}_{n,0})\right) \\
&= 2\left(\log \prod_{i=1}^{m_0 n} \frac{p_{\theta_0 + \widehat{\psi}_{n,1}/\sqrt{m_1 n}}}{p_{\theta_0}}(X_{i,0}) - \sup_{h \in H_{m_0 n}} \log \prod_{i=1}^{m_0 n} \frac{p_{\theta_0 + h/\sqrt{m_0 n}}}{p_{\theta_0}}(X_{i,0})\right) \\
&= 2\left(\sqrt{\frac{m_0}{m_1}} \widehat{\psi}_{n,1}^T G_{n,0} - \frac{m_0}{2m_1} \widehat{\psi}_{n,1}^T I(\theta_0) \widehat{\psi}_{n,1} \right. \\
&\qquad \left. - \sup_{h \in H_{m_0 n} \cap B(M_n)} \left(h^T G_{n,0} - \frac{1}{2} h^T I(\theta_0) h\right)\right) + o_{P_{\theta_0}}(1) \\
&= \|I(\theta_0)^{-1/2} G_{n,0} - I(\theta_0)^{1/2}[H_{m_0 n} \cap B(M_n)]\|^2 \\
&\qquad - \|I(\theta_0)^{-1/2} G_{n,0} - \sqrt{\frac{m_0}{m_1}} I(\theta_0)^{1/2} \widehat{\psi}_{n,1}\|^2 + o_{P_{\theta_0}}(1).
\end{aligned}
$$

By Theorem 5.39 in [7], the MLE $\hat{\theta}_{n,1}$ is asymptotically linear with $\widehat{\psi}_{n,1} = I(\theta_0)^{-1}G_{n,1} + o_{P_{\theta_0}}(1)$. Hence,

$$\Lambda_n = \|I(\theta_0)^{-1/2}G_{n,0} - I(\theta_0)^{1/2}[H_{m_0 n} \cap B(M_n)]\|^2 \tag{5}$$
$$- \|I(\theta_0)^{-1/2}G_{n,0} - \sqrt{\frac{m_0}{m_1}}I(\theta_0)^{-1/2}G_{n,1}\|^2 + o_{P_{\theta_0}}(1).$$

Now we use Le Cam's Lemmas to show contiguity of P_{θ_0} and P_{θ_n}. Applying (4),

$$\log \frac{\mathrm{d}P_{\theta_n}^{\otimes n}}{\mathrm{d}P_{\theta_0}^{\otimes n}} = \log \left(\prod_{i=1}^{m_0 n} \frac{p_{\theta_0+h/\sqrt{n}}}{p_{\theta_0}}(X_{i,0}) \prod_{i=1}^{m_1 n} \frac{p_{\theta_0+h/\sqrt{n}}}{p_{\theta_0}}(X_{i,1}) \right)$$
$$= \frac{1}{\sqrt{n}} \left(\sum_{i=1}^{m_0 n} h^T \dot{\ell}_{\theta_0}(X_{i,0}) + \sum_{i=1}^{m_1 n} h^T \dot{\ell}_{\theta_0}(X_{i,1}) \right) - \frac{1}{2}h^T I(\theta_0)h + o_{P_{\theta_0}}(1).$$

The central limit theorem yields that under P_{θ_0},

$$\begin{pmatrix} G_{n,0} \\ G_{n,1} \\ \log \frac{\mathrm{d}P_{\theta_n}^{\otimes n}}{\mathrm{d}P_{\theta_0}^{\otimes n}} \end{pmatrix} \xrightarrow{\mathcal{D}} \mathcal{N}_{2d+1}\left(\mu, \Sigma\right),$$

where

$$\mu := \begin{pmatrix} 0 \\ 0 \\ -\frac{1}{2}h^T I(\theta_0)h \end{pmatrix}, \qquad \Sigma := \begin{bmatrix} I(\theta_0) & 0 & \sqrt{m_0}I(\theta_0)h \\ 0 & I(\theta_0) & \sqrt{m_1}I(\theta_0)h \\ \sqrt{m_0}I(\theta_0)h & \sqrt{m_1}I(\theta_0)h & h^T I(\theta_0)h \end{bmatrix}.$$

By Le Cam's first lemma, the probability measures P_{θ_0} and P_{θ_n} are thus mutually contiguous and therefore $o_{P_{\theta_0}}(1)$ and $o_{P_{\theta_n}}(1)$ interchangeable. By Le Cam's third lemma, it follows that under P_{θ_n} we have

$$\begin{pmatrix} G_{n,0} \\ G_{n,1} \end{pmatrix} \xrightarrow{\mathcal{D}} \mathcal{N}_{2d}\left(\begin{pmatrix} \sqrt{m_0}I(\theta_0)h \\ \sqrt{m_1}I(\theta_0)h \end{pmatrix}, \begin{bmatrix} I(\theta_0) & 0 \\ 0 & I(\theta_0) \end{bmatrix} \right). \tag{6}$$

We may now use this joint convergence in (5) and arrive at our claim by observing that for any converging sequence of random vectors $X_n \xrightarrow{\mathcal{D}} X$ and any sequence of converging sets $H_n \to H$ it holds that

$$\|X_n - H_n\| \xrightarrow{\mathcal{D}} \|X - H\|;$$

see Lemma 7.13 in [7]. Indeed, our assumption (A3) implies the convergence $H_{m_0 n} \cap B(M_n) \to H_0$ and our claim follows. $\qquad\square$

Remark 3.2. Throughout this work we use the MLE $\widehat{\theta}_{n,1}$ to solve the estimation task on data set D_1. Suppose we employ instead a suboptimal, asymptotically linear estimator $\tilde{\theta}_{n,1}$, that is,

$$\sqrt{m_1 n}(\tilde{\theta}_{n,1} - \theta_0) = \frac{1}{\sqrt{m_1 n}}I(\theta_0)^{-1} \sum_{i=1}^{m_1 n} \tilde{g}(X_{i,1}) + o_{P_{\theta_0}}(1),$$

with $\text{Var}[\tilde{g}(X_{i,1})] \succeq I(\theta_0)$, where $\succeq$ denotes the Loewner order. Tracing the proof of Theorem 3.1, one obtains that

$$\tilde{\Lambda}_n \xrightarrow{\mathcal{D}} \|X + \sqrt{m_0}I(\theta_0)^{1/2}h - I(\theta_0)^{1/2}H_0\|^2 - \|X - \sqrt{\tfrac{m_0}{m_1}}\tilde{Y}\|^2, \qquad (7)$$

where $X, \tilde{Y}$ are independent, $X \sim \mathcal{N}_d(0, \text{Id})$ and $\tilde{Y} \sim \mathcal{N}_d(\mu_{\tilde{Y}}, V_{\tilde{Y}})$ with $\mu_{\tilde{Y}} \neq 0$ and $V_{\tilde{Y}} \succeq \text{Id}$. In fact, we may assume that $V_{\tilde{Y}}$ is diagonal with diagonal entries $(V_{\tilde{Y}})_{jj} \geq 1$. (Otherwise, apply an orthogonal transformation to $X - \sqrt{\tfrac{m_0}{m_1}}\tilde{Y}$ to diagonalize $V_{\tilde{Y}}$.) The second part of the representation of the asymptotic distribution is thus

$$\|X - \sqrt{\tfrac{m_0}{m_1}}\tilde{Y}\|^2 \overset{\mathcal{D}}{=} \sum_{i=1}^{d} \left(1 + \tfrac{m_0}{m_1}(V_{\tilde{Y}})_{ii}\right) Z_i^2 \geq \left(1 + \tfrac{m_0}{m_1}\right) \sum_{i=1}^{d} Z_i^2,$$

where $Z_i \sim \mathcal{N}(-\sqrt{\tfrac{m_0}{m_1}}(\mu_{\tilde{Y}})_i, 1)$ are independent. Now, $\sum_{i=1}^{d} Z_i^2 \sim \chi_d^2(\lambda)$ with noncentrality parameter $\lambda > 0$. Since $\chi_d^2(\lambda)$ is stochastically larger than a (central) χ_d^2-distribution, we obtain from (7) that the limiting distribution when using a suboptimal estimator is stochastically smaller than the limit distribution of Theorem 3.1 with the MLE. Using a suboptimal estimator on D_1 thus leads to a decrease in power for the SLRT.

In the sequel, we investigate properties of the limiting distribution of the split likelihood ratio statistic in the smooth case, where the original null hypothesis is a k-dimensional smooth manifold and the limiting set H_0 is thus a k-dimensional tangent space. We start by introducing the arising new class of distributions, *noncentral split chi-square distributions*, that depends on four parameters, the dimension of the parameter space, the dimension of the null hypothesis, the splitting ratio, and a noncentrality parameter.

Definition 3.3 (Noncentral split chi-square distribution)**.** Let $d \in \mathbb{N}$, $p \in \{0, \dots, d\}$, $\delta \geq 0$, and $m_0 \in (0, 1)$. The d-dimensional *noncentral split chi-square distribution* with p degrees of freedom, noncentrality parameter δ and splitting ratio m_0, denoted $\text{split}_{m_0}\text{-}\chi_{p,d}^2(\delta)$, is the distribution of

$$\|X_{[p]} + \sqrt{m_0}h\|^2 - \|X - \sqrt{\tfrac{m_0}{1-m_0}}Y\|^2 \ \sim \ \text{split}_{m_0}\text{-}\chi_{p,d}^2(\delta),$$

where $X, Y \sim \mathcal{N}_d(0, \text{Id})$ independent and $h \in \mathbb{R}^p$ such that $h^T h = \delta$.

We emphasize that the noncentral split chi-square distribution is well-defined in that it depends on the vector h only through its norm δ. This follows from the invariance of the standard normal distribution of X and Y under orthogonal rotations analogously to the classical noncentral chi-square distribution.

While the classical noncentral chi-square distribution is the distribution of the squared distance from a standard normal vector to some fixed point in the space, the noncentral split chi-square distribution is the distribution of the difference of two squared distances. The first part is the squared distance from

a standard normal vector to a fixed point in the space. However, a second part arises from splitting the data into two subsets, namely, the squared distance of two independent standard normal vectors scaled according to the splitting ratio. Notice that the two parts are not independent.

In the following we calculate the first moments of this new class of distributions. In Section 3.2 we use the calculated moments to approximate the noncentral split chi-square distribution and thus the asymptotic behavior of the SLRT.

Corollary 3.4 (Moments of the noncentral split chi-square distribution). *Let* $Z \sim \text{split}_{m_0}\text{-}\chi^2_{p,d}(\delta)$. *Then*

1. $\mathbb{E}[Z] = p - d - d\frac{m_0}{1-m_0} + m_0\delta$,
2. $\text{Var}[Z] = 2(d-p) + 4d\frac{m_0}{1-m_0} + 2d\frac{m_0^2}{(1-m_0)^2} + 4m_0\delta$.

Proof. Let

$$\epsilon \sim \mathcal{N}_{2d}\left(0, \begin{bmatrix} m_0^{-1}\text{Id} & 0 \\ 0 & m_1^{-1}\text{Id} \end{bmatrix}\right),$$

and define

$$\mu := \begin{pmatrix} h \\ 0 \\ h \\ 0 \end{pmatrix}, \quad A := \begin{bmatrix} 0 & 0 & \text{Id}_\text{p} & 0 \\ 0 & -\text{Id}_\text{k} & 0 & \text{Id}_\text{k} \\ \text{Id}_\text{p} & 0 & -\text{Id}_\text{p} & 0 \\ 0 & \text{Id}_\text{k} & 0 & -\text{Id}_\text{k} \end{bmatrix},$$

with $h \in \mathbb{R}^p$ such that $h^T h = \delta$. Then the quadratic form $m_0(\epsilon + \mu)^T A(\epsilon + \mu)$ follows a $\text{split}_{m_0}\text{-}\chi^2_{p,d}(\delta)$ distribution and we can use properties of quadratic forms to calculate moments of the noncentral split chi-square distribution.

Using $\mathbb{E}[(\epsilon+\mu)^T A(\epsilon+\mu)] = \text{tr}[A\Sigma] + \mu^T A\mu$, a short calculation yields the claim for the expectation and the claimed variance follows via $\text{Var}[(\epsilon+\mu)^T A(\epsilon+\mu)] = 2\,\text{tr}[A\Sigma A\Sigma] + 4\mu^T A\Sigma A\mu$. □

Remark 3.5. Higher moments can be calculated via the cumulants $\kappa_n(\epsilon^T A\epsilon) = 2^{n-1}(n-1)!\,\text{tr}[A^n]$ with the following formulas for moments of quadratic forms:

1. $\mathbb{E}[(\epsilon^T A\epsilon)^1] = \kappa_1$.
2. $\mathbb{E}[(\epsilon^T A\epsilon)^2] = \kappa_1^2 + \kappa_2$.
3. $\mathbb{E}[(\epsilon^T A\epsilon)^3] = \kappa_1^3 + 3\kappa_1\kappa_2 + \kappa_3$.
4. $\mathbb{E}[(\epsilon^T A\epsilon)^4] = \kappa_1^4 + 6\kappa_1^2\kappa_2 + 3\kappa_2^2 + 4\kappa_1\kappa_3 + \kappa_4$.

Formulas for moments up to order ten can be found in [4].

Due to the rotational invariance of the standard normal distribution, we may study the limit of the SLRT in the smooth case, where the limiting hypothesis is a k-dimensional tangent space, by simply assuming that $I(\theta_0)^{1/2}H_0$ is a coordinate subspace, i.e., $I(\theta_0)^{1/2}H_0 = \{0\}^p \times \mathbb{R}^k$ with $d = p + k$.

Corollary 3.6. *If the rotated limiting hypothesis* $I(\theta_0)^{1/2}H_0 = \{0\}^p \times \mathbb{R}^k$ *is a coordinate subspace, then the asymptotic distribution from Theorem 3.1 follows a*

d-dimensional noncentral split chi-square distribution with p degrees of freedom, noncentrality parameter $\tilde{h}^T\tilde{h}$ and splitting ratio m_0. That is

$$\Lambda_\infty \overset{\mathcal{D}}{=} \|X_{[p]} + \sqrt{m_0}\tilde{h}_{[p]}\|^2 - \|X - \sqrt{\tfrac{m_0}{m_1}}Y\|^2 \ \sim \ \mathrm{split}_{m_0}\text{-}\chi^2_{p,d}(\tilde{h}^T_{[p]}\tilde{h}_{[p]}),$$

with $X, Y \sim \mathcal{N}_d(0, \mathrm{Id})$ independent and $\tilde{h} = [I(\theta_0)^{1/2}h]_{[p]}$.

Proof. We look at the first part of the limiting distribution from Theorem 3.1. With $X \sim \mathcal{N}_d(\sqrt{m_0}I(\theta_0)^{1/2}h, \mathrm{Id})$ we have

$$\|X - I(\theta_0)^{1/2}H_0\|^2 = \inf_{\theta \in \mathbb{R}^k} \left(X - \begin{pmatrix} 0 \\ \theta \end{pmatrix}\right)^T \left(X - \begin{pmatrix} 0 \\ \theta \end{pmatrix}\right) = X^T_{[p]}X_{[p]},$$

and the claim follows immediately. $\qquad\square$

Remark 3.7. Under the null hypothesis, the limiting distribution of the split likelihood ratio test statistic reduces to the following difference of dependent (scaled) chi-square distributions

$$\Lambda_n \overset{\mathcal{D}}{\to} \|X_{[p]}\|^2 - \|X - \sqrt{\tfrac{m_0}{m_1}}Y\|^2, \tag{8}$$

with $X, Y \sim \mathcal{N}_d(0, \mathrm{Id})$ independent.

The limiting null distribution in (8) clearly shows the asymptotic difference between the LRT and the SLRT. For the SLRT, a new second term arises in the limit that behaves like a scaled chi-square distributed random variable where the scaling factor depends only on the chosen splitting ratio. Furthermore, looking at Corollary 3.4, the limiting distribution has a negative expectation under the null hypothesis.

The SLRT uses the conservative critical value $-2\log(\alpha)$ that is universally valid but whose adoption may come with a loss of power. In Figures 1 and 2 we illustrate the source of this loss of power in different settings by comparing the universal threshold (`SLRT`) of the SLRT with (simulated) quantiles of a split chi-square distribution (`Asym`), the limiting distribution under the null hypothesis (8). The difference between the universal threshold and the quantile of the limiting distribution is smaller for lower significance level α and thus, the power loss, which stems from employing an universal threshold, is less noteworthy for low significance levels. Furthermore, the universal threshold is asymptotically more accurate for smaller splitting ratios. Moreover, we observe that using the universal threshold is asymptotically less precise for higher dimensions of the parameter space and for higher dimensional hypotheses. In Section 4.1 we further analyze the power loss from using the universal threshold in simulations.

3.2. *Optimal splitting ratio*

The main advantage of the SLRT over classical likelihood methods is its flexibility for settings where asymptotic distributions are difficult to obtain. This
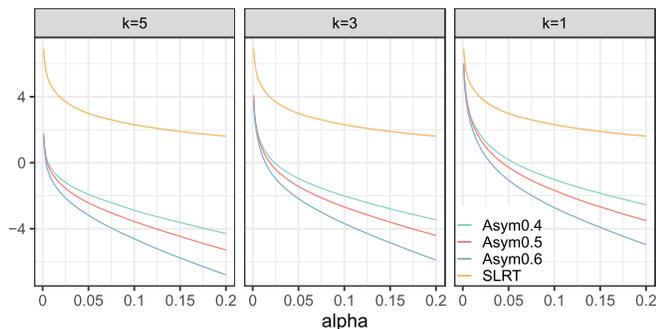
FIG 1. *Quantile of* split_{m_0}-$\chi^2_{p,6}$ *compared to the universal threshold.*
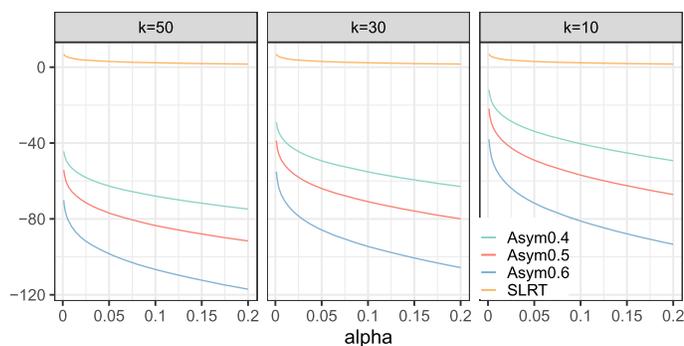

FIG 2. *Quantile of* split_{m_0}-$\chi^2_{p,60}$ *compared to the universal threshold.*

flexibility that stems from using only the general Markov inequality to control the type I error comes at the price of a potential loss of power. In the smooth setting of Theorem 3.1, we could improve the asymptotic power of the SLRT by using quantiles from the calculated asymptotic distribution, but such an asymptotic SLRT is not of practical relevance as the testing problem could then be better solved using the standard LRT, see Section 4.1.

Instead, our focus will remain on the SLRT with its conservative critical value $-2\log(\alpha)$, and our goal is to provide a new method for choosing the splitting ratio m_0 that helps retain power. The idea behind our proposed method is simple. Having access to the asymptotic distribution of the split likelihood ratio, the noncentral split chi-square distribution, we choose the splitting ratio that achieves the highest (asymptotic) power. Given both the dimensions of the null and alternative hypotheses and a significance level α, we minimize the cumulative distribution function of the noncentral split chi-square distribution at $-2\log(\alpha)$ with respect to the splitting ratio. To achieve a meaningful and comparable power, we propose to scale the unknown noncentrality parameter such that the best-performing method achieves a power of 0.8. We note that this
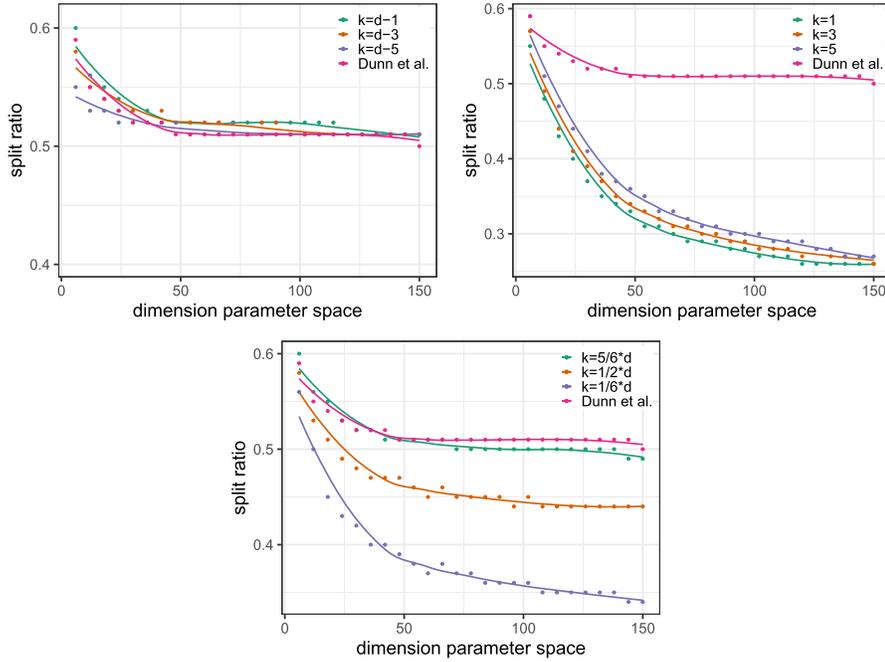
Fɪɢ 3. *Optimal splitting ratio against dimension of parameter space.*

tuning parameter of our method can easily be adapted if practitioners prefer to choose the best-performing method in different power levels. However, in our experience the effect of this tuning parameter on the performance is negligible across reasonable power levels.

As we are still lacking dedicated numerical routines to evaluate the cumulative distribution function of the noncentral split chi-square distribution, we use Monte Carlo approximations via repeated sampling from a noncentral split chi-square distribution. Using these approximations, we then choose the best-performing splitting ratio over a fine grid.

Figure 3 displays our results for the optimal choice of the splitting ratio against the dimension of the parameter space for different regimes of the null hypothesis space k. We observe that the underlying dimension of the null hypothesis k is crucial for the optimal choice of the splitting ratio. For comparison Figure 3 additionally shows the difference to the existing proposal for the choice of the splitting ratio by Dunn et al. [2]. To obtain a high power they propose to use the split ratio

$$m_0 = 1 - \frac{\sqrt{4d^2 + 8d\log(1/\alpha)} - 2d}{4\log(1/\alpha)},\tag{9}$$

which minimizes the squared radius of the universal inference confidence set for the mean vector of a Gaussian distribution. In contrast to what our results

---

Algorithm 1 Optimal splitting ratio

Input: d, p, α
initialize δ small
power $\leftarrow 0.5$
while power< 0.8 **do**
    $\exp(m_0) \leftarrow p - d - d\frac{m_0}{1-m_0} + m_0\delta$ $\triangleright$ Define functions in split m_0
    $\text{var}(m_0) \leftarrow 2(d-p) + 4d\frac{m_0}{1-m_0} + 2d\frac{m_0^2}{(1-m_0)^2} + 4m_0\delta$
    $\text{target}(m_0) \leftarrow \texttt{pnorm}(-2\log(\alpha), \exp(m_0), \sqrt{\text{var}(m_0)})$
    (min, value) $\leftarrow$ **minimize** $\text{target}(m_0)$ **for** m_0 **in** $(0,1)$ $\triangleright$ Optimization
    power $\leftarrow$ 1-value
    increase δ
end while
return min

---

suggest, their proposed split does not depend on the dimension of the null hypothesis. This dimensionality is, however, incorporated in our proposed choice of the splitting ratio.

Note that analytically the split (9), proposed by Dunn et al. [2], converges to 0.5 for high dimensions d. In a similar fashion, our proposed splitting ratio converges to 0.5 for high dimensions when testing a fixed number of parameters to be constant, that is $d - k$ is fixed. However, we can see that the optimal splitting ratio varies with the dimension of the tested hypothesis space, thus, e.g. in settings where the number of tested, fixed parameters in the hypothesis grows proportional with the dimension, a splitting ratios below 0.5 is beneficial, even in the limit.

Further, due to the impact of the noncentrality parameter and the complexity of the limit distribution, it is difficult to determine the optimal splitting ratio analytically. To avoid extensive Monte Carlo approximation for the calculation of the optimal splitting ratio, we additionally propose the following computationally fast alternative. Instead of using extensive simulations to approximate the noncentral split chi-square distribution, we employ normal approximations with the expectation and variance calculated based on Corollary 3.4. This then leads to Algorithm 1, which very quickly determines an optimal splitting ratio based on the dimension of the null hypothesis k, the dimension of the parameter space d, and the significance level α via repeated minimization of values of the standard Gaussian cdf (the `pnorm` function in `R`).

Our simple algorithm constitutes a fast solution to obtain an optimal splitting ratio based on normal approximations of the limit distribution. Without considering the variance, our algorithm can be interpreted as maximizing the expectation of the limit distribution of the SLRT, see Corollary 3.4. However, this expectation depends on the distance of the local alternative δ. Thus, to achieve a meaningful choice between the different splitting ratios, we first scaled δ to represent a setting where the test still has power against the local alternative but the task is nevertheless non-trivial. Furthermore, to get some more intuition for the behaviour of our proposed splitting ratio, Figure 4 provides a visual rule of thumb for the choice of the optimal splitting ratio based on the ratio of the
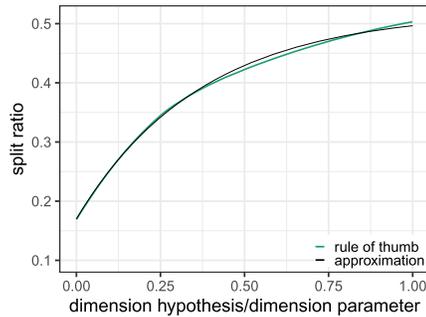
FIG 4. *Rule of thumb for the choice of the split ratio in high dimensions.*

dimension of the hypothesis space and the dimension of the parameter space derived in a high-dimensional setting. We found the impact of the significance level α in this high-dimensional setting to be negligible. For further simplicity and easy computation, this visual rule of thumb can be approximated with the function $m_0 = -\exp(-2.7\frac{k}{d} - 1.05) + 0.52$.

In Section 4.2 we analyze the performance of both proposed methods, the Monte Carlo method and the normal approximation variant numerically and show that it outperforms the existing proposal by Dunn et al. [2] in regular settings. Further, in Section 4.3 we show that this splitting ratio leads to optimal power even in the investigated irregular settings.

4. Simulations

We present the results of a simulation study that investigates the asymptotic behavior of the SLRT and compare its performance in different model settings, namely, regular and irregular settings, different dimensions of the parameter space d, different dimensions of the null hypothesis space k, and different splitting ratios m_0. All reported quantities are computed from simulations with 100,000 replications and if not stated otherwise we use the significance level $\alpha = 0.05$.

4.1. Power of the SLRT in regular setting

How much does using the universal threshold cost in terms of power asymptotically? In the following, we explore this question by comparing the power of the SLRT using the two different critical values, the standard universal threshold (`SLRT`) and the quantile of the asymptotic distribution (`Asym`). To this end we consider samples from a d-dimensional multivariate standard normal distribution with mean vector $(\theta, \dots, \theta)$ with $\theta = 0.1$ and test the hypothesis that the first $d - k$ entries of θ equal zero.

Figures 5 and 6 display the (simulated) power of both variants as well as that of the classical LRT against the sample size. In this regular setting where the
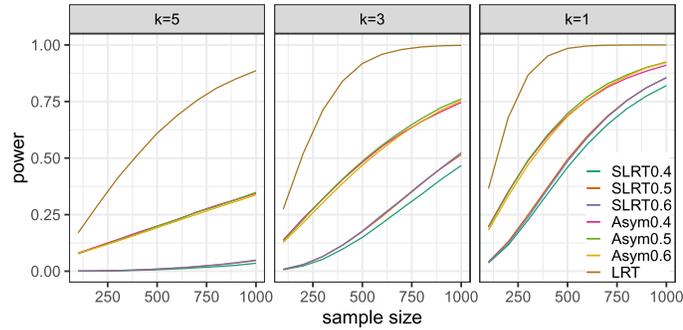
FIG 5. *Power against sample size in Gaussian setting with $\theta = 0.1$, $d = 6$.*
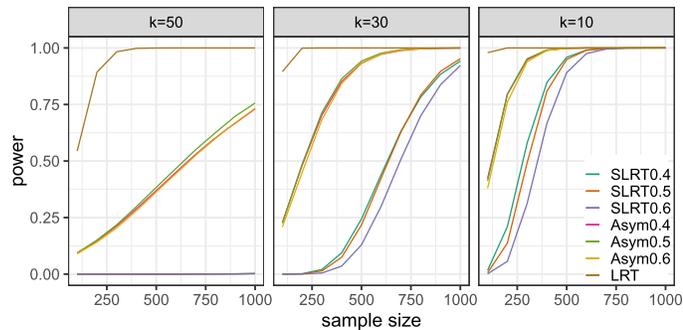

FIG 6. *Power against sample size in Gaussian setting with $\theta = 0.1$, $d = 60$.*

classical asymptotic distribution theory holds, the LRT outperforms the SLRT also when using the asymptotically correct quantiles. Furthermore, we see again that the power loss from using the universal threshold is larger in higher dimensions and higher dimensional null hypothesis settings. The simulations show that the choice of the splitting ratio plays an important role in the performance of the SLRT, especially in higher dimensional settings. In the following, we further examine the impact of the optimal choice of the splitting ratio in simulations.

4.2. Influence of the splitting ratio

In the following experiments, we analyze the influence of the splitting ratio on the asymptotic power of the SLRT. To this end, we sample data from a noncentral split chi-square distribution and calculate the power for testing the hypothesis of a zero noncentrality parameter δ. Figures 7 and 8 show the (simulated) power against the splitting ratio for the two different critical values, the universal threshold (SLRT) and the asymptotic quantile (Asym). We can see that in the lower dimensional setting a splitting ratio above 0.5 performs best
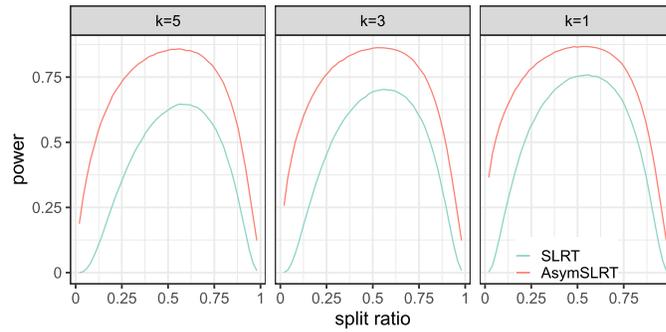
FIG 7. *Power of SLRT against splitting ratio, $d = 6$, $\delta = 40$.*
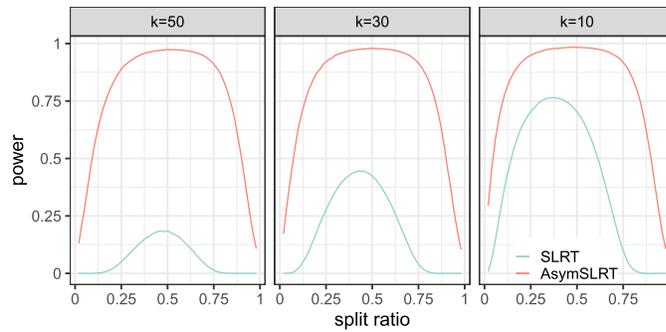
FIG 8. *Power of SLRT against splitting ratio, $d = 60$, $\delta = 180$.*

while in the higher dimensional setting a smaller splitting ratio below 0.5 seems beneficial, especially for a lower dimensional null hypothesis.

Figure 9 quantifies the improvement in power that can be achieved with our proposed (empirical) optimal splitting ratio (`emp.optim`) and the fast estimation routine (`est.optim`) that uses a normal approximation instead of extensive simulations to approximate the power of the SLRT. We compare the power for different noncentrality parameters $\delta \in \{100, 250\}$ plotted as 'dashed' and 'solid' lines respectively. Figure 9 displays that our fast estimation routine of the optimal split leads to valid approximations with a similarly good performance as the empirical optimal splitting ratio and further that there is a notable gain in power by using our new proposed optimal splitting ratios compared to the split (9) by Dunn et al. [2], especially in higher dimensions.

This is even more apparent in Figure 10, where we calculated the power for two different regimes of increasing noncentrality parameter δ. For each dimension of the parameter space, we chose the smallest δ such that the test with our new proposed optimal splitting ratio achieves a power of 0.8 and 0.65 respectively. While our methods, therefore, keep the power level, the split from (9) leads to a rapid loss of power in higher dimensions.
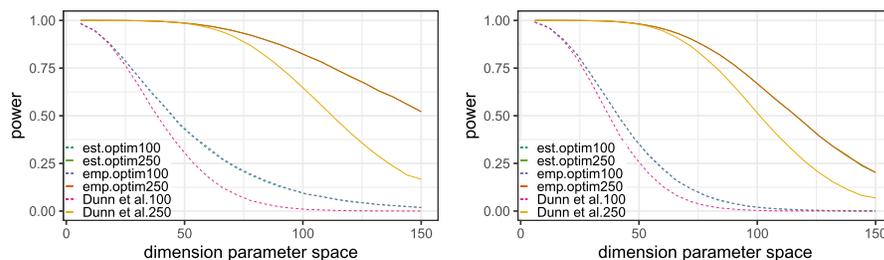
FIG 9. *Power for fixed noncentrality parameter and $k = 5$ (left); $k = \frac{1}{6} d$ (right).*
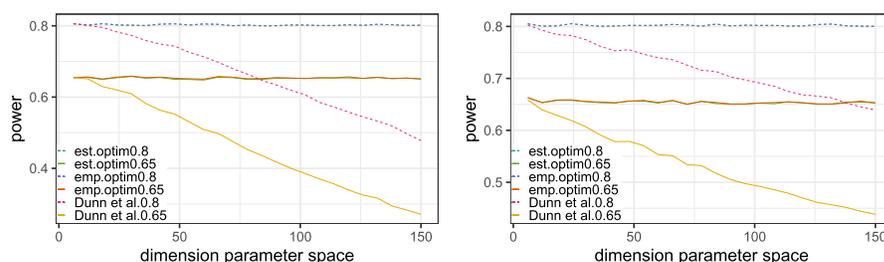

FIG 10. *Power for increasing noncentrality parameter and $k = 5$ (left); $k = \frac{1}{6} d$ (right)*

4.3. Power of the SLRT in irregular setting

Our proposed optimal splitting ratio is based on the limit distribution of the SLRT obtained under regularity conditions. In the following, we show that this optimal choice of the splitting ratio leads to an improvement in power even in irregular settings. To this end, we investigate the performance of the SLRT in the one-factor analysis setting with 12 observed variables. Assuming zero means, the one-factor model is given by the family of normal distributions $\mathcal{N}_{12}(0, \Sigma)$ with $\Sigma \in F_{12,1} := \{\Omega + \Gamma\Gamma^T : \Omega \in \mathbb{R}^{12 \times 12}_{>0} \text{ diagonal}, \Gamma \in \mathbb{R}^{12}\}$. In the following experiment, we consider testing the one-factor model against the saturated alternative, that is, the entire cone of positive definite matrices $PD(12)$. Note that the hypothesis defines a 24-dimensional subset of the 78-dimensional parameter space, thus, our algorithm suggests using the optimal splitting ratio 0.41 while (9) suggests using the splitting ratio 0.51. Drton [1] shows that in this one-factor analysis setting the hypothesis space has singularities at loadings Γ with less than 3 nonzero values and thus, at those points, classic asymptotic distribution theory for likelihood ratio tests is not valid.

Figure 11 displays the power of the SLRT under alternatives around irregular and regular points using the two different splitting ratios. More specifically, we set Ω as the diagonal matrix with all diagonal entries equal $1/5$ and $\Gamma = (5, 5, 0, ..., 0)$ for the irregular and $\Gamma = (5, 5, 5, 0, ..., 0)$ for the regular setting, respectively. Then we sampled $n = 2000$ data points from a two-factor alterna-
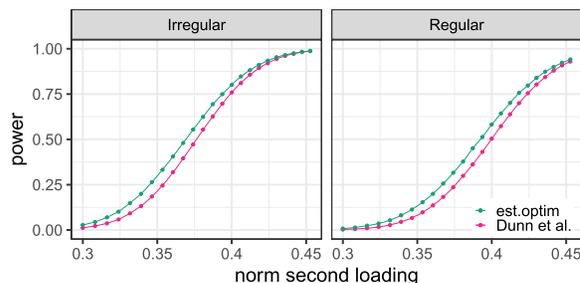
FIG 11. *Power of SLRT in one-factor analysis.*

tive $\mathcal{N}_{12}(0, \Sigma)$ with $\Sigma = \Omega + \Gamma\Gamma^T + \Gamma_2\Gamma_2^T$, where $\Gamma_2 = (h/\sqrt{12}, ..., h/\sqrt{12})$ with varying values of the norm h of the second factor loading Γ_2. We observe, that the optimality of our proposed splitting ratio extends to this irregular setting and our splitting ratio outperforms the splitting ratio suggested by Dunn et al. [2] in all settings.

5. Extension to the cross-fit SLRT

The SLRT and the universal inference framework starts with randomly splitting the available data. Thus, the value of the split likelihood ratio statistic still varies given a fixed data set, depending on the random split. It is natural to think about reducing this randomness by aggregating the results of different splits at the cost of performing more computations. Already in their initial work Wasserman, Ramdas and Balakrishnan [8] propose the cross-fit SLRT, a variant of the SLRT where the test statistic is calculated twice with alternating roles of the two data sets. Then, both results are averaged to obtain the cross-fit split likelihood ratio statistic $\frac{1}{2}(\Lambda_n + \Lambda_n^{swap})$, where Λ_n^{swap} is defined by (2) with the roles of D_0 and D_1 swapped.

Similar to the proof of Theorem 3.1 we can derive the asymptotic distribution of the cross-fit split likelihood ratio statistic.

Corollary 5.1. *Under assumptions (A1)-(A3), the cross-fit split likelihood ratio statistic satisfies under P_{θ_n} with $\theta_n = \theta_0 + h/\sqrt{n}$*

$$\Lambda_n + \Lambda_n^{swap} \xrightarrow{\mathcal{D}} \|X + \sqrt{m_0}I(\theta_0)^{1/2}h - I(\theta_0)^{1/2}H_0\|^2 - \|X - \sqrt{\tfrac{m_0}{m_1}}Y\|^2$$

$$+ \|Y + \sqrt{m_1}I(\theta_0)^{1/2}h - I(\theta_0)^{1/2}H_0\|^2 - \|Y - \sqrt{\tfrac{m_1}{m_0}}X\|^2,$$

where $X, Y \sim \mathcal{N}_d(0, \mathrm{Id})$ independent and $\|x - H_0\| = \inf_{h \in H_0} \|x - h\|$.

Since this cross-fit variant of the SLRT only splits the data once and then uses the same data sets twice with alternating roles, the limit distribution is still generated by two independent random variables. Analogous to the SLRT, we can thus calculate the expectation and variance of the arising limit distribution in

the smooth setting of Corollary 3.6, where the limiting hypothesis is a coordinate subspace, in the same way as the proof of Corollary 3.4 by exploiting properties of quadratic forms.

Corollary 5.2. *In the smooth setting of Corollary 3.6 the expectation and variance of the limit distribution of the cross-fit split likelihood ratio test statistics is given by*

1. $\mathbb{E}[\frac{1}{2}(\Lambda_\infty + \Lambda_\infty^{swap})] = p - d - \frac{1}{2}d\big(\frac{m_0}{1-m_0} + \frac{1-m_0}{m_0}\big) + \frac{1}{2}\delta$,
2. $\mathrm{Var}[\frac{1}{2}(\Lambda_\infty + \Lambda_\infty^{swap})] = (d-p)\big(1 + \frac{m_0}{1-m_0} + \frac{1-m_0}{m_0}\big) + d\big(2 + \frac{m_0}{1-m_0} + \frac{1-m_0}{m_0}\big)$
   $\quad + \frac{1}{2}d\big(\frac{m_0^2}{(1-m_0)^2} + \frac{(1-m_0)^2}{m_0^2}\big) + \delta$,

where $\delta = [I(\theta_0)^{1/2}h]_{[p]}^T[I(\theta_0)^{1/2}h]_{[p]}$.

Furthermore, analogously to Section 3.2, we can employ the limit distribution to determine the optimal splitting ratio and obtain the intuitive result that for the cross-fit SLRT an even split of $m_0 = 0.5$ is optimal in all dimensions.

Remark 5.3. For an equal splitting ratio $m_0 = 0.5$ the cross-fit split likelihood ratio statistic has the same expectation but a lower variance than the split likelihood ratio statistic in the limit. Thus, in situations where the method has power, the cross-fit SLRT further improves upon the SLRT.

The cross-fit SLRT employs equal weights $w_0 = 0.5$ to combine both test statistics and obtain the cross-fit split likelihood ratio test statistic, that is $w_0\Lambda_n + (1-w_0)\Lambda_n^{swap}$. Using equal weights is intuitive for an even splitting ratio $m_0 = 0.5$, since both test statistics are expected to perform similarly. However, considering the results from the previous section, for different splitting ratios, it might be beneficial to vary the weights and emphasize the better-performing test statistic. Furthermore, the idea of the cross-fit SLRT of swapping the roles of the two data sets allows us to derive properties of the limit distribution and is conceptually simple. Nevertheless, in view of our previous analysis, subsampling provides a promising alternative. Instead of using the same split data set with swapped roles to calculate the second test statistic, we repeat the process of randomly splitting the available data. Such a method has a similar computational burden as the cross-fit SLRT, however, we can make use of the optimal splitting ratio for both test statistics. Note that this subsampling procedure can be extended to multiple repeats and thus, further decrease the randomness of the splits at the cost of computation time.

Figure 12 compares the performance of the different extensions of the SLRT in the one-factor analysis setting introduced in Section 4.3. We display the power of the cross-fit SLRT with different splitting ratios m_0 and different weights w_0 as well as the subsampling alternative with our proposed optimal splitting ratio $m_0 = 0.41$. As previously mentioned, under the assumption of fixed, equal weights $w_0 = 0.5$, the best-performing splitting ratio for the cross-fit SLRT is an even split $m_0 = 0.5$. However, for an uneven splitting ratio, we can emphasize the better-performing test statistic by adjusting the weights to achieve similar performance. Furthermore, Figure 12 shows that using our proposed optimal
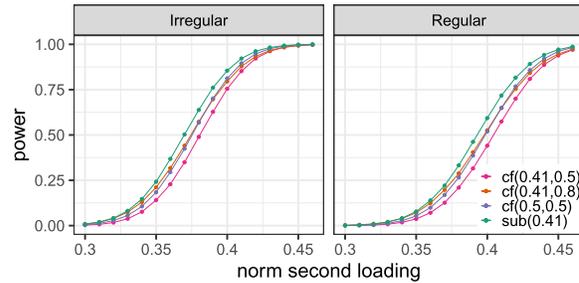
FIG 12. *Power of cross-fit(m_0, w_0) and subsampling SLRT in one-factor analysis.*

splitting ratio for both test statistics in a subsampling procedure outperforms the competition in all settings.

6. Conclusion

The split likelihood ratio test (SLRT) is a flexible tool that provides valid level α tests in finite samples even when classical regularity conditions are not satisfied. The underlying universal approach of splitting the data allows one to conduct rather simple analyses even in complicated settings. In general, this flexibility leads to a rather conservative method and, thus, it is of interest to carefully choose the splitting ratio in order to mitigate possible loss of power.

In order to provide new insights about the performance of the SLRT we studied its asymptotic behavior in the setting of smooth hypotheses. Our study gives rise to a new class of distributions, noncentral split chi-square distributions, that appear as limiting distributions of the SLRT. The split chi-square distribution depends on the dimensions of both null and alternative hypotheses and not only the difference of the dimensions. Naturally, it also depends on the chosen data splitting ratio. Using the new class of distributions, we analyzed the power of the SLRT in extensive simulations, and we proposed a new routine for calculating the optimal splitting ratio for the SLRT that significantly boosts power, especially in higher dimensions.

References

[1] DRTON, M. (2009). Likelihood ratio tests and singularities. *Ann. Statist.* **37** 979–1012. MR2502658

[2] DUNN, R., RAMDAS, A., BALAKRISHNAN, S. and WASSERMAN, L. (2021). Gaussian Universal Likelihood Ratio Testing.

[3] HARTIGAN, J. A. (1985). A failure of likelihood asymptotics for normal mixtures. In *Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer, Vol. II (Berkeley, Calif., 1983). Wadsworth Statist./Probab. Ser.* 807–810. Wadsworth, Belmont, CA. MR822066

[4] KENDALL, M. G. and STUART, A. (1969). *The advanced theory of statistics. Vol. 1: Distribution theory*, Third ed. Hafner Publishing Co., New York. MR0246399 MR0243648

[5] STRIEDER, D., FREIDLING, T., HAFFNER, S. and DRTON, M. (2021). Confidence in causal discovery with linear causal models. In *Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence* (C. DE CAMPOS and M. H. MAATHUIS, eds.). *Proceedings of Machine Learning Research* **161** 1217–1226. PMLR.

[6] TSE, T. and DAVISON, A. C. A Note on Universal Inference. *Stat* e501.

[7] VAN DER VAART, A. W. (1998). *Asymptotic statistics. Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge University Press, Cambridge. MR1652247 MR1652247

[8] WASSERMAN, L., RAMDAS, A. and BALAKRISHNAN, S. (2020). Universal inference. *Proceedings of the National Academy of Sciences* **117** 16880–16890.