May 6, 2003

3.9 A likelihood ratio test for nested composite hypotheses: Wilks's theorem. Let Θ be a d-dimensional parameter space, specifically, an open set in $\mathbb{R}^d$. Let H_0 be a k-dimensional subset of Θ, in a sense to be made more precise below, for some $k < d$. For example, H_0 could be the intersection with Θ of a k-dimensional flat hyperplane. Let $\{P_\theta,\ \theta \in \Theta\}$ be an equivalent family of laws on a sample space $(X, \mathcal{B})$ with a likelihood function $f(\theta, x) > 0$ for all $\theta \in \Theta$ and $x \in X$.

Assume that observations $X_1, \ldots, X_n$ are i.i.d. P_θ for some $\theta \in \Theta$. We want to test the hypothesis that $\theta \in H_0$. S. S. Wilks proposed the following test: let $L(\theta, x) := \log f(\theta, x)$ be the log likelihood. For n observations, let the maximum log likelihoods over Θ and H_0 be respectively

$$MLL_d := \sup_{\theta\in\Theta} \sum_{j=1}^{n} L(\theta, X_j), \qquad MLL_k := \sup_{\theta\in H_0} \sum_{j=1}^{n} L(\theta, X_j).$$

Let $W := 2(MLL_d - MLL_k)$. Wilks found that if the hypothesis H_0 is true, then the distribution of W converges as $n \to \infty$ to a χ^2 distribution with $d - k$ degrees of freedom, not depending on the true $\theta = \theta_0 \in H_0$. Thus, H_0 would be rejected if W is too large in terms of the tabulated χ^2_{d-k} distribution.

It turns out that Wilks's conclusion can be proved under the same assumptions as are used to prove the lower bounds on asymptotic efficiency of estimators in Section 3.7 and efficiency of maximum likelihood estimators in Section 3.8. It will be said that H_0 is a k-dimensional C^2 imbedded submanifold of Θ for some $k < d$ if for each $\theta \in H_0$, after a translation of coordinates taking θ to 0 and a suitable rotation of coordinates, H_0 has a tangent hyperplane K_0 at 0 given by $\theta_{k+1} = \cdots = \theta_d = 0$, meaning that the intersection of H_0 with a neighborhood $\mathcal{V}$ of 0 is given by $\theta_j = f_j(\{\theta_i\}_{i=1}^k)$ for $j = k+1, \ldots, d$, where f_j are C^2 functions defined on an open neighborhood $\mathcal{W}$ of 0 in $\mathbb{R}^k$ with $f_j(0) = 0$ and $\nabla f_j(0) = 0 \in \mathbb{R}^k$ for each $j = k+1, \ldots, d$.

We have the following:

3.9.1 Theorem (Wilks's theorem). Assume (AC-1) through (AC-5) in Section 3.7 and (EML-1) through (EML-5) in Section 3.8 for Θ where in (EML-3), T_n are maximum likelihood estimators of $\theta \in \Theta$. Let H_0 be a k-dimensional C^2 imbedded submanifold of Θ containing θ_0 for some $k < d$. Let H_0 be parametrized in a neighborhood of θ_0 by $\eta := \{\eta_i\}_{i=1}^k$ in the open set $\mathcal{W} \subset \mathbb{R}^k$ in the given definition of imbedded submanifold. Let U_n be maximum likelihood estimators of η in $\mathcal{W}$, assumed to exist and be unique with probability converging to 1 as $n \to \infty$. Assume also that $U_n \to \eta_0 = 0$ in probability as $n \to \infty$.

Then as $n \to \infty$, the distribution of W converges to a χ^2_{d-k} distribution.

Proof. By the way, (EML-1) implies (AC-1), and (AC-2) and (AC-3) imply (EML-2).

(AC-6) follows from (EML-1) through (EML-5) by way of Theorem 3.8.1. As in the definition of submanifold, we can assume by translation and rotation of coordinates that $\theta_0 = 0$ and H_0 has the tangent hyperplane K_0 at 0. Then

$$\theta \mapsto \Big(\theta_1, \ldots, \theta_k, \big\{\theta_j - f_j\left(\{\theta_i\}_{i=1}^k\right)\big\}_{j=1}^d\Big)$$

is a C^2 map of the open neighborhood $\mathcal{V}$ of 0 in $\mathbb{R}^d$ onto another such neighborhood $\mathcal{U}$, with a C^2 inverse given by

$$\phi \mapsto \left(\phi_1, \ldots, \phi_k, \left\{\phi_j + f_j\left(\{\phi_i\}_{i=1}^k\right)\right\}_{j=1}^d\right).$$

Thus the map is what is called a C^2 diffeomorphism. It takes $H_0 \cap \mathcal{V}$ onto $K_0 \cap \mathcal{U}$. Thus we can assume that H_0 is the flat hyperplane K_0, replacing $\mathcal{V}$ by $\mathcal{U}$.

Now, make another rotation of coordinates so that the Fisher information matrix $I(\theta_0) = I(0)$ is diagonalized. Let its diagonal entries be $a_1, \ldots, a_d$, all > 0. Then by a linear change of parameters, replacing θ_j by $\sqrt{a_j}\theta_j$, $I(0)$ becomes the identity matrix. Since K_0 is still a k-dimensional linear subspace after these transformations, by another rotation we can assume $H_0 = K_0$ is (as before) the hyperplane $\theta_{k+1} = \cdots = \theta_d = 0$. All the transformations made, and their inverses, have been C^2 with bounded first and second partial derivatives for their coordinates on a neighborhood of 0. Thus, by the chain rule, all the assumptions still hold in the new coordinates.

Now, it's easily verified that since $\theta_0 \in H_0$, all the assumptions (AC-1) through (AC-5) and (EML-1) through (EML-5) imply their counterparts for $K_0 \cap \mathcal{U}$ in place of Θ except that for (EML-3), consistency of the MLEs $U_n \in K_0 \cap \mathcal{U}$ has been separately assumed.

Let $V_n := \sqrt{n}T_n$ and $W_n := \sqrt{n}U_n \in K_0$. Then by Theorem 3.8.1, as $n \to \infty$ we have convergence in distribution

$$\mathcal{L}(V_n) \to N_d(0, I), \qquad \mathcal{L}(W_n) \to N_k(0, I) \tag{3.9.2}$$

on $\mathbb{R}^d$ and K_0 respectively, where $N_r(0, I)$ is the r-dimensional standard normal distribution for $r = k, d$.

A multivariate Taylor expansion of $L(\theta, x)$ around $\theta_0 = 0$ is given by

$$L(\theta, x) \ = \ L(0, x) + \nabla L(0, x) \cdot \theta + \frac{1}{2}\theta' \mathcal{H}_d(0, x)\theta + R(\theta, x) \tag{3.9.3}$$

where $\mathcal{H}_d(0, x)$ is the matrix $\partial^2 L(0, x)/\partial\theta_r\partial\theta_s|_{r,s=1}^d$ and for each x, the remainder $R(\theta, x) = o(|\theta|^2)$ as $\theta \to 0$. Let $S_n := \sum_{j=1}^n \nabla_\theta(0, X_j)$. We have $E_0 \nabla L(0, x) = 0$ by assumption (AC-3). Thus by the central limit theorem (RAP, 9.5.6) and assumption (EML-2), the distribution of $S_n/\sqrt{n}$ converges as $n \to \infty$ to $N_d(0, I)$.

By a Taylor expansion of $\nabla_\theta L(\theta, x)$ around $\theta = 0$ we get for each x and for θ close enough to 0

$$\nabla_\theta L(\theta, x) = \nabla_\theta(0, x) + \mathcal{H}_d(0, x) \cdot \theta + r(\theta, x)$$

where the remainder term $r(\theta, x) = o(|\theta|)$ as $\theta \to 0$ for x fixed. Thus

$$\sum_{j=1}^n \nabla_\theta L(\theta, X_j) = S_n + \sum_{j=1}^n \mathcal{H}_d(0, X_j) \cdot \theta + \sum_{j=1}^n r(\theta, X_j).$$

Substituting $\theta = T_n$, where the left side is 0, and dividing by $\sqrt{n}$ gives

$$\frac{S_n}{\sqrt{n}} = -\frac{1}{n}\sum_{j=1}^n \mathcal{H}_d(0, X_j) \cdot V_n + \sqrt{n}o_p(T_n), \tag{3.9.4}$$

if $\sum_{j=1}^n r(T_n, X_j) = no_P(T_n)$, as will be shown after the main proof is completed. By (AC-4), $E_0\mathcal{H}_d(0,x) = -I(0) = -I$ in $\mathbb{R}^{d^2}$. Thus by the law of large numbers (for each of the d^2 matrix entries), we have $(-1/n)\sum_{j=1}^n \mathcal{H}_d(0,X_j) = I + o_p(1)$. Since $V_n = O_p(1)$ and $\sqrt{n}o_p(T_n) = o_p(1)$ by (3.9.2), we get by (3.9.4)

$$V_n \;=\; S_n/\sqrt{n} + o_p(1). \tag{3.9.5}$$

Analogously, define $S_n^{(k)} := \sum_{j=1}^n \nabla_{\theta^{(k)}} L(0,X_j)$ where $\nabla_{\theta^{(k)}} := (\partial/\partial\theta_1, \dots, \partial/\partial\theta_k)$. Then $S_n^{(k)}$ consists of just the first k coordinates of S_n. In the same way as in (3.9.5) we then get

$$W_n \;=\; S_n^{(k)}/\sqrt{n} + o_p(1). \tag{3.9.6}$$

By the definitions of MLL_d and T_n and (3.9.3), if $\sum_{j=1}^n R(T_n, X_j) = no_p(|T_n|^2)$ as will be shown in (3.9.14) below, it follows that
(3.9.7)

$$MLL_d = \sum_{j=1}^n L(T_n, X_j) = \sum_{j=1}^n L(0,X_j) + S_n \cdot T_n + \frac{1}{2}T_n' \sum_{j=1}^n \mathcal{H}_d(0,X_j)T_n + no_p(|T_n|^2).$$

We have $no_p(|T_n|^2) = o_p(1)$ by (3.9.2). By (3.9.5), we have $S_n \cdot T_n = |V_n|^2 + o_p(1)$. In the term of (3.9.7) with $\mathcal{H}_d$, we can replace T_n by V_n twice, dividing the sum by n, and then as in the proof of (3.9.5) we see that the term is $-(1/2)|V_n|^2 + o_p(1)$. Thus (3.9.7) yields

$$MLL_d = \sum_{j=1}^n L(0,X_j) + \frac{1}{2}|V_n|^2 + o_p(1).$$

Proceeding in the same way for MLL_k we get

$$MLL_k = \sum_{j=1}^n L(0,X_j) + \frac{1}{2}|W_n|^2 + o_p(1).$$

For the Wilks statistic W, applying (3.9.5) and (3.9.6), we get

$$W = 2(MLL_d - MLL_k) = (|S_n|^2 - |S_n^{(k)}|^2)/n + o_p(1) = |Y^{(d-k)}|^2/n + o_p(1)$$

where $Y^{(d-k)}$ is the projection of S_n onto the $d-k$ coordinates $\theta_{k+1}, \dots, \theta_d$. Since $S_n/\sqrt{n}$ converges in distribution to $N_d(0,I)$, $Y^{(d-k)}/\sqrt{n}$ converges in distribution to $N_{d-k}(0,I)$. It follows that the distribution of W converges to χ^2_{d-k}.

Proof of (3.9.4) and (3.9.14). For (3.9.4) we need to show that $\sum_{j=1}^n r(T_n, X_j) = no_P(T_n)$. For each $r = 1, \dots, d$, we have

$$\frac{\partial L(\theta,x)}{\partial\theta_r} - \frac{\partial L(0,x)}{\partial\theta_r} = \int_0^1 \frac{d}{dt}\frac{\partial L(t\theta,x)}{\partial\theta_r}dt$$

$$= \sum_{s=1}^{d} \theta_s \int_0^1 \frac{\partial^2 L(t\theta, x)}{\partial\theta_r \partial\theta_s} dt. \tag{3.9.8}$$

Now, for each $r, s = 1, \ldots, d$,

$$\int_0^1 \frac{\partial^2 L(t\theta, x)}{\partial\theta_r \partial\theta_s} dt = \frac{\partial^2 L(0, x)}{\partial\theta_r \partial\theta_s} + \zeta_{rs}(\theta, x)$$

where

$$|\zeta_{rs}(\theta, x)| \leq \varepsilon_{rs}(\theta, x) \ := \ \sup_{|\phi| \leq |\theta|} \left| \frac{\partial^2 L(\phi, x)}{\partial\theta_r \partial\theta_s} - \frac{\partial^2 L(0, x)}{\partial\theta_r \partial\theta_s} \right|. \tag{3.9.9}$$

It follows by (3.9.8) that the remainder

$$|r(T_n, X_j)| \leq \sum_{r,s=1}^{d} \varepsilon_{rs}(T_n, X_j). \tag{3.9.10}$$

By (AC-5), we have $\varepsilon_{rs}(\theta, x) \leq 2M(x)$ for all x and all θ in a small enough neighborhood $\mathcal{U}_1$ of 0. We can assume that $M(x) \geq 1$ for all x. Also, by the C^2 property of f in θ (AC-2) and (AC-1), $L(\theta, x)$ is C^2 in θ, so $\varepsilon_{rs}(\theta, x) \to 0$ as $\theta \to 0$. So, for any given $\varepsilon > 0$ and for all x, there is a positive integer $k = k(x, \varepsilon)$ such that if $|\theta| < 1/k$ then

$$\varepsilon(\theta, x) \ := \ \sum_{r,s=1}^{d} \varepsilon_{rs}(\theta, x) \ < \ \varepsilon. \tag{3.9.11}$$

Since $E_0 M < \infty$, by dominated convergence there is a $\gamma > 0$ small enough so that if $P_0(A) < \gamma$, then $\int_A d^2 M dP_0 < \varepsilon/2$. For $k \geq k(\varepsilon)$ large enough, the set A of x such that (3.9.11) fails has $P_0(A) < \gamma$ and so $\int_A M dP_0 < \varepsilon/2$. Thus $P_0(A) < \varepsilon/2$ since $M \geq 1$. For $|\theta| < 1/k(\varepsilon)$ we have $\varepsilon(\theta, x) < \varepsilon$ for $x \notin A$ and $\leq 2d^2(M1_A)(x)$ otherwise. By the strong law of large numbers, we have almost surely for n large enough, by choice of A, $n^{-1}\sum_{j=1}^{n} 2d^2(M1_A)(X_j) < \varepsilon$ and then

$$\frac{1}{n}\sum_{j=1}^{n} \varepsilon(\theta, X_j) \leq n\frac{\varepsilon}{n} + \frac{2}{n}\sum_{j=1}^{n}(M1_A)(X_j) \ \leq \ 2\varepsilon. \tag{3.9.12}$$

As $n \to \infty$, since $T_n \to 0$ in probability by (3.9.2), we have

$$P(|T_n| < 1/k(\varepsilon)) \to 1. \tag{3.9.13}$$

It follows by (3.9.10) that $\sum_{j=1}^{n} r(T_n, X_j) = n o_p(T_n)$ as desired, so (3.9.4) is proved.

It remains to prove

$$\sum_{j=1}^{n} R(T_n, X_j) \ = \ n o_p(|T_n|^2). \tag{3.9.14}$$

By a form of Taylor's theorem with integral remainder we have

$$R(\theta, x) \;=\; \theta' \int_0^1 (1-u)[\mathcal{H}_d(u\theta, x) - \mathcal{H}_d(0, x)]du \cdot \theta.$$

Thus for $\varepsilon_{rs}(\theta, x)$ as defined in (3.9.10) and $\varepsilon(\theta, x)$ in (3.9.12),

$$|R(\theta, x)| \;\leq\; \sum_{r,s=1}^{d} |\theta_r||\theta_s|\varepsilon_{rs}(\theta, x) \;\leq\; |\theta|^2 \varepsilon(\theta, x).$$

Then by (3.9.12) and (3.9.13), (3.9.14) follows, completing the proof of the theorem. □

PROBLEM

1. Let $\Theta = \mathbb{R}^d$, let $P_\theta \;:=\; N(\mu, I)$ for $\mu \in \mathbb{R}^d$ and for some $k < d$ let $H_0 \;:=\; K_0 \;:=$ $\{\mu : \; \mu_{k+1} = \cdots = \mu_d = 0$. Show that in this case the Wilks statistic $W \;:=\; 2(MLL_d - MLL_k)$ has exactly a χ^2_{d-k} distribution for all n, not only asymptotically as $n \to \infty$.

NOTES

Wilks first published his theorem in a paper, Wilks (1938), then gave an exposition of it in his book, Wilks (1962, §13.8). Chernoff (1954) gave another proof. Van der Vaart (1998, Chapter 16) gives a more recent exposition. The Notes by van der Vaart (1998, p. 240) suggest that Wilks's original proof was not rigorous. The bibliography in van der Vaart's book includes Wilks's 1938 paper but not the 1962 book. The proof in the 1962 book seems rather long.

REFERENCES

Chernoff, Herman (1954). On the distribution of the likelihood ratio statistic. *Ann. Math. Statist.* **25**, 573-578.

van der Vaart, A. W. (1998). *Asymptotic Statistics.* Cambridge University Press.

Wilks, S. S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann. Math. Statist.* **9**, 60-62.

Wilks, S. S. (1962). *Mathematical Statistics.* Wiley, New York; 2d printing, corrected, 1963.