# 557: MATHEMATICAL STATISTICS II
# LARGE SAMPLE AND ASYMPTOTIC RESULTS

We now assess the properties of statistical procedures when the sample size $n$ becomes **large** (*large sample theory*), or in the limit as $n$ **tends to infinity** (*asymptotic theory*).

**5.1 Point Estimators**

Consider a potentially infinite sequence of random variables $X_1, X_2, \ldots, X_n, \ldots$, and a corresponding sequence of estimators $\{T_n, n \geq 1\}$ of parameter $\tau(\theta)$, where, for each $n$,

$$T_n \equiv T_n(X_1, \ldots, X_n).$$

**Consistency and Asymptotic Unbiasedness**

The sequence $\{T_n, n \geq 1\}$ is **consistent** for $\tau(\theta)$ if

$$T_n \longrightarrow \tau(\theta) \qquad \forall\, \theta$$

in probability (*weak consistency*), almost surely (*strong consistency*), or in $r$th mean for some $r$ (for $r = 2$, *mean-square consistency*). The sequence of estimators is **asymptotically unbiased** for $\tau(\theta)$ if

$$\lim_{n \longrightarrow \infty} \mathrm{E}_{f_{T_n|\theta}}[T_n|\theta] = \tau(\theta).$$

Recall that $X_n \overset{r=2}{\longrightarrow} X \Longrightarrow X_n \overset{p}{\longrightarrow} X$, so that mean-square consistency implies weak consistency. Then

$$\mathrm{E}_{f_{T_n|\theta}}[(T_n - \tau(\theta))^2|\theta] = \mathrm{Var}_{f_{T_n|\theta}}[T_n|\theta] + \Big(\mathrm{E}_{f_{T_n|\theta}}[T_n - \tau(\theta)|\theta]\Big)^2$$

so mean-square consistency follows if $T_n$ is asymptotically unbiased and has variance converging to zero. The *asymptotic variance* of $T_n$ is $\sigma^2$ if, for some sequence of constants $\{k_n\}$,

$$\lim_{n \longrightarrow \infty} k_n \mathrm{Var}_{f_{T_n|\theta}}[T_n|\theta] = \sigma^2 < \infty$$

**Efficiency**

A sequence of asymptotically unbiased estimators $T_n = T_n(\underset{\sim}{X})$ of $\tau(\theta)$ is **efficient** if the variance of $\sqrt{n}\,(T_n - \tau(\theta))$ converges to the lower bound on variance dictated by the Cramér-Rao result, that is

$$\lim_{n \longrightarrow \infty} n\, \mathrm{E}_{f_{T_n|\theta}}[(T_n - \tau(\theta))^2|\theta] = (\dot{\tau}(\theta))^2\, \mathcal{I}(\theta)^{-1}$$

Note: For finite $n$, an unbiased estimator $T$ is sometimes termed efficient if its variance attains the Cramér-Rao lower bound; the efficiency, $e_T(\theta)$, of an unbiased estimator of $\theta$ is defined by

$$e_T(\theta) = \frac{\mathcal{I}(\theta)^{-1}}{\mathrm{Var}_{f_{T|\theta}}[T|\theta]}.$$

These definitions can be extended to the multivariate case.

**Asymptotic Relative Efficiency**

Consider two estimators $\tau(\theta)$, $T_{1n} = T_{1n}(\underset{\sim}{X})$ and $T_{2n} = T_{2n}(\underset{\sim}{X})$. The **Asymptotic Relative Efficiency** (ARE) of $T_{1n}$ with respect to $T_{2n}$ is defined as the ratio of their asymptotic mean-square errors (AMSE)

$$\mathrm{ARE}_\theta(T_{1n}, T_{2n}) = \frac{\mathrm{AMSE}_\theta(T_{2n})}{\mathrm{AMSE}_\theta(T_{1n})} = \frac{\lim\limits_{n \longrightarrow \infty} \mathrm{E}_{f_{T_{2n}|\theta}}[(T_{2n} - \tau(\theta))^2]}{\lim\limits_{n \longrightarrow \infty} \mathrm{E}_{f_{T_{1n}|\theta}}[(T_{1n} - \tau(\theta))^2]}.$$

For two asymptotically unbiased estimators, the ARE is the ratio of the asymptotic variances.

**Asymptotic Behaviour Of The Maximum Likelihood Estimator**

Consider a random sample $x_1, \ldots, x_n$ from a probability model indexed by parameter $\utilde{\theta} \in \Theta \subseteq \mathbb{R}^d$, with density denoted $f_{X|\utilde{\theta}}$ with support $\mathbb{X}$. Denote the true value of $\utilde{\theta}$ by $\utilde{\theta}_0$. Denote by $L(\utilde{\theta}|\utilde{x})$ and $l(\utilde{\theta}|\utilde{x})$ the likelihood and log likelihood respectively, and denote by

$$\dot{l}_j(\utilde{\theta}) = \frac{\partial l(\utilde{\theta}|x)}{\partial \theta_j} \qquad \ddot{l}_{jk}(\utilde{\theta}|x) = \frac{\partial^2 l(\utilde{\theta}|x)}{\partial \theta_j \partial \theta_k} \qquad \dddot{l}_{jkl}\ (\utilde{\theta}|x) = \frac{\partial^3 l(\utilde{\theta}|x)}{\partial \theta_j \partial \theta_k \partial \theta_l}$$

the partial derivatives up to order three of $l(\utilde{\theta}|x) = \log f_{X|\utilde{\theta}}(x|\utilde{\theta})$. Note that

$$l_n(\utilde{\theta}) = l(\utilde{\theta}|x_1, \ldots, x_n) = \sum_{i=1}^{n} l(\utilde{\theta}|x_i)$$

and, for the

$$\dot{l}_n(\theta) = \dot{\utilde{l}}(\utilde{\theta}|x_1, \ldots, x_n) = \sum_{i=1}^{n} \dot{l}(\utilde{\theta}|x_i) = \sum_{i=1}^{n} \frac{\dot{f}_{X|\utilde{\theta}}(x_i|\utilde{\theta})}{f_{X|\utilde{\theta}}(x_i|\utilde{\theta})} \tag{1}$$

with similar results for the other derivatives. Under mild regularity conditions, we prove that a solution to the equation found by equating (1) to zero provides an estimate for which the corresponding estimator that is weakly consistent for $\utilde{\theta}_0$.

**Regularity Conditions:**

A1. **Identifiability :** $f_{X|\utilde{\theta}_1}(x|\utilde{\theta}_1) = f_{X|\utilde{\theta}_2}(x|\utilde{\theta}_2)\ \forall\, x \in \mathbb{X} \quad \Longleftrightarrow \quad \utilde{\theta}_1 = \utilde{\theta}_2$

A2. $\mathbb{X}$ **does not depend on** $\utilde{\theta}$.

A3. $\Theta$ contains an **open neighbourhood**, $\Theta_0 \subset \mathbb{R}^d$, of $\utilde{\theta}_0$

To find the maximum likelihood estimate, we solve the system of **likelihood equations**

$$\dot{\utilde{l}}_n(\utilde{\theta}) = \dot{\utilde{l}}(\utilde{\theta}|\utilde{x}) = 0 \tag{LE}$$

that is, a system of $d$ equations based on the first partial derivative vector $\dot{\utilde{l}}$.

First, note that if $\utilde{\theta} \neq \utilde{\theta}_0$,

$$T_n(\utilde{x}, \utilde{\theta}_0, \utilde{\theta}) = \frac{1}{n} \frac{l_n(\utilde{\theta})}{l_n(\utilde{\theta}_0)} = \frac{1}{n} \sum_{i=1}^{n} \log \left\{ \frac{f_{X|\theta}(x|\utilde{\theta})}{f_{X|\utilde{\theta}}(x|\utilde{\theta}_0)} \right\}$$

then, as $n \longrightarrow \infty$, by the weak law of large numbers (WLLN), say,

$$T_n(\utilde{X}, \utilde{\theta}_0, \utilde{\theta}) \stackrel{p}{\longrightarrow} \mathrm{E}_{f_{X|\utilde{\theta}}} \left[ \log \frac{f_{X|\utilde{\theta}}(X|\utilde{\theta})}{f_{X|\utilde{\theta}}(X|\utilde{\theta}_0)} \right] = \int \log \left\{ \frac{f_{X|\theta}(x|\utilde{\theta})}{f_{X|\utilde{\theta}}(x|\utilde{\theta}_0)} \right\} f_{X|\utilde{\theta}}(x|\utilde{\theta}_0)\, dx = -K(\utilde{\theta}_0, \utilde{\theta}) < 0$$

where $K(\utilde{\theta}_0, \utilde{\theta})$ is the Kullback-Leibler divergence between the pdfs with parameters $\utilde{\theta}_0$ and $\utilde{\theta}$. Hence, by A1, $T_n(\utilde{X}, \utilde{\theta}_0, \utilde{\theta})$ converges to something negative. Thus, for all $\utilde{\theta} \neq \utilde{\theta}_0$,

$$\Pr[L(\utilde{\theta}_0|\utilde{X}) > L(\utilde{\theta}|\utilde{X})|\utilde{\theta}_0] \longrightarrow 1 \tag{2}$$

as $n \longrightarrow \infty$; with probability converging to 1, the likelihood at $\utilde{\theta}_0$ is greater than the likelihood elsewhere in $\Theta$.

**Consistency and Asymptotic Normality: Univariate Case**

In the case $d = 1$, it is now straightforward to show that a solution - not necessarily the maximum likelihood solution - to the equation (LE) is weakly consistent for $\theta_0$, under additional regularity conditions: provided that the log-likelihood is suitably **differentiable** with respect to $\theta$ on $\Theta_0$.

A3. $\Theta$ contains an **open neighbourhood**, $\Theta_0 \subset \mathbb{R}$, of $\theta_0$ on which

(i) $l(\theta|x)$ is **twice continuously differentiable with respect to** $\theta$ for all $x \in \mathbb{X}$.

(ii) Third derivatives of $l(\theta|x)$ **exist** and are **absolutely bounded**, that is for $\theta \in \Theta_0$

$$\left|\dddot{l}\ (\theta|x)\right| \leq M(x) \qquad \text{where} \qquad \mathrm{E}_{f_{X|\theta}}\left[M(X)|\,\theta_0\right] < m < \infty$$

A4.

$$E_{f_{X|\theta}}\left[\dot{l}(\theta_0|X)\right] = 0 \qquad\qquad E_{f_{X|\theta}}\left[(\dot{l}(\theta_0|X))^2\right] < \infty.$$

**Consistency:** Let $a > 0$ and consider the set

$$B_a \equiv \left\{\underset{\sim}{x} : L(\theta_0 - a|\underset{\sim}{x}) < L(\theta_0|\underset{\sim}{x}) \text{ and } L(\theta_0 + a|\underset{\sim}{x}) < L(\theta_0|\underset{\sim}{x})\right\} \subset \Theta_0$$

By equation (2), $\Pr(B_a) \longrightarrow 1$ as $n \longrightarrow \infty$. Therefore, with probability tending to one,

$$L(\theta_0 - a|\underset{\sim}{x}) < L(\theta_0|\underset{\sim}{x}) > L(\theta_0 + a|\underset{\sim}{x}).$$

As the log-likelihood is differentiable in a neighbourhood of $\theta_0$, $L(\theta|\underset{\sim}{x})$ has a local maximum, $\widetilde{\theta}_n(a)$, in the set $(\theta_0 - a, \theta_0 + a)$, at which

$$\dot{l}_n(\widetilde{\theta}_n(a)) = 0.$$

Hence, for $a$ arbitrarily small

$$\Pr[|\widetilde{\theta}_n(a) - \theta_0| < a|\theta_0] \longrightarrow 1$$

as $n \longrightarrow \infty$, so therefore the sequence of estimators $\{\widetilde{\theta}_n(a), n \geq 1\}$ converges in probability to $\theta_0$. To obtain the required result independent of $a$, let $\widetilde{\theta}_n$ be the root of the likelihood equations closest to $\theta_0$.

*Note that this portion of the proof only requires differentiability of $f_{X|\theta}(x|\theta)$ on an open neighbourhood $\Theta_0$, and not the remaining parts of A3 and A4.*

**Asymptotic Normality:** Consider a Taylor expansion of $\dot{l}_n(\theta)$ around $\theta_0$

$$\dot{l}_n(\theta) = \dot{l}_n(\theta_0) + (\theta - \theta_0)\ddot{l}_n(\theta_0) + \frac{1}{2}(\theta - \theta_0)^2\ \dddot{l}_n\ (\theta^\star)$$

where $\theta^\star$ lies between $\theta_0$ and $\theta$. Evaluating this at $\theta = \widetilde{\theta}_n$, a root of the likelihood equation, we have

$$0 = \dot{l}_n(\widetilde{\theta}_n) = \dot{l}_n(\theta_0) + (\widetilde{\theta}_n - \theta_0)\ddot{l}_n(\theta_0) + \frac{1}{2}(\widetilde{\theta}_n - \theta_0)^2\ \dddot{l}_n\ (\theta_n^\star)$$

so that on rearrangement

$$\sqrt{n}(\widetilde{\theta}_n - \theta_0) = \frac{\dot{l}_n(\theta_0)/\sqrt{n}}{-(1/n)\ddot{l}_n(\theta_0) - (1/2n)(\widetilde{\theta}_n - \theta_0)\ \dddot{l}_n\ (\theta_n^\star)}$$

Now, in terms of $X_1, \ldots, X_n$ as $n \longrightarrow \infty$, by the Central Limit Theorem

$$\frac{1}{\sqrt{n}}\,\dot{l}_n(\theta_0) = \sqrt{n}\,\frac{1}{n}\left\{\sum_{i=1}^{n}\frac{\dot{f}_{X|\theta}(X_i|\theta_0)}{f_{X|\theta}(X_i|\theta_0)}\right\} = \sqrt{n}S(\underset{\sim}{X};\theta_0) \stackrel{d}{\longrightarrow} Z \sim \text{Normal}(0, V(\theta_0))$$

where

$$V(\theta_0) = \text{Var}_{f_{X|\theta}}[S(X;\theta_0)] = \mathcal{I}(\theta_0).$$

Similarly, by the Weak Law of Large Numbers, as $n \longrightarrow \infty$,

$$-\frac{1}{n}\ddot{l}_n(\theta_0) = \frac{1}{n}\sum_{i=1}^{n}\Psi(\theta_0;X_i) \stackrel{p}{\longrightarrow} \mathcal{I}(\theta_0).$$

Finally, with probability tending to 1,

$$\left|\frac{1}{n}\stackrel{\dots}{l}_n(\theta_n^\star)\right| = \left|\frac{1}{n}\sum_{i=1}^{n}\stackrel{\dots}{l}(\theta_n^\star;X_i)\right| < \frac{1}{n}\sum_{i=1}^{n}M(X_i) \stackrel{p}{\longrightarrow} \text{E}_{f_{X|\theta}}[M(X)|\theta_0].$$

Hence, as $\widetilde{\theta}_n \longrightarrow \theta_0$, $(\widetilde{\theta}_n - \theta_0) \stackrel{p}{\longrightarrow} 0$, and

$$\frac{1}{n}(\widetilde{\theta}_n - \theta_0)\stackrel{\dots}{l}_n(\theta_n^\star) \stackrel{p}{\longrightarrow} 0.$$

Thus, by Slutsky's Theorem

$$\sqrt{n}(\widetilde{\theta}_n - \theta_0) \stackrel{d}{\longrightarrow} \text{Normal}(0, \mathcal{I}(\theta_0)^{-1})$$

**Extension to the Multivariate Case**

With extensions to the regularity conditions, we can provide a similar result in the multivariate case.

**Extended Regularity Conditions:**

A3. $\Theta$ contains an **open neighbourhood**, $\Theta_0 \subset \mathbb{R}^d$, of $\underset{\sim}{\theta}_0$ on which

(i) $l(\underset{\sim}{\theta}|x)$ is **twice continuously differentiable with respect to** $\underset{\sim}{\theta}$ for all $x \in \mathbb{X}$.

(ii) Third derivatives of $l(\underset{\sim}{\theta}|x)$ **exist** and are **absolutely bounded**, that is

$$\left|\stackrel{\dots}{l}_{jkl}(\underset{\sim}{\theta}|x)\right| \leq M_{jkl}(x) \qquad \underset{\sim}{\theta} \in \Theta_0$$

for all $j, k, l$, for some function $M_{jkl}(x)$ where

$$E_{f_{X|\underset{\sim}{\theta}_0}}[M_{jkl}(X)|\underset{\sim}{\theta}_0] < m_{jkl} < \infty$$

A4. (i) $E_{f_{X|\underset{\sim}{\theta}_0}}\left[\dot{l}_j(\underset{\sim}{\theta}_0|X)\right] = 0$ for $j = 1, \dots, d$.

(ii) $E_{f_{X|\underset{\sim}{\theta}_0}}\left[(\dot{l}_j(\underset{\sim}{\theta}_0|X))^2\right] < \infty$ for $j = 1, \dots, d$.

(iii) The $k \times k$ Fisher information matrix $\mathcal{I}(\underset{\sim}{\theta}_0)$ with $(j,k)^{\text{th}}$ entry

$$\mathcal{I}_{jk}(\underset{\sim}{\theta}_0) = E_{f_{X|\underset{\sim}{\theta}_0}}\left[-\ddot{l}_{jk}(\underset{\sim}{\theta}_0|X)\right]$$

is **positive definite**.

**Existence, Consistency and Asymptotic Normality of a Root of the Likelihood Equations**
Suppose that conditions A1 to A4 hold. Then, as $n \longrightarrow \infty$, with probability converging to 1, there exist solutions $\tilde{\underset{\sim}{\theta}}_n$ of the likelihood equations (LE) such that

$$\tilde{\underset{\sim}{\theta}}_n \stackrel{p}{\longrightarrow} \underset{\sim}{\theta}_0.$$

In addition

$$\sqrt{n}(\tilde{\underset{\sim}{\theta}}_n - \underset{\sim}{\theta}_0) \stackrel{d}{\longrightarrow} \text{Normal}(\underset{\sim}{0}, \mathcal{I}(\underset{\sim}{\theta}_0)^{-1})$$

**Proof (NOT EXAMINABLE)**
Let $a > 0$, and define $Q_a$ such that $Q_a = \{\underset{\sim}{\theta} \in \Theta : \|\underset{\sim}{\theta} - \underset{\sim}{\theta}_0\| = a\}$. Consider a third order Taylor expansion of $\underset{\sim}{l}_n(\underset{\sim}{\theta})$ of around $\underset{\sim}{\theta}_0$. Rearranging, and dividing by $n$, we have

$$\begin{aligned}
\frac{1}{n}(l_n(\underset{\sim}{\theta}) - l_n(\underset{\sim}{\theta}_0)) &= \frac{1}{n}\sum_{j=1}^{k} A_j(\underset{\sim}{x})(\theta_j - \theta_{0j}) + \frac{1}{2n}\sum_{j=1}^{d}\sum_{k=1}^{d} B_{jk}(\underset{\sim}{x})(\theta_j - \theta_{0j})(\theta_k - \theta_{0k}) \\
&+ \frac{1}{6n}\sum_{j=1}^{d}\sum_{k=1}^{d}\sum_{l=1}^{d}(\theta_j - \theta_{j0})(\theta_k - \theta_{k0})(\theta_l - \theta_{l0})\left\{\sum_{i=1}^{n}\gamma_{jkl}(x_i)M_{jkl}(x_i)\right\} \qquad (3) \\
&= s_1 + s_2 + s_3
\end{aligned}$$

say, where $0 \leq |\gamma_{jkl}(x)| \leq 1$, and, for $j, k = 1, \ldots, d$,

$$A_j(\underset{\sim}{x}) = \dot{l}_j(\underset{\sim}{\theta}_0|\underset{\sim}{x}) \qquad\qquad B_{jk}(\underset{\sim}{x}) = \ddot{l}_{jk}(\underset{\sim}{\theta}_0|\underset{\sim}{x})$$

Let $S_1, S_2$ and $S_3$ be the random variables corresponding to the quantities $s_1, s_2$ and $s_3$. We aim to show that the supremum of $(l_n(\underset{\sim}{\theta}) - l_n(\underset{\sim}{\theta}_0))/n$ on $Q_a$ is negative with probability tending to 1 if $a$ is sufficiently small; to do this, we show that the supremum of $S_2$ is negative, while $S_1$ and $S_2$ are negligible compared to $S_2$. Now, by the WLLN and assumption A3(i),

$$\frac{1}{n}A_j(\underset{\sim}{X}) = \frac{1}{n}\,\dot{l}_j(\underset{\sim}{\theta}_0|\underset{\sim}{X}) \xrightarrow{p} \mathrm{E}_{f_{X|\underset{\sim}{\theta}_0}}[\dot{l}_j(\underset{\sim}{\theta}_0|X)] = 0 \qquad (4)$$

and by the WLLN

$$\frac{1}{n}B_{jk}(\underset{\sim}{X}) = \frac{1}{n}\,\ddot{l}_{jk}(\underset{\sim}{\theta}_0|\underset{\sim}{x}) \xrightarrow{p} \mathrm{E}_{f_{X|\underset{\sim}{\theta}_0}}[\ddot{l}_{jk}(\underset{\sim}{\theta}_0|X)] = -\mathcal{I}_{jk}(\underset{\sim}{\theta}_0) \qquad (5)$$

On $Q_a$, we have

$$|S_1| \leq \frac{1}{n}a\sum_{j=1}^{d}|A_j(\underset{\sim}{X})|$$

so that for any $a$, as $n \longrightarrow \infty$, from equation (4), with probability tending to 1,

$$\frac{1}{n}|A_j(\underset{\sim}{X})| < a^2 \qquad \therefore \qquad |S_1| < sa^3$$

Secondly,

$$\begin{aligned}
2S_2 &= \frac{1}{n}\sum_{j=1}^{d}\sum_{k=1}^{d} B_{jk}(\underset{\sim}{X})(\theta_j - \theta_{0j})(\theta_k - \theta_{0k}) \\
&= \sum_{j=1}^{d}\sum_{k=1}^{d}\left(\frac{1}{n}B_{jk}(\underset{\sim}{X}) - (-\mathcal{I}_{jk}(\underset{\sim}{\theta}_0))\right)(\theta_j - \theta_{0j})(\theta_k - \theta_{0k}) + \sum_{j=1}^{d}\sum_{k=1}^{d}(-\mathcal{I}_{jk}(\underset{\sim}{\theta}_0))(\theta_j - \theta_{0j})(\theta_k - \theta_{0k})
\end{aligned}$$

As before, as $n \longrightarrow \infty$, from equation (5), with probability tending to 1,

$$\left|\sum_{j=1}^{d}\sum_{k=1}^{d}\left(\frac{1}{n}B_{jk}(\underset{\sim}{X}) - (-\mathcal{I}_{jk}(\underset{\sim}{\theta}_0))\right)(\theta_j - \theta_{0j})(\theta_k - \theta_{0k})\right| < s^2a^3 \qquad (6)$$

whereas the second term is the constant quadratic form

$$\sum_{j=1}^{d}\sum_{k=1}^{d}(-\mathcal{I}_{jk}(\underset{\sim}{\theta}_0))(\theta_j - \theta_{0j})(\theta_k - \theta_{0k}) = -(\underset{\sim}{\theta} - \underset{\sim}{\theta}_0)^{\mathsf{T}}\mathcal{I}(\underset{\sim}{\theta}_0)(\underset{\sim}{\theta} - \underset{\sim}{\theta}_0).$$

Now, as $\mathcal{I}(\underset{\sim}{\theta}_0)$ is positive definite, it has a singular value decomposition $\mathcal{I}(\underset{\sim}{\theta}_0) = V^\mathsf{T} D V$, where $D$ is the diagonal eigenvalue matrix $D = \text{diag}(\lambda_1, \ldots, \lambda_d)$, with $0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_d$, and $V$ is the matrix of eigenvectors, with $V^\mathsf{T} V = I_d$. Thus

$$-(\underset{\sim}{\theta} - \underset{\sim}{\theta}_0)^\mathsf{T} \mathcal{I}(\underset{\sim}{\theta}_0)(\underset{\sim}{\theta} - \underset{\sim}{\theta}_0) = -\sum_{j=1}^{d} \lambda_j \, \xi_j(\underset{\sim}{\theta}_0, \underset{\sim}{\theta})^2$$

where $\underset{\sim}{\xi}(\underset{\sim}{\theta}_0, \underset{\sim}{\theta}) = V(\underset{\sim}{\theta} - \underset{\sim}{\theta}_0)$, so that

$$\sum_{j=1}^{d} \xi_j(\underset{\sim}{\theta}_0, \underset{\sim}{\theta})^2 = \underset{\sim}{\xi}(\underset{\sim}{\theta}_0, \underset{\sim}{\theta})^\mathsf{T} \underset{\sim}{\xi}(\underset{\sim}{\theta}_0, \underset{\sim}{\theta}) = (\underset{\sim}{\theta} - \underset{\sim}{\theta}_0)^\mathsf{T} V^\mathsf{T} V (\underset{\sim}{\theta} - \underset{\sim}{\theta}_0) = (\underset{\sim}{\theta} - \underset{\sim}{\theta}_0)^\mathsf{T} (\underset{\sim}{\theta} - \underset{\sim}{\theta}_0) = \sum_{j=1}^{d} (\theta_j - \theta_{0j})^2$$

Now, on the surface of the hypersphere $Q_a$, $\|(\underset{\sim}{\theta} - \underset{\sim}{\theta}_0)\| = a$ so

$$\sum_{j=1}^{d} \xi_j(\underset{\sim}{\theta}_0, \underset{\sim}{\theta})^2 = a^2 \geq \lambda_1 \sum_{j=1}^{d} \xi_j(\underset{\sim}{\theta}_0, \underset{\sim}{\theta})^2 \geq \lambda_1 \, a^2 \qquad \therefore \qquad -(\underset{\sim}{\theta} - \underset{\sim}{\theta}_0)^\mathsf{T} \mathcal{I}(\underset{\sim}{\theta}_0)(\underset{\sim}{\theta} - \underset{\sim}{\theta}_0) \leq -\lambda_1^2 a^2 \tag{7}$$

Hence, combining equations (6) and (7), with probability tending to 1, for $a$ small enough, $S_2 < -ca^2$. Finally, for $S_3$, with probability tending to 1,

$$\left| \frac{1}{n} \sum_{i=1}^{n} M_{jkl} \right| < 2m_{jkl} \qquad \therefore \qquad |S_3| < \frac{1}{6} s^3 a^3 \sum_{j=1}^{d} \sum_{k=1}^{d} \sum_{l=1}^{d} m_{jkl} = ba^3$$

say. Thus, combining results we have

$$\sup_{\underset{\sim}{\theta} \in Q_a} (S_1 + S_2 + S_3) \leq \sup_{\underset{\sim}{\theta} \in Q_a} S_2 + \sup_{\underset{\sim}{\theta} \in Q_a} \|S_1 + S_3\| < -ca^2 + (b+s)a^2$$

which is **negative** if $a < c/(b+s)$. Thus, $l$ has a local maximum inside $Q_a$, as for $n$ large enough, with probability at least $1 - \epsilon$ that is, as $(l_n(\underset{\sim}{\theta}) - l_n(\underset{\sim}{\theta}_0))/n < 0$, or equivalently,

$$\Pr\left[\ l_n(\underset{\sim}{\theta}) < l_n(\underset{\sim}{\theta}_0) \text{ for all } \underset{\sim}{\theta} \in Q_a \mid \underset{\sim}{\theta}_0 \right] \longrightarrow 1 \qquad \text{as} \qquad n \longrightarrow \infty.$$

Therefore, as the likelihood equations (LE) are satisfied at local maxima, it follows that (with probability converging to 1 as $n \longrightarrow \infty$) there **exists** a solution, $\tilde{\underset{\sim}{\theta}}_n(a)$, inside $Q_a$, for any $a$ small enough. Thus the result follows as

$$\lim_{n \longrightarrow \infty} \Pr[\ \|\tilde{\underset{\sim}{\theta}}_n(a) - \underset{\sim}{\theta}_0\| < a\ ] = 1 \qquad \therefore \qquad \tilde{\underset{\sim}{\theta}}_n(a) \xrightarrow{p} \underset{\sim}{\theta}_0.$$

The proof of asymptotic normality proceeds in a similar fashion to the univariate case; by multivariate Taylor's Theorem in the $d \times 1$ system of equations

$$\frac{1}{\sqrt{n}} \dot{\underset{\sim}{l}}_n(\underset{\sim}{\theta}_0) = -\frac{1}{\sqrt{n}} \ddot{\underset{\sim}{l}}_n(\underset{\sim}{\theta}_0)(\tilde{\underset{\sim}{\theta}}_n - \underset{\sim}{\theta}_0) - \frac{1}{2\sqrt{n}} (\widetilde{\theta}_n - \underset{\sim}{\theta}_0)^\mathsf{T} \dddot{\underset{\sim}{l}}_n(\underset{\sim}{\theta}^\star_n)(\widetilde{\theta}_n - \underset{\sim}{\theta}_0)$$

The left hand side converges in probability (and in distribution) to $Z \sim \text{Normal}(\underset{\sim}{0}, \mathcal{I}(\underset{\sim}{\theta}_0))$, and for the right hand side,

$$-\frac{1}{n} \left[ \ddot{\underset{\sim}{l}}_n(\underset{\sim}{\theta}_0) + \frac{1}{2} (\widetilde{\theta}_n - \underset{\sim}{\theta}_0)^\mathsf{T} \dddot{\underset{\sim}{l}}_n(\underset{\sim}{\theta}^\star_n) \right] \xrightarrow{p} \mathcal{I}(\underset{\sim}{\theta}_0)$$

by analogy with the univariate case. Hence by Slutsky's Theorem, for large $n$,

$$-\frac{1}{\sqrt{n}} \left[ \ddot{\underset{\sim}{l}}_n(\underset{\sim}{\theta}_0) + \frac{1}{2} (\widetilde{\theta}_n - \underset{\sim}{\theta}_0)^\mathsf{T} \dddot{\underset{\sim}{l}}_n(\underset{\sim}{\theta}^\star_n) \right] (\widetilde{\theta}_n - \underset{\sim}{\theta}_0) = \mathcal{I}(\underset{\sim}{\theta}_0) \sqrt{n} (\widetilde{\theta}_n - \underset{\sim}{\theta}_0) + \mathrm{o}_P(1)$$

where $\mathrm{o}_P(1)$ represents a term that converges in probability to zero. Hence

$$\sqrt{n}(\widetilde{\theta}_n - \underset{\sim}{\theta}_0) \xrightarrow{p} Z \sim \text{Normal}(\underset{\sim}{0}, \mathcal{I}(\underset{\sim}{\theta}_0)^{-1})$$

and the result is proved. ∎