**STA2212: Inference and Likelihood**

**A. Notation**

**One random variable**: Given a model for $X$ which assumes $X$ has a density $f(x;\theta), \quad \theta \in \Theta \subset \mathbb{R}^k$, we have the following definitions:

| | | |
|---|---|---|
| likelihood function | $L(\theta; x) = c(x) f(x;\theta)$ | |
| log-likelihood function | $\ell(\theta; x) = \log L(\theta; x) = \log f(x;\theta) + a(x)$ | |
| score function | $u(\theta) = \partial \ell(\theta; x)/\partial\theta$ | $s(x;\theta)$ |
| observed information function | $j(\theta) = -\partial^2 \ell(\theta; x)/\partial\theta\partial\theta^T$ | |
| expected information (in one observation) | $i(\theta) = \mathrm{E}_\theta\{U(\theta)U(\theta)^T\}$[1] | $I(\theta)$ (9.11) |

**Independent observations**: When we have $X_i$ independent, identically distributed from $f(x_i;\theta)$, then, denoting the observed sample $\boldsymbol{x} = (x_1, \ldots, x_n)$ we have:

| | | |
|---|---|---|
| likelihood function | $L(\theta; \boldsymbol{x}) = \prod_{i=1}^n f(x_i;\theta)$ | $\mathcal{L}_n(\theta)$ |
| log-likelihood function | $\ell(\theta) = \ell(\theta; \boldsymbol{x}) = \sum_{i=1}^n \ell(\theta; x_i)$ | $\ell_n(\theta)$ |
| maximum likelihood estimate | $\hat\theta = \hat\theta(\boldsymbol{x}) = \arg\sup_\theta \ell(\theta)$ | $\hat\theta_n$ |
| score function | $U(\theta) = \ell'(\theta) = \sum U_i(\theta)$ | $\Sigma_i s(X_i;\theta)$ |
| observed information function | $j(\theta) = -\ell''(\theta) = -\ell''(\theta; \boldsymbol{x})$ | $-H(\theta)$ p.133 |
| observed (Fisher) information | $j(\hat\theta)$ | lazy notation |
| expected (Fisher) information | $i(\theta) = \mathrm{E}_\theta\{U(\theta)U(\theta)^T\} = n i_1(\theta)$ | $I_n(\theta) = nI(\theta)$ (Th. 9.17) |

**Comments:**

1. the maximum likelihood estimate $\hat\theta_n$ is usually obtained by solving the *score equation* $\ell'(\theta; \boldsymbol{x}) = 0$.

2. It doesn't really matter for the definitions above if the observations are independent and identically distributed (i.i.d.), or only independent, but the theorems that are proved in Ch. 9 do assume i.i.d. for simplicity.

3. AoS does not have separate notation for the *observed* Fisher information, which is the negative second derivative at the maximum. But Theorem 9.17 shows that $\mathrm{E}_\theta\{-\ell''(\theta; X)\} = \mathrm{E}_\theta\{j(\theta)\} = I(\theta)$, in models for which we can interchange differentiation and integration in $\int f(x;\theta)dx = 1$.

4. There are important distinctions to be careful about in the notation for likelihood and its quantities:

   (a) Are we working with a single observation or $n$ observations?

   (b) Is the variable $x$, or the vector $\boldsymbol{x} = (x_1, \ldots, x_n)$, random ($X$) or fixed ($x$)?

   (c) Do we want to find the distribution of something ($X$ is random) or calculate data summaries ($x$ is fixed)?

[1] $U(\theta) = u(\theta; X)$

**B. First order asymptotic theory** AoS §9.3-9.7

1. $\theta$ **is a scalar**

If the components of $\boldsymbol{X}$ are i.i.d., then the score function $U(\theta; \boldsymbol{X})$ is a sum of i.i.d. random variables, and we can show that it has expected value 0 and variance $I_n(\theta)$ (or $i(\theta)$ in my notation). Under some regularity conditions on the density $f(x_i; \theta)$, the central limit theorem gives

$$\frac{U(\theta)}{I_n^{1/2}(\theta)} \xrightarrow{d} N(0,1). \qquad \rightsquigarrow \tag{1}$$

Almost everything else follows from this result and Slutsky's theorem. For example, we can show that

$$(\hat{\theta} - \theta) I_n^{1/2}(\theta) = U(\theta)/I_n^{1/2}(\theta) + o_p(1),$$

where $o_p(1)$ means a remainder term that goes to 0 in probability as $n \to \infty$, so we have the second result

$$(\hat{\theta} - \theta) I_n^{1/2}(\theta) \xrightarrow{d} N(0,1). \tag{2}$$

These limit theorems give us two corresponding approximations to use with $n$ fixed:

$$U(\theta) \dot{\sim} N\left(0, I_n(\theta)\right), \qquad \approx \tag{3}$$

and

$$\hat{\theta} - \theta \dot{\sim} N\left(0, 1/I_n(\theta)\right). \tag{4}$$

The notation $\dot{\sim}$ is read as "is approximately distributed as".

Having the unknown quantity $\theta$ in the variance in (3) and (4) is inconvenient, but to the same order of approximation, we can replace $I_n(\theta)$ by $I_n(\hat{\theta})$ or by $j(\hat{\theta})$. In AoS, $I_n^{-1/2}(\theta)$ is called `se` and $I_n^{-1/2}(\hat{\theta})$ is called $\widehat{\texttt{se}}$, but the use of $j(\hat{\theta}) = -\ell''(\hat{\theta}; \boldsymbol{x})$ is not mentioned. It should be, because careful study of the remainder term (the $o_p(1)$ term above) indicates that of all the choices, $j(\hat{\theta})$ gives the best approximation for fixed $n$. It is also readily available in software that finds maximum likelihood estimates using Newton's method to solve $\ell'(\hat{\theta}) = 0$; see AoS p.143. In Theorem 9.19, $1/I_n^{1/2}(\hat{\theta})$ is used in (4) to define an approximate confidence interval for the unknown parameter $\theta$.

2. $\theta$ **is a vector of length** $k$ AoS 9.10

The results above all generalize directly to a vector $\theta$ of unknown parameters. The notation on p.1 already includes this case. The score function is a $k \times 1$ vector and the observed and expected Fisher information are $k \times k$ matrices. The limit theorems corresponding to (1) and (2) are

$$I_n^{-1/2}(\theta) U_n(\theta) \xrightarrow{d} N_k(0, \mathcal{I}_k), \quad I_n^{1/2}(\theta)(\hat{\theta} - \theta) \xrightarrow{d} N_k(0, \mathcal{I}_k), \tag{5}$$

where $N_k(0, \mathcal{I}_k)$ is the multivariate standard normal distribution and $\mathcal{I}_k$ is the $k \times k$ identity matrix. Because this limit statement involves taking the square root of the matrix $I_n$, the results in (5) are rarely used in this form. That is why AoS, Theorem 9.27 simply gives the analogue of (3):

$$\hat{\theta} - \theta \dot{\sim} N_k \left(0, I_n^{-1}(\hat{\theta})\right) \tag{6}$$

(Actually, AoS doesn't distinguish in Theorem 9.7 between $I_n^{-1}(\theta)$ and $I_n^{-1}(\hat{\theta})$ but it should. In Theorem 9.28 the result correctly uses $I_n^{-1}(\hat{\theta}) \equiv J_n(\hat{\theta}) \equiv \hat{J}_n$.)

The approximation in (6) is for the whole vector $\hat{\theta}$ but that's not so useful in practice. However we can specialize the result to a single component, giving, for example,

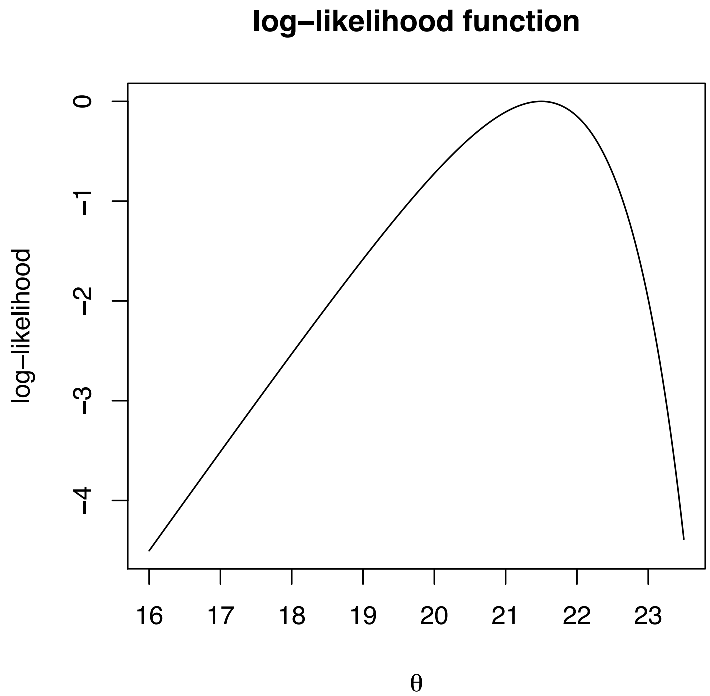
$$\hat{\theta}_j - \theta_j \dot{\sim} N\left(0, J_n(\hat{\theta})_{jj}\right), \tag{7}$$

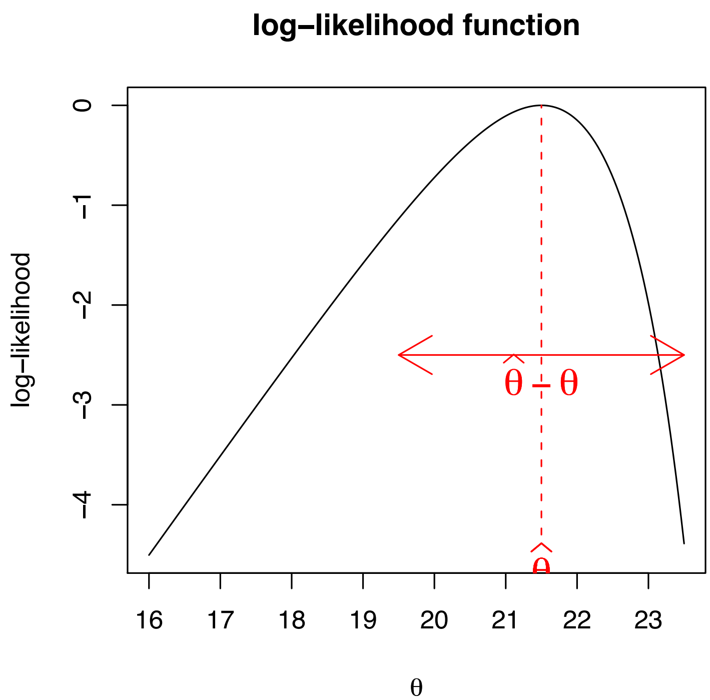
i.e. the $j$th diagonal element of the inverse matrix is the approximate variance of the $j$th component of the vector $\hat{\theta}$. We also have that $J_n(\hat{\theta})_{jk}$ is the asymptotic covariance of $\hat{\theta}_j, \hat{\theta}_k$.

Result (7) corresponds to the standard output from the `R` command `glm`. The following is a logistic regression model from the Final Homework in Applied Stats I. Each line in the table of coefficients is an application of (7). The matrix $\hat{J}_n$ is obtained with the command `vcov(Boston.glm)`.

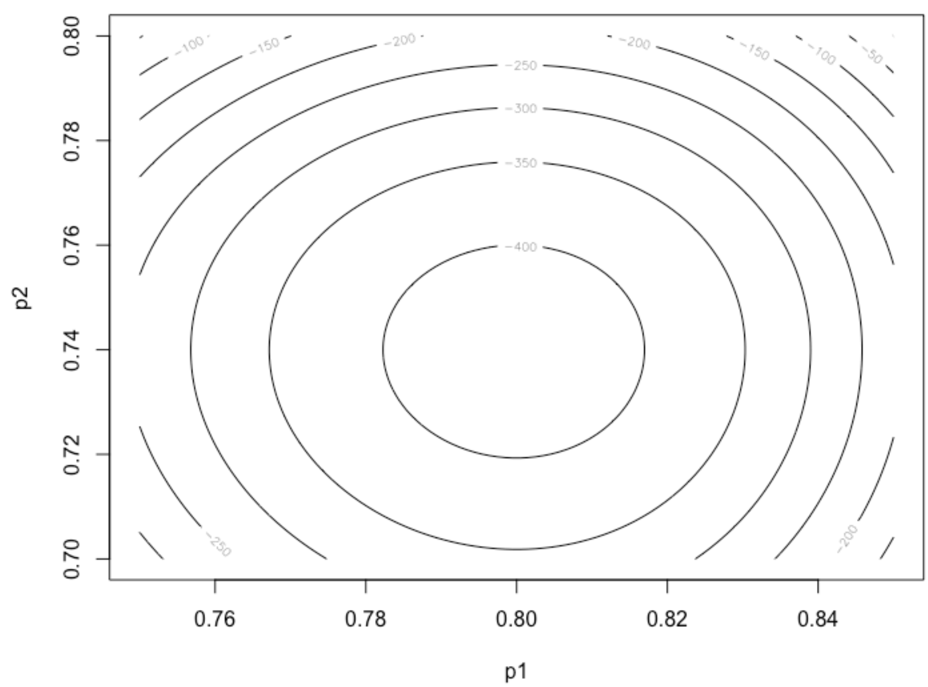
```
 library(MASS)
 data(Boston)
 Boston$crim2 <- Boston$crim > median(Boston$crim) # define binary response
 Boston.glm <- glm(crim2 ~ . - crim, family = binomial,
 data = Boston) #fit logistic regression
 summary(Boston.glm)
...
 Coefficients:
               Estimate Std. Error z value Pr(>|z|)
 (Intercept) -34.103704   6.530014  -5.223 1.76e-07 ***
 zn           -0.079918   0.033731  -2.369  0.01782 *
 indus        -0.059389   0.043722  -1.358  0.17436
 chas          0.785327   0.728930   1.077  0.28132
 nox          48.523782   7.396497   6.560 5.37e-11 ***
 rm           -0.425596   0.701104  -0.607  0.54383
 age           0.022172   0.012221   1.814  0.06963 .
 dis           0.691400   0.218308   3.167  0.00154 **
 rad           0.656465   0.152452   4.306 1.66e-05 ***
 tax          -0.006412   0.002689  -2.385  0.01709 *
 ptratio       0.368716   0.122136   3.019  0.00254 **
 black        -0.013524   0.006536  -2.069  0.03853 *
 lstat         0.043862   0.048981   0.895  0.37052
 medv          0.167130   0.066940   2.497  0.01254 *
```

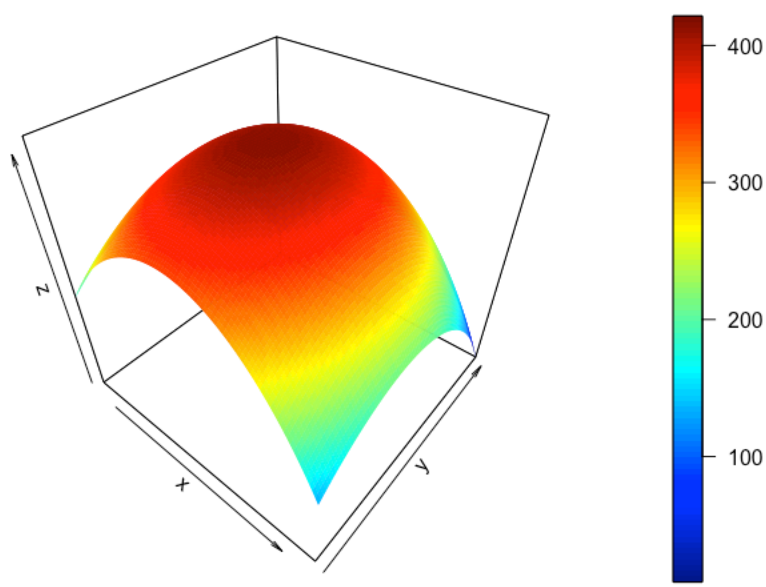
Here is the log-likelihood function for $\theta$ with a single observation from the density $f(x;\theta) = \exp\{-(x-\theta) - e^{(x-\theta)}\}$.



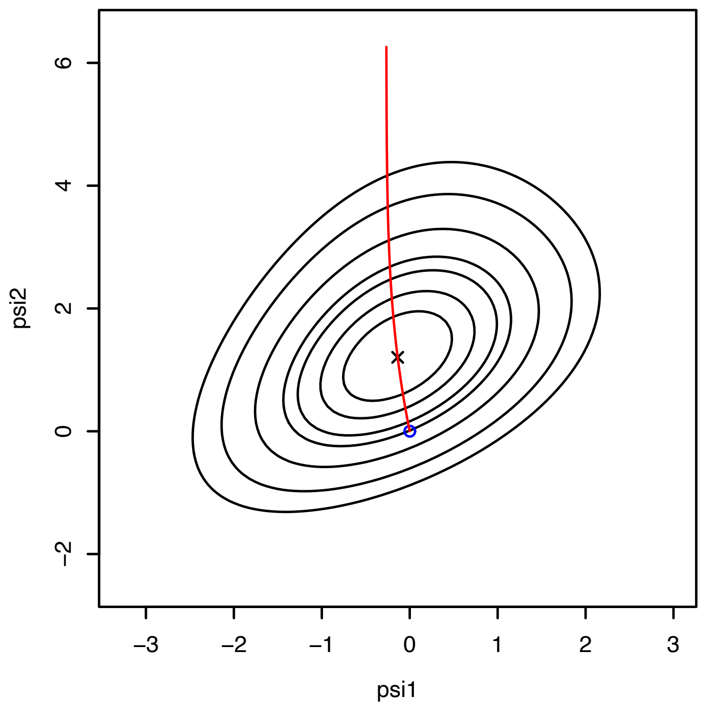



Here is the log-likelihood function for $(p_1, p_2)$ in a model for two independent binomial observations, using the data given in AoS Exercise 9.7 (d) .



Another example




**C. Profile log-likelihoods**: $\theta = (\psi, \lambda)$

Very often there are a small number of *parameters of interest*, but the model has additional parameters to improve the modelling of the observed data. We can treat it as a multivariate problem, as in B2 above, but it is sometimes convenient to work instead with the profile log-likelihood function:

$$\ell_{\mathrm{p}}(\psi; \boldsymbol{x}) = \ell(\psi, \hat{\lambda}_\psi; \boldsymbol{x}), \tag{8}$$

and in lazy notation we drop the dependence on the observed data and write $\ell_{textp}(\psi)$. In (8) $\hat{\lambda}_\psi$ is the maximum likelihood estimate of $\lambda$, when $\psi$ is fixed.

Likelihood functions are just (proportional to) density functions with the arguments switched.[2] Profile likelihood functions are not proportional to the density of an observable random variable; the maximization gets in the way. But inference based on $\ell_{\mathrm{p}}(\psi)$ has some similarities to inference based on the log-likelihood function. In particular:

[2]i.e. the data is fixed and the parameter varies

1. $\hat{\psi} = \arg\sup_\psi \ell_{\rm p}(\psi)$ i.e. you can compute the MLE in steps
2. $\text{a.var}(\hat{\psi}) = j_{\rm p}^{-1}(\hat{\psi}) \equiv \{-\ell_{\rm p}''(\hat{\psi})\}^{-1}$ as with the full likelihood
3. $(\hat{\psi} - \psi) j_{\rm p}^{1/2}(\hat{\psi}) \xrightarrow{d} N(0,1)$ a version of the usual limit theorem
4. $\hat{\psi} \pm 1.96 j_{\rm p}^{-1/2}(\hat{\psi})$ is an approximate 95% confidence interval for $\psi$ as in the Figure on p.4
5. in AoS notation, $j_p^{-1/2}(\hat{\psi}) = \widehat{\mathsf{se}}(\hat{\psi})$ Thm.9.28

*Example* Suppose $x_1, \ldots, x_n$ are i.i.d. observations from the gamma distribution, with density function

$$f(x_i; \alpha, \beta) = \frac{1}{\Gamma(\alpha)\beta^\alpha} x_i^{\alpha-1} e^{-x/\beta},$$

where $\alpha$ is the shape parameter and $\beta$ is the scale parameter.[3] Then

$$\ell(\alpha, \beta; \boldsymbol{x}) = -n \log\{\Gamma(\alpha)\} - n\alpha \log(\beta) + (\alpha - 1)\Sigma \log(x_i) - \Sigma x_i/\beta.$$

Note that the sufficient statistics for $(\alpha, \beta)$ are $(\Sigma x_i, \Sigma \log(x_i))$.

We'll assume $\alpha$ is the parameter of interest and $\beta$ is the nuisance parameter. The constrained maximum likelihood estimate of $\beta$ solves

$$\frac{\partial \ell(\alpha, \beta)}{\partial \beta} = 0,$$

which leads to an explicit expression

$$\hat{\beta}_\alpha = \frac{1}{\alpha n} \Sigma x_i.$$

The profile log-likelihood function is then

$$\begin{aligned} \ell_{\rm p}(\alpha) = \ell(\alpha, \hat{\beta}_\alpha) &= -n \log\{\Gamma(\alpha)\} - n\alpha \log(\hat{\beta}_\alpha) + (\alpha - 1)\Sigma \log(x_i) - \Sigma x_i/(\hat{\beta}_\alpha) \\ &= -n \log\{\Gamma(\alpha)\} - n\alpha \log(\bar{x}) + n\alpha \log \alpha + \alpha \Sigma \log(x_i) - n\alpha, \end{aligned}$$

where in the second line I dropped functions only of $\boldsymbol{x}$. We can now find $\hat{\alpha}$ as the solution to $\ell_{\rm p}(\alpha) = 0$, and its asymptotic variance is estimated by $\{-\ell_{\rm p}''(\hat{\alpha})\}^{-1}$.

**D. Your friend the delta method**

Maximum likelihood estimates are asymptotically normally distributed, when the model for the data is "well-behaved". In the same setting, smooth functions of maximum likelihood estimates are also asymptotically normally distributed. These functions don't need to be one-to-one, but they need to be differentiable.

[3] There are several other ways to parametrize the gamma distribution.

On the annotated slides for Jan.13, I defined $g(\theta)$ as a mapping from $\mathbb{R}^k$ to $\mathbb{R}^m$, with $m \leq k$. In Thm 9.28, $m = 1$. The delta method uses a simple Taylor series expansion to derive the expected value and variance of $g(\hat{\theta})$:

$$\begin{aligned} \mathrm{E}_\theta\{g(\hat{\theta})\} &\doteq g(\theta) \\ \mathrm{var}_\theta\{g(\hat{\theta})\} &\doteq \left(\frac{\partial g(\theta)}{\partial \theta^T}\right) I_n^{-1}(\theta) \left(\frac{\partial g(\theta)}{\partial \theta}\right). \end{aligned}$$

In these expressions $g$ is an $m \times k$ vector and $I_n^{-1}$ is a $k \times k$ matrix, so the expected value is $m \times 1$ and the variance-covariance matrix is $m \times m$. Since as written the variance depends on the unknown parameter $\theta$, we would estimate it as either

$$\left(\frac{\partial g(\hat{\theta})}{\partial \theta}\right)^T I_n^{-1}(\hat{\theta}) \left(\frac{\partial g(\hat{\theta})}{\partial \theta}\right)$$

or

$$\left(\frac{\partial g(\hat{\theta})}{\partial \theta}\right)^T j_n^{-1}(\hat{\theta}) \left(\frac{\partial g(\hat{\theta})}{\partial \theta}\right).$$

See Example 9.29 on p.134. The text uses $\nabla g$ as shorthand for $\partial g(\theta)/\partial\theta$.

**E. Likelihood Ratio Statistic** The likelihood ratio statistic, sometimes called Wilks' statistics or Wilks' Lambda is defined as

$$\begin{aligned} W(\theta) &= 2\log\left(\frac{\sup_\theta f(\boldsymbol{x};\theta)}{f(\boldsymbol{x};\theta)}\right) \\ &= 2\{\ell(\hat{\theta}) - \ell(\theta)\}. \end{aligned}$$

To derive its asymptotic distribution, we write

$$\begin{aligned} W(\theta) &= 2\{\ell(\hat{\theta}) - [\ell(\hat{\theta}) + (\theta-\hat{\theta})\ell'(\hat{\theta}) + \frac{1}{2}(\theta-\hat{\theta})^2\ell''(\hat{\theta}) + ...\} \\ &= (\hat{\theta}-\theta)^2 j_n(\hat{\theta}) + ... \\ &= (\hat{\theta}-\theta)^2 I_n(\theta) + ..., \end{aligned}$$

where I am trying to be careful about noting that the information quantities (observed, $j$, and expected, $I$) are based on a sample of size $n$. I have written this as if $\theta$ is scalar, but if $\theta \in \mathbb{R}^k$ we simply have

$$W(\theta) = (\hat{\theta}-\theta)^T I(\theta)(\hat{\theta}-\theta) + ...,$$

a quadratic form. As long as we can ensure that ... converges to 0 in probability, we get

$$W(\theta) \xrightarrow{d} \chi^2_k, \quad n \to \infty,$$

from $\sqrt{n}(\hat{\theta}-\theta) \xrightarrow{d} N_k(0, I_1^{-1}(\theta))$.

If $\theta = (\psi, \lambda)$, with $\psi \in \mathbb{R}^d$ the parameter of interest, the likelihood ratio statistic is defined using the profile likelihood:

$$\begin{aligned} W(\psi) &= 2\{\ell_{\mathrm{p}}(\hat{\psi}) - \ell_{\mathrm{p}}(\psi)\} \\ &= 2\{\ell(\hat{\psi}, \hat{\lambda}) - \ell(\psi, \hat{\lambda}_\psi)\} \\ &= 2 \log \left\{ \frac{\sup_{\psi,\lambda} L(\psi, \lambda; \boldsymbol{x})}{\sup_\lambda L(\psi, \lambda; \boldsymbol{x})} \right\}. \end{aligned}$$

Under regularity conditions on the underlying model $f(x;\theta)$, it can be shown that

$$W(\psi) \xrightarrow{d} \chi^2_d, \quad n \to \infty;$$

see SM §4.5 (p.138,9) for the proof. A very slightly more general definition is given in AoS Definition 10.21: in the context of testing a composite null hypothesis $H_0 : \theta \in \Theta_0$ against $H_1 : \theta \notin \Theta_0$ as

$$W = 2 \log \left\{ \frac{\sup_{\theta \in \Theta} L(\theta; \boldsymbol{x})}{\sup_{\theta \in \Theta_0} L(\theta; \boldsymbol{x})} \right\} = 2 \log \left\{ \frac{L(\hat{\theta}; \boldsymbol{x})}{L(\hat{\theta}_0; \boldsymbol{x})} \right\}.$$