MATH 557 – ASYMPTOTIC THEORY

Suppose that

- data $x_{1:n} = (x_1, \ldots, x_n)$ are realizations of independent and identically distributed (i.i.d.) random variables $X_1, \ldots, X_n$ drawn from distribution with pdf $f_0(x)$. We term this model the *true* model.
- we wish to represent the data using a parametric pdf $f_X(x;\theta_0)$, where θ_0 is k dimensional parameter. We may term this model the *working* model.

We wish to understand how to estimate θ_0, and what happens to the estimator of θ_0 when n becomes large. In a standard analysis, we can use maximum likelihood estimation, and standard asymptotic theory. However, this analysis assume that $f_0(x) \equiv f_X(x;\theta_0)$, that is, the parametric model is *correctly specified*; if $f_0(x) \neq f_X(x;\theta_0)$, the model is *incorrectly specified*, and the theory needs to be reconsidered.

1. **Interpreting θ_0 in the working model:** Recall that we define the 'true' value of θ_0 as

$$\theta_0 = \arg\min_{\theta} KL(f_0, f_X(X;\theta)) \tag{1}$$

Note that

$$KL(f_0, f_X(\theta)) = \int \log f_0(x) f_0(x)\, dx - \int \log f_X(x;\theta) f_0(x)\, dx$$

or equivalently, denoting $\log f_X(x;\theta)$ by $\ell(x;\theta)$,

$$\theta_0 = \arg\max_{\theta} \mathbb{E}_{f_0}\left[\ell(X;\theta)\right]. \tag{2}$$

2. **Maximum likelihood:** We use a random sample $x_1, \ldots, x_n$ and aim to maximize the sample-based expectation (or sample mean) to produce an estimator. Specifically, the estimator based on (2) will be

$$\widehat{\theta}_n = \arg\max_{\theta} \frac{1}{n}\sum_{i=1}^{n} \ell(X_i;\theta).$$

The justification for this is the *weak law of large numbers*; this says that sample means converge in probability to expected values, and here that implies

$$\frac{1}{n}\sum_{i=1}^{n} \ell(X_i;\theta) \xrightarrow{p} \mathbb{E}_{f_0}\left[\ell(X;\theta)\right] \tag{3}$$

as $n \longrightarrow \infty$ for any fixed θ, provided the expectation exists.

We will assume that the log density $\ell(y;\theta)$ is at least three times differentiable with respect to θ; under this assumption, the estimate is defined as the solution to the *score equations*, the system of k equations given by

$$\frac{\partial}{\partial\theta}\left\{\frac{1}{n}\sum_{i=1}^{n} \ell(x_i;\theta)\right\} = \mathbf{0}_k$$

or equivalently,

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial\theta}\left\{\ell(x_i;\theta)\right\} = \frac{1}{n}\sum_{i=1}^{n} U(x_i;\theta) = \mathbf{0}_k \tag{4}$$

say, where $U(x;\theta) = \dot{\ell}(x;\theta) = \partial\ell_1(x;\theta)/\partial\theta$. Denote the solution of (4) by $\widehat{\theta}_n \equiv \widehat{\theta}_n(x_{1:n})$.

3. **Taylor expansion:** We consider a Taylor expansion of the function $\ell(x;\theta)$ with respect to θ around θ_0. We have any value of θ

$$\ell(x;\theta) = \ell(x;\theta_0) + \dot{\ell}(x;\theta_0)(\theta-\theta_0) + \frac{1}{2}(\theta-\theta_0)^\top \ddot{\ell}(x;\theta_0)(\theta-\theta_0) + \mathcal{R}_3(x;\theta^*) \tag{5}$$

where

$$\ddot{\ell}(x;\theta) = \frac{\partial^2 \ell(x;\theta)}{\partial\theta\partial\theta^\top} \qquad (k \times k).$$

and $\mathcal{R}_3(x;\theta^*)$ is a remainder term, for some θ^* such that $\|\theta_0 - \theta^*\| \leqslant \|\theta_0 - \theta\|$. Evaluating (5) for each of $x_1,\ldots,x_n$ and summing the result, we have

$$\ell_n(\theta) = \ell_n(\theta_0) + \dot{\ell}_n(\theta_0)^\top(\theta-\theta_0) + \frac{1}{2}(\theta-\theta_0)^\top \ddot{\ell}_n(\theta_0)(\theta-\theta_0) + \mathcal{R}_3(x_{1:n};\theta^*). \tag{6}$$

Evaluating this expression at $\theta = \widehat{\theta}_n$ and rearranging we have

$$\ell_n(\widehat{\theta}_n) - \ell_n(\theta_0) = \dot{\ell}_n(\theta_0)^\top(\widehat{\theta}_n-\theta_0) + \frac{1}{2}(\widehat{\theta}_n-\theta_0)^\top \ddot{\ell}_n(\theta_0)(\widehat{\theta}_n-\theta_0) + \mathcal{R}_3(x_{1:n};\theta^*) \tag{7}$$

where $\|\theta_0 - \theta^*\| \leqslant \|\theta_0 - \widehat{\theta}_n\|$.

4. **Asymptotic behaviour:** Consider now the previous equation (7) written in terms of random variables, with $\widehat{\theta}_n = \widehat{\theta}_n(X_{1:n})$:

$$\ell_n(\widehat{\theta}_n) - \ell_n(\theta_0) \quad = \quad \dot{\ell}_n(\theta_0)^\top(\widehat{\theta}_n-\theta_0) + \frac{1}{2}(\widehat{\theta}_n-\theta_0)^\top \ddot{\ell}_n(\theta_0)(\widehat{\theta}_n-\theta_0) + \mathcal{R}_3(X_{1:n};\theta^*)$$

First consider the behaviour, for arbitrary θ, of the quantity

$$\frac{1}{n}\left(\ell_n(\theta) - \ell_n(\theta_0)\right) = \frac{1}{n}\sum_{i=1}^{n}\left(\ell(X_i;\theta) - \ell(X_i;\theta_0)\right).$$

We may rewrite this expression with terms involving the true density f_0 that cancel :

$$\frac{1}{n}\sum_{i=1}^{n}\ell(X_i;\theta) - \frac{1}{n}\sum_{i=1}^{n}\ell(X_i;\theta_0) = \frac{1}{n}\sum_{i=1}^{n}(\ell(X_i;\theta) - \ell_0(X_i)) - \frac{1}{n}\sum_{i=1}^{n}(\ell(X_i;\theta_0) - \ell_0(X_i)) \tag{8}$$

where $\ell_0(x) = \log f_0(x)$. For any θ, as $n \longrightarrow \infty$, we have by the weak law of large numbers that

$$\frac{1}{n}\sum_{i=1}^{n}(\ell(X_i;\theta) - \ell_0(X_i)) \xrightarrow{p} \mathbb{E}_{f_0}\left[\log\left(\frac{f_X(X;\theta)}{f_0(X)}\right)\right] = -KL(f_0, f_X(.;\theta))$$

as $X_1,\ldots,X_n \sim f_0$. Therefore

$$\frac{1}{n}\sum_{i=1}^{n}\ell(X_i;\theta) - \frac{1}{n}\sum_{i=1}^{n}\ell(X_i;\theta_0) \xrightarrow{p} KL(f_0, f_X(\theta_0)) - KL(f_0, f_X(.;\theta))$$

By definition of θ_0 via (1), $KL(f_0, f_X(\theta))$ attains its minimum value at $\theta = \theta_0$, so

$$KL(f_0, f_X(.;\theta_0)) - KL(f_0, f_X(.;\theta)) \leqslant 0$$

and the random variable on the left hand side of (8) converges in probability to a non-positive constant. Therefore, we have that

$$\Pr_{f_0}[\ell_n(\theta_0) \geqslant \ell_n(\theta)] \longrightarrow 1 \tag{9}$$

as $n \longrightarrow \infty$. That is, with probability tending to 1, the log likelihood $\ell_n(\theta_0)$ is not less than $\ell_n(\theta)$ for any other θ. If we make an **identifiability** assumption, this statement may be strengthened: the model $f_X(x;\theta)$ is *identifiable* if, for two parameter values $\theta^\dagger = \theta^\ddagger$,

$$f_X(x;\theta^\dagger) = f_X(x;\theta^\ddagger) \text{ for all } x \quad \Longrightarrow \quad \theta^\dagger = \theta^\ddagger.$$

If the model is identifiable, then the "true" value θ_0 is uniquely defined, and we have

$$\Pr_{f_0}[\ell_n(\theta_0) > \ell_n(\theta)] \longrightarrow 1 \qquad \theta \neq \theta_0. \tag{10}$$

The theory holds for fixed θ. However, in equation (7), the first term is $\ell_n(\widehat{\theta}_n(X_{1:n}))$, that is, where the parameter at which the log-likelihood is evaluated is itself a random variable, namely the estimator $\widehat{\theta}_n(X_{1:n})$. To determine the behaviour of $\ell_n(\widehat{\theta}_n(X_{1:n}))$ under identifiability and the differentiability assumption above:

- Fix $a > 0$ and consider the set in the data sample space

  $$B_n(a) \equiv \{x_{1:n} : \|\theta - \theta_0\| = a, \ell_n(\theta) < \ell_n(\theta_0)\}$$

  that is, $B_n(a)$ is the set of all data configurations such that, for all θ lying on the ball of radius a centered at θ_0, the log-likelihood at θ_0 exceeds that at θ.
- With probability tending to one, $\ell_n(\theta) < \ell_n(\theta_0)$ for all θ by (10), so

  $$\Pr_{f_0}(B_n(a)) \longrightarrow 1 \qquad \text{as } n \longrightarrow \infty.$$

- As the log-likelihood is differentiable with respect to θ, $\ell_n(\theta)$ must have a local maximum inside $B_n(a)$; denote the maximizing value by $\widehat{\theta}_a(x_{1:n})$, and note that

  $$\dot{\ell}_n(\widehat{\theta}_a(x_{1:n})) = \mathbf{0}_k$$

  so that the maximizing value is a solution to the usual likelihood estimating equation. This proves the **existence** of a local maximum.

  Note: Strictly, $\widehat{\theta}_a(x_{1:n})$ is not necessarily the mle, as it is only guaranteed to be a local maximum of the likelihood in the neighbourhood of the true θ_0.

- Hence, as $n \longrightarrow \infty$,

  $$\Pr_{f_0}[\|\widehat{\theta}_a(X_{1:n}) - \theta_0\| < a] \longrightarrow 1$$

  so therefore the sequence of estimators $\{\widehat{\theta}_a(X_{1:n}), n \geqslant 1\}$ **converges in probability** to θ_0. This holds for a arbitrarily small.

- For any a, there is at least one local maximum in the neighbourhood of θ_0. Let $\widehat{\theta}_n(x_{1:n})$ be the root of the likelihood equations closest to θ_0; this does not depend on the choice of a.

Therefore $\widehat{\theta}_n(X_{1:n}) \xrightarrow{p} \theta_0$ and $\widehat{\theta}_n(X_{1:n})$ is **consistent** for θ_0, and by "continuous mapping" (as $\ell_n(\theta)$ is a continuous function in θ)

$$\left| \frac{1}{n} \left\{ \ell_n(\widehat{\theta}_n(X_{1:n})) - \ell_n(\theta_0) \right\} \right| \xrightarrow{p} 0$$

so that, from (3), as $n \longrightarrow \infty$

$$\frac{1}{n} \sum_{i=1}^{n} \ell(X_i; \widehat{\theta}_n(X_{1:n})) \xrightarrow{p} \mathbb{E}_{f_0}\left[\ell(Y;\theta_0)\right] \tag{11}$$

5. **Asymptotic Normality:** The next result uses the Mean Value Theorem. For a continuous function such as $\dot{\ell}_n(\theta)$, with defined second derivative $\ddot{\ell}_n(\theta)$, it is guaranteed that there exists an 'intermediate' $\widetilde{\theta} = c\widehat{\theta}_n + (1-c)\theta_0$ for some c, $0 < c < 1$, such that

$$\dot{\ell}_n(\widehat{\theta}_n) = \dot{\ell}_n(\theta_0) + \ddot{\ell}_n(\widetilde{\theta})(\widehat{\theta}_n - \theta_0)$$

The left hand side is zero as $\widehat{\theta}_n$ is the mle. Provided $\ddot{\ell}_n(\widetilde{\theta})$ is non-singular, we may write after rescaling and rearrangement that

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) = \left\{-\frac{1}{n}\ddot{\ell}_n(\widetilde{\theta})\right\}^{-1}\left\{\sqrt{n}\left(\frac{1}{n}\dot{\ell}_n(\theta_0)\right)\right\} \tag{12}$$

- In its random variable form, second term on the right hand side of (12) is

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} U(X_i;\theta_0)\right)$$

that is, a sample average quantity scaled by $\sqrt{n}$. But by definition of θ_0,

$$\mathbb{E}_{f_0}[U(X_i;\theta_0)] = \int \dot{\ell}(y;\theta_0) f_0(y)\, dy = \mathbf{0}_k$$

as, by definition θ_0 minimizes $KL(f_0, f_X(X;\theta))$, and therefore must be a solution of this equation. Therefore, by the Central Limit Theorem

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=1}^{n} U(X_i;\theta_0)\right) \stackrel{d}{\longrightarrow} \text{Normal}_k(\mathbf{0}_k, \mathcal{J}_{f_0}(\theta_0)) \tag{13}$$

where

$$\mathcal{J}_{f_0}(\theta_0) = \mathbb{E}_{f_0}[U(X;\theta_0)U(X;\theta_0)^\top] \equiv \text{Var}_{f_0}[U(X;\theta_0)] \qquad (k \times k \times k).$$

- For the first term on the right hand side of (12), as $\widehat{\theta}_n \stackrel{p}{\longrightarrow} \theta_0$, we have that

$$-\frac{1}{n}\ddot{\ell}_n(\widetilde{\theta}) \stackrel{a.s.}{\longrightarrow} \mathcal{I}_{f_0}(\theta_0).$$

Therefore we write for an asymptotic approximation to (12)

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) = \left\{-\frac{1}{n}\ddot{\ell}_n(\theta_0)\right\}\left\{\frac{1}{\sqrt{n}}\dot{\ell}_n(\theta_0)\right\} + \text{o}_p(1)$$

where the distribution of the second term given by (13), and where $\text{o}_p(1)$ denotes a term that converges in probability to zero. We therefore have that

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \stackrel{d}{\longrightarrow} \text{Normal}_k(\mathbf{0}_k, \{\mathcal{I}_{f_0}(\theta_0)\}^{-1}\mathcal{J}_{f_0}(\theta_0)\{\mathcal{I}_{f_0}(\theta_0)\}^{-1}).$$

6. **Correct specification:** Under correct specification, $f_0(x) \equiv f_X(x;\theta_0)$, and we have from earlier results that

$$\mathcal{J}_{\theta_0}(\theta_0) = \mathcal{I}_{\theta_0}(\theta_0)$$

and hence from the general result we deduce that

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) \stackrel{d}{\longrightarrow} \text{Normal}_k(\mathbf{0}_k, \{\mathcal{I}_{\theta_0}(\theta_0)\}^{-1}).$$