PSYCHOMETRIKA—VOL. 85, NO. 4, 996–1012
DECEMBER 2020
https://doi.org/10.1007/s11336-020-09735-0

THEORY AND METHODS

A NOTE ON LIKELIHOOD RATIO TESTS FOR MODELS WITH LATENT VARIABLES

YUNXIAO CHEN AND IRINI MOUSTAKI
LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

HAORAN ZHANG
FUDAN UNIVERSITY

The likelihood ratio test (LRT) is widely used for comparing the relative fit of nested latent variable models. Following Wilks' theorem, the LRT is conducted by comparing the LRT statistic with its asymptotic distribution under the restricted model, a χ^2 distribution with degrees of freedom equal to the difference in the number of free parameters between the two nested models under comparison. For models with latent variables such as factor analysis, structural equation models and random effects models, however, it is often found that the χ^2 approximation does not hold. In this note, we show how the regularity conditions of Wilks' theorem may be violated using three examples of models with latent variables. In addition, a more general theory for LRT is given that provides the correct asymptotic theory for these LRTs. This general theory was first established in Chernoff (J R Stat Soc Ser B (Methodol) 45:404–413, 1954) and discussed in both van der Vaart (Asymptotic statistics, Cambridge, Cambridge University Press, 2000) and Drton (Ann Stat 37:979–1012, 2009), but it does not seem to have received enough attention. We illustrate this general theory with the three examples.

Key words: Wilks' theorem, χ^2 distribution, latent variable models, random effects models, dimensionality, tangent cone.

1. Introduction

1.1. Literature on Likelihood Ratio Test

The likelihood ratio test (LRT) is one of the most popular methods for comparing nested models. When comparing two nested models that satisfy certain regularity conditions, the p-value of an LRT is obtained by comparing the LRT statistic with a χ^2 distribution with degrees of freedom equal to the difference in the number of free parameters between the two nested models. This reference distribution is suggested by the asymptotic theory of LRT that is known as Wilks' theorem (Wilks 1938).

However, for the statistical inference of models with latent variables (e.g., factor analysis, item factor analysis for categorical data, structural equation models, random effects models, finite mixture models), it is often found that the χ^2 approximation suggested by Wilks' theorem does not hold. There are various published studies showing that the LRT is not valid under certain violations/conditions (e.g., small sample size, wrong model under the alternative hypothesis, large number of items, non-normally distributed variables, unique variances equal to zero, lack of identifiability), leading to over-factoring and over-rejections; see, e.g., Hakstian et al. (1982),

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s11336-020-09735-0) contains supplementary material, which is available to authorized users .

Correspondence should be made to Yunxiao Chen, Department of Statistics, London School of Economics and Political Science, London, UK. Email: y.chen186@lse.ac.uk

© 2020 The Author(s)

https://doi.org/10.1007/s11336-020-09735-0 Published online by Cambridge University Press

Liu and Shao (2003), Hayashi et al. (2007), Asparouhov and Muthén (2009), Wu and Estabrook (2016), Deng et al. (2018), Shi et al. (2018), Yang et al. (2018) and Auerswald and Moshagen (2019). There is also a significant amount of the literature on the effect of testing at the boundary of parameter space that arise when testing the significance of variance components in random effects models as well as in structural equation models (SEM) with linear or nonlinear constraints (see Stram and Lee 1994, 1995; Dominicus et al. 2006; Savalei and Kolenikov 2008; Davis-Stober 2009; Wu and Neale 2013; Du and Wang 2020).

Theoretical investigations have shown that certain regularity conditions of Wilks' theorem are not always satisfied when comparing nested models with latent variables. Takane et al. (2003) and Hayashi et al. (2007) were among the ones who pointed out that models for which one needs to select dimensionality (e.g., principal component analysis, latent class, factor models) have points of irregularity in their parameter space that in some cases invalidate the use of LRT. Specifically, such issues arise in factor analysis when comparing models with different number of factors rather than comparing a factor model against the saturated model. The LRT for comparing a q-factor model against the saturated model does follow a χ^2 distribution under mild conditions. However, for nested models with different number of factors (q-factor model is the correct one against the one with $(q + k)$ factors), the LRT is likely not χ^2-distributed due to violation of one or more of the regularity conditions. This is in line with the two basic assumptions required by the asymptotic theory for factor analysis and SEM: the identifiability of the parameter vector and non-singularity of the information matrix (see Shapiro 1986 and references therein). More specifically, Hayashi et al. (2007) focus on exploratory factor analysis and on the problem that arises when the number of factors exceeds the true number of factors that might lead to rank deficiency and non-identifiability of model parameters. That corresponds to the violations of the two regularity conditions. Those findings go back to Geweke and Singleton (1980) and Amemiya and Anderson (1990). More specifically, Geweke and Singleton (1980) studied the behavior of the LRT in small samples and concluded that when the regularity conditions from Wilks' theorem are not satisfied the asymptotic theory seems to be misleading in all sample sizes considered.

1.2. Our Contributions

The contribution of this note is twofold. First, we provide a discussion about situations under which Wilks' theorem for LRT may fail. Via three examples, we provide a relatively more complete picture about this issue in models with latent variables. Second, we introduce a unified asymptotic theory for LRT that covers Wilks' theorem as a special case and provides the correct asymptotic reference distribution for LRT when Wilks' theorem fails. This unified theory does not seem to have received enough attention in psychometrics, even though it has been established in statistics for long (Chernoff 1954; van der Vaart 2000; Drton 2009). In this note, we provide a tutorial on this theory, by presenting the theorems in a more accessible way and providing illustrative examples.

1.3. Examples

To further illustrate the issue with the classical theory for LRT, we provide three examples. These examples suggest that the χ^2 approximation can perform poorly and give p-values that can be either more conservative or more liberal.

Example 1. **(Exploratory factor analysis)** Consider a dimensionality test in exploratory factor analysis (EFA). For ease of exposition, we consider two hypothesis testing problems: (a) testing a one-factor model against a two-factor model and (b) testing a one-factor model against a saturated multivariate normal model with an unrestricted covariance matrix. Similar examples have been considered in Hayashi et al. (2007) where similar phenomena have been studied.

https://doi.org/10.1007/s11336-020-09735-0 Published online by Cambridge University Press

1(a). Suppose that we have J mean-centered continuous indicators, $\mathbf{X} = (X_1, ..., X_J)^\top$, which follow a J-variate normal distribution $N(\mathbf{0}, \boldsymbol{\Sigma})$. The one-factor model parameterizes $\boldsymbol{\Sigma}$ as

$$\boldsymbol{\Sigma} = \mathbf{a}_1\mathbf{a}_1^\top + \boldsymbol{\Delta},$$

where $\mathbf{a}_1 = (a_{11}, ..., a_{J1})^\top$ contains the loading parameters and $\boldsymbol{\Delta} = diag(\delta_1, ..., \delta_J)$ is diagonal matrix with a diagonal entries $\delta_1, ..., \delta_J$. Here, $\boldsymbol{\Delta}$ is the covariance matrix for the unique factors. Similarly, the two-factor model parameterizes $\boldsymbol{\Sigma}$ as

$$\boldsymbol{\Sigma} = \mathbf{a}_1\mathbf{a}_1^\top + \mathbf{a}_2\mathbf{a}_2^\top + \boldsymbol{\Delta},$$

where $\mathbf{a}_2 = (a_{12}, ..., a_{J2})^\top$ contains the loading parameters for the second factor and we set $a_{12} = 0$ to ensure model identifiability. Obviously, the one-factor model is nested within the two-factor model. The comparison between these two models is equivalent to test

$$H_0 : \mathbf{a}_2 = \mathbf{0} \text{ versus } H_a : \mathbf{a}_2 \neq \mathbf{0}.$$

If Wilks' theorem holds, then under H_0 the LRT statistic should asymptotically follow a χ^2 distribution with $J - 1$ degrees of freedom.

We now provide a simulated example. Data are generated from a one-factor model, with $J = 6$ indicators and $N = 5000$ observations. The true parameter values are given in Table 1. We generate 5000 independent datasets. For each dataset, we compute the LRT for comparing the one- and two-factor models. Results are presented in panel (a) of Fig. 1. The black solid line shows the empirical cumulative distribution function (CDF) of the LRT statistic, and the red dashed line shows the CDF of the χ^2 distribution suggested by Wilks' Theorem. A substantial discrepancy can be observed between the two CDFs. Specifically, the χ^2 CDF tends to stochastically dominate the empirical CDF, implying that p-values based on this χ^2 distribution tend to be more liberal. In fact, if we reject H_0 at 5% significance level based on these p-values, the actual type I error is 10.8%. These results suggest the failure of Wilks' theorem in this example.

1(b). When testing the one-factor model against the saturated model, the LRT statistic is asymptotically χ^2 if Wilks' theorem holds. The degrees of freedom of the χ^2 distribution are $J(J+1)/2 - 2J$, where $J(J+1)/2$ is the number of free parameters in an unrestricted covariance matrix $\boldsymbol{\Sigma}$ and $2J$ is the number of parameters in the one-factor model. In panel (b) of Fig. 1, the black solid line shows the empirical CDF of the LRT statistic based on 5000 independent simulations, and the red dashed line shows the CDF of the χ^2 distribution with 9 degrees of freedom. As we can see, the two curves almost overlap with each other, suggesting that Wilks' theorem holds here.

TABLE 1.
Values of the true parameters for the simulations in Example 1.

a_{11}	a_{21}	a_{31}	a_{41}	a_{51}	a_{61}
1.17	1.87	1.42	1.71	1.23	1.78
δ_1	δ_2	δ_3	δ_4	δ_5	δ_6
1.38	0.85	1.46	0.78	1.24	0.60

https://doi.org/10.1007/s11336-020-09735-0 Published online by Cambridge University Press

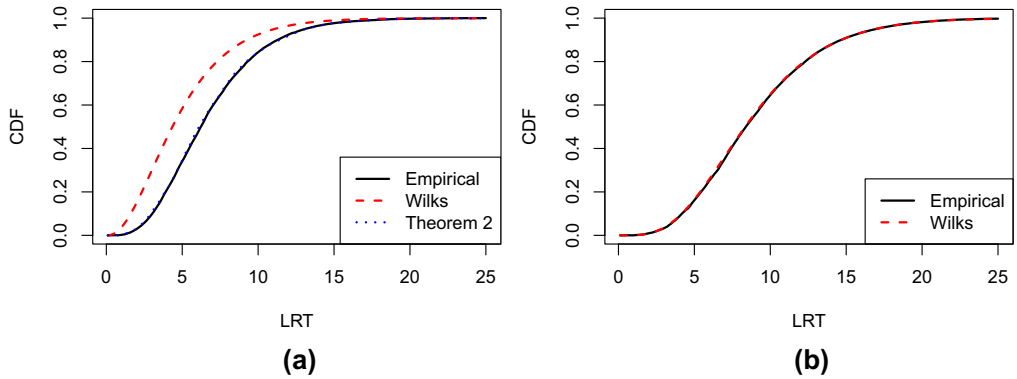


FIGURE 1.
a Results of Example 1(a). The black solid line shows the empirical CDF of the LRT statistic, based on 5000 independent simulations. The red dashed line shows the CDF of the χ^2 distribution with 5 degrees of freedom as suggested by Wilks' theorem. The blue dotted line shows the CDF of the reference distribution suggested by Theorem 2. **b** Results of Example 1(b). The black solid line shows the empirical CDF of the LRT statistic, and the red dashed line shows the CDF of the χ^2 distribution with 9 degrees of freedom as suggested by Wilks' theorem

Example 2. **(Exploratory item factor analysis)** We further give an example of exploratory item factor analysis (IFA) for binary data, in which similar phenomena as those in Example 1 are observed. Again, we consider two hypothesis testing problems: (a) testing a one-factor model against a two-factor model and (b) testing a one-factor model against a saturated multinomial model for a binary random vector.

2(a). Suppose that we have a J-dimensional response vector, $\mathbf{X} = (X_1, ..., X_J)^\top$, where all the entries are binary-valued, i.e., $X_j \in \{0, 1\}$. It follows a categorical distribution, satisfying

$$P(\mathbf{X} = \mathbf{x}) = \pi_{\mathbf{x}}, \mathbf{x} \in \{0, 1\}^J,$$

where $\pi_{\mathbf{x}} \geq 0$ and $\sum_{\mathbf{x}\in\{0,1\}^J} \pi_{\mathbf{x}} = 1$.

The exploratory two-factor IFA model parameterizes $\pi_{\mathbf{x}}$ by

$$\pi_{\mathbf{x}} = \int\int \prod_{j=1}^{J} \frac{\exp(x_j(d_j + a_{j1}\xi_1 + a_{j2}\xi_2))}{1 + \exp(d_j + a_{j1}\xi_1 + a_{j2}\xi_2)} \phi(\xi_1)\phi(\xi_2)d\xi_1 d\xi_2,$$

where $\phi(\cdot)$ is the probability density function of a standard normal distribution. This model is also known as a multidimensional two-parameter logistic (M2PL) model (Reckase 2009). Here, a_{jk}s are known as the discrimination parameters and d_js are known as the easiness parameters. We denote $\mathbf{a}_1 = (a_{11}, ..., a_{J1})^\top$ and $\mathbf{a}_2 = (a_{12}, ..., a_{J2})^\top$. For model identifiability, we set $a_{12} = 0$. When $a_{j2} = 0$, $j = 2, ..., J$, then the two-factor model degenerates to the one-factor model. Similar to Example 1(a), if Wilks' theorem holds, the LRT statistic should asymptotically follow a χ^2 distribution with $J - 1$ degrees of freedom.

Simulation results suggest the failure of this χ^2 approximation. In Fig. 2, we provide plots similar to those in Fig. 1, based on 5000 datasets simulated from a one-factor IFA model with sample size $N = 5000$ and $J = 6$. The true parameters of this IFA model are given in Table 2. The result is shown in panel (a) of Fig. 2, where a similar pattern is observed as that in panel (a) of Fig. 1 for Example 1(a).

https://doi.org/10.1007/s11336-020-09735-0 Published online by Cambridge University Press

TABLE 2.
Values of the true parameters for the simulations in Example 2.

d_1	d_2	d_3	d_4	d_5	d_6
−0.23	−0.12	0.07	0.31	−0.29	0.19
a_{11}	a_{21}	a_{31}	a_{41}	a_{51}	a_{61}
0.83	1.22	0.96	0.91	1.02	1.25

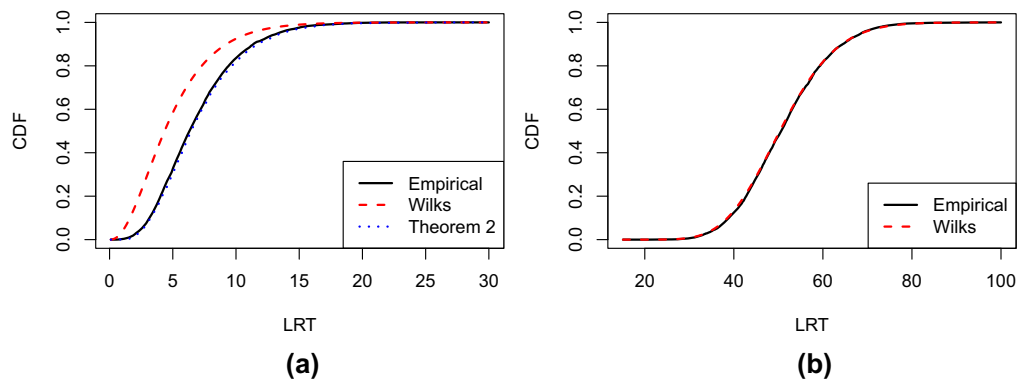


FIGURE 2.
a Results of Example 2(a). The black solid line shows the empirical CDF of the LRT statistic, based on 5000 independent simulations. The red dashed line shows the CDF of the χ^2 distribution with 5 degrees of freedom as suggested by Wilks' theorem. The blue dotted line shows the CDF of the reference distribution suggested by Theorem 2. **b** Results of Example 2(b). The black solid line shows the empirical CDF of the LRT statistic, and the red dashed line shows the CDF of the χ^2 distribution with 51 degrees of freedom as suggested by Wilks' theorem

2(b). When testing the one-factor IFA model against the saturated model, the LRT statistic is asymptotically χ^2 if Wilks' theorem holds, for which the degree of freedom is $2^J - 1 - 2J$. Here, $2^J - 1$ is the number of free parameters in the saturated model, and $2J$ is the number of parameters in the one-factor IFA model. The result is given in panel (b) of Fig. 2. Similar to Example 1(b), the empirical CDF and the CDF implied by Wilks' theorem are very close to each other, suggesting that Wilks' theorem holds here.

Example 3. **(Random effects model)** Our third example considers a random intercept model. Consider two-level data with individuals at level 1 nested within groups at level 2. Let X_{ij} be data from the jth individual from the ith group, where $i = 1, ..., N$ and $j = 1, ..., J$. For simplicity, we assume all the groups have the same number of individuals. Assume the following random effects model,

$$X_{ij} = \beta_0 + \mu_i + \epsilon_{ij},$$

where β_0 is the overall mean across all the groups, $\mu_i \sim N(0, \sigma_1^2)$ characterizes the difference between the mean for group i and the overall mean, and $\epsilon_{ij} \sim N(0, \sigma_2^2)$ is the individual level residual.

To test for between-group variability under this model is equivalent to test

$$H_0 : \sigma_1^2 = 0 \text{ against } H_a : \sigma_1^2 > 0.$$

https://doi.org/10.1007/s11336-020-09735-0 Published online by Cambridge University Press

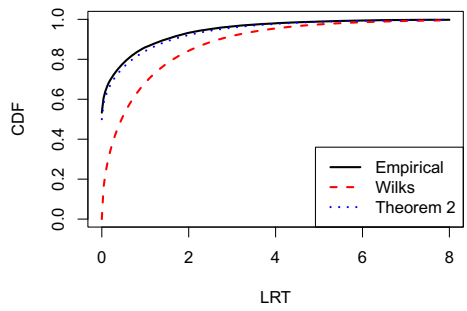


FIGURE 3.
The black solid line shows the empirical CDF of the LRT statistic, based on 5000 independent simulations. The red dashed line shows the CDF of the χ^2 distribution with one degree of freedom as suggested by Wilks' theorem. The blue dotted line shows the CDF of the mixture of χ^2 distribution suggested by Theorem 2 (Color figure online)

If Wilks' theorem holds, then the LRT statistic should follow a χ^2 distribution with one degree of freedom. We conduct a simulation study and show the results in Fig. 3. In this figure, the black solid line shows the empirical CDF of the LRT statistic, based on 5000 independent simulations from the null model with $N = 200$, $J = 20$, $\beta_0 = 0$, and $\sigma_2^2 = 1$. The red dashed line shows the CDF of the χ^2 distribution with one degree of freedom. As we can see, the two CDFs are not close to each other, and the empirical CDF tends to stochastically dominate the theoretical CDF suggested by Wilks' theorem. It suggests the failure of Wilks' theorem in this example.

This kind of phenomenon has been observed when the null model lies on the boundary of the parameter space, due to which the regularity conditions of Wilks' theorem do not hold. The LRT statistic has been shown to often follow a mixture of χ^2 distribution asymptotically (e.g., Shapiro 1985; Self and Liang 1987), instead of a χ^2 distribution. As it will be shown in Sect. 2, such a mixture of χ^2 distribution can be derived from a general theory for LRT.

We now explain why Wilks' theorem does not hold in Examples 1(a), 2(a), and 3. We define some generic notations. Suppose that we have i.i.d. observations $\mathbf{X}_1, ..., \mathbf{X}_N$, from a parametric model $\mathcal{P}_\Theta = \{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta \subset \mathbb{R}^k\}$, where $\mathbf{X}_i = (X_{i1}, ..., X_{iJ})^\top$. We assume that the distributions in $\mathcal{P}_\Theta$ are dominated by a common σ-finite measure ν with respect to which they have probability density functions $p_{\boldsymbol{\theta}} : \mathbb{R}^J \to [0, \infty)$. Let $\Theta_0 \subset \Theta$ be a submodel and we are interested in testing

$$H_0 : \boldsymbol{\theta} \in \Theta_0 \text{ versus } H_a : \boldsymbol{\theta} \in \Theta \backslash \Theta_0.$$

Let $p_{\boldsymbol{\theta}^*}$ be the true model for the observations, where $\boldsymbol{\theta}^* \in \Theta_0$.

The likelihood function is defined as

$$l_N(\boldsymbol{\theta}) = \sum_{i=1}^{N} \log p_{\boldsymbol{\theta}}(\mathbf{X}_i),$$

and the LRT statistic is defined as

$$\lambda_N = 2 \left(\sup_{\boldsymbol{\theta} \in \Theta} l_N(\boldsymbol{\theta}) - \sup_{\boldsymbol{\theta} \in \Theta_0} l_N(\boldsymbol{\theta}) \right).$$

Under suitable regularity conditions, Wilks' theorem suggests that the LRT statistic λ_N is asymptotically χ^2.

https://doi.org/10.1007/s11336-020-09735-0 Published online by Cambridge University Press

Wilks' theorem for LRT requires several regularity conditions; see, e.g., Theorem 12.4.2, Lehmann and Romano (2006). Among these conditions, there are two conditions that the previous examples do not satisfy. First, it is required that $\boldsymbol{\theta}^*$ is an interior point of Θ. This condition is not satisfied for Example 3, when Θ is taken to be $\{(\beta_0, \sigma_1^2, \sigma_2^2) : \beta_0 \in \mathbb{R}, \sigma_1^2 \in [0, \infty), \sigma_2^2 \in [0, \infty)\}$, as the null model lies on the boundary of the parameter space. Second, it is required that the expected Fisher information matrix at $\boldsymbol{\theta}^*$, $I(\boldsymbol{\theta}^*) = E_{\boldsymbol{\theta}^*}[\nabla l_N(\boldsymbol{\theta}^*)\nabla l_N(\boldsymbol{\theta}^*)^\top]/N$ is strictly positive definite. As we summarize in Lemma 1, this condition is not satisfied in Examples 1(a) and 2(a), when Θ is taken to be the parameter space of the corresponding two-factor model. However, interestingly, when comparing the one-factor model with the saturated model, the Fisher information matrix is strictly positive definite in Examples 1(b) and 2(b), for both simulated examples.

Lemma 1. *(1) For the two-factor model given in Example 1(a), choose the parameter space to be*

$$\Theta = \left\{(\delta_1, ..., \delta_J, a_{11}, ..., a_{J1}, a_{22}, ..., a_{J2})^\top \in \mathbb{R}^{3J-1} : \delta_j > 0,\ j = 1, ..., J\right\}.$$

If the true parameters satisfy $a_{j2}^* = 0,\ j = 2, ..., J$, then $I(\boldsymbol{\theta}^*)$ is non-invertible.

(2) For the two-factor IFA model given in Example 2(a), choose the parameter space to be $\Theta = \mathbb{R}^{3J-1}$. If the true parameters satisfy $a_{j2}^* = 0,\ j = 2, ..., J$, then $I(\boldsymbol{\theta}^*)$ is non-invertible.

We remark on the consequences of having a non-invertible information matrix. The first consequence is computational. If the information matrix is non-invertible, then the likelihood function does not tend to be strongly convex near the MLE, resulting in slow convergence. In the context of Examples 1(a) and 2(a), it means that computing the MLE for the corresponding two-factor models may have convergence issue. When convergence issue occurs, the obtained LRT statistic is below its actual value, due to the log likelihood for the two-factor model not achieving the maximum. Consequently, the p-value tends to be larger than its actual value, and thus, the decision based on the p-value tends to be more conservative than the one without convergence issue. This convergence issue is observed when conducting simulations for these examples. To improve the convergence, we use multiple random starting points when computing MLEs. The second consequence is a poor asymptotic convergence rate for the MLE. That is, the convergence rate is typically much slower than the standard parametric rate $N^{-1/2}$, even though the MLE is still consistent; see Rotnitzky et al. (2000) for more theoretical results on this topic.

We further provide some remarks on the LRT in Examples 1(b) and 2(b) that use a LRT for comparing the fitted model with the saturated model. Although Wilks' theorem holds asymptotically in example 2(b), the χ^2 approximation may not always work well as in our simulated example. This is because, when the number of items becomes larger and the sample size is not large enough, the contingency table for all 2^J response patterns may be sparse, and thus, the saturated model cannot be accurately estimated. In that case, it is better to use a limited information inference method (e.g., Maydeu-Olivares and Joe 2005, 2006) as a goodness-of-fit test statistic. Similar issues might also occur to Example 1(b).

2. General Theory for Likelihood Ratio Test

The previous discussions suggest that Wilks' theorem does not hold for Examples 1(a), 2(a), and 3, due to the violation of regularity conditions. It is then natural to ask: What asymptotic distribution does λ_N follow in these situations? Is there asymptotic theory characterizing such

https://doi.org/10.1007/s11336-020-09735-0 Published online by Cambridge University Press

irregular situations? The answer to these questions is "yes." In fact, a general theory characterizing these less regular situations has already been established in Chernoff (1954). In what follows, we provide a version of this general theory that is proved in van der Vaart (2000), Theorem 16.7. It is also given in Drton (2009), Theorem 2.6. Two problems will be considered, (1) comparing a submodel with the saturated model as in Examples 1(b) and 2(b), and (2) comparing two submodels as in Examples 1(a), 2(a), and 3.

2.1. *Testing Submodel Against Saturated Model*

We first introduce a few notations. We use $\mathbb{R}_{pd}^{J\times J}$ and $\mathbb{R}_{d}^{J\times J}$ to denote the spaces of $J \times J$ strictly positive definite matrices and diagonal matrices, respectively. In addition, we define a one-to-one mapping ρ: $\mathbb{R}_{pd}^{J\times J} \mapsto \mathbb{R}^{J(J+1)/2}$ that maps a positive definite matrix to a vector containing all its upper triangular entries (including the diagonal entries). That is, $\rho(\Sigma) = (\sigma_{11}, \sigma_{12}..., \sigma_{1J}, \sigma_{22}, ..., \sigma_{2J}, ..., \sigma_{JJ})^\top$, for $\Sigma = (\sigma_{ij})_{J\times J} \in \mathbb{R}_{pd}^{J\times J}$. We also define a one-to-one mapping μ: $\mathbb{R}_{d}^{J\times J} \mapsto \mathbb{R}^{J}$ that maps a diagonal matrix to a vector containing all its diagonal entries.

We consider to compare a submodel versus the saturated model. Let Θ_0 and Θ be the parameter spaces of the submodel and the saturated model, respectively, satisfying $\Theta_0 \subset \Theta \subset \mathbb{R}^k$. Also let $\boldsymbol{\theta}^* \in \Theta_0$ be the true parameter vector. The asymptotic theory of the LRT for comparing Θ_0 versus Θ requires regularity conditions C1-C5.

C1. The true parameter vector $\boldsymbol{\theta}^*$ is in the interior of Θ.

C2. There exists a measurable map $\dot{l}_{\boldsymbol{\theta}} : \mathbb{R}^J \to \mathbb{R}^k$ such that

$$\lim_{\mathbf{h}\to\mathbf{0}} \frac{1}{\|\mathbf{h}\|^2} \int_{\mathbb{R}^J} \left(\sqrt{p_{\boldsymbol{\theta}+\mathbf{h}}(\mathbf{x})} - \sqrt{p_{\boldsymbol{\theta}}(\mathbf{x})} - \frac{1}{2}\mathbf{h}^\top \dot{l}_{\boldsymbol{\theta}}(\mathbf{x}) \sqrt{p_{\boldsymbol{\theta}}(\mathbf{x})} \right)^2 d\nu(\mathbf{x}) = 0, \qquad (1)$$

and the Fisher information matrix $I(\boldsymbol{\theta}^*)$ for $\mathcal{P}_\Theta$ is invertible.

C3. There exists a neighborhood of $\boldsymbol{\theta}^*$, $U_{\boldsymbol{\theta}^*} \subset \Theta$, and a measurable function $\dot{l} : \mathbb{R}^J \to \mathbb{R}$, square integrable as $\int_{\mathbb{R}^J} \dot{l}(\mathbf{x})^2 dP_{\boldsymbol{\theta}^*}(\mathbf{x}) < \infty$, such that

$$|\log p_{\boldsymbol{\theta}_1}(\mathbf{x}) - \log p_{\boldsymbol{\theta}_2}(\mathbf{x})| \le \dot{l}(\mathbf{x}) \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|, \quad \forall \boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in U_{\boldsymbol{\theta}^*}.$$

C4. The maximum likelihood estimators (MLEs)

$$\hat{\boldsymbol{\theta}}_{N,\Theta} = \arg\max_{\boldsymbol{\theta}\in\Theta} l_N(\boldsymbol{\theta})$$

and

$$\hat{\boldsymbol{\theta}}_{N,\Theta_0} = \arg\max_{\boldsymbol{\theta}\in\Theta_0} l_N(\boldsymbol{\theta})$$

are consistent under $P_{\boldsymbol{\theta}^*}$.

The asymptotic distribution of λ_N depends on the local geometry of the parameter space Θ_0 at $\boldsymbol{\theta}^*$. This is characterized by the tangent cone $T_{\Theta_0}(\boldsymbol{\theta}^*)$, to be defined below.

Definition 1. The tangent cone $T_{\Theta_0}(\boldsymbol{\theta}^*)$ of the set $\Theta_0 \subset \mathbb{R}^k$ at the point $\boldsymbol{\theta}^* \in \mathbb{R}^k$ is the set of vectors in $\mathbb{R}^k$ that are limits of sequences $\alpha_n(\boldsymbol{\theta}_n - \boldsymbol{\theta}^*)$, where α_n are positive reals and $\boldsymbol{\theta}_n \in \Theta_0$ converge to $\boldsymbol{\theta}^*$.

https://doi.org/10.1007/s11336-020-09735-0 Published online by Cambridge University Press

The following regularity is required for the tangent cone $T_{\Theta_0}(\boldsymbol{\theta}^*)$ that is known as the Chernoff regularity.

C5. For every vector $\boldsymbol{\tau}$ in the tangent cone $T_{\Theta_0}(\boldsymbol{\theta}^*)$, there exist $\epsilon > 0$ and a map $\boldsymbol{\alpha} : [0, \epsilon) \to \Theta_0$ with $\boldsymbol{\alpha}(0) = \boldsymbol{\theta}^*$ such that $\boldsymbol{\tau} = \lim_{t \to 0+} [\boldsymbol{\alpha}(t) - \boldsymbol{\alpha}(0)]/t$.

Under the above regularity conditions, Theorem 1 holds and explains the phenomena in Examples 1(b) and 2(b).

Theorem 1. *Suppose that conditions C1-C5 are satisfied for comparing nested models $\Theta_0 \subset \Theta \subset \mathbb{R}^k$, with $\boldsymbol{\theta}^* \in \Theta_0$ being the true parameter vector. Then as N grows to infinity, the likelihood ratio statistic λ_N converges to the distribution of*

$$\min_{\boldsymbol{\tau} \in T_{\Theta_0}(\boldsymbol{\theta}^*)} \|\mathbf{Z} - I(\boldsymbol{\theta}^*)^{\frac{1}{2}} \boldsymbol{\tau}\|^2, \tag{2}$$

where $\mathbf{Z} = (Z_1, ..., Z_k)^\top$ is a random vector consisting of i.i.d. standard normal random variables.

Remark 1. We give some remarks on the regularity conditions. Conditions C1-C4 together ensure the asymptotic normality for $\sqrt{N}(\hat{\boldsymbol{\theta}}_{N,\Theta} - \boldsymbol{\theta}^*)$. Condition C1 depends on both the true model and the saturated model. As will be shown below, this condition holds for the saturated models in Examples 1(b) and 2(b). Equation (1) in C2 is also known as the condition of "differentiable in quadratic mean" for $\mathcal{P}_\Theta$ at $\boldsymbol{\theta}^*$. If the map $\boldsymbol{\theta} \mapsto \sqrt{p_{\boldsymbol{\theta}}(\mathbf{x})}$ is continuously differentiable for every $\mathbf{x}$, then C2 holds with $\dot{l}_{\boldsymbol{\theta}}(\mathbf{x}) = \frac{\partial}{\partial \boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\mathbf{x})$ (Lemma 7.6, van der Vaart (2000)). Furthermore, C3 holds if $\dot{l}(\mathbf{x}) = \sup_{\boldsymbol{\theta} \in U_{\boldsymbol{\theta}^*}} \dot{l}_{\boldsymbol{\theta}}(\mathbf{x})$ is square integrable with respect to the measure $P_{\boldsymbol{\theta}^*}$. Specifically, if $\dot{l}(\mathbf{x})$ is a bounded function, then C3 holds. C4 holds for our examples by Theorem 10.1.6, Casella and Berger (2002). C5 requires certain regularity on the local geometry of $T_{\Theta_0}(\boldsymbol{\theta}^*)$, which also holds for our examples below.

Remark 2. By Theorem 1, the asymptotic distribution for λ_N depends on the tangent cone $T_{\Theta_0}(\boldsymbol{\theta}^*)$. If $T_{\Theta_0}(\boldsymbol{\theta}^*)$ is a linear subspace of $\mathbb{R}^k$ with dimension k_0, then one can easily show that the asymptotic reference distribution of λ_N is χ^2 with degrees of freedom $k - k_0$. As we explain below, Theorem 1 directly applies to Examples 1(b) and 2(b). If $T_{\Theta_0}(\boldsymbol{\theta}^*)$ is a convex cone, then λ_N converges to a mixture of χ^2 distribution (Shapiro 1985; Self and Liang 1987). That is, for any $x > 0$, $P(\lambda_N \leq x)$ converges to $\sum_{i=0}^{k} w_k P(\xi_i \leq x)$, as N goes to infinity, where $\xi_0 \equiv 0$ and ξ_i follows a χ^2 distribution with i degrees of freedom for $i > 0$. Moreover, the weights sum up to 1/2 for the components with even degrees of freedom, and so do the weights for the components with odd degrees of freedom (Shapiro 1985).

Example 4. **(Exploratory factor analysis, revisited)** Now we consider Example 1(b). As the saturated model is a J-variate normal distribution with an unrestricted covariance matrix, its parameter space can be chosen as

$$\Theta = \{\rho(\boldsymbol{\Sigma}) : \boldsymbol{\Sigma} \in \mathbb{R}_{pd}^{J \times J}\} \subset \mathbb{R}^{J(J+1)/2},$$

and the parameter space for the restricted model is

$$\Theta_0 = \left\{\rho(\boldsymbol{\Sigma}) : \boldsymbol{\Sigma} = \mathbf{a}_1 \mathbf{a}_1^\top + \boldsymbol{\Delta},\ \mathbf{a}_1 \in \mathbb{R}^J, \boldsymbol{\Delta} \in \mathbb{R}_{pd}^{J \times J} \cap \mathbb{R}_{d}^{J \times J}\right\}.$$

https://doi.org/10.1007/s11336-020-09735-0 Published online by Cambridge University Press

Suppose $\boldsymbol{\theta}^* = \rho(\boldsymbol{\Sigma}^*) \in \Theta_0$, where $\boldsymbol{\Sigma}^* = \mathbf{a}_1^* \mathbf{a}_1^{*\top} + \boldsymbol{\Delta}^*$. It is easy to see that C1 holds with the current choice of Θ. The tangent cone $T_{\Theta_0}(\boldsymbol{\theta}^*)$ takes the form:

$$T_{\Theta_0}(\boldsymbol{\theta}^*) = \left\{ \rho(\boldsymbol{\Sigma}) : \boldsymbol{\Sigma} = \mathbf{a}_1^* \mathbf{b}_1^\top + \mathbf{b}_1 \mathbf{a}_1^{*\top} + \mathbf{B},\ \mathbf{b}_1 \in \mathbb{R}^J, \mathbf{B} \in \mathbb{R}_d^{J \times J} \right\},$$

which is a linear subspace of $\mathbb{R}^{J(J+1)/2}$ with dimension $2J$, as long as $a_{j1}^* \neq 0,\ j = 1, ..., J$. By Theorem 1, λ_N converges to the χ^2 distribution with degrees of freedom $J(J+1)/2 - 2J$.

Example 5. **(Exploratory item factor analysis, revisited)** Now we consider Example 2(b). As the saturated model is a 2^J-dimensional categorical distribution, its parameter space can be chosen as

$$\Theta = \left\{ \boldsymbol{\theta} = \{\theta_{\mathbf{x}}\}_{\mathbf{x} \in \Gamma_J} : \theta_{\mathbf{x}} \geq 0, \sum_{\mathbf{x} \in \Gamma_J} \theta_{\mathbf{x}} \leq 1 \right\} \subset \mathbb{R}^{2^J - 1},$$

where $\Gamma_J := \{0, 1\}^J \backslash \{(0, ..., 0)^\top\}$. Then, the parameter space for the restricted model is

$$\Theta_0 = \left\{ \boldsymbol{\theta} \in \Theta : \theta_{\mathbf{x}} = \int \prod_{j=1}^{J} \frac{\exp(x_j(d_j + a_{j1}\xi_1))}{1 + \exp(d_j + a_{j1}\xi_1)} \phi(\xi_1) d\xi_1,\ \mathbf{a}_1, \mathbf{d} \in \mathbb{R}^J \right\}. \tag{3}$$

Let $\boldsymbol{\theta}^* \in \Theta_0$ that corresponds to true item parameters $\mathbf{a}_1^* = (a_{j1}^*, ..., a_{J1}^*)^\top$ and $\mathbf{d}^* = (d_1^*, ..., d_J^*)^\top$. By the form of Θ_0, $\boldsymbol{\theta}^*$ is an interior point of Θ.

For any $\mathbf{x} \in \Gamma_J$, we define $\mathbf{f}_{\mathbf{x}} = (f_1(\mathbf{x}), ..., f_J(\mathbf{x}))^\top$ and $\mathbf{g}_{\mathbf{x}} = (g_1(\mathbf{x}), ..., g_J(\mathbf{x}))^\top$, where

$$f_l(\mathbf{x}) = \int \prod_{j=1}^{J} \frac{\exp(x_j(d_j^* + a_{j1}^*\xi_1))}{1 + \exp(d_j^* + a_{j1}^*\xi_1)} \left[x_l - \frac{\exp(d_l^* + a_{l1}^*\xi_1)}{1 + \exp(d_l^* + a_{l1}^*\xi_1)} \right] \phi(\xi_1) d\xi_1,$$

and

$$g_l(\mathbf{x}) = \int \prod_{j=1}^{J} \frac{\exp(x_j(d_j^* + a_{j1}^*\xi_1))}{1 + \exp(d_j^* + a_{j1}^*\xi_1)} \left[x_l - \frac{\exp(d_l^* + a_{l1}^*\xi_1)}{1 + \exp(d_l^* + a_{l1}^*\xi_1)} \right] \xi_1 \phi(\xi_1) d\xi_1,$$

for $l = 1, ..., J$. Then the tangent cone $T_{\Theta_0}(\boldsymbol{\theta}^*)$ has the form

$$T_{\Theta_0}(\boldsymbol{\theta}^*) = \left\{ \boldsymbol{\theta} = \{\theta_{\mathbf{x}}\}_{\mathbf{x} \in \Gamma_J} : \theta_{\mathbf{x}} = \mathbf{b}_0^\top \mathbf{f}_{\mathbf{x}} + \mathbf{b}_1^\top \mathbf{g}_{\mathbf{x}},\ \mathbf{b}_0, \mathbf{b}_1 \in \mathbb{R}^J \right\},$$

which is a linear subspace of $\mathbb{R}^{2^J - 1}$ with dimension $2J$. By Theorem 1, λ_N converges to the distribution of χ^2 with degrees of freedom $2^J - 1 - 2J$.

https://doi.org/10.1007/s11336-020-09735-0 Published online by Cambridge University Press

2.2. Comparing Two Nested Submodels

Theorem 1 is not applicable to Example 3, because $\boldsymbol{\theta}^*$ is on the boundary of Θ if Θ is chosen to be $\{(\beta_0, \sigma_1^2, \sigma_2^2) : \beta_0 \in \mathbb{R}, \sigma_1^2 \in [0, \infty), \sigma_2^2 \in [0, \infty)\}$, and thus, C1 is violated. Theorem 1 is also not applicable to Examples 1(a) and 2(a), because the Fisher information matrix is not invertible when Θ is chosen to be the parameter space of the two-factor EFA and IFA models, respectively, in which case condition C2 is violated.

To derive the asymptotic theory for such problems, we view them as a problem of testing nested submodels under a saturated model for which $\boldsymbol{\theta}^*$ is an interior point of Θ and the information matrix is invertible. Consider testing

$$H_0 : \boldsymbol{\theta} \in \Theta_0 \text{ versus } H_a : \boldsymbol{\theta} \in \Theta_1 \backslash \Theta_0,$$

where Θ_0 and Θ_1 are two nested submodels of a saturated model Θ, satisfying $\Theta_0 \subset \Theta_1 \subset \Theta \subset \mathbb{R}^k$. Under this formulation, Theorem 2 provides the asymptotic theory for the LRT statistic $\lambda_N = 2\left(\sup_{\boldsymbol{\theta} \in \Theta_1} l_N(\boldsymbol{\theta}) - \sup_{\boldsymbol{\theta} \in \Theta_0} l_N(\boldsymbol{\theta})\right)$.

To obtain the asymptotic distribution of λ_N, regularity conditions C1-C5 are still required for $\Theta_0 \subset \Theta$. Two additional conditions are needed for Θ_1, which are satisfied for Examples 6, 7 and 8.

C6. The MLE under Θ_1, $\hat{\boldsymbol{\theta}}_{N,\Theta_1} = \arg\max_{\boldsymbol{\theta} \in \Theta_1} l_N(\boldsymbol{\theta})$, is consistent under $P_{\boldsymbol{\theta}^*}$.

C7. Let $T_{\Theta_1}(\boldsymbol{\theta}^*)$ be the tangent cone for Θ_1, defined the same as in Definition 1, but with Θ_0 replaced by Θ_1. $T_{\Theta_1}(\boldsymbol{\theta}^*)$ satisfies Chernoff regularity. That is, for every vector $\boldsymbol{\tau}$ in the tangent cone $T_{\Theta_1}(\boldsymbol{\theta}^*)$ there exist $\epsilon > 0$ and a map $\boldsymbol{\alpha} : [0, \epsilon) \to \Theta_1$ with $\boldsymbol{\alpha}(0) = \boldsymbol{\theta}^*$ such that $\boldsymbol{\tau} = \lim_{t \to 0+} [\boldsymbol{\alpha}(t) - \boldsymbol{\alpha}(0)]/t$.

Theorem 2. *Let $\boldsymbol{\theta}^* \in \Theta_0$ be the true parameter vector. Suppose that conditions C1-C7 are satisfied. As N grows to infinity, the likelihood ratio statistic λ_N converges to the distribution of*

$$\min_{\boldsymbol{\tau} \in T_{\Theta_0}(\boldsymbol{\theta}^*)} \|\mathbf{Z} - I(\boldsymbol{\theta}^*)^{\frac{1}{2}} \boldsymbol{\tau}\|^2 - \min_{\boldsymbol{\tau} \in T_{\Theta_1}(\boldsymbol{\theta}^*)} \|\mathbf{Z} - I(\boldsymbol{\theta}^*)^{\frac{1}{2}} \boldsymbol{\tau}\|^2, \tag{4}$$

where $\mathbf{Z} = (Z_1, ..., Z_k)^\top$ is a random vector consisting of i.i.d. standard normal random variables, and $I(\boldsymbol{\theta}^*)^{\frac{1}{2}}$ satisfies $I(\boldsymbol{\theta}^*)^{\frac{1}{2}} (I(\boldsymbol{\theta}^*)^{\frac{1}{2}})^\top = I(\boldsymbol{\theta}^*)$ that can be obtained by eigenvalue decomposition.

Example 6. **(Random effects model, revisited)** Now we consider Example 3. Let $\mathbf{1}_n$ denote a length-n vector whose entries are all 1 and $\mathbf{I}_n$ denote the $n \times n$ identity matrix. As $\mathbf{X}_i = (X_{i1}, ..., X_{iJ})^\top$ from the random effects model is multivariate normal with mean $\beta_0 \mathbf{1}_J$ and covariance matrix $\sigma_1^2 \mathbf{1}_J \mathbf{1}_J^\top + \sigma_2^2 \mathbf{I}_J$, the saturated parameter space can be taken as

$$\Theta = \{(\rho(\boldsymbol{\Sigma})^\top, \beta_0)^\top : \boldsymbol{\Sigma} \in \mathbb{R}_{pd}^{J \times J}, \beta_0 \in \mathbb{R}\}.$$

The parameter space for restricted models are

$$\Theta_0 = \{(\rho(\boldsymbol{\Sigma})^\top, \beta_0)^\top : \boldsymbol{\Sigma} = \sigma_2^2 \mathbf{I}_J,\ \sigma_2^2 > 0, \beta_0 \in \mathbb{R}\},$$

and

$$\Theta_1 = \{(\rho(\boldsymbol{\Sigma})^\top, \beta_0)^\top : \boldsymbol{\Sigma} = \sigma_1^2 \mathbf{1}_J \mathbf{1}_J^\top + \sigma_2^2 \mathbf{I}_J, \sigma_1^2 \geq 0, \sigma_2^2 > 0, \beta_0 \in \mathbb{R}\}.$$

https://doi.org/10.1007/s11336-020-09735-0 Published online by Cambridge University Press

Let $\boldsymbol{\theta}^* = (\rho(\boldsymbol{\Sigma}^*), \beta_0^*) \in \Theta_0$, where $\boldsymbol{\Sigma}^* = \sigma_2^{*2}\mathbf{I}_J$. Then, C1 holds. The tangent cones for Θ_0 and Θ_1 are

$$T_{\Theta_0}(\boldsymbol{\theta}^*) = \{(\rho(\boldsymbol{\Sigma})^\top, b_0)^\top : \boldsymbol{\Sigma} = b_2\mathbf{I}_J,\ b_0, b_2 \in \mathbb{R}\}$$

and

$$T_{\Theta_1}(\boldsymbol{\theta}^*) = \{(\rho(\boldsymbol{\Sigma})^\top, b_0)^\top : \boldsymbol{\Sigma} = b_1\mathbf{1}_J\mathbf{1}_J^\top + b_2\mathbf{I}_J,\ b_1 \geq 0, b_0, b_2 \in \mathbb{R}\}.$$

By Theorem 2, λ_N converges to the distribution of (4).

In this example, the form of (4) can be simplified, thanks to the forms of $T_{\Theta_0}(\boldsymbol{\theta}^*)$ and $T_{\Theta_1}(\boldsymbol{\theta}^*)$. We denote

$$\mathbf{c}_0 = (0, ..., 0, 1),\ \mathbf{c}_1 = (\rho(\mathbf{1}_J\mathbf{1}_J^\top)^\top, 0)^\top,\ \mathbf{c}_2 = (\rho(\mathbf{I}_J)^\top, 0)^\top \in \mathbb{R}^{J(J+1)/2+1}.$$

It can be seen that $T_{\Theta_0}(\boldsymbol{\theta}^*)$ is a two-dimensional linear subspace spanned by $\{\mathbf{c}_0, \mathbf{c}_2\}$, and $T_{\Theta_1}(\boldsymbol{\theta}^*)$ is a half three-dimensional linear subspace defined as $\{\alpha_0\mathbf{c}_0 + \alpha_1\mathbf{c}_1 + \alpha_2\mathbf{c}_2 : \alpha_1 \geq 0, \alpha_0, \alpha_2 \in \mathbb{R}\}$. Let $\mathbf{P}_0$ denote the projection onto $T_{\Theta_0}(\boldsymbol{\theta}^*)$. Define

$$\mathbf{v} = \frac{\mathbf{c}_1 - \mathbf{P}_0\mathbf{c}_1}{\|\mathbf{c}_1 - \mathbf{P}_0\mathbf{c}_1\|},$$

and then, (4) has the form

$$\|\mathbf{v}^\top\mathbf{Z}\|^2 1_{\{\mathbf{v}^\top\mathbf{Z}\geq 0\}}. \tag{5}$$

It is easy to see that $\mathbf{v}^\top\mathbf{Z}$ follows standard normal distribution. Therefore, λ_N converges to the distribution of $w^2 1_{\{w\geq 0\}}$, where w is a standard normal random variable. This is known as a mixture of χ^2 distribution. The blue dotted line in Fig. 3 shows the CDF of this mixture χ^2 distribution. This CDF is very close to the empirical CDF of the LRT, confirming our asymptotic theory.

Example 7. **(Exploratory factor analysis, revisited)** Now we consider Example 1(a). Let Θ, Θ_0, $\boldsymbol{\theta}^*$ and $T_{\Theta_0}(\boldsymbol{\theta}^*)$ be the same as those in Example 4. In addition, we define

$$\Theta_1 = \left\{\rho(\boldsymbol{\Sigma}) : \boldsymbol{\Sigma} = \mathbf{a}_1\mathbf{a}_1^\top + \mathbf{a}_2\mathbf{a}_2^\top + \boldsymbol{\Delta},\ \mathbf{a}_1, \mathbf{a}_2 \in \mathbb{R}^J, a_{12} = 0, \boldsymbol{\Delta} \in \mathbb{R}_{pd}^{J\times J} \cap \mathbb{R}_d^{J\times J}\right\}.$$

The tangent cone of Θ_1 at $\boldsymbol{\theta}^*$ becomes

$$T_{\Theta_1}(\boldsymbol{\theta}^*) = \left\{\rho(\boldsymbol{\Sigma}) : \boldsymbol{\Sigma} = \mathbf{a}_1^*\mathbf{b}_1^\top + \mathbf{b}_1\mathbf{a}_1^{*\top} + \mathbf{b}_2\mathbf{b}_2^\top + \mathbf{B},\ \mathbf{b}_1, \mathbf{b}_2 \in \mathbb{R}^J, b_{12} = 0, \mathbf{B} \in \mathbb{R}_d^{J\times J}\right\}.$$

Note that $T_{\Theta_1}(\boldsymbol{\theta}^*)$ is not a linear subspace, due to the $\mathbf{b}_2\mathbf{b}_2^\top$ term. Therefore, by Theorem 2, the asymptotic distribution of λ_N is not χ^2 . See the blue dotted line in Panel (a) of Fig. 1 for the CDF of this asymptotic distribution. This CDF almost overlaps with the empirical CDF of the LRT, suggesting that Theorem 2 holds here.

https://doi.org/10.1007/s11336-020-09735-0 Published online by Cambridge University Press

Example 8. **(Exploratory item factor analysis, revisited)** Now we consider Example 2(a). Let Θ, Θ_0, $\boldsymbol{\theta}^*$ and $T_{\Theta_0}(\boldsymbol{\theta}^*)$ be the same as those in Example 5. Let

$$\Theta_1 = \left\{ \boldsymbol{\theta} \in \Theta : \theta_{\mathbf{x}} = \int\int \prod_{j=1}^{J} \frac{\exp(x_j(d_j + a_{j1}\xi_1 + a_{j2}\xi_2))}{1+\exp(d_j + a_{j1}\xi_1 + a_{j2}\xi_2)} \phi(\xi_1)\phi(\xi_2) d\xi_1 d\xi_2, a_{12} = 0, \mathbf{x} \in \Gamma_J \right\}$$

be the parameter space for the two-factor model. Recall $\mathbf{f}_{\mathbf{x}}$ and $\mathbf{g}_{\mathbf{x}}$ as defined in Example 5. For any $\mathbf{x} \in \Gamma_J$, we further define $\mathbf{H}_{\mathbf{x}} = (h_{rs}(\mathbf{x}))_{J\times J}$, where

$$\begin{aligned} h_{rs}(\mathbf{x}) = \int \prod_{j=1}^{J} \frac{\exp(x_j(d_j^* + a_{j1}^*\xi_1))}{1+\exp(d_j^* + a_{j1}^*\xi_1)} \left[x_r - \frac{\exp(d_r^* + a_{r1}^*\xi_1)}{1+\exp(d_r^* + a_{r1}^*\xi_1)} \right] \\ \times \left[x_s - \frac{\exp(d_s^* + a_{s1}^*\xi_1)}{1+\exp(d_s^* + a_{s1}^*\xi_1)} \right] \phi(\xi_1) d\xi_1 \end{aligned}$$

for $r \neq s$, and

$$\begin{aligned} h_{rr}(\mathbf{x}) = \int \prod_{j=1}^{J} \frac{\exp(x_j(d_j^* + a_{j1}^*\xi_1))}{1+\exp(d_j^* + a_{j1}^*\xi_1)} \left\{ \left[x_r - \frac{\exp(d_r^* + a_{r1}^*\xi_1)}{1+\exp(d_r^* + a_{r1}^*\xi_1)} \right]^2 \right. \\ \left. - \frac{\exp(d_r^* + a_{r1}^*\xi_1)}{(1+\exp(d_r^* + a_{r1}^*\xi_1))^2} \right\} \phi(\xi_1) d\xi_1. \end{aligned}$$

Then, the tangent cone of Θ_1 at $\boldsymbol{\theta}^*$ is

$$T_{\Theta_1}(\boldsymbol{\theta}^*) = \left\{ \boldsymbol{\theta} = \{\theta_{\mathbf{x}}\}_{\mathbf{x}\in\Gamma_J} : \theta_{\mathbf{x}} = \mathbf{b}_0^\top \mathbf{f}_{\mathbf{x}} + \mathbf{b}_1^\top \mathbf{g}_{\mathbf{x}} + \mathbf{b}_2^\top \mathbf{H}_{\mathbf{x}} \mathbf{b}_2,\ \mathbf{b}_0, \mathbf{b}_1, \mathbf{b}_2 \in \mathbb{R}^J, b_{12} = 0 \right\}. \quad (6)$$

Similar to Example 7, $T_{\Theta_1}(\boldsymbol{\theta}^*)$ is not a linear subspace, and thus, λ_N is not asymptotically χ^2. In Panel (a) of Fig. 2, the asymptotic CDF suggested by Theorem 2 is shown as the blue dotted line. Similar to the previously examples, this CDF is very close to the empirical CDF of the LRT.

3. Discussion

In this note, we point out how the regularity conditions of Wilks' theorem may be violated, using three examples of models with latent variables. In these cases, the asymptotic distribution of the LRT statistic is no longer χ^2, and therefore, the test may no longer be valid. It seems that the regularity conditions of Wilks' theorem, especially the requirement on a non-singular Fisher information matrix, have not received enough attention. As a result, the LRT is often misused. Although we focus on LRT, it is worth pointing out that other testing procedures, including the Wald and score tests, as well as limited information tests (e.g., tests based on bivariate information), require similar regularity conditions and thus may also be affected.

We present a general theory for LRT first established in Chernoff (1954) that is not widely known in psychometrics and related fields. As we illustrate by the three examples, this theory applies to irregular cases not covered by Wilks' theorem. There are other examples for which this general theory is useful. For example, Examples 1(a) and 2(a) can be easily generalized to

https://doi.org/10.1007/s11336-020-09735-0 Published online by Cambridge University Press

the comparison of factor models with different numbers of factors, under both confirmatory and exploratory settings. This theory can also be applied to model comparison in latent class analysis that also suffers from a non-invertible information matrix. To apply the theorem, the key is to choose a suitable parameter space and then characterize the tangent cone at the true model.

There are alternative inference methods for making statistical inference under such irregular situations. One method is to obtain a reference distribution for LRT via parametric bootstrap. Under the same regularity conditions as in Theorem 2, we believe that the parametric bootstrap is still consistent. The parametric bootstrap may even achieve better approximation accuracy for finite sample data than the asymptotic distributions given by Theorems 1 and 2. However, for complex latent variable models (e.g., IFA models with many factors), the parametric bootstrap may be computationally intensive, due to the high computational cost of repeatedly computing the marginal maximum likelihood estimators. On the other hand, Monte Carlo simulation of the asymptotic distribution in Theorem 2 is computationally much easier, even though there are still optimizations to be solved. Another method is the split likelihood ratio test recently proposed by Wasserman et al. (2020) that is computationally fast and does not suffer from singularity or boundary issues. By making use of a sample splitting trick, this split LRT is able to control the type I error at any pre-specified level. However, it may be quite conservative sometimes.

This paper focuses on the situations where the true model is exactly a singular or boundary point of the parameter space. However, the LRT can also be problematic when the true model is near a singular or boundary point. A recent article by Mitchell et al. (2019) provides a treatment of this problem, where a finite sample approximating distribution is derived for LRT.

Besides the singularity and boundary issues, the asymptotic distribution may be inaccurate when the dimension of the parameter space is relatively high comparing with the sample size. This problem has been intensively studied in statistics and a famous result is the Bartlett correction which provides a way to improve the χ^2 approximation (Bartlett 1937; Bickel and Ghosh 1990; Cordeiro 1983; Box 1949; Lawley 1956; Wald 1943). When the regularity conditions do not hold, the classical form of Bartlett correction may no longer be suitable. A general form of Bartlett correction remains to be developed, which is left for future investigation.

Acknowledgments

The authors thank the editor, associate editor, and three reviewers for their supportive and insightful comments. Yunxiao Chen acknowledges the support from the National Academy of Education/Spencer Postdoctoral Fellowship.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

https://doi.org/10.1007/s11336-020-09735-0 Published online by Cambridge University Press

Appendix

Proof of Lemma 1. Denote the (i, j)-entry of the Fisher information matrix $I(\boldsymbol{\theta}^*)$ as q_{ij}. In both cases, we show that $q_{ij} = 0$ for $i \geq 2J+1$, or $j \geq 2J+1$, and therefore, $I(\boldsymbol{\theta}^*)$ is non-invertible. Since

$$q_{ij} = \int \frac{\partial \log p_{\boldsymbol{\theta}}(\mathbf{x})}{\partial \theta_i}\Big|_{\boldsymbol{\theta}^*} \frac{\partial \log p_{\boldsymbol{\theta}}(\mathbf{x})}{\partial \theta_j}\Big|_{\boldsymbol{\theta}^*} p_{\boldsymbol{\theta}^*}(\mathbf{x}) d\mathbf{x},$$

it suffices to show that

$$\frac{\partial \log p_{\boldsymbol{\theta}}(\mathbf{x})}{\partial \theta_i}\Big|_{\boldsymbol{\theta}^*} = 0, \quad j \geq 2J+1.$$

In the case of two-factor model, it suffices to show that

$$\frac{\partial \log p_{\boldsymbol{\theta}}(\mathbf{x})}{\partial a_{l2}}\Big|_{\boldsymbol{\theta}^*} = 0,$$

for $l = 2, ..., J$. Let σ_{ij} be the (i, j)-entry of the covariance matrix Σ and it is easy to see that $\sigma_{ij} = a_{i1}a_{j1} + a_{i2}a_{j2} + 1_{\{i=j\}}\delta_i$, where $a_{12} = 0$. By the chain rule,

$$\frac{\partial \log p_{\boldsymbol{\theta}}(\mathbf{x})}{\partial a_{l2}} = \sum_{i \leq j} \frac{\partial \log p_{\boldsymbol{\theta}}(\mathbf{x})}{\partial \sigma_{ij}} \frac{\partial \sigma_{ij}}{\partial a_{l2}}.$$

Since

$$\begin{aligned} \frac{\partial \sigma_{ij}}{\partial a_{l2}}\Big|_{\boldsymbol{\theta}^*} &= 1_{\{l=i\}} a^*_{j2} + 1_{\{l=j\}} a^*_{i2} \\ &= 0, \end{aligned}$$

then $I(\boldsymbol{\theta}^*)$ is non-invertible in the case of two-factor model.
In the case of two-factor IFA model, since

$$\frac{\partial \log p_{\boldsymbol{\theta}}(\mathbf{x})}{\partial \theta_i} = \frac{1}{p_{\boldsymbol{\theta}}(\mathbf{x})} \frac{\partial p_{\boldsymbol{\theta}}(\mathbf{x})}{\partial a_{l2}}$$

it suffices to show that

$$\frac{\partial p_{\boldsymbol{\theta}}(\mathbf{x})}{\partial a_{l2}}\Big|_{\boldsymbol{\theta}^*} = 0,$$

for $l = 2, ..., J$. Since

$$\begin{aligned} \frac{\partial p_{\boldsymbol{\theta}}(\mathbf{x})}{\partial a_{l2}}\Big|_{\boldsymbol{\theta}^*} &= \int\int \prod_{j=1}^{J} \frac{\exp(x_j(d^*_j + a^*_{j1}\xi_1))}{1+\exp(d^*_j + a^*_{j1}\xi_1)} \left[x_l - \frac{\exp(d^*_l + a^*_{l1}\xi_1)}{1+\exp(d^*_l + a^*_{l1}\xi_1)}\right] \xi_2 \phi(\xi_1)\phi(\xi_2) d\xi_1 d\xi_2 \\ &= \int \xi_2 \phi(\xi_2) d\xi_2 \times \int \prod_{j=1}^{J} \frac{\exp(x_j(d^*_j + a^*_{j1}\xi_1))}{1+\exp(d^*_j + a^*_{j1}\xi_1)} \left[x_l - \frac{\exp(d^*_l + a^*_{l1}\xi_1)}{1+\exp(d^*_l + a^*_{l1}\xi_1)}\right] \phi(\xi_1) d\xi_1 \\ &= 0, \end{aligned}$$

https://doi.org/10.1007/s11336-020-09735-0 Published online by Cambridge University Press

then $I(\boldsymbol{\theta}^*)$ is non-invertible in the case of two-factor IFA model. □

References

Amemiya, Y., & Anderson, T. W. (1990). Asymptotic chi-square tests for a large class of factor analysis models. *The Annals of Statistics*, *18*, 1453–1463.

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling*, *16*, 397–438.

Auerswald, M., & Moshagen, M. (2019). How to determine the number of factors to retain in exploratory factor analysis: A comparison of extraction methods under realistic conditions. *Psychological Methods*, *24*, 468–491.

Bartlett, M. S. (1937). Properties of sufficiency and statistical tests. *Proceedings of the Royal Society of London. Series A-Mathematical and Physical Sciences*, *160*, 268–282.

Bickel, P. J., & Ghosh, J. (1990). A decomposition for the likelihood ratio statistic and the Bartlett correction-a Bayesian argument. *The Annals of Statistics*, *18*, 1070–1090.

Box, G. E. (1949). A general distribution theory for a class of likelihood criteria. *Biometrika*, *36*, 317–346.

Casella, G., & Berger, R. L. (2002). *Statistical inference*. Belmont, CA: Duxbury.

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

Cordeiro, G. M. (1983). Improved likelihood ratio statistics for generalized linear models. *Journal of the Royal Statistical Society: Series B (Methodological)*, *45*, 404–413.

Davis-Stober, C. P. (2009). Analysis of multinomial models under inequality constraints: Applications to measurement theory. *Journal of Mathematical Psychology*, *53*, 1–13.

Deng, L., Yang, M., & Marcoulides, K. M. (2018). Structural equation modeling with many variables: A systematic review of issues and developments. *Frontiers in Psychology*, *9*, 580.

Dominicus, A., Skrondal, A., Gjessing, H. K., Pedersen, N. L., & Palmgren, J. (2006). Likelihood ratio tests in behavioral genetics: Problems and solutions. *Behavior Genetics*, *36*, 331–340.

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

Du, H., & Wang, L. (2020). Testing variance components in linear mixed modeling using permutation. *Multivariate Behavioral Research*, *55*, 120–136.

Geweke, J. F., & Singleton, K. J. (1980). Interpreting the likelihood ratio statistic in factor models when sample size is small. *Journal of the American Statistical Association*, *75*, 133–137.

Hakstian, A. R., Rogers, W. T., & Cattell, R. B. (1982). The behavior of number-of-factor rules with simulated data. *Multivariate Behavioral Research*, *17*, 193–219.

Hayashi, K., Bentler, P. M., & Yuan, K.-H. (2007). On the likelihood ratio test for the number of factors in exploratory factor analysis. *Structural Equation Modeling: A Multidisciplinary Journal*, *14*, 505–526.

Lawley, D. N. (1956). A general method for approximating to the distribution of likelihood ratio criteria. *Biometrika*, *43*, 295–303.

Lehmann, E. L., & Romano, J. P. (2006). *Testing statistical hypotheses*. New York, NY: Springer.

Liu, X., & Shao, Y. (2003). Asymptotics for likelihood ratio tests under loss of identifiability. *The Annals of Statistics*, *31*, 807–832.

Maydeu-Olivares, A., & Joe, H. (2005). Limited-and full-information estimation and goodness-of-fit testing in 2^n contingency tables: A unified framework. *Journal of the American Statistical Association*, *100*, 1009–1020.

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika*, *71*, 713–732.

Mitchell, J. D., Allman, E. S., & Rhodes, J. A. (2019). Hypothesis testing near singularities and boundaries. *Electronic Journal of Statistics*, *13*, 2150–2193.

Reckase, M. (2009). *Multidimensional item response theory*. New York: Springer.

Rotnitzky, A., Cox, D. R., Bottai, M., & Robins, J. (2000). Likelihood-based inference with singular information matrix. *Bernoulli*, *6*, 243–284.

Savalei, V., & Kolenikov, S. (2008). Constrained versus unconstrained estimation in structural equation modeling. *Psychological Methods*, *13*, 150–170.

Self, S. G., & 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.

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

Shapiro, A. (1986). Asymptotic theory of overparameterized structural models. *Journal of the American Statistical Association*, *81*, 142–149.

Shi, D., Lee, T., & Terry, R. A. (2018). Revisiting the model size effect in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*, 21–40.

Stram, D. O., & Lee, J. W. (1994). Variance components testing in the longitudinal mixed effects model. *Biometrics*, *50*, 1171–1177.

Stram, D. O., & Lee, J. W. (1995). Correction to "variance components testing in the longitudinal mixed effects model". *Biometrics*, *51*, 1196.

Takane, Y., van der Heijden, P. G. M., & Browne, M. W. (2003). On likelihood ratio tests for dimensionality selection. In T. Higuchi, Y. Iba, & M. Ishiguro (Eds.), *Proceedings of science of modeling: The 30th anniversary meeting of the information criterion (AIC)* (pp. 348–349). Japan Institute of Statistical Mathematics: Tokyo.

van der Vaart, A. W. (2000). *Asymptotic statistics*. Cambridge: Cambridge University Press.

https://doi.org/10.1007/s11336-020-09735-0 Published online by Cambridge University Press

Wald, A. (1943). Tests of statistical hypotheses concerning several parameters when the number of observations is large. *Transactions of the American Mathematical Society*, *54*, 426–482.
Wasserman, L., Ramdas, A., & Balakrishnan, S. (2020). Universal inference. *Proceedings of the National Academy of Sciences*, *117*, 16880–16890.
Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *The Annals of Mathematical Statistics*, *9*, 60–62.
Wu, H., & Estabrook, R. (2016). Identification of confirmatory factor analysis models of different levels of invariance for ordered categorical outcomes. *Psychometrika*, *81*, 1014–1045.
Wu, H., & Neale, M. C. (2013). On the likelihood ratio tests in bivariate acde models. *Psychometrika*, *78*, 441–463.
Yang, M., Jiang, G., & Yuan, K.-H. (2018). The performance of ten modified rescaled statistics as the number of variables increases. *Structural Equation Modeling: A Multidisciplinary Journal*, *25*, 414–438.

Manuscript Received: 9 AUG 2020
Accepted: 24 NOV 2020
Published Online Date: 21 DEC 2020

https://doi.org/10.1007/s11336-020-09735-0 Published online by Cambridge University Press