似然方法

似然方法及其理论

似然方法是统计学中应用最广泛的方法，似然准则是统计学的最高准则之一。总体所有未知的信息都包含在似然函数中，参数的统计推断应该基于似然函数，这就是所谓的似然准则（likelihood principle）。R.A.Fisher 于 1920's 提出了极大似然方法，参数的极大似然估计使得似然函数达到极大，具有相合、渐近正态、渐近有效（最优）和不变性。除了估计参数，似然也可以用来进行假设检验、置信区间等其它统计推断，这些似然推断方法同样具有最优性。我们主要介绍经典的似然理论，即极大似然估计的性质，基于似然的检验方法包括似然比检验（LRT）、Rao's score 检验，Wald 检验的零分布。

0 Kullack-Leibler 距离

Kullack-Leibler 距离度量了两个概率密度的差异，在似然理论中有重要应用，特别地，模型选择的 AIC 准则的构建基于该度量。

Lemma 1 假设 f, g 是两个概率密度, Kullback-Leibler （KL）距离或散度（divergence）定义为

$$D(g\mid f) = E_f\left(\log\left(\frac{f}{g}\right)\right) = \int f(x)\log\left(\frac{g(x)}{f(x)}\right)$$

其中 E_f 表示对 f 求期望。则

$$D(g\mid f) \geq 0 \text{ 或 } E_f \log f \geq E_f \log g.$$

证明：因为 $-\log(x)$ 是凸函数，所以

$$E_f\left(-\log\left(\frac{g}{f}\right)\right) \geq -\log\left(E_f\left(\frac{g}{f}\right)\right) = -\log\left(\int \frac{g}{f} f\right) = -\log\left(\int g\right) = -\log(1) = 0$$

1 极大似然估计

本节给出一些基本定义，并从直观上讨论为什么 MLE 具有优良性质。为了简便，我们仅考虑样本是 iid 并存在概率密度的情形。独立但不同分布的情形类似。假设

$$X_1, \dots, X_n \text{ iid} \sim f(x\mid\theta), \theta \in \Theta \subset R^k, \quad f \text{ 是关于某个测度的密度函数.}$$

$X_1, \dots, X_n$ 的联合分布 $\prod_{i=1}^n f(x_i\mid\theta)$ 作为未知参数 θ 的函数称为似然函数:

$$L(\theta) = \prod_{i=1}^n f(x_i\mid\theta)$$

对数似然函数

$$l(\theta) = \log(L(\theta)) = \sum_{i=1}^n \log[f(x_i\mid\theta)]$$

似然方法的重要性不仅仅局限于经典的似然理论及其应用，多似然方法有众多拓展和应用，包括条件似然（conditional likelihood，消除冗余参数)、部分似然（partial likelihood, Cox 半参数模

型)、剖面似然（profile likelihood）、拟似然方法（quasi-likelihood，只假设矩函数形式而不假设具体分布)、估计方程和广义估计方程方法 (EE: EstimatingEquation, GEE)、经验似然或非参数似然等等。

将似然函数中的乘积变成加法, 易于计算和优化, 更为重要的是对数似然函数是随机变量的独立和，可直接应用大数律或中心极限定理，比如由大数律

$$l(\theta)/n \to E(\log f(X \mid \theta)) = \int \log f(x \mid \theta) \times f(x \mid \theta) dx, n \to \infty$$

而右端的极限与 KL 距离有关。极大似然估计 $\hat{\theta}$使得似然函数或等价地对数似然达到最大

$$\hat{\theta} = \arg\max l(\theta)$$

为什么极大似然估计是未知参数的良好估计？假设 θ是真参数，任取 $\theta^* \neq \theta$（即样本来自于总体 $f(x \mid \theta)$ 而不是 $f(x \mid \theta^*)$), 则由大数定律和引理 1
$\frac{l(\theta) - l(\theta^*)}{n} = \frac{\sum_{i=1}^{n} \log(f(x_i|\theta)/f(x_i|\theta^*)}{n} \to_{a.s.} E_\theta \log(\frac{f(x|\theta)}{f(x|\theta^*)}) = \int f(x \mid \theta) \log(\frac{f(x|\theta)}{f(x|\theta^*)}) dx > 0,$
其中 E_θ表示按照分布 $f(x \mid \theta)$ 求期望。上述事实说明当 $n \to \infty$ 时，几乎必然有 $l(\theta) > l(\theta^*)$，即真参数使得似然函数最大，而按照定义似然估计使得似然函数最大，所以似然估计应该等于或接近真参数。即使样本量有限，使得似然函数达到极大的点即 MLE 应该在真参数 θ附近，即 $\hat{\theta} \to \theta$（依概率或 a.s.），这就是所谓的相合性。

对数似然函数关于参数的一阶导数、二阶导数（Hessian 矩阵）是似然理论中最重要的量。一阶导数称为计分（score）函数，负的二阶导数称为（观测）Fisher 信息阵。

Definition 1 (计分函数和信息阵)

- 对数似然函数的一阶导数称为计分函数 (score):

$$U(\theta) = \dot{l} = \frac{\partial l(\theta)}{\partial \theta} = \frac{f'(x \mid \theta)}{f(x \mid \theta)}$$

- 观测信息阵 (observed Fisher information):

$$I(\theta) = -\ddot{l} = -\frac{\partial^2 l(\theta)}{\partial \theta \partial \theta^T}$$

- Fisher 信息阵：观测信息阵的期望

$$i(\theta) = E(I(\theta)) = E[-\frac{\partial^2 l(\theta)}{\partial \theta \partial \theta^T}]$$

若 $X_1, X_2, \ldots, X_n$ iid, 则 $i(\theta) = nE[-\frac{\partial^2 \log(f(x|\theta)}{\partial \theta \partial \theta^T}] \triangleq n i_1(\theta), i_1(\theta)$ 代表单个样本的信息。

Theorem 1 正则条件下,

$$EU(\theta) = 0, var(U(\theta)) = EUU^T = i(\theta)$$

即

$$E(\frac{\partial l(\theta)}{\partial \theta}) = 0, \quad E(\frac{\partial l(\theta)}{\partial \theta})(\frac{\partial l(\theta)}{\partial \theta})^T = -E\frac{\partial^2 l(\theta)}{\partial \theta \partial \theta^T}$$

其中的 E 是对 $X_1, \dots, X_n$ iid $\sim f(x \mid \theta)$ 求期望。

为了求解 MLE 使得对数似然函数 $l(\theta)$ 达到极大,通常求解如下似然方程

$$U(\theta) = \partial l / \partial \theta = 0.$$

该方程未必有唯一解。
Remark1: $i(\theta) = var(U)$ 说明 Fisher 信息阵是正定的,由大数定律,

$$\frac{1}{n} I(\theta) = \frac{1}{n} \sum_{i=1}^{n} \frac{\partial^2 \log(f(x_i \mid \theta))}{\partial\theta\partial\theta^T} \to E\left(\frac{\partial^2 \log(f(x_1 \mid \theta))}{\partial\theta\partial\theta^T}\right) = i_1(\theta) = i(\theta)/n > 0,$$

因此当 $n \to \infty$,几乎必然 $I(\theta) > 0$. 即对数似然函数的二阶导数负定,因此似然方程几乎必然有唯一解。

Remark2: 第一个性质 $E_\theta(U(\theta)) = 0$ 说明 $U(\theta)$ 在平均意义下在真参数 θ 处为 0,或者可认为 $U(\theta) \approx 0$, 这保证了似然方程 $U(\hat{\theta}) = 0$ 的解 $\hat{\theta} \approx \theta$。

Remark3: 由 Taylor 展开

$$0 = U(\hat{\theta}) \approx U(\theta) - I(\theta)(\hat{\theta} - \theta) \approx U(\theta) - i(\theta)(\hat{\theta} - \theta)$$

知

$$\hat{\theta} - \theta \approx i(\theta)^{-1} U(\theta),$$

因为 $U = U(\theta)$ 是独立随机变量之和,定理 1 求出了 U 的均值和方差,则由中心极限定理

$$var(U)^{-1/2}(U - E(U)) = i(\theta)^{-1/2} U(\theta) = i_1^{-1}(\theta) U(\theta)/\sqrt{n} \xrightarrow{d} N(0, I_k).$$

即

$$U(\theta)/\sqrt{n} \xrightarrow{d} N(0, i_1(\theta)).$$

由上述两式可得 $\hat{\theta}$ 服从渐近正态分布

$$\sqrt{n}(\hat{\theta} - \theta) = \sqrt{n} i(\theta)^{-1} U(\theta) = i_1^{-1} U(\theta)/\sqrt{n} \xrightarrow{d} N(0, i_1(\theta^{-1})).$$

以上各个 Remark 从直观上大概说明了 MLE 应该具有的优良性。下面将我们上述直观(稍微)严格化。

2 极大似然估计的渐近性质

正则条件 (Regularity Conditions)

i. f 的支撑不依赖于 θ.
ii. f 是可识别的 (若 $f(x \mid \theta_1) = f(x \mid \theta_2)$ 对任何 x 都成立,则必有 $\theta_1 = \theta_2$), $\Theta \subset R^k$ 是开集。
iii. $\left|\frac{\partial^3 \log(f(x \mid \theta))}{\partial\theta^3}\right| \le M(x), \quad E_\theta M(x) < \infty.$

将 $U(\theta)$ 在真参数 θ 处 Taylor 展开并利用 $I(\theta)/n \to i(\theta)$,

$$0 = U(\hat{\theta}) = U(\theta) + -I(\theta)(\hat{\theta} - \theta) + o_p(n^{-\frac{1}{2}}) = U(\theta) - n i_1(\theta)(\hat{\theta} - \theta) + o_p(n^{-\frac{1}{2}})$$

所以我们可将 MLE 表示为（为了简单，省略小 o 项）

$$\hat{\theta} - \theta = [n i_1(\theta)]^{-1} U(\theta) \tag{1}$$

or

$$\sqrt{n}(\hat{\theta} - \theta) = i_1^{-1}(\theta) \frac{U(\theta)}{\sqrt{n}} \tag{2}$$

下面将多次使用该表示。

Theorem 2 (相合性 consistency) 正则条件下，MLE $\hat{\theta}$是参数 θ的相合估计，即

$$\hat{\theta} \xrightarrow{Pr.} \theta$$

Proof. 由大数定律, $U(\theta)/n = \sum_{i=1}^{n} \frac{\partial \log f(x_i|\theta)}{\partial \theta} \to E\left(\frac{\partial \log f(x_1|\theta)}{\partial \theta}\right) = 0$, 所以 $\hat{\theta} = \theta + i^{-1}(\theta)\frac{U}{n} + o_p(1) \to 0$ a.s.

Theorem 3 (渐近正态 Asymptotic normality) 正则条件下,

$$\frac{U(\theta)}{\sqrt{n}} \to_d N(0, i_1(\theta)) \tag{3}$$

$$\sqrt{n}(\hat{\theta} - \theta) \to_d N(0, i_1(\theta)^{-1}) \tag{4}$$

Proof. 注意到 score 函数 $U(\theta)$ 是独立随机变量的和， $E(U(\theta)) = 0, var(U(\theta)) = i(\theta) = n i_1(\theta)$，由中心极限定理

$$var(U)^{-1/2}[U(\theta) - E\,U(\theta)] = (n i_1(\theta))^{-1/2} U(\theta) \xrightarrow{d} N(0, I_k)$$

即

$$U(\theta)/\sqrt{n} \xrightarrow{d} N(0, i_1(\theta)).$$

由（2）

$$\sqrt{n}(\hat{\theta} - \theta) = i_1^{-1}(\theta) \frac{U(\theta)}{\sqrt{n}} \xrightarrow{d} N(0, i_1(\theta)^{-1})$$

极大似然估计是渐近最优的，即在极限情况下，在所有 θ的无偏或渐近无偏估计中，极大似然估计的方差最小。

Theorem 4 (斩近最优 Asymptotic efficiency,Hajek) 若 $\tilde{\theta}$是任一 θ的无偏估计，则它可以分解为

$$\tilde{\theta} = \hat{\theta} + Z$$

其中 Z 与 $\hat{\theta}$不相关，这表明

$$var(\tilde{\theta}) = var(\hat{\theta}) + var(Z) \geq var(\hat{\theta})$$

Proof. $\dim(\theta) = k$, 对任何无偏检验 $\tilde{\theta}$, $\theta = E_\theta \tilde{\theta} = \int \tilde{\theta} L(\theta)$, 两边同时对 θ求导, 得（注意似然函数 L 是联合概率密度)

$$I_k = \int \tilde{\theta} L' = \int \tilde{\theta} L'/L \times L = E\,\tilde{\theta} L'/L = E(\tilde{\theta} U(\theta))^T,$$

其中 $L'/L = (\log L)' = U(\theta)$，所以

$$E(\tilde{\theta}-\hat{\theta})U^{T} = I_k - E(\hat{\theta}U^{T})U = I_k - E(i^{-1}UU^{T}) = I_k - I_k = 0,$$

所以 $Z = \tilde{\theta}-\hat{\theta}$ 与 $\perp\hat{\theta}$ 不相关,

$$\Rightarrow \tilde{\theta} = \hat{\theta} + Z,\ \text{where}\ Z \perp \hat{\theta}$$

下述定理可用于构造假设检验和置信区间
Theorem 5 正则条件下, 下述三个量渐近等价且都收玫到卡方分布

$$2(l(\hat{\theta}) - l(\theta)) \approx (\hat{\theta}-\theta)^{T}[ni_1(\theta)](\hat{\theta}-\theta) \approx U(\theta)^{T}[ni_1(\theta)]^{-1}U(\theta) \to_d \chi^2_k, k = \dim(\theta) \tag{5}$$

因为 $I(\theta)/n \to i_1(\theta)$, 由 Slusky 引理, 上式中 $i(\theta) = ni_1(\theta)$ 换成 $I(\theta)$ 仍然成立。
Proof. 二阶 Taylor 展开,
$l(\hat{\theta}) - l(\theta) = U(\theta)^{T}I(\hat{\theta}-\theta) - \frac{1}{2}(\hat{\theta}-\theta)^{T}I(\theta)(\hat{\theta}-\theta) + o_p(1) = U(\theta)^{T}(\hat{\theta}-\theta) - \frac{1}{2}\sqrt{n}(\hat{\theta}-\theta)^{T}i_1(\theta)\sqrt{n}(\hat{\theta}-$
由 (1), $U(\theta) = ni_1(\theta)(\hat{\theta}-\theta)$, 上述 Taylor 展开中以 $ni_1(\theta)(\hat{\theta}-\theta)$ 替代 $U(\theta)$, 得

$$l(\hat{\theta}) - l(\theta) = \frac{1}{2}(\hat{\theta}-\theta)^{T}[ni(\theta)](\hat{\theta}-\theta) + o_p(1)$$

Talyor 展开中以 $[ni_1(\theta)]^{-1}U(\theta)$ 替代 $\hat{\theta}-\theta$, 得

$$l(\hat{\theta}) - l(\theta) = \frac{1}{2}U(\theta)^{T}[n_1 i(\theta)]^{-1}U(\theta)$$

由定理 3 知近似地有 $U(\theta) \sim N(0, ni_1(\theta))$, $\sqrt{n}(\hat{\theta}-\theta) \sim N(0, (ni_1(\theta))^{-1})$，从而上述各个量都依分布收敛到 χ^2_k。

3 基于似然的检验：简单原假设情形

考虑简单原假设 $H_0 : \theta = \theta_0$。基于定理 5 的结论, 可构造基于似然的三种常见检验。

- 似然比检验（LRT）

$$LRT = -2\log(\lambda) = 2[l(\hat{\theta}) - l(\theta_0)] \sim_{H_0} \chi^2_k$$

其中似然比 $\lambda = L(\theta_0)/L(\hat{\theta})$。

- Wald 检验

$$W = (\hat{\theta}-\theta_0)^{T} I(\theta_0)(\hat{\theta}-\theta_0) \sim_{H_0} \chi^2_k$$

- Score 检验

$$S = U(\theta_0)^{T} I(\theta_0)^{-1} U(\theta_0) \sim_{H_0} \chi^2_k$$

当上述检验统计量大于 $\chi^2_k(\alpha)$ 时，在水平 α 下拒绝 H_0。三个统计量渐近等价。 W, S 中的 $I(\theta_0)$ 可换成 $ni_1(\theta_0)$ （后者需要更多的计算，需要计算期望，因而不太常用，事实上其功效略差于前者）。

Remark4：三个检验渐近等价，一般具有相近的功效，但它们在具体应用中还是有差异的。LRT 计算最简单，无需计算对数似然的导数；Wald 检验最直观，它直接对比 $\hat{\theta}$ 和 θ_0 的差异；Score 检验考察 $U(\theta_0)$（的模）是否接近 0，若接近 0，那么 $\hat{\theta} \approx \theta_0$（因为 $U(\hat{\theta}) = 0$），从从而不能拒绝原假设。

4 基于似然的检验：复合原假设情形

复合假设是更常见的情形。假设参数空间 $\Theta \subset R^k, \dim(\Theta) = k$。我们希望检验复合原假设

$$H_0: \theta \in \Theta_0 \subset \Theta, \quad \dim(\Theta_0) = m < k$$

假设存在二次可导的函数 g 使得 $\Theta_0 = g(R^m)$，则在正则条件下，当原假设成立时

$$-2\left[\log\left(\frac{L(\hat{\theta}_0)}{L(\hat{\theta})}\right)\right] \to_d \chi^2_{k-m} \quad \text{as } n \to \infty$$

其中

$$\hat{\theta}_0 = \arg\max_{\theta \in \Theta_0} l(\theta); \hat{\theta} = \arg\max_{\theta \in \Theta} l(\theta)$$

通过重新参数化，上述复合假设问题可转化为如下情形（实践中最常见的情形），即参数 θ 可划分为两部分：

$$\theta_{k\times 1} = (\psi_{(k-m)\times 1}, \lambda_{m\times 1})^T$$

其中我们感兴趣的或者需要检验的是参数 ψ，而 λ 称为冗余参数（nuisance parameter）。原假设为

$$H_0: \psi = \psi_0 \quad (\psi_0 \text{ 已知})$$

原假设成立时的参数空间为

$$\Theta_0 = \{(\psi_0, \lambda): \lambda \in R^m\}, \quad \dim(\Theta_0) = m$$

对数似然函数

$$l(\theta) = l(\psi, \lambda) = \sum_{i=1}^{n} \log f(x_i; \theta)$$

记 MLE $\hat{\theta} = (\hat{\psi}, \hat{\lambda})^T$，原假设成立时 ψ 的 MLE 为 $\hat{\psi}_0$，记 $\hat{\theta}_0 = (\psi_0, \hat{\lambda}_0)^T$。划分计分函数

$$U(\theta) = \frac{\partial l}{\partial \theta} = \begin{pmatrix} \frac{\partial l}{\partial \psi} \\ \frac{\partial l}{\partial \lambda} \end{pmatrix} \triangleq \begin{pmatrix} U_\psi \\ U_\lambda \end{pmatrix}$$

划分观测信息阵如下

$$I = -\frac{\partial^2 l}{\partial \theta \partial \theta^T} = \begin{pmatrix} I_{\psi\psi} & I_{\psi\lambda} \\ I_{\lambda\psi} & I_{\lambda\lambda} \end{pmatrix}$$

记其逆矩阵

$$I^{-1}=\begin{pmatrix} I^{\psi\psi} & I^{\psi\lambda} \\ I^{\lambda\psi} & I^{\lambda\lambda}\end{pmatrix}$$

其中$I^{\psi\psi}=I_{\psi\psi\cdot\lambda}=I_{\psi\psi}-I_{\psi\lambda}I_{\lambda\lambda}^{-1}I_{\lambda\psi}$。类似于简单假设情形，我们定义似然比检验（LRT），score 检验 S, Wald 检验 W 如下。

复合假设的似然检验

- 似然比检验

$$LRT=2[l(\hat{\theta})-l(\hat{\theta}_0)]$$

- Score 检验

$$S=U_{\psi}(\hat{\theta}_0)^{T}I_{\psi\psi}^{-1}(\hat{\theta}_0)I_{\psi}(\hat{\theta}_0)\quad(\hat{\theta}_0:\text{原假设下的 MLE})$$

- Wald 检验

$$W=(\hat{\psi}-\psi_0)^{T}I^{\psi\psi}(\hat{\theta})^{-1}(\hat{\psi}-\psi_0)\ (\hat{\theta}:\ \text{无假设下的 MLE})$$

三个检验在渐近等价，原假设成立时它们的极限分布都是χ^2_{k-m} (定理 6)。

Remark5: Remark4仍适用于这里。另外，许多著名的检验统计量都是复合情形下的 Score检验（比如列联表的 Pearson 卡方法检验）。

例1. 假设 I 个多项分布总体

$$x_i=(x_{i1},\dots,x_{iJ})^{T}\sim \text{Multinomial}(n_i;p_i),$$

其中 $p_i=(p_{i1},\dots,p_{iJ})^{T}$, $p_{i1}+\dots+p_{ij}=1, p_{ij}>0, i=1,\dots,J$，将所有计数组成 $I\times J$ 列联表 $X=(x_{ij},1\le i\le I,1\le j\le J)$。齐一性假设

$$H_0:p_1=\dots=p_I$$

则 H_0 的 score 检验为列联表的 Pearson 卡方检验。
Theorem 6 LRT,S,W 渐近等价, 原假设 $H_0:\psi=\psi_0$ 成立时它们的分布收玫到 χ^2_{k-m}, $n\to\infty$。

Appendix: Sketch proof of Theorem 6

We only prove LRT. Let $\hat{\theta}_0=(\psi_0,\hat{\lambda}_0)$, where $\hat{\lambda}_0$ is the MLE under the null, i.e, we have

$$\frac{\partial l(\theta)}{\partial\lambda}\Big|_{\psi=\psi_0,\lambda=\hat{\lambda}_0}=0$$

The global MLE is denoted by $\hat{\theta}=(\hat{\psi},\hat{\lambda})$, then by Taylor expansion, the log-likelihood ratio is (omitting high-order terms)

$$\begin{aligned} l(\hat{\theta}_0)-l(\hat{\theta}) &= \frac{\partial l(\theta)}{\partial\theta}\Big|_{\theta=\hat{\theta}}+\frac{1}{2}(\hat{\theta}_0-\hat{\theta})^{T}\left[\frac{\partial^2 l(\theta)}{\partial\theta\partial\theta^{T}}\Big|_{\theta=\hat{\theta}}\right](\hat{\theta}_0-\hat{\theta}) \\ &\approx -\frac{n}{2}(\hat{\theta}_0-\hat{\theta})^{T}[i(\hat{\theta})](\hat{\theta}_0-\hat{\theta}) \end{aligned} \tag{6}$$

where we have subsituted $\frac{\partial^2 l(\theta)}{\partial\theta\partial\theta^T}\Big|_{\theta=\hat{\theta}}$ by its expectation at true parameter $\theta_0 = (\psi_0, \lambda)$.

Next, we expand $U(\hat{\theta}_0)$ around $\hat{\theta}$.

$$U(\hat{\theta}_0) = \begin{pmatrix} \frac{\partial l(\theta)}{\partial\psi}\Big|_{(\psi_0,\hat{\lambda}_0)} \\ \frac{\partial l(\theta)}{\partial\lambda}\Big|_{(\psi_0,\hat{\lambda}_0)} \end{pmatrix} = \begin{pmatrix} \frac{\partial l(\theta)}{\partial\psi}\Big|_{(\psi_0,\hat{\lambda}_0)} \\ 0 \end{pmatrix}$$

$$= \frac{\partial U}{\partial\theta}(\hat{\theta}_0 - \hat{\theta}) = -ni(\theta_a)(\hat{\theta}_0 - \hat{\theta}) = -n\begin{pmatrix} i_{\psi\psi} & i_{\psi\lambda} \\ i_{\lambda\psi} & i_{\lambda\lambda} \end{pmatrix}\begin{pmatrix} \psi_0 - \hat{\psi} \\ \hat{\lambda}_0 - \hat{\lambda} \end{pmatrix}$$

$$= -n\begin{pmatrix} i_{\psi\psi}(\psi_0 - \hat{\psi}) + i_{\psi\lambda}(\hat{\lambda}_0 - \hat{\lambda}) \\ i_{\lambda\psi}(\psi_0 - \hat{\psi}) + i_{\lambda\lambda}(\hat{\lambda}_0 - \hat{\lambda}) \end{pmatrix}$$

So from the second component in the above equation we have

$$\hat{\lambda}_0 - \hat{\lambda} = -i_{\lambda\lambda}^{-1} i_{\lambda\psi}(\psi_0 - \hat{\psi})$$

Then,

$$\hat{\theta}_0 - \hat{\theta} = \begin{pmatrix} \psi_0 - \hat{\psi} \\ \hat{\lambda}_0 - \hat{\lambda} \end{pmatrix} = \begin{pmatrix} \psi_0 - \hat{\psi} \\ -i_{\lambda\lambda}^{-1} i_{\lambda\psi}(\psi_0 - \hat{\psi}) \end{pmatrix} = \begin{pmatrix} I \\ -i_{\lambda\lambda}^{-1} i_{\lambda\psi} \end{pmatrix}(\psi_0 - \hat{\psi}) \tag{7}$$

Then from (6) and (7), we have

$$l(\hat{\theta}_0) - l(\hat{\theta}) \approx -\frac{n}{2}(\psi_0 - \hat{\psi})^T (I, -i_{\psi\lambda} i_{\lambda\lambda}^{-1}) \begin{pmatrix} i_{\psi\psi} & i_{\psi\lambda} \\ i_{\lambda\psi} & i_{\lambda\lambda} \end{pmatrix}\begin{pmatrix} I \\ -i_{\lambda\lambda}^{-1} i_{\lambda\psi} \end{pmatrix}(\psi_0 - \hat{\psi})$$

$$= -\frac{n}{2}(\psi_0 - \hat{\psi})^T i_{\psi\psi\cdot\lambda}(\theta_0)(\psi_0 - \hat{\psi}) \tag{8}$$

where $i_{\psi\psi\cdot\lambda} = i_{\psi\psi} - i_{\psi\lambda} i_{\lambda\lambda}^{-1} i_{\lambda\lambda}$. Since under the null hypothesis, true parameter is $\theta_0 = (\psi_0, \lambda)^T$, $\sqrt{n}(\hat{\theta} - \theta_0) \to_d N(0, i^{-1}(\theta_0))$, then marginally

$$\sqrt{n}(\hat{\psi} - \psi_0) \to N(0, i_{\psi\psi\cdot\lambda}^{-1}(\theta_0))$$

combining with (8), we have

$$-2\{l(\hat{\theta}_0) - l(\hat{\theta})\} = [\sqrt{n}(\hat{\psi} - \psi_0)][i_{\psi\psi\lambda}(\theta_0)][\sqrt{n}(\hat{\psi} - \psi_0)] \to \chi^2_{k-m}$$

where $k - m$ is the length of ψ.

Reference:

Cox and Hinkley (1974) Theoretical Statistics. CRC.