Stat 8112 Lecture Notes

The Wilks, Wald, and Rao Tests

Charles J. Geyer

September 26, 2020

1 Introduction

One of the most familiar of results about maximum likelihood is that the likelihood ratio test statistic has an asymptotic chi-square distribution. Those who like eponyms call this the Wilks theorem and the hypothesis test using this test statistic the Wilks test.[1] Let $\hat{\theta}_n$ be the MLE for a model and θ_n^* the MLE for a smooth submodel. The Wilks test statistic is

$$T_n = 2\big[l_n(\hat{\theta}_n) - l_n(\theta_n^*)\big] \tag{1}$$

and the Wilks theorem says

$$T_n \xrightarrow{w} \text{ChiSq}(p - d),$$

where p is the dimension of the parameter space of the model and d is the dimension of the parameter space of the submodel.

It is sometimes easy to calculate one of $\hat{\theta}_n$ and θ_n^* and difficult or impossible to calculate the other. This motivates two other procedures that are asymptotically equivalent to the Wilks test. The Rao test statistic is

$$R_n = \big(\nabla l_n(\theta_n^*)\big)^T J_n(\theta_n^*)^{-1} \nabla l_n(\theta_n^*), \tag{2}$$

where

$$J_n(\theta) = -\nabla^2 l_n(\theta)$$

is the observed Fisher information matrix for sample size n. Under the conditions for the Wilks theorem, the Rao test statistic is asymptotically equivalent to the Wilks test statistic

$$R_n - T_n = o_p(1). \tag{3}$$

The test using the statistic R_n called the *Rao test* or the *score test* or the *Lagrange multiplier test*. An important point about the Rao test statistic is

[1] The name is Wilks with an "s" so if one wants a possessive form it would be Wilks's theorem

that, unlike the likelihood ratio test statistic, it only depends on the MLE for the null hypothesis θ_n^*.

The Wald test statistic is

$$W_n = g(\hat{\theta}_n)^T \big[\nabla g(\hat{\theta}_n) J_n(\hat{\theta}_n)^{-1} \big(\nabla g(\hat{\theta}_n) \big)^T \big]^{-1} g(\hat{\theta}_n), \tag{4}$$

where g is a vector-to-vector constraint function such that the submodel is the set of θ such that $g(\theta) = 0$. Under the conditions for the Wilks theorem, the Wald test statistic is asymptotically equivalent to the Wilks test statistic

$$W_n - T_n = o_p(1). \tag{5}$$

An important point about the Wald test statistic is that, unlike the likelihood ratio test statistic, it only depends on the MLE for the alternative hypothesis $\hat{\theta}_n$.

2 Setup

We work under the setup in Geyer (2013). Write

$$q_n(\delta) = l_n(\theta_0 + \tau_n^{-1}\delta) - l_n(\theta_0), \tag{6}$$

where θ_0 is the true unknown parameter value and τ_n is the rate of convergence of the MLE's, that is, we assume

$$\tau_n(\hat{\theta}_n - \theta_0) = O_p(1)$$
$$\tau_n(\theta_n^* - \theta_0) = O_p(1)$$

(our (6) is (14) in Geyer (2013) except that article allows the true unknown parameter to be different for each n). Geyer (2013) says: in "no-n" asymptotics τ_n plays no role and we can take $\tau_n = 1$, so all (6) does is shift the origin to the true unknown parameter value, so the null hypothesis is $\delta = 0$ in the new parameterization. In the "usual asymptotics" of maximum likelihood (appendix C of that article) we have $\tau_n = n^{1/2}$, where n is the sample size. In this handout we need $\tau_n \to \infty$ unless the submodel is flat.

We assume the random function q_n and the maximum likelihood estimators $\hat{\theta}_n$ and θ_n^* satisfy the conditions of Theorem 3.3 in Geyer (2013), which we call the LAMN conditions. These are implied by the "usual regularity conditions" for maximum likelihood Appendix C of that article.

3 Model and Submodel

3.1 Explicit Submodel Parameterization

There are two ways of specifying a submodel. The first is to specify a one-to-one vector-to-vector function h whose range is the parameter space of the submodel. Then we can write

$$\theta = h(\varphi) \tag{7}$$

and consider φ the parameter for the submodel, so the log likelihood for the parameter φ and for the submodel is

$$l_n^*(\varphi) = l_n(h(\varphi)).$$

As always, the reference distribution used for calculating P-values or α-levels assumes the truth is in the null hypothesis (submodel), that is, the true unknown parameter value for θ is

$$\theta_0 = h(\varphi_0), \tag{8}$$

where φ_0 is the true unknown parameter value for φ.

The corresponding "q" function for the submodel is

$$q_n^*(\varepsilon) = l_n^*(\varphi_0 + \tau_n^{-1}\varepsilon) - l_n^*(\varphi_0). \tag{9}$$

Maximizing q_n^* gives the MLE ε_n^* for the parameter ε. Then

$$\varphi_n^* = \varphi_0 + \tau_n^{-1}\varepsilon_n^*$$

is the MLE for the parameter φ, and

$$\theta_n^* = h(\varphi_n^*)$$

is the MLE in the submodel for the original parameter θ.

3.2 Lagrange Multipliers

Alternatively, the submodel can be given (as in our discussion of the Wald test) by a constraint function g, in which case the submodel parameter space is the set of θ such that $g(\theta) = 0$. The submodel MLE can be determined without explicit parameterization of the submodel using the method of Lagrange multipliers. The function

$$L_\lambda(\theta) = l_n(\theta) - \lambda^T g(\theta) \tag{10}$$

is called the *Lagrangian function* for the problem and λ is called the vector of *Lagrange multipliers* for the problem. The method of Lagrange multipliers seeks to maximize (10) while simultaneously satisfying the constraint, that is, we attempt to solve the simultaneous equations

$$\nabla l_n(\theta) - \lambda^T \nabla g(\theta) = 0 \tag{11a}$$

$$g(\theta) = 0 \tag{11b}$$

for the unknowns θ and λ. The dimension of the model is p and the dimension of the submodel is d, so the dimension of θ is p and the dimension of λ is $p - d$, and (11a) and (11b) constitute $2p - d$ nonlinear equations in $2p - d$ unknowns. If they have a solution, denote the solution $(\theta_n^*, \lambda_n^*)$.

If θ_n^* is a local maximizer of the Lagrangian, that is, if there is a neighborhood W of θ_n^* such that

$$l_n(\theta) - \lambda_n^* g(\theta) \le l_n(\theta_n^*) - \lambda_n^* g(\theta_n^*), \qquad \theta \in W,$$

then clearly θ_n^* is a local maximizer for the constrained problem, that is,

$$l_n(\theta) \le l_n(\theta_n^*), \qquad \theta \in W \text{ and } g(\theta) = 0$$

(and the same holds with "local" replaced by "global" in both places if $W = \mathbb{R}^p$).

Although either works, the method of Section 3.1 is more straightforward than the method of this section. Hence we will hereafter concentrate on that. Since none of the test statistics we deal with depend on the method by which θ_n^* is found (only on its value), all of our results apply to either method (assuming the method of Lagrange multipliers does manage to find the correct root of the equations (11a) and (11b)).

Our main reason for introducing the method of Lagrange multipliers (other than just to briefly cover them) is to make the name "Lagrange multiplier test" for the Rao test make sense. Obviously, from (11a) we have

$$\nabla l_n(\theta_n^*) = (\lambda_n^*)^T \nabla g(\theta_n^*)$$

so the left side, which appears in the Rao test statistic (2) can be computed using the Lagrange multipliers. The name "score test" comes from the name "score function" given to $\theta \mapsto \nabla l_n(\theta)$ by R. A. Fisher.

4 The Wilks Test

Theorem 1. *Under the assumptions of Theorem 3.3 of Geyer (2013) and the setup of the preceding section, assume the function h defining the submodel parameterization is twice continuously differentiable and*

$$H = \nabla h(\varphi_0)$$

is a full rank matrix, and define

$$q^*(\varepsilon) = q(H\varepsilon), \qquad \varepsilon \in \mathbb{R}^d,$$

where, from equation (16) in Geyer (2013),

$$q(\delta) = \delta^T Z + \tfrac{1}{2}\delta^T K \delta, \qquad \delta \in \mathbb{R}^p$$

satisfies the LAMN conditions (Z is a random vector, K is a random matrix, K is almost surely positive definite, and the conditional distribution of Z given K is Normal$(0, K)$*). Define*

$$\hat{\delta}_n = \tau_n(\hat{\theta}_n - \theta_0) \tag{12a}$$

$$\varepsilon_n^* = \tau_n(\varphi_n^* - \varphi_0) \tag{12b}$$

where $\hat{\theta}_n$ is the MLE for the model and φ_n^* is the MLE for the submodel (both produced by one-step-Newton or infinite-step-Newton updates of τ_n-consistent estimators). Assume $\tau_n \to \infty$ unless the function h is affine ($h(\varphi) = a + H\varphi$, where a is a constant vector), in which case τ_n can be any sequence. Then

$$\begin{pmatrix} q_n \\ q_n^* \\ \hat{\delta}_n \\ \varepsilon_n^* \end{pmatrix} \xrightarrow{w} \begin{pmatrix} q \\ q^* \\ K^{-1}Z \\ (H^T K H)^{-1} H^T Z \end{pmatrix}$$

Proof. The asymptotics of q_n and $\hat{\delta}_n$ come directly from Theorem 3.3 in Geyer (2013). By the Skorohod theorem there exist random elements $\tilde{q}_n$ and $\tilde{q}$ of $C^2(\mathbb{R}^p)$ such that $\tilde{q}_n$ and q_n have the same laws and $\tilde{q}$ and q have the same laws and $\tilde{q}_n \xrightarrow{\text{a.s.}} \tilde{q}$ in $C^2(\mathbb{R}^p)$. Clearly $\tilde{q}$ can have the form

$$\tilde{q}(\delta) = \delta^T \widetilde{Z} + \tfrac{1}{2}\delta^T \widetilde{K} \delta, \qquad \delta \in \mathbb{R}^p,$$

where the pair $(\widetilde{Z}, \widetilde{K})$ has the same law as the pair (Z, K).

The definition (6) can be rewritten

$$l_n(\theta) = l_n(\theta_0) + q_n(\tau_n(\theta - \theta_0)) \tag{13}$$

from which we get

$$\begin{aligned} q_n^*(\varepsilon) &= l_n(h(\varphi_0 + \tau_n^{-1}\varepsilon)) - l_n(\theta_0) \\ &= q_n(\tau_n(h(\varphi_0 + \tau_n^{-1}\varepsilon) - \theta_0)) \end{aligned}$$

Hence if we define $\tilde{q}_n^*$ by

$$\tilde{q}_n^*(\varepsilon) = \tilde{q}_n(\tau_n(h(\varphi_0 + \tau_n^{-1}\varepsilon) - \theta_0))$$

then $\tilde{q}_n^*$ will have the same law as q_n^*, and

$$\nabla\tilde{q}_n^*(\varepsilon) = \big(\nabla\tilde{q}_n(\tau_n(h(\varphi_0 + \tau_n^{-1}\varepsilon) - \theta_0))\big)^T \nabla h(\varphi_0 + \tau_n^{-1}\varepsilon)$$

and

$$\begin{aligned} &\nabla^2\tilde{q}_n^*(\varepsilon) = \\ &\quad \big[\nabla h(\varphi_0 + \tau_n^{-1}\varepsilon)\big]^T \big[\nabla^2\tilde{q}_n(\tau_n(h(\varphi_0 + \tau_n^{-1}\varepsilon) - \theta_0))\big]\big[\nabla h(\varphi_0 + \tau_n^{-1}\varepsilon)\big] \\ &\qquad + \tau_n^{-1}\nabla\tilde{q}_n(\tau_n(h(\varphi_0 + \tau_n^{-1}\varepsilon) - \theta_0))\nabla^2 h(\varphi_0 + \tau_n^{-1}\varepsilon). \end{aligned}$$

Suppose $\varepsilon_n \to \varepsilon$. If we assume $\tau_n \to \infty$, then

$$\tau_n(h(\varphi_0 + \tau_n^{-1}\varepsilon_n) - \theta_0) \to H\varepsilon. \tag{14}$$

If instead of assuming $\tau_n \to \infty$ we assume h is an affine function, then we have equality for all n in (14) and the limit holds trivially. And

$$\begin{aligned} \tilde{q}_n^*(\varepsilon_n) &\xrightarrow{\text{a.s.}} \tilde{q}(H\varepsilon) \\ &= \tilde{q}^*(\varepsilon) \\ \nabla\tilde{q}_n^*(\varepsilon_n) &\xrightarrow{\text{a.s.}} \big(\nabla\tilde{q}(H\varepsilon)\big)^T H \\ &= \nabla\tilde{q}^*(\varepsilon) \\ \nabla^2\tilde{q}_n^*(\varepsilon_n) &\xrightarrow{\text{a.s.}} H^T\nabla^2\tilde{q}(H\varepsilon)H \\ &= \nabla^2\tilde{q}^*(\varepsilon) \end{aligned}$$

which (since they hold for any sequence $\varepsilon_n \to \varepsilon$) imply $\tilde{q}_n^* \xrightarrow{\text{a.s.}} \tilde{q}^*$ in $C^2(\mathbb{R}^d)$. Since almost sure convergence implies weak convergence, we have the joint

convergence of q_n, $\hat{\delta}_n$, and q_n^* asserted by this theorem. The joint convergence of all four quantities asserted by this theorem now follows by applying Theorem 3.3 of Geyer (2013) to the submodel. We verify that the submodel satisfies the LAMN conditions by observing that

$$q^*(\varepsilon) = \varepsilon^T H^T Z - \frac{1}{2} \varepsilon^T H^T K H \varepsilon$$

is quadratic in ε, and $H^T K H$ is an almost surely positive definite matrix, because K is almost surely positive definite and H has full rank, and the distribution of $H^T K H$ does not depend on the parameter ε because the distribution of K does not depend on the parameter ε. □

Corollary 2. *Under the conditions of the theorem,*

$$2\big[q_n(\hat{\delta}_n) - q_n^*(\varepsilon_n^*)\big] \xrightarrow{w} Z^T \big[K^{-1} - H(H^T K H)^{-1} H^T\big] Z. \tag{15}$$

Proof. The continuous mapping theorem applied to the assertion of Theorem 1 gives

$$2\big[q_n(\hat{\delta}_n) - q_n^*(\varepsilon_n^*)\big] \xrightarrow{w} 2\big[q(K^{-1}Z) - q^*\big((H^T K H)^{-1} H^T Z\big)\big],$$

evaluation and subtraction being continuous operations.[2] Plugging in the formulas for q and q^* given by the theorem finishes the proof. □

Corollary 3. *Under the conditions of the theorem,*

$$2\big[l_n(\hat{\theta}_n) - l_n(\theta_n^*)\big] \xrightarrow{w} \text{ChiSq}(p - d).$$

Before proving this corollary and in aid of proving it we review some facts about orthogonal projections that are familiar from the theory of linear models.

Lemma 4. *If M is a matrix whose column dimension is equal to its rank, then $M^T M$ is invertible and $M(M^T M)^{-1} M^T$ is the orthogonal projection onto the subspace spanned by the columns of M, and the dimension of this subspace is the column dimension of M.*

Proof. We shift from thinking of M and M^T as matrices to thinking of them as the linear transformations represented by those matrices. Halmos (1974, Section 49) gives the relationship between the range and null space of

[2] That evaluation is continuous means $f_n \to f$ in $C(\mathbb{R}^p)$ and $x_n \to x$ in $\mathbb{R}^p$ imply $f_n(x_n) \to f(x)$ in $\mathbb{R}$, which is the defining property of $C(\mathbb{R}^p)$.

arbitrary linear operators and their adjoint operators on finite-dimensional vector spaces

$$\mathcal{R}(M^T) = \mathcal{N}(M)^\perp \tag{16a}$$
$$\mathcal{N}(M^T) = \mathcal{R}(M)^\perp \tag{16b}$$

where $V^\perp$ denotes the vector subspace comprising vectors orthogonal to vectors in V. The assumption that the matrix M has rank equal to its column dimension means that the corresponding linear transformation is one-to-one and its null space is the zero-dimensional subspace $\{0\}$.

Hence by (16a) M^T maps onto $\mathbb{R}^d$ but is not one-to-one. By (16b) if we denote the range space of M by V, then $V^\perp$ is the null space of M^T, and this means the restriction of M^T to V is one-to-one. In summary, $M : \mathbb{R}^d \to V$ is a vector space isomorphism as is $M^T|V : V \mapsto \mathbb{R}^d$. Thus $M^T M : \mathbb{R}^d \to \mathbb{R}^d$ is a vector space isomorphism, and the restriction of $M(M^T M)^{-1} M^T$ to V is a vector space isomorphism $V \to V$. Since the dimension of V is d, the corresponding matrix has rank d.

We shift back to thinking of M and M^T as matrices. The proof that $M(M^T M)^{-1} M^T$ is symmetric and idempotent is straightforward calculation using the rules for transposes and inverses of products. □

Lemma 5. *If P is an orthogonal projection, then $I - P$ is also an orthogonal projection. If P projects onto V, then $I - P$ projects onto $V^\perp$, and the ranks of P and $I - P$ sum to the dimension of P and $I - P$.*

Proof. If P is symmetric, then so is $I - P$. If P is idempotent ($P = P^2$), then $(I - P)^2 = I - 2P + P^2 = I - P$.

If x an element of the vector space on which P projects and y is an element of the vector space on which $I - P$ projects, then $x = Pu$ and $y = (I - P)v$ for some vectors u and v and

$$x^T y = u^T P(I - P)v = u^T (P - P^2)v = 0$$

so P and $I - P$ project on orthogonal subspaces.

Conversely, if x is any vector, then x can be decomposed

$$x = Px + (I - P)x$$

as the sum of vectors in the subspaces on which P and $I - P$ project. Thus the sum of the dimensions of these subspaces is the dimension of the whole space. □

Proof of Corollary 3. By definition $q_n(\hat{\delta}_n) = l_n(\hat{\theta}_n)$ and $q_n^*(\varepsilon_n^*) = l_n(\theta_n^*)$ so we only need to confirm that the right side of (15), call it X^2, has the asserted chi-square distribution. In aid of that write $Z = K^{1/2}Y$, where Y is a standard normal random vector (so the conditional distribution of Z given K is indeed Normal$(0, K)$, as the theorem asserts). Then

$$X^2 = Y^T K^{1/2} \bigl[K^{-1} - H(H^T K H)^{-1} H^T \bigr] K^{1/2} Y$$

Hence by Exercise 2 of Section 9 in Ferguson (1996) the conditional distribution of X^2 given K is chi-square if

$$P = K^{1/2} \bigl[K^{-1} - H(H^T K H)^{-1} H^T \bigr] K^{1/2}$$

is an orthogonal projection matrix and the degrees of freedom of this chi-square distribution is the rank of P. If the rank of P does not depend on K, then the conditional distribution of X^2 given K does not depend on K, hence the conditional distribution is the same as the marginal distribution. Thus we need only show that P is symmetric and idempotent and $\operatorname{tr}(P) = p - d$.

Note that

$$\begin{aligned} P &= I - K^{1/2} H (H^T K H)^{-1} H^T K^{1/2} \\ &= I - M(M^T M)^{-1} M^T \end{aligned} \tag{17}$$

where $M = K^{1/2} H$. Thus the result follows from Lemmas 4 and 5 if we show that the ranks of M and H are the same, and this follows from the fact that K is full rank, hence $K^{1/2}$ is full rank, hence the columns of M are linearly independent if and only if the columns of H are linearly independent. □

5 The Rao Test

Theorem 6. *Under the assumptions of Theorem 1 the test statistics* (1) *and* (2) *are asymptotically equivalent.*

Proof. What is to be shown is (3). From (13) we have

$$\begin{aligned} l_n(\theta) &= l_n(\theta_0) + q_n(\tau_n(\theta - \theta_0)) \\ \nabla l_n(\theta) &= \tau_n \nabla q_n(\tau_n(\theta - \theta_0)) \\ \nabla^2 l_n(\theta) &= \tau_n^2 \nabla^2 q_n(\tau_n(\theta - \theta_0)) \end{aligned}$$

From (12b) we have

$$\theta_n^* = h(\varphi_n^*) = h(\varphi_0 + \tau_n^{-1} \varepsilon_n^*).$$

Hence

$$\begin{aligned} R_n = {} & \big[\nabla q_n(\tau_n(h(\varphi_0 + \tau_n^{-1}\varepsilon_n^*) - \theta_0))\big]^T \\ & \big[-\nabla^2 q_n(\tau_n(h(\varphi_0 + \tau_n^{-1}\varepsilon_n^*) - \theta_0))\big]^{-1} \\ & \big[\nabla q_n(\tau_n(h(\varphi_0 + \tau_n^{-1}\varepsilon_n^*) - \theta_0))\big] \end{aligned}$$

and by Theorem 1 we have

$$R_n \xrightarrow{w} \big[\nabla q(H\varepsilon^*)\big]^T \big[-\nabla^2 q(H\varepsilon^*)\big]^{-1} \big[\nabla q(H\varepsilon^*)\big] \tag{18}$$

where

$$\varepsilon^* = (H^T K H)^{-1} H^T Z.$$

The right side of (18) is

$$Z^T [I - H(H^T K H)^{-1} H^T K] K^{-1} [I - K H (H^T K H)^{-1} H^T] Z$$

which simplifies to the same as the right side of (15). So this shows that T_n and R_n have the same asymptotic distribution, but, because T_n and R_n are defined on the same probability space and converge jointly to the same functions of the same random objects Z and K, this also shows that $T_n - R_n$ converges weakly to zero, which is what was to be shown. □

6 The Wald Test

In our discussion of the Wald test we assume there is a parameterization (7) of the submodel, even though the Wald test makes no use of it. This is for comparison with the Wilks and Rao tests.

Theorem 7. *Under the assumptions of Theorem 1 and under the additional assumptions that g is continuously differentiable and*

$$G = \nabla g(\theta_0)$$

is a full rank matrix, the test statistics (1) *and* (4) *are asymptotically equivalent.*

Proof. By assumption $g(\theta_0) = 0$, and this implies

$$\tau_n g(\theta_0 + \tau_n^{-1}\delta_n) \to G\delta, \qquad \text{whenever } \delta_n \to \delta,$$

if $\tau_n \to \infty$ or if g is an affine function (in which case the submodel is flat). Hence by the continuous mapping theorem

$$\tau_n g(\hat{\theta}_n) \xrightarrow{w} G K^{-1} Z, \tag{19a}$$

where $\hat{\theta}_n$, K, and Z are as defined in Theorem 1. By assumption g is continuously differentiable, and this implies (again the continuous mapping theorem)

$$\nabla g(\hat{\theta}_n) \xrightarrow{w} G. \tag{19b}$$

By Theorem 3.3 in Geyer (2013)

$$-\nabla^2 q_n(\hat{\delta}_n) = -\tau_n^{-2} \nabla^2 l_n(\hat{\theta}_n) \xrightarrow{w} K. \tag{19c}$$

Moreover (19a), (19b), and (19c) hold simultaneously. Hence by another application of the continuous mapping theorem (this is essentially the delta method applied to our setup)

$$W_n \xrightarrow{w} Z^T K^{-1} G^T \left(G K^{-1} G^T \right)^{-1} G K^{-1} Z. \tag{20}$$

By assumption $g(h(\varphi)) = 0$ for all φ, hence the chain rule says $GH = 0$, which means the rows of G are orthogonal to the columns of H. Since both G and H are assumed to be full rank, if the columns of H span a vector subspace of dimension d (the dimension of the submodel and the column dimension of H), then the rows of G span the subspace $V^\perp$, which has dimension $p - d$.

As in the proof of Corollary 3, write $Z = K^{1/2} Y$, where Y is a standard normal random vector, so the right side of (20) becomes $Y^T Q Y$, where

$$Q = K^{-1/2} G^T \left(G K^{-1} G^T \right)^{-1} G K^{-1/2} = N (N^T N)^{-1} N^T,$$

where

$$N = K^{-1/2} G^T.$$

By Lemma 4, the matrix Q is the orthogonal projection on the subspace U spanned by the columns of N. Recall from the proof of Corollary 3 that

$$T_n \xrightarrow{w} Y^T P Y,$$

P is given by (17) and $M = K^{1/2} H$. Since $N^T M = GH = 0$, every vector in the column space of N is orthogonal to every vector in the column space of H, thus Q and $I - P$ project onto complementary subspaces, and, since both G and H are full rank, $I - P$ projects onto $U^\perp$. Thus P is the orthogonal projection onto U, and $Q = I - P$.

This shows that T_n and W_n have the same asymptotic distribution, but, because T_n and W_n are defined on the same probability space and converge jointly to the same functions of the same random objects Z and K, this also shows that $T_n - W_n$ converges weakly to zero, which is what was to be shown. □

7 The No-N View

When we are discussing hypothesis tests comparing a model and smooth submodel, the "no-n" view of asymptotics propounded by Geyer (2013) becomes a little more complicated. We need not only the "quadraticity" assumptions about the log likelihood discussed that article, but we also need the submodel parameter space to be "nearly flat" in the neighborhood of the true parameter value that contains most of the asymptotic distribution of θ_n^*. This is because we are essentially approximating the submodel parameter space by the *flat* affine subspace

$$\{\, H\varepsilon : \varepsilon \in \mathbb{R}^d \,\},$$

and if this approximation is bad, then the asymptotic distribution of $\hat{\theta}_n^*$ and the test statistics containing it will also be bad.

This seems to leave out the Wald statistic, but it has a slightly different but related problem. In the asymptotics of the Wald statistic, we are essentially using the delta method, which replaces the nonlinear function g by its best affine approximation

$$\theta \mapsto g(\theta_0) + G(\theta - \theta_0)$$

(best near θ_0). And if this approximation is bad, then the asymptotic distribution of the Wald test statistic will also be bad. Hence for all three test statistics we need not only the stochastic regularity condition that the log likelihood is nearly quadratic in the LAMN sense but also the non-stochastic, geometric regularity condition that the submodel parameter space is nearly flat.

8 Expected Fisher Information

We now leave "no-n" territory and move to the classical regularity conditions for independent and identically distributed data described in Geyer (2013, Appendix C). There we have, by differentiability under the integral sign

$$K = E_{\theta_0}\{J_1(\theta_0)\}$$

and

$$I_1(\theta) = E_\theta\{J_1(\theta)\}$$

is a continuous function of θ (this is the statement following equation (37) in Geyer (2013)). Hence by the continuous mapping theorem

$$I_1(\hat{\theta}_n) \xrightarrow{P} I_1(\theta_0) = K.$$

As always we have $I_n(\theta) = nI_1(\theta)$ because the expectation of a sum is the sum of the expectations. Thus we conclude that we can replace $J_n(\,\cdot\,)$ by $I_n(\,\cdot\,)$ in either of the test statistics (2) or (4) without changing the conclusions of any of the theorems or corollaries in this handout.

9 Power

9.1 Uniform Integrability

Recall the following definition from the handout on weak convergence in Polish spaces. A sequence X_n of random variables is *uniformly integrable* if

$$\lim_{c \to \infty} \sup_{n \in \mathbb{N}} E\{I_{(c,\infty)}(|X_n|)|X_n|\} = 0.$$

The following two theorems about uniform integrability are used here. They are Theorems 3.5 and 3.6 in Billingsley (1999).

Theorem 8. *If the sequence X_n is uniformly integrable and $X_n \xrightarrow{w} X$, then $E(X_n) \to E(X)$.*

Theorem 9. *If the random variables X_n are nonnegative, $X_n \xrightarrow{w} X$, and $E(X_n) \to E(X)$, then the sequence X_n is uniformly integrable.*

9.2 Le Cam's Third Lemma

The following is a special case of what is sometimes called "Le Cam's third lemma" (see Chapter 3 of Le Cam and Yang (2000), where their Proposition 1 is the general form).

Theorem 10. *Under the assumptions of Theorem 3.3 of Geyer (2013) and Theorem 1 above, if X_n is any sequence of random elements of a Polish space with X_n defined on the same probability space as q_n, and*

$$(X_n, q_n) \xrightarrow{w} (X, q), \tag{21}$$

where the law of (X_n, q_n) has parameter value $\delta = 0$ and the law of (X, q) also has parameter value $\delta = 0$, then (21) also holds when the law of (X_n, q_n) has parameter value δ_n and the law of (X, q) has parameter value δ_∞, where $\delta_n \to \delta_\infty$.

Proof. First some terminology. We will say "under the null" to refer to the laws for parameter value zero and "under the alternatives" to refer to the

laws for parameter value δ_n or δ_∞. Note that $e^{q_n(\delta_n)}$ is a likelihood ratio and hence the density of the law for the "alternative" parameter value δ_n with respect to the law for the "null" parameter value zero. And similarly for $e^{q(\delta_\infty)}$ in the asymptotic LAMN model. Thus we can refer distributions under alternatives to the null distribution, that is, for any bounded measurable function f, we have

$$E_{\delta_n}\{f(X_n, q_n)\} = E_0\big\{f(X_n, q_n)e^{q_n(\delta_n)}\big\} \tag{22a}$$

and similarly

$$E_{\delta_\infty}\{f(X, q)\} = E_0\big\{f(X, q)e^{q(\delta_\infty)}\big\}. \tag{22b}$$

Because they are probability densities, likelihood ratios integrate to one, that is, $E_0(e^{q_n(\delta_n)}) = 1$ and $E_0(e^{q(\delta_\infty)}) = 1$. Thus we have $E_0(e^{q_n(\delta_n)}) \to E_0(e^{q(\delta_\infty)})$. We also have $e^{q_n(\delta_n)} \xrightarrow{w} e^{q(\delta_\infty)}$ under the null by Theorem 3.3 in Geyer (2013) and the continuous mapping theorem. Thus by Theorem 9 the sequence $e^{q_n(\delta_n)}$ is uniformly integrable under the null. From the definition of uniform integrability, this obviously implies that the sequence $f(X_n, q_n)e^{q_n(\delta_n)}$ is also uniformly integrable under the null for any bounded f. Thus by Theorem 8 we have

$$E_0\big\{f(X_n, q_n)e^{q_n(\delta_n)}\big\} \to E_0\big\{f(X, q)e^{q(\delta_\infty)}\big\}$$

and by (22a) and (22b) this implies

$$E_{\delta_n}\{f(X_n, q_n)\} \to E_{\delta_\infty}\{f(X, q)\},$$

and the latter holding for all bounded continuous f says that (21) holds under the alternatives, which is what was to be proved. □

Corollary 11. *All the conclusions of Theorem 3.3 in Geyer (2013) and all theorems and corollaries in this handout hold under alternatives $\delta_n \to \delta_\infty$ as well as under the null, except for Corollary 3 which calculates under the null. In particular, the asymptotic equivalences asserted by Theorems 6 and 7 hold under the alternatives as well as under the null, and the limiting distribution of the Wilks, Rao, and Wald test statistics given by Corollary 2 holds under the alternatives as well as under the null. On the right side of* (15) *the distribution of the random matrix K does not depend on the parameter δ_∞ by the LAMN assumptions, but the the conditional distribution of the random vector Z given K is Normal$(K\delta_\infty, K)$ and hence does depend on the parameter δ_∞. And similarly where Z and K appear in conclusions of other theorems and corollaries in Geyer (2013) or this handout.*

Corollary 12. *Under alternatives $\delta_n \to \delta_\infty$ the asymptotic distribution of the Wilks, Rao, and Wald statistics is noncentral chi-square with $p - d$ degrees of freedom and noncentrality parameter*

$$\lambda = \delta_\infty^T K^{1/2} P K^{1/2} \delta_\infty, \tag{23}$$

where P is given by (17)*. If K is random (LAMN conditions), then this asymptotic noncentral chi-square distribution is conditional on K, since the noncentrality parameter depends on K.*

Proof. As in the proof of Corollary 3 the asymptotic distribution of these test statistics is the distribution of $Y^T P Y$, where $Y = K^{-1/2} Z$ and P is the orthogonal projection matrix with rank $p - d$ given by (17). The only difference is that now Z has mean $K\delta_\infty$ rather than mean zero, and hence Y has mean $\mu_\infty = K^{1/2}\delta_\infty$.

Since P is idempotent,

$$Y^T P Y = Y^T P^2 Y = X^T X,$$

where $X = PY$ is normally distributed with mean $PK^{1/2}\delta_\infty$ and variance matrix $P^2 = P$. Now apply the lemma in Chapter 10 in Ferguson (1996), with our P being the Σ in Ferguson's lemma. The lemma says $X^T X$ has a noncentral chi-square distribution with rank(P) degrees of freedom and noncentrality parameter $E(X)^T E(X)$, which agrees with the right side of (23). □

References

Billingsley, P. (1999). *Convergence of Probability Measures*, second edition. New York: Wiley.

Ferguson, T. S. (1996). *A Course in Large Sample Theory*. London: Chapman & Hall.

Geyer, C. J. (2013). Asymptotics of maximum likelihood without the LLN or CLT or sample size going to infinity. In *Advances in Modern Statistical Theory and Applications: A Festschrift in honor of Morris L. Eaton*, G. L. Jones and X. Shen eds. IMS Collections, Vol. 10, pp. 1–24. Institute of Mathematical Statistics: Hayward, CA.

Halmos, P. (1974). *Finite Dimensional Vector Spaces*. New York: Springer-Verlag. Reprint of second edition (1958), originally published by Van Nostrand

Le Cam, L. and Yang, G. L. (2000). *Asymptotics in Statistics: Some Basic Concepts*, second ed. New York: Springer-Verlag.