



Asymptotic normality of test statistics under alternative hypotheses

Alexander Shapiro

School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, Georgia 30332-0205, USA

ARTICLE INFO

Article history:

Received 12 January 2008

Available online 23 September 2008

AMS subject classifications:

primary 62F05

secondary 62H25

62E20

Keywords:

Stochastic optimization

Likelihood ratio test statistic

Asymptotic normality

Asymptotic bias

Nonnested models

Moment (covariance) structures

Discrepancy functions

ABSTRACT

The aim of this paper is to present a framework for asymptotic analysis of likelihood ratio and minimum discrepancy test statistics. First order asymptotics are presented in a general framework under minimal regularity conditions and for not necessarily nested models. In particular, these asymptotics give sufficient and in a sense necessary conditions for asymptotic normality of test statistics under alternative hypotheses. Second order asymptotics, and their implications for bias corrections, are also discussed in a somewhat informal manner. As an example, asymptotics of test statistics in the analysis of covariance structures are discussed in detail.

© 2008 Elsevier Inc. All rights reserved.

1. Introduction

Classical result of Wilks [11] says that the large sample distribution of the log-likelihood ratio statistic for testing nested models is approximately chi-square under the null hypothesis and provided certain regularity conditions are satisfied. Moreover, under a sequence of local alternatives (often referred to as¹ Pitman drift), the log-likelihood ratio statistic has asymptotically a noncentral chi-square distribution. These results are routinely employed for testing nested models. Large sample properties of log-likelihood ratio statistics under alternative hypotheses and for nonnested models were studied, e.g., by Vuong [9] (for a more recent discussion of that topic see, e.g., Golden [2] and the references therein). Recently it was argued in Yuan et al. [12] that in some cases normal approximations can give better asymptotics, for misspecified models, than the noncentral chi-square for the large sample distribution of the log-likelihood ratio statistic in the analysis of moment (covariance) structures.

The aim of this paper is to present a very general framework for asymptotic analysis of likelihood ratio test statistics. The analysis is based on somewhat old results which seem to be little known in the statistics literature. These results allow to simplify and generalize the analysis of [9,12] in a significant way. In particular, we show in detail how it can be applied to the analysis of covariance structures.

This paper will be organized as follows. In the next section we describe asymptotics of the optimal value of a stochastic optimization problem. We present two approaches to such analysis which could be convenient in different situations. In Section 3 we discuss examples and applications of the general theory. In particular, we discuss applications to the analysis of covariance structures. We use the following notation and terminology throughout the paper. By A^T we denote the transpose of matrix (vector) A , and by $\text{tr}(B)$ the trace of (square) matrix B . The gradient vector (of partial derivatives) of a function

E-mail address: ashapiro@isye.gatech.edu.

¹ See McManus [3] for a historical overview of who invented local power analysis.

$g : \mathbb{R}^m \rightarrow \mathbb{R}$, at a point $\theta \in \mathbb{R}^m$, is denoted either by $\nabla g(\theta)$ or $\partial g(\theta)/\partial \theta$. The Hessian matrix (of second order partial derivatives) of $g(\cdot)$, at θ , is denoted by $\nabla^2 g(\theta)$ or $\partial^2 g(\theta)/\partial \theta \partial \theta^T$. For a random vector X having probability distribution P , we denote by $\mathbb{E}_P[h(X)]$ the expected value of function $h(X)$. When there will be no ambiguity of what distribution is used, we omit the subscript P and simply write $\mathbb{E}[h(X)]$. The notation “ \Rightarrow ” stands for convergence in distribution. By writing $P \stackrel{d}{=} f(\cdot)$ we mean that probability distribution P , of a random vector X , has density function $f(x)$, and the notation $X \sim N(\mu, \Sigma)$ means that random vector X has multivariate normal distribution with mean vector μ and covariance matrix Σ . We use notation “ $:=$ ” to denote “equal by definition”. By $\text{dist}(\theta, \Theta) := \inf_{\theta' \in \Theta} \|\theta - \theta'\|$ we denote the distance from point θ to set Θ .

2. Basic asymptotics

In this section we discuss some basic results for asymptotics of a very general class of statistics given by optimal value of a stochastic problem. We describe these results in two frameworks which could be convenient in different situations.

2.1. Framework of stochastic optimization

Let X be a random vector, whose probability distribution P is supported on set $\mathfrak{X} \subset \mathbb{R}^p$, $\Theta \subset \mathbb{R}^m$ be a (nonempty) parameter set and $V : \mathfrak{X} \times \Theta \rightarrow \mathbb{R}$ be a real valued function. The set \mathfrak{X} is supposed to be equipped with its Borel sigma algebra. We make the following assumptions about function $V(x, \theta)$.

- (A1) For every $\theta \in \Theta$ the function $V(\cdot, \theta)$ is measurable and P -integrable, and hence the expectation $v(\theta) := \mathbb{E}[V(X, \theta)]$ is well defined and finite valued.
- (A2) For every $\theta \in \Theta$, the expectation $\mathbb{E}[V(X, \theta)^2]$ is finite.
- (A3) There exists a measurable function $\kappa : \mathfrak{X} \rightarrow \mathbb{R}$ such that $\mathbb{E}[\kappa(X)^2]$ is finite and

$$|V(x, \theta') - V(x, \theta)| \leq \kappa(x) \|\theta' - \theta\| \tag{2.1}$$

for all $\theta, \theta' \in \Theta$ and a.e. $x \in \mathfrak{X}$.

The above assumptions are reasonably simple regularity conditions which could be verified in particular applications. Assumption (A1) postulates that the optimization problem considered below is well defined. Existence of second order moments, postulated in assumptions (A2), is a natural and in a sense minimal requirement for derivation of Central Limit Theorem type results. Assumption (A3) implies, of course, that for every $x \in \mathfrak{X}$ the function $V(x, \cdot)$ is continuous (in fact, even Lipschitz continuous) on Θ . Note that if $V(x, \cdot)$ is differentiable and the set Θ is convex, then (2.1) holds with $\kappa(x) = \sup_{\theta \in \Theta} \|\nabla_{\theta} V(x, \theta)\|$. Note also that assumption (A3) implies that

$$V(x, \theta)^2 \leq 2V(x, \bar{\theta})^2 + 2\kappa(x)^2 \|\bar{\theta} - \theta\|^2, \quad \forall \theta, \bar{\theta} \in \Theta,$$

and hence if $\mathbb{E}[V(X, \bar{\theta})^2]$ is finite at some point $\bar{\theta} \in \Theta$, then it is finite for all $\theta \in \Theta$.

In order to simplify the presentation we also assume that the set Θ is compact. This assumption can be relaxed by replacing it by some other conditions, we will discuss this later.

Now let X_1, \dots, X_n be an iid random sample of n realizations of the random vector X . Consider the sample average function $\widehat{V}_n(\theta) := \frac{1}{n} \sum_{i=1}^n V(X_i, \theta)$, and the optimization problem

$$\min_{\theta \in \Theta} \widehat{V}_n(\theta). \tag{2.2}$$

By assumption (A3) the function $\widehat{V}_n(\cdot)$ is continuous on (compact) set Θ . Therefore, the optimal value of problem (2.2), denoted \hat{v}_n , is finite and this problem has a nonempty (and compact) set $\arg \min_{\theta \in \Theta} \widehat{V}_n(\theta)$ of optimal solutions. Under the above assumptions the optimal value \hat{v}_n (considered as a function of the random sample) is measurable and there exists a measurable selection $\hat{\theta}_n \in \arg \min_{\theta \in \Theta} \widehat{V}_n(\theta)$ (see, e.g., Rockafellar and Wets [4, Chapter 14]).

By the (strong) Law of Large Numbers we have that $\widehat{V}_n(\theta)$ converges (pointwise) w.p.1 to $v(\theta)$ as $n \rightarrow \infty$. In fact it is possible to show that, under the above assumptions, this convergence is uniform on Θ . It follows that \hat{v}_n and $\hat{\theta}_n$ converge w.p.1 to their counterparts of the limiting (expected value) optimization problem

$$\min_{\theta \in \Theta} v(\theta). \tag{2.3}$$

That is, $\hat{v}_n \rightarrow v^*$ and $\text{dist}(\hat{\theta}_n, \Theta^*) \rightarrow 0$ w.p.1, where $v^* := \inf_{\theta \in \Theta} v(\theta)$ and

$$\Theta^* := \arg \min_{\theta \in \Theta} v(\theta) \tag{2.4}$$

denote the optimal value and the set of optimal solutions, respectively, of the problem (2.3). The minimizer $\hat{\theta}_n$ is often referred to as an M estimator. Note that the assertion “ $\Theta^* = \{\theta^*\}$ is a singleton” simply means that the optimization problem (2.3) has unique optimal solution θ^* .

We can formulate now a basic asymptotic result [7, Theorems 3.2 and 3.3].

Theorem 1. Suppose that assumptions (A1)–(A3) hold and the set Θ is compact. Then $\hat{\vartheta}_n$ converges w.p.1 to ϑ^* , and

$$\hat{\vartheta}_n = \inf_{\theta \in \Theta^*} \widehat{V}_n(\theta) + o_p(n^{-1/2}). \tag{2.5}$$

If, moreover, $\Theta^* = \{\theta^*\}$ is a singleton, then

$$\hat{\vartheta}_n = \widehat{V}_n(\theta^*) + o_p(n^{-1/2}), \tag{2.6}$$

and $n^{1/2}(\hat{\vartheta}_n - \vartheta^*)$ converges in distribution to normal $N(0, \sigma^2(\theta^*))$, where

$$\sigma^2(\theta) := \text{Var}[V(X, \theta)] = \mathbb{E}[V(X, \theta)^2] - v(\theta)^2. \tag{2.7}$$

Remark 1. By the definition of the set Θ^* we have that $v(\theta) = \vartheta^*$ for any $\theta \in \Theta^*$. Therefore Eq. (2.5) can be written in the following equivalent form

$$n^{1/2}(\hat{\vartheta}_n - \vartheta^*) = \inf_{\theta \in \Theta^*} Y_n(\theta) + o_p(1), \tag{2.8}$$

where $Y_n(\theta) := n^{1/2}(\widehat{V}_n(\theta) - v(\theta))$. Consequently, $n^{1/2}(\hat{\vartheta}_n - \vartheta^*)$ has the same limit (asymptotic) distribution as the term $\inf_{\theta \in \Theta^*} Y_n(\theta)$ on the right-hand side of (2.8). Note also that for any set of points $\theta_1, \dots, \theta_k \in \Theta$, by the Central Limit Theorem (CLT), random vector $(Y_n(\theta_1), \dots, Y_n(\theta_k))$ converges in distribution to multivariate normal. This leads to a very general asymptotic result. In particular, if $\Theta^* = \{\theta^*\}$ is a singleton, it follows that $n^{1/2}(\hat{\vartheta}_n - \vartheta^*) \Rightarrow N(0, \sigma^2(\theta^*))$. If Θ^* is not a singleton, then the limiting distribution of $n^{1/2}(\hat{\vartheta}_n - \vartheta^*)$ is given by the minimum of (correlated) normally distributed random variables. That is, uniqueness of the minimizer θ^* is a sufficient and “almost necessary” condition for asymptotic normality of $\hat{\vartheta}_n$. ■

Proof of the above theorem is quite sophisticated, it is based on a *first order* expansion of the optimal value function, an infinite-dimensional Delta Theorem and functional CLT. Under considerably stronger assumptions, it is also possible to derive a second order term in an asymptotic expansion of the optimal value statistic $\hat{\vartheta}_n$. We discuss below a particular case of a general formula of [8, Theorem 4.4], which is sufficient for many statistical applications.

We make the following additional assumptions.

- (i) The set $\Theta^* = \{\theta^*\}$ is a singleton.
- (ii) For every $x \in \mathcal{X}$ the function $V(x, \cdot)$ is continuously differentiable.
- (iii) The expectation function $v(\theta)$ is twice continuously differentiable in neighborhood of the point θ^* .
- (iv) In a neighborhood $\mathcal{Y} \subset \mathbb{R}^m$ of the point θ^* the set Θ is defined by equality constraints, i.e.,

$$\Theta \cap \mathcal{Y} := \{\theta \in \mathcal{Y} : c_j(\theta) = 0, j = 1, \dots, k\}, \tag{2.9}$$

where $c_j(\cdot)$ are twice continuously differentiable functions.

- (v) Gradient vectors $\nabla c_j(\theta^*)_{j=1, \dots, k}$ are linearly independent.

Because of assumption (v), by the first order optimality conditions, there exist (uniquely defined) Lagrange multipliers $\bar{\lambda}_j, j = 1, \dots, k$, such that

$$\nabla v(\theta^*) + \sum_{j=1}^k \bar{\lambda}_j \nabla c_j(\theta^*) = 0. \tag{2.10}$$

Consider the Hessian matrix

$$H := \nabla^2 v(\theta^*) + \sum_{j=1}^k \bar{\lambda}_j \nabla^2 c_j(\theta^*) \tag{2.11}$$

and the linear space (of dimension $m - k$)

$$\mathcal{T} := \{h \in \mathbb{R}^m : h^\top \nabla c_j(\theta^*) = 0, j = 1, \dots, k\}. \tag{2.12}$$

Note that \mathcal{T} represents the tangent space to Θ at θ^* . Assume, further, that:

- (vi) Matrix H is positive definite on the linear space \mathcal{T} , i.e., $h^\top H h > 0$ for any $h \in \mathcal{T}, h \neq 0$.

Note that by second order necessary conditions we have that $h^T H h \geq 0$ for any $h \in \mathcal{T}$. Therefore assumption (vi) is, in a sense, a minimal requirement for nondegeneracy of the matrix H . In particular, assumption (vi) holds if the matrix H is nonsingular.

Under some mild additional conditions the following second order expansion of $\hat{\vartheta}_n$ holds

$$\hat{\vartheta}_n = \widehat{V}_n(\theta^*) + \frac{1}{2} \varphi(\zeta_n) + o_p(n^{-1}), \tag{2.13}$$

where $\zeta_n := \nabla \widehat{V}_n(\theta^*) - \nabla v(\theta^*)$ and

$$\varphi(z) := \inf_{h \in \mathcal{T}} \{2h^T z + h^T H h\} \tag{2.14}$$

(see [8] for technical details). The term $\widehat{V}_n(\theta^*)$, on the right-hand side of (2.13), is exactly the same as the corresponding term in the first order expansion (2.6). The second term $\frac{1}{2} \varphi(\zeta_n)$ is quadratic. That is, the function $\varphi(\cdot)$ is quadratic and can be written as

$$\varphi(z) = -z^T [A(A^T H A)^{-1} A^T] z, \tag{2.15}$$

where A is a matrix generating the linear space \mathcal{T} , i.e., A is an $m \times (m - k)$ matrix of full column rank $m - k$ such that $A^T [\nabla C_j(\theta^*)] = 0, j = 1, \dots, k$. The generating matrix A is defined up to a transformation $A \mapsto AC$, where C can be any nonsingular $(m - k) \times (m - k)$ matrix. Such transformation of the generating matrix does not change the right-hand side part of Eq. (2.15). Note also that $\varphi(\cdot) \leq 0$ (just take $h = 0$ on the right-hand side of (2.14)). In other words the matrix $A^T H A$ is positive definite. This should be not surprising in view of the second order optimality conditions discussed after the assumption (vi).

By the CLT, $n^{1/2} \zeta_n$ converges in distribution to a multivariate normal $N(0, \Psi)$, with Ψ equal to the covariance matrix of $\nabla_{\theta} V(X, \theta^*)$, i.e.,

$$\Psi = \mathbb{E} \left[(\nabla_{\theta} V(X, \theta^*) - \nabla v(\theta^*)) (\nabla_{\theta} V(X, \theta^*) - \nabla v(\theta^*))^T \right], \tag{2.16}$$

and hence $\zeta_n = O_p(n^{-1/2})$. Therefore, the additional (second order) term $\varphi(\zeta_n)$ in (2.13) is of order $O_p(n^{-1})$. We have that $\mathbb{E}[\widehat{V}_n(\theta^*)] = v(\theta^*) = \vartheta^*$, and the mean of limiting (asymptotic) distribution of $n \zeta_n^T [A(A^T H A)^{-1} A^T] \zeta_n$ is equal to $\text{tr} [A(A^T H A)^{-1} A^T \Psi]$. Therefore,

$$-\frac{1}{2} n^{-1} \text{tr} [A(A^T H A)^{-1} A^T \Psi] \tag{2.17}$$

can be viewed as the asymptotic bias of $\hat{\vartheta}_n$. By the above discussion the asymptotic bias (2.17) is always less than or equal to zero and typically is negative. This should be not surprising since $\mathbb{E}[\hat{\vartheta}_n] \leq \vartheta^*$ and $\mathbb{E}[\widehat{V}_n(\theta^*)] = \vartheta^*$. The asymptotic bias (2.17) is of order $O(n^{-1})$. It is interesting to note that if the set Θ^* is not a singleton, then the asymptotic distribution of the first term on the right-hand side of (2.8) typically has a negative mean, and hence in that case the asymptotic bias of $\hat{\vartheta}_n$ is of order $O(n^{-1/2})$.

In particular, suppose that there are no equality constraints in the definition of Θ , and θ^* is an interior point of the set Θ . Then $\mathcal{T} = \mathbb{R}^m$ and formulas (2.13)–(2.16) can be applied with $H = \nabla^2 v(\theta^*)$ and $\nabla v(\theta^*) = 0$. In that case $\varphi(z) = -z^T H^{-1} z$ and the asymptotic bias is equal to $-\frac{1}{2} n^{-1} \text{tr} [H^{-1} \Psi]$.

2.2. Framework of moment structures

Let $\hat{\lambda}_n$ be an estimate of an unknown (population) vector $x_0 \in \mathbb{R}^d$. For example, $\hat{\lambda}_n$ can be the sample average and/or sample covariance matrix, based on a sample of size n , viewed as an estimate of the corresponding population mean and/or population covariance matrix. We assume that $\hat{\lambda}_n$ and x_0 vary in a convex open set \mathcal{X} . For example, if $\hat{\lambda}_n$ is represented by the sample covariance matrix, then \mathcal{X} is formed by positive definite matrices of the corresponding dimension. We make the following assumptions.

(B1) As n tends to infinity, $n^{1/2}(\hat{\lambda}_n - x_0)$ converges in distribution to normal $N(0, \Omega)$.

The above assumption implies, of course, that $\hat{\lambda}_n$ converges in probability to x_0 .

As before, we also assume that Θ is a nonempty compact parameter set. Furthermore, let $q : \mathcal{X} \times \Theta \rightarrow \mathbb{R}$ be a given function, and let

$$v^* := \inf_{\theta \in \Theta} q(x_0, \theta) \quad \text{and} \quad \Theta^* := \arg \min_{\theta \in \Theta} q(x_0, \theta) \tag{2.18}$$

be the optimal value and the set of optimal solutions, respectively, of the optimization problem associated with the population vector x_0 , while

$$\hat{v}_n := \inf_{\theta \in \Theta} q(\hat{\lambda}_n, \theta) \quad \text{and} \quad \hat{\theta}_n \in \arg \min_{\theta \in \Theta} q(\hat{\lambda}_n, \theta) \tag{2.19}$$

are the corresponding estimators.

We make the following assumption about function $q(x, \theta)$.

(B2) For every $\theta \in \Theta$ the function $q(\cdot, \theta)$ is differentiable and $\nabla_x q(x, \theta)$ is continuous, jointly in x and θ , on the set $\mathcal{X} \times \Theta$. It follows, of course, that the function $q(x, \theta)$ itself is continuous on $\mathcal{X} \times \Theta$, and hence is measurable with respect to the Borel sigma algebra of $\mathcal{X} \times \Theta$. Consider

$$Z_n(\theta) := n^{1/2} [\nabla_x q(x_0, \theta)]^T (\hat{x}_n - x_0). \tag{2.20}$$

Because of the condition (B1), $Z_n(\theta) \Rightarrow N(0, \sigma(\theta)^2)$, where

$$\sigma(\theta)^2 := [\nabla_x q(x_0, \theta)]^T \Omega [\nabla_x q(x_0, \theta)]. \tag{2.21}$$

We have the following result about asymptotics of \hat{v}_n [6, Theorem 5.3].

Theorem 2. Suppose that assumptions (B1)–(B2) hold and the set Θ is compact, and consider $Z_n(\theta)$ and $\sigma(\theta)^2$ defined in (2.20) and (2.21), respectively. Then \hat{v}_n converges in probability to v^* and

$$n^{1/2}(\hat{v}_n - v^*) = \inf_{\theta \in \Theta^*} Z_n(\theta) + o_p(1). \tag{2.22}$$

If, moreover, $\Theta^* = \{\theta^*\}$ is a singleton, then $n^{1/2}(\hat{v}_n - v^*) \Rightarrow N(0, \sigma(\theta^*)^2)$.

Results of Theorems 1 and 2 give, in a sense, first order asymptotics of the corresponding optimal value statistics and can be applied in different situations. In applications of these results to studying asymptotics of likelihood ratio and minimum discrepancy test statistics, discussed in the next section, it is natural to assume that the set Θ^* of optimal solutions is a singleton. However, these results go beyond these applications and there are situations where this assumption does not hold. As an example of the case where Θ^* essentially is not a singleton we may refer to asymptotics of the so-called minimum trace factor analysis (cf., [5]).

The assumption of compactness of Θ can be replaced by the condition that for all n large enough, $\hat{\theta}_n$ stays in a compact subset of Θ wp.1. This condition, in turn, can be often verified by ad hoc methods.

Under stronger assumptions it is also possible to add a second order term in an expansion of \hat{v}_n . We assume in the remainder of this section that $\Theta^* = \{\theta^*\}$ is a singleton, i.e., $q(x_0, \cdot)$ has unique minimizer θ^* over the parameter set Θ . Suppose, further, that near the point θ^* the parameter set is defined by equality constraints in the form (2.9) and assume the following.

(B3) The function $q(\cdot, \cdot)$ is twice continuously differentiable.

(B4) The constraint functions $c_j(\cdot), j = 1, \dots, k$, are twice continuously differentiable and gradient vectors $\nabla c_j(\theta^*)_{j=1, \dots, k}$ are linearly independent.

It follows by the first order optimality conditions that there exist (uniquely defined) Lagrange multipliers $\bar{\lambda}_j, j = 1, \dots, k$, associated with the minimizer θ^* of $q(x_0, \cdot)$, such that

$$\nabla_\theta q(x_0, \theta^*) + \sum_{j=1}^k \bar{\lambda}_j \nabla c_j(\theta^*) = 0. \tag{2.23}$$

Consider the following Hessian matrices

$$H_{xx} := \frac{\partial^2 q(x_0, \theta^*)}{\partial x \partial x^T}, \quad H_{x\theta} := \frac{\partial^2 q(x_0, \theta^*)}{\partial x \partial \theta^T} \quad \text{and} \quad H_{\theta\theta} := \frac{\partial^2 q(x_0, \theta^*)}{\partial \theta \partial \theta^T} + \sum_{j=1}^k \bar{\lambda}_j \frac{\partial^2 c_j(\theta^*)}{\partial \theta \partial \theta^T}, \tag{2.24}$$

of order $d \times d, d \times m$ and $m \times m$, respectively. By the second order necessary conditions we have that $h^T H_{\theta\theta} h \geq 0$ for any $h \in \mathcal{T}$, where \mathcal{T} is the linear (tangent) space defined in (2.12). We assume the following second order sufficient conditions.

(B5) For any $h \in \mathcal{T}, h \neq 0$, it holds that $h^T H_{\theta\theta} h > 0$.

Consider, further,

$$\psi(z) := \inf_{h \in \mathcal{T}} \{z^T H_{xx} z + 2z^T H_{x\theta} h + h^T H_{\theta\theta} h\}. \tag{2.25}$$

The function $\psi(\cdot)$ is quadratic and can be written as $\psi(z) = z^T Q z$, where

$$Q := H_{xx} - H_{x\theta} A (A^T H_{\theta\theta} A)^{-1} A^T H_{x\theta}^T, \tag{2.26}$$

and A is an $m \times (m - k)$ matrix of full column rank generating the linear space \mathcal{T} . If $\mathcal{T} = \mathbb{R}^m$, i.e., θ^* is an interior point of Θ , then we can take A as the identity matrix and in that case $Q = H_{xx} - H_{x\theta} H_{\theta\theta}^{-1} H_{x\theta}^T$. We have the following result (cf., [6, Theorem 5.4]).

Theorem 3. Suppose that $\Theta^* = \{\theta^*\}$ is a singleton, assumptions (B1), (B3)–(B5) hold and Θ is compact. Then

$$\hat{v}_n - v^* = [\nabla_x q(x_0, \theta^*)]^\top (\hat{x}_n - x_0) + \frac{1}{2} (\hat{x}_n - x_0)^\top Q (\hat{x}_n - x_0) + o_p(n^{-1}), \tag{2.27}$$

where matrix Q is defined in (2.26).

If \hat{x}_n is an unbiased estimator of x_0 , then the expectation of the first term on the right-hand side of (2.27) is zero. In that case $\frac{1}{2}n^{-1}\text{tr}[\Omega Q]$ can be viewed as an asymptotic bias of the estimator \hat{v}_n of v^* .

3. Applications and examples

As first application we consider the classical maximum likelihood method. Let X be a random vector whose true, but unknown, probability distribution P is modelled by density $f(x, \theta)$ depending on parameter vector $\theta \in \Theta$. Let X_1, \dots, X_n be an iid random sample of X and $L_n^f(\theta) := \prod_{i=1}^n f(X_i, \theta)$ be the corresponding likelihood function. The ML estimator is obtained by maximizing $L_n^f(\theta)$ over $\theta \in \Theta$. This can be formulated in the framework of problem (2.2) by taking $V(x, \theta) := -\log f(x, \theta)$ and hence getting $v(\theta) = -\mathbb{E}_P[\log f(X, \theta)]$. Consider the statistic

$$T_n^f := \sup_{\theta \in \Theta} \log L_n^f(\theta). \tag{3.1}$$

Note that

$$-n^{-1}T_n^f = \inf_{\theta \in \Theta} \left\{ \frac{1}{n} \sum_{i=1}^n [-\log f(X_i, \theta)] \right\},$$

that is, $-n^{-1}T_n^f$ is the optimal value of the corresponding minimization problem (2.2) for $V(x, \theta) = -\log f(x, \theta)$.

Consider $T_*^f := \sup_{\theta \in \Theta} \mathbb{E}_P [\log f(X, \theta)]$ and the set

$$\Theta^* := \arg \max_{\theta \in \Theta} \mathbb{E}_P [\log f(X, \theta)] = \arg \min_{\theta \in \Theta} \mathbb{E}_P [-\log f(X, \theta)]. \tag{3.2}$$

By Theorem 1 we have here, under the respective assumptions (A1)–(A3) and compactness of Θ , the following asymptotics

$$n^{-1/2}T_n^f = n^{-1/2} \sup_{\theta \in \Theta^*} \log L_n^f(\theta) + o_p(1), \tag{3.3}$$

and that $n^{-1}T_n^f$ converges to T_*^f w.p.1. Suppose, further, that $\Theta^* = \{\theta^*\}$ is a singleton, and near θ^* the set Θ is defined by equality constraints in the form (2.9). Then under appropriate regularity conditions, in particular ensuring that derivatives with respect to θ can be taken inside the expectation, we have by (2.13) the following second order expansion

$$T_n^f = \log L_n^f(\theta^*) - \frac{1}{2}n\zeta_n [A(A^\top HA)^{-1}A^\top] \zeta_n + o_p(1), \tag{3.4}$$

where A is a matrix generating the linear space \mathcal{T} and

$$\zeta_n := \frac{1}{n} \frac{\partial \log L_n^f(\theta^*)}{\partial \theta} - \mathbb{E}_P \left[\frac{\partial \log f(X, \theta^*)}{\partial \theta} \right] \quad \text{and} \quad H := \mathbb{E}_P \left[\frac{\partial^2 \log f(X, \theta^*)}{\partial \theta \partial \theta^\top} \right] + \sum_{j=1}^k \bar{\lambda}_j \frac{\partial^2 c(\theta^*)}{\partial \theta \partial \theta^\top}. \tag{3.5}$$

Remark 2. Suppose that the model is correct, i.e., there is $\theta_0 \in \Theta$ such that the true distribution P is defined by the density $f(\cdot, \theta_0)$, written $P \stackrel{d}{=} f(\cdot, \theta_0)$. Then, as is well known, θ_0 is a maximizer of $\mathbb{E}_P[\log f(X, \cdot)]$, and hence $\theta_0 \in \Theta^*$. Consequently in that case

$$n^{-1/2}T_n^f = n^{-1/2} \log L_n^f(\theta_0) + o_p(1). \tag{3.6}$$

Note that if there are several values of the parameter vector defining the same distribution P , then formula (3.6) still holds with θ_0 being any point of the parameter set Θ such that $P \stackrel{d}{=} f(\cdot, \theta_0)$.

Also in that case, under appropriate conditions ensuring that derivatives with respect to θ can be taken inside the expectation, we have in (3.5) that the term $\mathbb{E}_P \left[\frac{\partial \log f(X, \theta_0)}{\partial \theta} \right] = 0$, the Lagrange multipliers $\bar{\lambda}_j = 0, j = 1, \dots, k$, and $H = -I(\theta_0)$, where

$$I(\theta_0) = -\mathbb{E}_P \left[\frac{\partial^2 \log f(X, \theta_0)}{\partial \theta \partial \theta^\top} \right] = \mathbb{E}_P \left[\frac{\partial \log f(X, \theta_0)}{\partial \theta} \frac{\partial \log f(X, \theta_0)}{\partial \theta^\top} \right] \tag{3.7}$$

is Fisher’s information matrix. ■

Now let $g(x, \gamma)$ be an alternative density model parameterized by vector $\gamma \in \Gamma$. Consider the corresponding statistic $T_n^g := \sup_{\gamma \in \Gamma} \log L_n^g(\gamma)$, where $L_n^g(\gamma) := \prod_{i=1}^n g(X_i, \gamma)$, and the associated log-likelihood ratio test statistic

$$T_n := T_n^g - T_n^f. \tag{3.8}$$

Denote $\Gamma^* := \arg \max_{\gamma \in \Gamma} \mathbb{E}_P[\log g(X, \gamma)]$ and

$$T^* := \sup_{\gamma \in \Gamma} \mathbb{E}_P[\log g(X, \gamma)] - \sup_{\theta \in \Theta} \mathbb{E}_P[\log f(X, \theta)].$$

By Theorem 1 and Eq. (3.3) we have the following result.

Theorem 4. Suppose that the functions $\log f(x, \theta), \theta \in \Theta$, and $\log g(x, \gamma), \gamma \in \Gamma$, satisfy the respective assumptions (A1)–(A3), and that the sets Θ and Γ are compact. Then

$$n^{-1/2}T_n = n^{-1/2} \sup_{\gamma \in \Gamma^*} \log L_n^g(\gamma) - n^{-1/2} \sup_{\theta \in \Theta^*} \log L_n^f(\theta) + o_p(1), \tag{3.9}$$

and $n^{-1}T_n$ converges to T^* w.p.1. If, moreover, $\Theta^* = \{\theta^*\}$ and $\Gamma^* = \{\gamma^*\}$ are singletons, then $n^{1/2} (n^{-1}T_n - T^*)$ converges in distribution to normal $N(0, \omega^*)$ with variance

$$\omega^* = \text{Var} \left[\log \frac{f(X, \theta^*)}{g(X, \gamma^*)} \right] = \mathbb{E}_P \left[\left(\log \frac{f(X, \theta^*)}{g(X, \gamma^*)} \right)^2 \right] - \left(\mathbb{E}_P \left[\log \frac{f(X, \theta^*)}{g(X, \gamma^*)} \right] \right)^2. \tag{3.10}$$

In case where Θ^* and Γ^* are singletons, the above convergence result and formula (3.10) were obtained in Vuong [9, Theorem 3.3] under considerably stronger regularity conditions. In particular, it was assumed there that θ^* and γ^* are interior points of the respective parameter sets.

Suppose now that we consider a parameterized model $f(x, \theta)$ and are interested in testing the (not necessarily nested) alternatives

$$H_0 : \theta \in \Theta_0 \quad \text{against} \quad H_1 : \theta \in \Theta_1. \tag{3.11}$$

In that case the log-likelihood ratio test statistic is

$$T_n = \sup_{\theta \in \Theta_1} \log L_n(\theta) - \sup_{\theta \in \Theta_0} \log L_n(\theta), \tag{3.12}$$

where $L_n(\theta) = L_n^f(\theta)$ is the corresponding likelihood function. By (3.9) we have here, under appropriate regularity conditions, that

$$n^{-1/2}T_n = n^{-1/2} \sup_{\theta \in \Theta_1^*} \log L_n(\theta) - n^{-1/2} \sup_{\theta \in \Theta_0^*} \log L_n(\theta) + o_p(1), \tag{3.13}$$

where

$$\Theta_j^* = \arg \max_{\theta \in \Theta_j} \mathbb{E}_P[\log f(X, \theta)], \quad j = 0, 1.$$

If, moreover, the function $\mathbb{E}_P[\log f(X, \cdot)]$ has unique maximizers θ_0^* and θ_1^* over the sets Θ_0 and Θ_1 , respectively, then

$$n^{-1/2}T_n = n^{-1/2} \log \frac{L_n(\theta_1^*)}{L_n(\theta_0^*)} + o_p(1). \tag{3.14}$$

Remark 3. Of course, if $\theta_0^* = \theta_1^*$, then the right-hand side of (3.14) is reduced to $o_p(1)$ and Eq. (3.14) simply says that $n^{-1/2}T_n$ converges in probability to zero. This happens if there exists a point $\theta_0 \in \Theta_1 \cap \Theta_0$ such that $P \stackrel{d}{=} f(\cdot, \theta_0)$, in which case $\theta_0^* = \theta_1^* = \theta_0$ (see Remark 2). In that case, in order to get a meaningful asymptotics of T_n , a second order expansion of the form (3.4) is required. By the classical result (cf., Wilks [11], Wald [10]), under H_0 (and certain regularity conditions) the test statistic $2T_n$ converges in distribution to a chi-square if parameter sets Θ_0 and Θ_1 are defined by (smooth) equality constraints and the hypotheses are nested (i.e., $\Theta_0 \subset \Theta_1$). In general, for not necessarily nested models, we have by (3.4) that if $P \stackrel{d}{=} f(\cdot, \theta_0)$, where $\theta_0 \in \Theta_0 \cap \Theta_1$, then (under appropriate regularity conditions)

$$2T_n \Rightarrow Y^T A_1 (A_1^T I(\theta_0) A_1)^{-1} A_1^T Y - Y^T A_0 (A_0^T I(\theta_0) A_0)^{-1} A_0^T Y, \tag{3.15}$$

where $I(\theta_0)$ is Fisher’s information matrix (defined in (3.7)), Y is a random vector having normal $N(0, I(\theta_0))$ distribution, and A_0 and A_1 are matrices generating the tangent spaces \mathcal{T}_0 and \mathcal{T}_1 to Θ_0 and Θ_1 , respectively, at θ_0 . If $\mathcal{T}_0 \subset \mathcal{T}_1$, i.e., the models are nested, then we can take matrix A_1 of the form $A_1 = [A_0, B]$, where B is an $m \times \nu$ matrix of full column rank $\nu = \dim(\mathcal{T}_1) - \dim(\mathcal{T}_0)$ such that $A_0^T I(\theta_0) B = 0$. Then the right-hand side of (3.15) is equal to $W^T (B^T I(\theta_0) B)^{-1} W$, where $W := B^T Y \sim N(0, B^T I(\theta_0) B)$. Consequently, in that case $W^T (B^T I(\theta_0) B)^{-1} W \sim \chi_\nu^2$, and hence $2T_n$ converges in distribution to (central) chi-square distribution with $\nu = \dim(\mathcal{T}_1) - \dim(\mathcal{T}_0)$ degrees of freedom.

3.1. Covariance structures' analysis

Let us discuss now the following example of covariance structures' analysis. Suppose that the probability distribution P , of random vector X , is hypothesized to be multivariate normal, $X \sim N(\mu, \Sigma(\theta))$, with covariance matrix $\Sigma(\theta)$ being function of parameter vector $\theta \in \mathbb{R}^m$ and mean vector μ treated as nuisance parameter. Suppose, further, that we are interested in testing alternative hypotheses of the form (3.11), where Θ_0 and Θ_1 are (not necessarily nested) subsets of \mathbb{R}^m . The log-likelihood function here (up to a constant) is

$$L_n(\theta) = -\frac{n}{2} \log |\Sigma(\theta)| - \frac{1}{2} \sum_{i=1}^n (X_i - \mu)^\top \Sigma(\theta)^{-1} (X_i - \mu). \tag{3.16}$$

The corresponding (two times) log-likelihood ratio test statistic can be written in the form

$$T_n = n \left[\inf_{\theta \in \Theta_0} F_{ML}(S, \Sigma(\theta)) - \inf_{\theta \in \Theta_1} F_{ML}(S, \Sigma(\theta)) \right], \tag{3.17}$$

where $S := n^{-1} \sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})^\top$ is the ML estimator² of the covariance matrix and

$$F_{ML}(S, \Sigma) := \log |\Sigma| + \text{tr}(S \Sigma^{-1}) - \log |S| - p \tag{3.18}$$

is the so-called ML discrepancy function.³

By (3.13) we have here, under appropriate regularity conditions, that

$$n^{-1/2} T_n = n^{1/2} \left[\inf_{\theta \in \Theta_0^*} F_{ML}(S, \Sigma(\theta)) - \inf_{\theta \in \Theta_1^*} F_{ML}(S, \Sigma(\theta)) \right] + o_p(1), \tag{3.19}$$

where

$$\Theta_j^* := \arg \min_{\theta \in \Theta_j} F_{ML}(S_0, \Sigma(\theta)), \quad j = 0, 1,$$

and $S_0 := \mathbb{E}_P[(X - \mu)(X - \mu)^\top]$ is the population covariance matrix. We also have that $n^{-1} T_n$ converges w.p.1 to

$$T^* := \inf_{\theta \in \Theta_0} F_{ML}(S_0, \Sigma(\theta)) - \inf_{\theta \in \Theta_1} F_{ML}(S_0, \Sigma(\theta)). \tag{3.20}$$

Moreover, if $\Theta_0^* = \{\theta_0^*\}$ and $\Theta_1^* = \{\theta_1^*\}$ are singletons, then

$$n^{1/2} (n^{-1} T_n - T^*) = n^{1/2} \text{tr} \left[(\Sigma(\theta_0^*)^{-1} - \Sigma(\theta_1^*)^{-1}) (S - S_0) \right] + o_p(1). \tag{3.21}$$

Denote $s := \text{vec}(S)$ and $\sigma_0 := \text{vec}(S_0)$, where $\text{vec}(S)$ operator stacks columns of $p \times p$ matrix S into $p^2 \times 1$ column vector. Assuming that components of the random vector X have fourth order moments, we have by the CLT that $n^{1/2}(s - \sigma_0)$ converges in distribution to normal with zero mean vector and $p^2 \times p^2$ covariance matrix denoted Ω . Then it follows from (3.21) that $n^{1/2} (n^{-1} T_n - T^*)$ converges in distribution to normal with zero mean and variance

$$\omega^* = [\text{vec}(\Sigma(\theta_0^*)^{-1} - \Sigma(\theta_1^*)^{-1})]^\top \Omega [\text{vec}(\Sigma(\theta_0^*)^{-1} - \Sigma(\theta_1^*)^{-1})]. \tag{3.22}$$

If vector X has multivariate normal distribution $N(\mu, \Sigma_0)$, then

$$\Omega = 2M_p(\Sigma_0 \otimes \Sigma_0), \tag{3.23}$$

with M_p being a certain $p^2 \times p^2$ symmetric idempotent matrix of rank $p(p + 1)/2$ (cf., Browne [1]), and formula (3.22) for the asymptotic variance takes the form

$$\omega^* = 2 \text{tr} \left\{ [(\Sigma(\theta_0^*)^{-1} - \Sigma(\theta_1^*)^{-1}) \Sigma_0]^2 \right\}. \tag{3.24}$$

The F_{ML} discrepancy function has the following properties: for any covariance matrices S and Σ , $F_{ML}(S, \Sigma) \geq 0$ and $F_{ML}(S, \Sigma) = 0$ iff $S = \Sigma$. It follows that if there is $\theta_0 \in \Theta_0 \cap \Theta_1$ such that $S_0 = \Sigma(\theta_0)$ (i.e., both hypotheses H_0 and H_1 hold), then θ_0 is a minimizer of $F_{ML}(S_0, \Sigma(\cdot))$, over both parameter sets, and $F_{ML}(S_0, \Sigma(\theta_0)) = 0$. In that case $T^* = 0$ and Eq. (3.19) reduces to the statement that $n^{-1/2} T_n$ converges in probability to zero. As it was mentioned earlier, in that case second order expansions are needed to get a meaningful asymptotics.

² The standard (unbiased) estimator of the covariance matrix differs from the ML estimator by the factor of $n/(n - 1)$. We denote here by S the ML estimator to simplify notation. Of course, $n/(n - 1)$ tends to one as $n \rightarrow \infty$, and this factor does not change the corresponding asymptotics.

³ The term $-\log |S| - p$ here does not depend on θ and can be omitted. This term appears while testing the parameterized model against the corresponding saturated model. This definition of the ML discrepancy function is standard in the theory of covariance structures.

In the present case it is also possible to employ the machinery of [Theorem 2](#) to derive the above asymptotics. We can view $F_{ML}(S, \Sigma(\theta))$ as a function of vector $s = \text{vec}(S)$ and parameter vector θ . We have that

$$\left. \frac{\partial F_{ML}(s, \sigma(\theta))}{\partial s} \right|_{s=\sigma_0} = \text{vec}(\Sigma(\theta)^{-1} - \Sigma_0^{-1}). \tag{3.25}$$

By using [Theorem 2](#) together with formula (3.25) it is straightforward to derive the above asymptotics (3.19)–(3.22). The required regularity conditions here are quite simple.

- (C1) The function $\Sigma(\cdot)$ is continuous.
- (C2) The parameter sets Θ_0 and Θ_1 are compact.
- (C3) $n^{1/2}(s - \sigma_0)$ converges in distribution to normal $N(0, \Omega)$.
- (C4) The sets $\Theta_0^* = \{\theta_0^*\}$ and $\Theta_1^* = \{\theta_1^*\}$ are singletons.

For formula (3.19) we only need to assume (C1)–(C3). If, moreover, (C4) holds, then (3.21) and (3.22) follow.

In the analysis of covariance (moment) structures it is more convenient to use [Theorem 2](#) rather than [Theorem 1](#). Sometimes other discrepancy functions, than the ML discrepancy function, are used. Consider, for example, the Generalized Least Squares discrepancy function

$$F_{GLS}(S, \Sigma) := \frac{1}{2} \text{tr} \left\{ [(S - \Sigma)S^{-1}]^2 \right\}, \tag{3.26}$$

and the corresponding test statistic T_n of the form (3.17) with F_{ML} replaced by F_{GLS} . This test statistic cannot be written as a likelihood ratio statistic and therefore [Theorem 1](#) cannot be applied. On the other hand [Theorem 2](#) can be applied in a straightforward way. We have that

$$\left. \frac{\partial F_{GLS}(s, \sigma(\theta))}{\partial s} \right|_{s=\sigma_0} = \text{vec} \left[\Sigma_0^{-1} (\Sigma(\theta) - \Sigma(\theta)\Sigma_0^{-1}\Sigma(\theta)) \Sigma_0^{-1} \right]. \tag{3.27}$$

Assume (C1)–(C4), with θ_0^* and θ_1^* being (unique) minimizers of $F_{GLS}(\Sigma_0, \Sigma(\theta))$ over $\theta \in \Theta_0$ and $\theta \in \Theta_1$, respectively. Then

$$n^{1/2} (n^{-1}T_n - T^*) = n^{1/2} \text{tr} \left[G(\theta_0^*, \theta_1^*)(S - \Sigma_0) \right] + o_p(1), \tag{3.28}$$

where

$$G(\theta_0^*, \theta_1^*) := \Sigma_0^{-1} \left[\Sigma(\theta_0^*) - \Sigma(\theta_0^*)\Sigma_0^{-1}\Sigma(\theta_0^*) - \Sigma(\theta_1^*) + \Sigma(\theta_1^*)\Sigma_0^{-1}\Sigma(\theta_1^*) \right] \Sigma_0^{-1}. \tag{3.29}$$

Consequently $n^{1/2} (n^{-1}T_n - T^*)$ converges in distribution to normal with zero mean and variance

$$\omega^* = \left[\text{vec} \left(G(\theta_0^*, \theta_1^*) \right) \right]^T \Omega \left[\text{vec} \left(G(\theta_0^*, \theta_1^*) \right) \right]. \tag{3.30}$$

If, moreover, formula (3.23) holds, then

$$\omega^* = 2 \text{tr} \left\{ \left[\Sigma_0^{-1}\Sigma(\theta_0^*) - \Sigma_0^{-1}\Sigma(\theta_0^*)\Sigma_0^{-1}\Sigma(\theta_0^*) - \Sigma_0^{-1}\Sigma(\theta_1^*) + \Sigma_0^{-1}\Sigma(\theta_1^*)\Sigma_0^{-1}\Sigma(\theta_1^*) \right]^2 \right\}. \tag{3.31}$$

If the alternative hypothesis (hypothesis H_1) is saturated, i.e., under H_1 the covariance matrix is unconstrained, then formulas (3.21), (3.22), (3.24) and (3.31) hold with matrix $\Sigma(\theta_1^*)$ replaced by matrix Σ_0 .

It is also possible to make bias corrections by using the second order term specified in [Theorem 3](#). Let us finally remark that, as it was mentioned before, if $\theta_0^* = \theta_1^*$, then the first order term in the expansion of the test statistic vanishes and the asymptotic variance ω^* degenerates into 0. This happens if there is $\theta_0 \in \Theta_0 \cap \Theta_1$ such that $\Sigma_0 = \Sigma(\theta_0)$. In that case a meaningful asymptotics of T_n is obtained by employing the corresponding second order expansions of the discrepancy function. This is the classical situation which in the case of nested models leads to a chi-square asymptotic distribution.

Acknowledgment

The author was supported in part by the National Science Foundation awards DMS-0510324 and DMI-0619977.

References

- [1] M.W. Browne, Generalized least squares estimators in the analysis of covariance structures, *South African Statistical Journal* 8 (1974) 1–24.
- [2] R.M. Golden, Discrepancy risk model selection test theory for comparing possibly misspecified or nonnested models, *Psychometrika* 68 (2003) 229–249.
- [3] D.A. McManus, Who invented local power analysis? *Econometric Theory* 7 (1991) 265–268.
- [4] R.T. Rockafellar, R.J.-B. Wets, *Variational Analysis*, Springer, Berlin, 1998.
- [5] A. Shapiro, Rank-reducibility of a symmetric matrix and sampling theory of minimum trace factor analysis, *Psychometrika* 47 (1982) 187–199.
- [6] A. Shapiro, Asymptotic distribution theory in the analysis of covariance structures (a unified approach), *South African Statistical Journal* 17 (1983) 33–81.
- [7] A. Shapiro, Asymptotic analysis of stochastic programs, *Annals of Operations Research* 30 (1991) 169–186.

- [8] A. Shapiro, Statistical inference of stochastic optimization problems, in: S. Uryasev (Ed.), *Probabilistic Constrained Optimization: Methodology and Applications*, Kluwer Academic Publishers, 2000, pp. 282–304.
- [9] Q.H. Vuong, Likelihood ratio tests for model selection and non-nested hypotheses, *Econometrica* 57 (1989) 307–333.
- [10] A. Wald, Tests of statistical hypotheses concerning several parameters when the number of observations is large, *Transaction of the American Mathematical Society* 54 (1943) 426–482.
- [11] S.S. Wilks, The Large sample distribution of the likelihood ratio for testing composite hypotheses, *Annals of Mathematical Statistics* 9 (1938) 60–62.
- [12] K.H. Yuan, K. Hayashi, P.M. Bentler, Normal theory likelihood ratio statistic for mean and covariance structure analysis under alternative hypotheses, *Journal of Multivariate Analysis* 98 (2007) 1262–1282.