Point Estimation

- Suppose that we have a random sample $X_1, \dots, X_n$ from a distribution with pdf or pmf $f(x \mid \boldsymbol{\theta})$ for some unknown parameter vector $\boldsymbol{\theta} = (\theta_1, \dots, \theta_p)^t$.

- One of the primary issues in statistical inference involves estimation of $\boldsymbol{\theta}$ based on the observed sample $x_1, \dots, x_n$.

- In this section of the course we shall examine point estimation of $\boldsymbol{\theta}$ based on a sample $\boldsymbol{x}$.

- We shall concentrate here on standard (Frequentist) methods and return to consider Bayesian methods later.

Point Estimators

Definition 2.1

A point estimator of a parameter $\boldsymbol{\theta}$ *is a function* $W(X_1, \ldots, X_n)$ *which is used to estimate the unknown parameter* $\boldsymbol{\theta}$*.*

- $*$ Note that a point estimator is a random variable.
- $*$ For a given observed sample $x_1, \ldots, x_n$, the observed value $W(x_1, \ldots, x_n)$ is called a point estimate.
- $*$ Note that the definition is very broad in that any statistic can be used to estimate the parameter but some estimators will have better properties than others.

Sampling Distributions

- Since a point estimator is a random variable it will have a probability distribution.

- The probability distribution associated with a point estimator is called its sampling distribution.

- The (frequentist) interpretation of the sampling distribution of an estimator is that it describes how the value of the estimator varies over an infinite number of repeated samples of the same size n, taken from the same population in the same manner.

- By examining the sampling distribution of estimators we can see their properties over repeated samples and compare them.

Bias of Estimators

Definition 2.2

The bias of a point estimator W of a parameter θ is defined to be

$$\textit{Bias}_{\theta}(W) \;=\; \mathrm{E}_{\theta}(W) - \theta$$

- ∗ The expectation in the above definition is taken relative to the sampling distribution of W when the true parameter value is θ.

- ∗ In general the bias will depend on θ.

Definition 2.3

A point estimator W of a parameter θ is said to be an unbiased estimator if, and only if,

$$\mathrm{E}_{\theta}(W) \;=\; \theta \quad \textit{for all possible } \theta$$

Variability of Estimators

- ∗ As well as the expected value of an estimator we are interested in its variability.

- ∗ The sampling variability of an estimator is simply $\text{Var}_\theta(W)$ where the expectations are taken relative to the sampling distribution of W when the true parameter value is θ.

- ∗ As for the bias, this will generally be a function of θ.

- ∗ If W_1 and W_2 are two unbiased estimators of the same parameter θ and

$$\text{Var}_\theta(W_1) \leqslant \text{Var}_\theta(W_2) \quad \text{for all possible } \theta$$

then we would generally prefer to use W_1 as our estimator since it varies less over repeated samples.

Mean Squared Error

Definition 2.4

Suppose that $W = W(\boldsymbol{X})$ is a point estimator of a scalar parameter θ. The **Mean Squared Error (MSE)** *of the estimator is*

$$\mathsf{MSE}(W,\theta) = \mathsf{E}_\theta\left((W-\theta)^2\right).$$

- ∗ $\mathsf{MSE}(W,\theta) = \mathsf{Var}_\theta(W) + \left(\mathsf{Bias}_\theta(W)\right)^2$.
- ∗ If W is unbiased then $\mathsf{MSE}(W,\theta) = \mathsf{Var}_\theta(W)$.
- ∗ Since the MSE takes into account both bias and variance it is often used to compare sets of estimators, not all of which are unbiased.

Method of Moments Estimation

- Probably the oldest point estimation method.

- It relies on equating population quantities to corresponding sample quantities which is intuitively very appealing.

- A very simple example is estimation of a population mean μ by a sample mean $\overline{X}$.

- Although this seems very reasonable, estimators produced using this method often do not have very good properties.

- They are, however, usually easy to find.

Moments

Definition 2.5

Suppose that a random variable X has probability density or mass function $f(x \mid \boldsymbol{\theta})$ then the (non-central) moments of X are defined as

$$\mu_r' = \mathsf{E}_{\boldsymbol{\theta}}(X^r) = \begin{cases} \displaystyle\int_{-\infty}^{\infty} x^r f(x \mid \boldsymbol{\theta})\, dx & \textit{for } X \textit{ continuous} \\ \displaystyle\sum_x x^r f(x \mid \boldsymbol{\theta}) & \textit{for } X \textit{ discrete} \end{cases}$$

Definition 2.6

Let $X_1, \dots, X_n$ be a random sample then the **sample moments** *are defined as*

$$m_r' = \frac{1}{n} \sum_{i=1}^{n} X_i^r$$

Method of Moments Estimation

Definition 2.7

Suppose that $X_1, \ldots, X_n$ is a random sample from a distribution with parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$. Let μ_r' be the moments of X and assume that μ_p' exists and is finite. Let m_r' be the sample moments of $X_1, \ldots, X_n$.

The **method of moments estimator** *of $\boldsymbol{\theta}$ is given by the solution to the p simultaneous equations*

$$m_r' = \mu_r'(\theta_1, \ldots, \theta_p) \qquad r = 1, \ldots, p$$

Central Method of Moments Estimation

- In some cases, better estimators can be obtained by matching the central population moments about the mean $\mu = \mathrm{E}_{\boldsymbol{\theta}}(X)$

$$\mu_r \;=\; \mathrm{E}_{\boldsymbol{\theta}}\left((X-\mu)^r\right)$$

  to the central sample moments

$$m_r \;=\; \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^r$$

- One instance where this is true is in situations where the population mean does not depend on $\boldsymbol{\theta}$.

- In that case we would not use the first moments and use subsequent central moments which adjust for the discrepancy between the sample mean and the true population mean.

Best Unbiased Estimators

- Method of Moments Estimators are often biased.

- There may be many possible unbiased estimators so we would like to find a *"best"* estimator in some sense.

- If we have two unbiased estimators $\widehat{\theta}_1$ and $\widehat{\theta}_2$ of an unknown θ such that

$$\mathrm{Var}_\theta(\widehat{\theta}_1) \leqslant \mathrm{Var}_\theta(\widehat{\theta}_2) \quad \text{for every possible } \theta$$

then we would prefer $\widehat{\theta}_1$.

Best Unbiased Estimators

Definition 2.8

An estimator T *is a* **best unbiased estimator** *of a scalar parameter* $\tau(\theta)$ *if it satisfies* $\mathsf{E}_\theta(T) = \tau(\theta)$ *for every* θ *and*

$$\mathsf{Var}_\theta(T) \leqslant \mathsf{Var}_\theta(T^*) \quad \textit{for every } \theta.$$

for any other estimator T^* *such that* $\mathsf{E}_\theta(T^*) = \tau(\theta)$ *for all* θ.

T *is also called a* **Minimum Variance Unbiased Estimator (MVUE)** *of* $\tau(\theta)$.

Challenges in Finding an MVUE

1. An MVUE may not exist!

2. In order to prove that our estimator is an MVUE we need to know the variance of other unbiased estimators.

3. For some estimators this may be a complex calculation and we need to do it for every possible unbiased estimator.

4. Under some conditions, however, we can make progress without needing to do all of these calculations.

Some Regularity Conditions

Definition 2.9

Suppose that $f(\boldsymbol{x} \mid \theta)$ is the joint density function of a random vector $\boldsymbol{X}$, f is said to be a regular density if it satisfies the conditions

1. *The support $\mathcal{X} = \{\boldsymbol{x} : f(\boldsymbol{x} \mid \theta) > 0\}$ does not depend on θ.*
2. *The set, Θ of possible values of θ is an open subset of $\mathbb{R}$.*
3. *The joint density function is such that*

$$\frac{\partial}{\partial \theta} \log f(\boldsymbol{x} \mid \theta)$$

   exists and is finite for all θ.

4. *For any scalar function $h(\boldsymbol{X})$*

$$\frac{\partial}{\partial \theta} \int h(\boldsymbol{x}) f(\boldsymbol{x} \mid \theta) d\boldsymbol{x} = \int h(\boldsymbol{x}) \frac{\partial}{\partial \theta} f(\boldsymbol{x} \mid \theta) dx.$$

The Cramér–Rao Inequality

Theorem 2.1

Let $\boldsymbol{X} = (X_1, \ldots, X_n)$ *be a random vector with probability density function* $f(\boldsymbol{x} \mid \theta)$ *which satisfies the regularity conditions described in Definition 2.9.*

Suppose that $T(\boldsymbol{X})$ *is such that* $\mathsf{Var}_\theta(T) < \infty$ *and* $\mathsf{E}_\theta(T) = \tau(\theta)$ *where* $\tau(\theta)$ *is a differentiable function of* θ *then*

$$\mathsf{Var}_\theta(T) \geqslant \frac{\left(\tau'(\theta)\right)^2}{\mathsf{E}_\theta\left[\left(\frac{\partial}{\partial\theta} \ln f(\boldsymbol{X} \mid \theta)\right)^2\right]} = R_\tau(\theta).$$

Utility of the Cramér–Rao Inequality

* Suppose we have an estimator T which is unbiased for $\tau(\theta)$ and we can find $\mathsf{Var}_\theta(T)$.

* Then if

$$\mathsf{Var}_\theta(T) \;=\; \frac{(\tau'(\theta))^2}{\mathsf{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\ln f(\boldsymbol{X} \mid \theta)\right)^2\right]} = R_\tau(\theta).$$

  we know that it must be a minimum variance unbiased estimator without needing to consider any other estimators at all!

* It can be shown that many best unbiased estimators do actually achieve this bound and so can be found in this way.

* Unfortunately, in some situations, even the the best unbiased estimator does not achieve the bound.

The Cramér–Rao Inequality for Random Samples

Corollary 2.1.1

Suppose that $X_1, \ldots, X_n$ is a random sample from a population with probability density $f(x \mid \theta)$ which is such that the regularity conditions of Definition 2.9 hold. Let $T(\boldsymbol{X})$ be such that $\mathsf{Var}_\theta(T) < \infty$ *and* $\mathsf{E}_\theta(T) = \tau(\theta)$ *where $\tau(\theta)$ is a differentiable function of θ then*

$$\mathsf{Var}_\theta(T) \geqslant \frac{\left(\tau'(\theta)\right)^2}{n\,\mathsf{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\ln f(X \mid \theta)\right)^2\right]} = R_\tau(\theta).$$

The Fisher Information

Definition 2.10

Let $\boldsymbol{X} = (X_1, \dots, X_n)^t$ be a random vector with probability density function $f(\boldsymbol{x} \mid \theta)$. The quantity

$$I(\theta) = \mathrm{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\ln f(\boldsymbol{X} \mid \theta)\right)^2\right]$$

is the **(Expected) Fisher Information**.

Lemma 2.1

If $f(\boldsymbol{x} \mid \theta)$ satisfies

$$\frac{d}{d\theta}\int \frac{\partial \ln f(\boldsymbol{x} \mid \theta)}{\partial\theta} f(\boldsymbol{x} \mid \theta) d\boldsymbol{x} = \int \frac{\partial}{\partial\theta}\left[\frac{\partial \ln f(\boldsymbol{x} \mid \theta)}{\partial\theta} f(\boldsymbol{x} \mid \theta)\right] d\boldsymbol{x}$$

Then

$$I(\theta) = \mathrm{E}_\theta\left[\left(\frac{\partial}{\partial\theta}\ln f(\boldsymbol{x} \mid \theta)\right)^2\right] = -\,\mathrm{E}_\theta\left[\frac{\partial^2}{\partial\theta^2}\ln f(\boldsymbol{x} \mid \theta)\right]$$

Efficient Estimators

Definition 2.11

Suppose that $X_1, \ldots, X_n$ is a random sample satisfying the conditions of Theorem 2.1 and T is an unbiased estimator of $\tau(\theta)$. T is said to be an **efficient estimator** *of $\tau(\theta)$ if, and only if, its variance achieves the Cramér–Rao lower bound, $R_\tau(\theta)$.*

If $\text{Var}_\theta(T) > R_\tau(\theta)$ *then the* **efficiency** *of the estimator T is the ratio*

$$\text{eff}(T) \; = \frac{R_\tau(\theta)}{\text{Var}_\theta(T)}.$$

The Rao–Blackwell Theorem

Theorem 2.2

Suppose that $X_1, \ldots, X_n$ is a random sample and that $T(\boldsymbol{X})$ is an unbiased estimator of $\tau(\theta)$. Let $S(\boldsymbol{X})$ be a sufficient statistic for θ and define the estimator

$$T^*(\boldsymbol{X}) = \mathsf{E}[T(\boldsymbol{X}) \mid S(\boldsymbol{X})]$$

Then T^* is an unbiased estimator of $\tau(\theta)$ whose variance is uniformly (in θ) no larger than that of T

The Rao–Blackwell Theorem

- The Rao-Blackwell Theorem says that we will have reduced variance of an unbiased estimator when that estimator is a function of a sufficient statistic.

- There may be many possible unbiased estimators that are functions of a sufficient statistic, however.

- The best estimator, however, is unique.

Theorem 2.3

If T is a best unbiased estimator of $\tau(\theta)$ then T is unique.

A Characterisation of Best Unbiased Estimators

Theorem 2.4

Suppose that T is an unbiased estimator of $\tau(\theta)$. T is the minimum variance unbiased estimator of $\tau(\theta)$ if, and only if, for any U satisfying $E_\theta(U) = 0$ for all θ (U is called an unbiased estimator of 0) we have

$$\mathrm{Cov}_\theta(T, U) \;=\; 0 \quad \textit{for every } \theta.$$

A Characterisation of Best Unbiased Estimators

- Use of this theorem means that we need to be able to characterize every unbiased estimator of 0 and make sure that our unbiased estimator of $\tau(\theta)$ is uncorrelated with all of them.

- If no unbiased estimators of 0 exist then it is clear that T must be the best unbiased estimator.

- The non-existence of unbiased estimators of 0 relates to the concept of completeness.

Completeness

Definition 2.12

Suppose T is a statistic with pdf or pmf $f(t \mid \theta)$ indexed by an unknown parameter θ. The family of probability distributions characterized by $f(t \mid \theta)$ is said to be complete if

$$\mathsf{E}_\theta\Big(g(T)\Big) = 0 \;\Rightarrow\; \mathsf{P}_\theta\Big(g(T) = 0\Big) = 1$$

∗ This definition says that, for a complete family, the only unbiased estimator of 0 is 0 itself!

∗ If the family of distributions for the statistic T is complete then we will refer to T as a complete statistic.

The Lehmann–Scheffé Theorem

Theorem 2.5

Let $T(\boldsymbol{X})$ *be an unbiased estimator of* $\tau(\theta)$ *and let* $S(\boldsymbol{X})$ *be a complete sufficient statistic for* θ*. Then the estimator* $T^*(X) =$ $\mathsf{E}(T \mid S)$ *is the unique minimum variance unbiased estimator of* $\tau(\theta)$*.*

Criticisms of Unbiased Estimation

1. It may be possible to find another estimator which has smaller MSE than the best unbiased estimator for every value of the parameter.

2. It may be impossible to find any unbiased estimator for a parameter.

3. Unbiased estimation is not invariant to changes in parametrization.

4. The idea of unbiasedness is an artificial and irrelevant property in finite samples. Asymptotic unbiasedness and consistency is what is really important.

Maximum Likelihood Estimation

Definition 2.13

Suppose $\boldsymbol{x} = (x_1, \ldots, x_n)^t$ *are the observed values of* n iid *random variables from a family of distributions indexed by the unknown parameter vector* $\boldsymbol{\theta} \in \Theta$ *and let the likelihood be* $L(\boldsymbol{\theta} \mid \boldsymbol{x})$. *Then the* **maximum likelihood estimate** *of* $\boldsymbol{\theta}$ *is a value* $\widehat{\boldsymbol{\theta}}(\boldsymbol{x})$ *such that*

$$L\left(\widehat{\boldsymbol{\theta}}(\boldsymbol{x}) \mid \boldsymbol{x}\right) \geqslant L(\boldsymbol{\theta}; \boldsymbol{x}) \quad \forall\ \boldsymbol{\theta} \in \Theta$$

The **maximum likelihood estimator** *of* $\boldsymbol{\theta}$ *is the random variable* $\widehat{\boldsymbol{\theta}}(\boldsymbol{X})$.

Maximum Likelihood Estimation

∗ If the likelihood $L(\boldsymbol{\theta} \mid \boldsymbol{x})$ is differentiable in $\boldsymbol{\theta}$, then an interior maximum of L can be given by solving the p equations

$$\frac{\partial}{\partial \theta_i} L(\theta_1, \ldots, \theta_p; \boldsymbol{x}) = 0 \quad i = 1, \ldots, p.$$

∗ The maximum value found this way is **not** necessarily the mle.

∗ Usually easier to maximize the *log-likelihood*

$$l(\boldsymbol{\theta} \mid \boldsymbol{x}) = \log L(\boldsymbol{\theta} \mid \boldsymbol{x})$$

∗ In the *iid* case

$$l(\boldsymbol{\theta} \mid \boldsymbol{x}) = \sum_{i=1}^{n} \log f_X(x_i \mid \boldsymbol{\theta})$$

Properties of Maximum Likelihood Estimators

Theorem 2.6

If $\widehat{\theta}$ is the maximum likelihood estimate of a parameter θ and $\eta = g(\theta)$ is a transformation of the parameter then the maximum likelihood estimate of η is

$$\widehat{\eta} = g(\widehat{\theta}).$$

Theorem 2.7

Let $X_1, \ldots, X_n$ be a sample form a population with density $f(x \mid \theta)$ and let $T(\boldsymbol{X})$ be the minimal sufficient statistic for θ. Then the maximum likelihood estimator of θ depends on the sample only through the value of $T(\boldsymbol{X})$.

Computation of the MLE

- It is not always possible to write down a simple formula for the maximum likelihood estimator.

- The derivative of the log-likelihood function is very often non-linear in the parameter.

- There can be multiple roots to the likelihood equations.

- Missing data can make the log-likelihood very complicated and hard to maximize.

- In many of these cases, however, we can numerically maximize the log-likelihood for our given data and hence find the value of the maximum likelihood estimate.

Newton–Raphson Method

- The mle is a solution to

$$U(\widehat{\theta}) = \left.\frac{\partial l(\theta; \boldsymbol{x})}{\partial \theta}\right|_{\theta=\widehat{\theta}} = 0$$

- Observed information matrix

$$J(\widehat{\theta}) = -\left.\frac{\partial l(\theta; \boldsymbol{x})}{\partial \theta \partial \theta^T}\right|_{\theta=\widehat{\theta}}$$

- From Taylor's Theorem

$$U(\widehat{\theta}) \approx U(\theta^*) - J(\theta^*)(\widehat{\theta} - \theta^*)$$

- Hence

$$\widehat{\theta} \approx \theta^* + J^{-1}(\theta^*)U(\theta^*)$$

The Newton–Raphson Method

1. Choose a reasonable starting value $\widehat{\theta}^{(0)}$.

2. Update your estimate

$$\widehat{\theta}^{(k+1)} = \widehat{\theta}^{(k)} + J^{-1}(\widehat{\theta}^{(k)})U(\widehat{\theta}^{(k)})$$

3. Terminate the algorithm when

$$\left\| \widehat{\theta}^{(k)} - \widehat{\theta}^{(k+1)} \right\| < \varepsilon.$$

for some pre-determined tolerance $\varepsilon > 0$.

Fisher Scoring Method

Replace $J(\widehat{\theta}^{(k)})$ with the expected Fisher information matrix evaluated at the current step

$$I(\widehat{\theta}^{(k)}) = E_\theta \left[-\frac{\partial^2 l(\theta; \boldsymbol{x})}{\partial\theta\partial\theta^T} \right] \Bigg|_{\theta=\widehat{\theta}^{(k)}}$$

- Choice of starting value is very important.
- It is possible that there are multiple solutions to the likelihood equation including minima and saddlepoints.
- It is often useful to start from multiple starting points to ensure that a global maximum is found.

The EM Algorithm

- Useful when there is missing data.

- $\boldsymbol{y}$ is the observed (incomplete) data with log likelihood $l(\theta; \boldsymbol{y})$.

- $\boldsymbol{x}$ is the complete data with log likelihood $l_c(\theta; \boldsymbol{x})$.

- Define the quantity

$$Q(\theta, \theta^*) = \mathsf{E}_{\theta^*}\left[l_c(\theta; \boldsymbol{X}) \mid \boldsymbol{y}\right]$$

- Usually easier to maximize $Q(\theta, \theta^*)$.

The EM Algorithm

1. Define the complete data $\boldsymbol{X}$ and its log likelihood $l_c(\theta; \boldsymbol{X})$

2. Choose an initial estimate $\widehat{\theta}^{(0)}$.

3. **E Step:** Calculate

$$Q(\theta, \widehat{\theta}^{(k)}) = \mathsf{E}_{\widehat{\theta}^{(k)}} \left[l_c(\theta; \boldsymbol{X}) \mid \boldsymbol{y} \right]$$

4. **M Step:** Choose $\widehat{\theta}^{(k+1)}$ to maximize $Q(\theta, \widehat{\theta}^{(k)})$.

5. Iterate the E and M steps until

$$L(\widehat{\theta}^{(k+1)}) - L(\widehat{\theta}^{(k)}) < \varepsilon.$$

EM Algorithm in the Full Exponential Family

∗ Suppose that the density of the complete data $\boldsymbol{X}$ is

$$f_{\boldsymbol{X}}(\boldsymbol{X};\theta) = h(\boldsymbol{x})c(\theta)\,\mathsf{exp}\left\{\sum_{i=1}^{d}\theta_i t_i(\boldsymbol{x})\right\}.$$

∗ Then we can write

$$Q(\theta,\widehat{\theta}^{(k)}) = \log c(\theta) + \sum_{i=1}^{d}\theta_i\,\mathsf{E}_{\widehat{\theta}^{(k)}}\left[t_i(\boldsymbol{X}) \mid \boldsymbol{y}\right].$$

∗ Hence $\widehat{\theta}^{(k+1)}$ satisfies

$$\begin{aligned}\mathsf{E}_{\widehat{\theta}^{(k)}}\left[\boldsymbol{t}(\boldsymbol{X}) \mid \boldsymbol{y}\right] &= -\left.\frac{\partial \log(c(\theta))}{\partial\theta}\right|_{\theta=\widehat{\theta}^{(k+1)}} \\ &= \mathsf{E}_{\widehat{\theta}^{(k+1)}}\left[\boldsymbol{t}(\boldsymbol{X})\right].\end{aligned}$$

Monotonicity of the EM algorithm

Theorem 2.8

If $\widehat{\theta}^{(k)}$ is a sequence of iterates such that

$$Q(\widehat{\theta}^{(k+1)}, \widehat{\theta}^{(k)}) \geqslant Q(\widehat{\theta}^{(k)}, \widehat{\theta}^{(k)})$$

Then the incomplete data likelihood is monotone increasing

$$L(\widehat{\theta}^{(k+1)}; \boldsymbol{y}) \geqslant L(\widehat{\theta}^{(k)}; \boldsymbol{y}).$$

Proof relies on Jensen's Inequality

Lemma 2.2 (Jensen's Inequality)

If X is a random variable with finite mean and g is a concave function such that $\mathsf{E}[|g(X)|] < \infty$ then

$$\mathsf{E}[g(X)] \leqslant g(\mathsf{E}[X]).$$

EM Gradient Algorithm

- $*$ Replace maximization with a single Newton–Raphson step.

$$\widehat{\theta}^{(k+1)} = \widehat{\theta}^{(k)} - \left[\frac{\partial^2 Q(\theta, \widehat{\theta}^{(k)})}{\partial\theta\partial\theta^T}\right]^{-1}_{\theta=\widehat{\theta}^{(k)}} \left[\frac{\partial Q(\theta, \widehat{\theta}^{(k)})}{\partial\theta}\right]_{\theta=\widehat{\theta}^{(k)}}.$$

- $*$ Not guaranteed to be monotone

- $*$ Alternative is to let $0 < a^{(k)} \leqslant 1$ and let

$$\widehat{\theta}^{(k+1)} = \widehat{\theta}^{(k)} - a^{(k)} \left[\frac{\partial^2 Q(\theta, \widehat{\theta}^{(k)})}{\partial\theta\partial\theta^T}\right]^{-1}_{\theta=\widehat{\theta}^{(k)}} \left[\frac{\partial Q(\theta, \widehat{\theta}^{(k)})}{\partial\theta}\right]_{\theta=\widehat{\theta}^{(k)}}.$$

where $a^{(k)}$ is chosen to ensure that

$$Q(\widehat{\theta}^{(k+1)}, \widehat{\theta}^{(k)}) \geqslant Q(\widehat{\theta}^{(k)}, \widehat{\theta}^{(k)})$$

Monte Carlo EM Algorithm

- ∗ Cannot calculate $Q(\theta, \widehat{\theta}^{(k)})$ in closed form.
- ∗ Suppose that we can sample $\boldsymbol{x}_1^{(k)}, \ldots, \boldsymbol{x}_R^{(k)}$ from the conditional distribution $f_{\boldsymbol{X}|\boldsymbol{Y}}(\cdot \mid \boldsymbol{y}; \widehat{\theta}^{(k)})$.
- ∗ Define

$$\widehat{Q}(\theta, \widehat{\theta}^{(k)}) = \frac{1}{R} \sum_{j=1}^{R} l_c(\theta; \boldsymbol{x}_j^{(k)})$$

- ∗ Now choose $\widehat{\theta}^{(k+1)}$ to maximize $\widehat{Q}(\theta, \widehat{\theta}^{(k)})$.

Monte Carlo EM Algorithm

- ∗ MCEM Algorithm will follow the EM "path" with random noise.

- ∗ Assessment of convergence is harder since convergence behaves like a monotone function with random noise.

- ∗ Variability of the random noise will depend on the value of R.

- ∗ It is common to take R small for the first steps and then increase R for later iterations.