Proof of Wilks' Theorem on LRT

This handout is intended to supply full details and re-cap of carefully defined notations of the argument given in class proving the asymptotic distributional convergence of the likelihood ratio test statistic of general null hypotheses restricting the values of a subset of parameter components to a chi-square with degrees of freedom equal to the number of components restricted. Throughout, we assume that the data $\mathbf{X} = \{X_i\}_{i=1}^n$ constitute an *iid* sample (of values in some Euclidean data-space) from a density $f(x, \vartheta)$ known except for unknown parameter $\vartheta \subset \Theta \subset \mathbf{R}^k$. We assume that the density f satisfies all of the regularity conditions previously needed to ensure that maximum likelihood estimators are locally unique, consistent, and asymptotically normal. These conditions include the restriction that Θ contain an open neighborhood of the true value ϑ_0 governing the data, and that Θ lies in some sufficiently small neighborhood of ϑ_0 not depending upon n. The likelihood for the data $\mathbf{X}$ is denoted by $L(\mathbf{X}, \vartheta)$ and the unrestricted Maximum Likelihood Estimator (MLE) for ϑ is $\hat{\vartheta}$.

Now consider the null hypothesis $H_0 : \vartheta_{0,j} = 0$ for $1 \le j \le r$, where $0 < r < k$ is fixed. Define the *restricted MLE* $\hat{\vartheta}^{res}$ as the maximizer of $L(\mathbf{X}, \vartheta)$ over parameter vectors $\vartheta \subset \Theta$ such that $\vartheta_{0,j} = 0$ for $1 \le j \le r$. We require a detailed set of notations designed to partition parameters, estimators, gradients, score statistics, and information matrices into parts respectively reflecting the first r and the last $k - r$ components. Under the null hypothesis, the parameter vector ϑ_0 and restricted MLE have the form

$$\vartheta_0 \;=\; \begin{pmatrix} \mathbf{0} \\ \vartheta_* \end{pmatrix} \quad , \qquad \hat{\vartheta}^{res} \;=\; \begin{pmatrix} \mathbf{0} \\ \hat{\vartheta}_* \end{pmatrix} \quad , \qquad \vartheta_{0*},\ \hat{\vartheta}_* \in \mathbf{R}^{k-r}$$

Next, denote by ∇_A, ∇_C respectively the gradient operator with respect to the first r and last $k - r$ component of ϑ. It is clear that the MLE definitions are equivalent to

$$\nabla \log L(\mathbf{X}, \hat{\vartheta}) \;\equiv\; \begin{pmatrix} \nabla_A \\ \nabla_C \end{pmatrix} \log L(\mathbf{X}, \hat{\vartheta}) \;=\; \begin{pmatrix} \mathbf{0} \\ \mathbf{0} \end{pmatrix} \quad , \quad \nabla_C \log L(\mathbf{X}, \hat{\vartheta}^{res}) \;=\; \mathbf{0}$$

Similarly, we can partition the Fisher Information Matrix

$$I(\vartheta_0) \;=\; \frac{-1}{n}\, E\left(\nabla^{\otimes 2} \log L(\mathbf{X}, \vartheta_0)\right) \;=\; \frac{1}{n}\, E\left(\nabla \log L(\mathbf{X}, \vartheta_0)\right)^{\otimes 2} \;=\; \begin{pmatrix} A & B \\ B^t & C \end{pmatrix}$$

where A, B, C are respectively $r \times r$, $r \times (k-r)$, $(k-r) \times (k-r)$ matrices. The *score statistic* S is given the partitioned notation

$$\frac{1}{\sqrt{n}}\, S(\vartheta_0) \;\equiv\; \frac{1}{\sqrt{n}}\, \nabla \log L(\mathbf{X}, \vartheta_0) \;\equiv\; \begin{pmatrix} \xi \\ \eta \end{pmatrix}$$

Now we begin to assemble background results, already derived in class, related to Taylor series expansion of $\nabla \log L(\mathbf{X}, \vartheta)$ as a vector function of ϑ around the points ϑ_0, $\hat{\vartheta}$, $\hat{\vartheta}^{res}$, all of which lie with high probability for large n in a tiny neighborhood of ϑ_0. Throughout, we will not be explicit about the remainder terms, but rather write $\approx$ to indicate that the left and right hand sides differ by a (usually random) quantity which converges under H_0 to 0 in probability as $n \to \infty$. Recall that (under H_0)

$$\begin{pmatrix} \xi \\ \eta \end{pmatrix} \;\approx\; I(\vartheta_0)\, \sqrt{n}\, (\hat{\vartheta} - \vartheta_0) \quad , \qquad \eta \;\approx\; C\, \sqrt{n}\, (\hat{\vartheta}_* - \vartheta_{0*}) \tag{1}$$

where the second part of (1) is really the same fact as the first, applied to the parametric Maximum-Likelihood estimation problem (under H_0) for the data $\mathbf{X}$ and the $(k-r)$-dimensional parameter ϑ_{0*}.

The problem considered in Wilks' Theorem is the asymptotic distribution of the Likelihood Ratio Statistic

$$-2 \log \Lambda \;\equiv\; \log(T_1/T_2) \quad , \quad T_1 \;\equiv\; \frac{L(\mathbf{X}, \hat{\vartheta})}{L(\mathbf{X}, \vartheta_0)} \quad , \quad T_2 \;\equiv\; \frac{L(\mathbf{X}, \hat{\vartheta}^{res})}{L(\mathbf{X}, \vartheta_0)} \tag{2}$$

However, Taylor expansion around ϑ_0 showed us that

$$\begin{aligned} 2 \log(T_1) \;&\approx\; \sqrt{n}\, (\hat{\vartheta} - \vartheta_0)^t\, \frac{1}{n}\, \nabla^{\otimes 2} \log L(\mathbf{X}, \vartheta_0)\, \sqrt{n}\, (\hat{\vartheta} - \vartheta_0) \\ &\approx\; \begin{pmatrix} \xi \\ \eta \end{pmatrix}^t I^{-1}(\vartheta_0) \begin{pmatrix} \xi \\ \eta \end{pmatrix} \end{aligned} \tag{3}$$

and similarly, using the fact that under H_0 the same data-sample $\mathbf{X}$ is governed by a $k-r$ dimensional parameter with associated information matrix C, we have

$$2 \log(T_2) \;\approx\; \eta^t\, C^{-1}\, \eta \tag{4}$$

Consider next another Taylor's expansion around ϑ_0:

$$\frac{1}{\sqrt{n}}\,\nabla_A \log L(\mathbf{X},\hat{\vartheta}^{res}) \;\approx\; \frac{1}{n}\,\nabla_A \log L(\mathbf{X},\vartheta_0)$$

$$+\,\frac{1}{n}\,\nabla_A\,\nabla^t\,\log L(\mathbf{X},\vartheta_0)\,\begin{pmatrix}\mathbf{0}\\ \sqrt{n}\,(\hat{\vartheta}_*-\vartheta_{0*})\end{pmatrix}$$

Now we use the Law of Large Numbers result that $n^{-1}\,\nabla^{\otimes 2}\log L(\mathbf{X},\vartheta_0)\approx -I(\vartheta_0)$, as we have so often done before, to conclude from the last displayed equation that

$$\frac{1}{\sqrt{n}}\,\nabla_A \log L(\mathbf{X},\hat{\vartheta}^{res}) \;\approx\; \xi \,-\, B\,\sqrt{n}\,(\hat{\vartheta}_*-\vartheta_{0*}) \;\approx\; \xi \,-\, B\,C^{-1}\,\eta \qquad (5)$$

where in the last step we have substituted the second part of (1).

Finally, we expand the difference between the right-hand sides of formulas (3) and (4) to obtain via (2):

$$-2\log\Lambda \;\approx\; \begin{pmatrix}\xi\\ \eta\end{pmatrix}^t I^{-1}(\vartheta_0)\,\begin{pmatrix}\xi\\ \eta\end{pmatrix} \,-\, \eta^t\,C^{-1}\,\eta \qquad (6)$$

But the definition of the matrix inverse implies that

$$\begin{pmatrix}B^t & C\end{pmatrix}\,I^{-1}(\vartheta_0) \;=\; \begin{pmatrix}B^t & C\end{pmatrix}\,\begin{pmatrix}A & B\\ B^t & C\end{pmatrix}^{-1} \;=\; (\mathbf{0}\;\; Id)$$

and therefore

$$\begin{pmatrix}B\,C^{-1}\,\eta\\ \eta\end{pmatrix}^t I^{-1}(\vartheta_0) \;=\; (C^{-1}\,\eta)^t\,\begin{pmatrix}B^t & C\end{pmatrix}\,I^{-1}(\vartheta_0) \;=\; (C^{-1}\,\eta)^t\,(\mathbf{0}\;\; Id)$$

This implies that the cross-terms in the expression

$$\begin{pmatrix}\xi\\ \eta\end{pmatrix}^t I^{-1}(\vartheta_0)\,\begin{pmatrix}\xi\\ \eta\end{pmatrix} \;=\; \left[\begin{pmatrix}\xi - BC^{-1}\eta\\ \mathbf{0}\end{pmatrix} + \begin{pmatrix}BC^{-1}\eta\\ \eta\end{pmatrix}\right]^t \cdot$$

$$I^{-1}(\vartheta_0)\,\left[\begin{pmatrix}\xi - BC^{-1}\eta\\ \mathbf{0}\end{pmatrix} + \begin{pmatrix}BC^{-1}\eta\\ \eta\end{pmatrix}\right]$$

are $\mathbf{0}$, implying that

$$\begin{pmatrix} \xi \\ \eta \end{pmatrix}^t I^{-1}(\vartheta_0) \begin{pmatrix} \xi \\ \eta \end{pmatrix} = (\xi - BC^{-1}\eta)^t \cdot$$

$$(A - BC^{-1}B^t)^{-1}(\xi - BC^{-1}\eta) + \eta^t C^{-1}\eta$$

Here we have used the fact that the upper-left block of $I^{-1}(\vartheta_0)$ is equal to $(A - B\,C^{-1}\,B^t)^{-1}$, a linear algebra fact which can readily be proved by solving for $u \in \mathbf{R}^r$, $v \in \mathbf{R}^{k-r}$ in terms of $x \in \mathbf{R}^r$, $y \in \mathbf{R}^{k-r}$ by elimination in the equations:

$$\begin{pmatrix} A & B \\ B^t & C \end{pmatrix} \begin{pmatrix} u \\ v \end{pmatrix} = \begin{pmatrix} x \\ y \end{pmatrix}$$

Now, substituting into (6), we find

$$-2\log\Lambda \approx (\xi - BC^{-1}\eta)^t (A - BC^{-1}B^t)^{-1} (\xi - BC^{-1}\eta) \tag{7}$$

We conclude that the last expression is distributed asymptotically as χ^2_r upon recalling that $(\xi^t\ \eta^t)^t$ is asymptotically $\mathcal{N}(\mathbf{0},\, I(\vartheta_0))$, which implies that $\xi - BC^{-1}\eta$ is also multivariate normal (r-dimensional) nondegenerate with mean $\mathbf{0}$ and variance

$$E\Big((\xi - BC^{-1}\eta)\,(\xi - BC^{-1}\eta)^t\Big) \approx A - BC^{-1}B^t$$

For future reference, we remark that the development given here shows (as in equation 5) among other things that the score statistic

$$\frac{1}{\sqrt{n}}\,\nabla_A \log L(\mathbf{X}, \hat{\vartheta}^{res})$$

is asymptotically the same is the 'adjusted' score statistic $\xi - B\,C^{-1}\,\eta$, which is asymptotically distributed as $\mathcal{N}(\mathbf{0},\, A - B\,C^{-1}\,B^t)$.