7. Point Estimation

Method of Moments Estimation

Definition 7.1

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

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

- The method of moments estimators arise from equating the sample moments m_r with the population moments $\mu_r' = \mathsf{E}[X^r]$.

- In most situations we can write the population moments in terms of the parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$.

Definition 7.2

Suppose that $X_1, \ldots, X_n$ is a random sample from a distribution with parameter vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$. Let μ'_r be the moments of X and assume that μ'_k 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 k simultaneous equations*

$$m_r = \mu'_r(\theta_1, \ldots, \theta_k) \qquad r = 1, \ldots, k$$

Maximum Likelihood Estimation

Definition 7.3

Suppose $\boldsymbol{x} = (x_1, \ldots, x_n)$ *are the observed values of* n iid *random variables from a family of distributions indexed by the unknown, possibly vector, parameter* $\boldsymbol{\theta} \in \Theta$ *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*

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

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

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

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

- 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

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

Theorem 7.1

If $\hat{\theta}$ 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 $\boldsymbol{\eta}$ is

$$\hat{\boldsymbol{\eta}} = g(\hat{\boldsymbol{\theta}}).$$

Theorem 7.2

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

Newton–Raphson Method

- The mle is a solution to

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

- Observed information matrix

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

- From Taylor's Theorem

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

- Hence

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

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

2. Update your estimate

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

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

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

- 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 $l(\theta \mid \boldsymbol{y})$.
- $\boldsymbol{x}$ is the complete data with log likelihood $l_c(\theta \mid \boldsymbol{x})$.
- Define the quantity

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

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

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

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

3. **E Step:** Calculate

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

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

5. Iterate the E and M steps until

$$L\left(\widehat{\theta}^{(k+1)} \mid \boldsymbol{y}\right) - L\left(\widehat{\theta}^{(k)} \mid \boldsymbol{y}\right) < \varepsilon.$$

Theorem 7.3

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

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

Then the incomplete data likelihood is monotone increasing

$$L\left(\theta^{(k)} \mid \boldsymbol{y}\right) \geqslant L\left(\widehat{\theta}^{(k)} \mid \boldsymbol{y}\right).$$

Proof relies on 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 Algorithm in the Full Exponential Family

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

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

- Then we can write

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

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

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

EM Gradient Algorithm

- Do not maximize Q but take one step of a Newton–Raphson algorithm.

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

- Not guaranteed to be monotone

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

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

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

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

Monte Carlo EM Algorithm

- Cannot calculate $Q\left(\theta, \widehat{\theta}^{(k)}\right)$ in closed form.

- Suppose that we can sample $\boldsymbol{x}_1^{(k)}, \ldots, \boldsymbol{x}_R^{(k)}$ from the conditional distribution $f_{\boldsymbol{X}|\boldsymbol{Y}}\left(\boldsymbol{x} \mid \boldsymbol{y}, \widehat{\theta}^{(k)}\right)$.

- Define

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

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

- 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.

Bayes Estimation

- In Bayesian statistics, the parameter θ is considered a random variable.

- The marginal distribution for θ is the probability distribution before any data is collected and so is called the prior distribution for θ and is commonly denoted $\pi(\theta)$.

- The model is now considered to be a conditional distribution of the data for a given value of the parameter.

- Hence the joint distribution of the data and parameter is the product of the likelihood and the prior.

- The conditional distribution of the parameter given the observed data is then

$$\pi(\theta \mid \boldsymbol{x}) = \frac{f(\boldsymbol{x} \mid \theta)\pi(\theta)}{\int f(\boldsymbol{x} \mid \theta)\pi(\theta)\, d\theta}$$

- This is called the posterior distribution for θ.

- The posterior distribution is then used to make inference about the random quantity θ.

- One obvious point estimator for θ is the mean of the posterior distribution

$$\widehat{\theta} = \mathsf{E}[\theta \mid \boldsymbol{x}] = \frac{\int \theta f(\boldsymbol{x} \mid \theta)\pi(\theta)\, d\theta}{\int f(\boldsymbol{x} \mid \theta)\pi(\theta)\, d\theta}$$

Conjugate Priors

Definition 7.4

Given a family $\mathcal{F}$ of pdf's (or pmf's) $f(x \mid \theta)$ indexed by a parameter θ, then a family, $\prod$ of prior distributions is said to be **conjugate** *for the family $\mathcal{F}$ if the posterior distribution of θ is in the family $\prod$ for all $f \in \mathcal{F}$, all priors $\pi(\theta) \in \prod$ and all possible data sets $\boldsymbol{x}$.*

- Conjugate families often make the mathematics of Bayesian statistics easier.

- They may not adequately describe the prior knowledge about the parameter.

Non-informative Priors

- How do we express prior ignorance about a parameter?

- If the set of possible values of θ is a finite interval then we may use a uniform distribution.

- Most parameter spaces, however are infinite.

Definition 7.5

Suppose that θ is a parameter with prior distribution $\pi(\theta)$. The prior distribution is called improper if

$$\int \pi(\theta)\, d\theta = \infty.$$

Definition 7.6

Suppose that $\boldsymbol{X} \sim f_{\boldsymbol{X}}(\boldsymbol{x} \mid \boldsymbol{\theta})$. *The Fisher Information Matrix is*

$$I(\boldsymbol{\theta}) = -\mathsf{E}\left[\frac{\partial^2}{\partial\theta\partial\theta^T} \log f_{\boldsymbol{X}}(\boldsymbol{x} \mid \boldsymbol{\theta})\right]$$

The **Jeffrey's Prior** *for* $\boldsymbol{\theta}$ *is*

$$\pi(\boldsymbol{\theta}) \propto \sqrt{|I(\boldsymbol{\theta})|}.$$

- Attempt to provide a general noninformative prior.
- The Jeffrey's Prior is often improper
- Usually, however,

$$\int f_{\boldsymbol{X}}(\boldsymbol{x} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta} < \infty$$