Chapter 2

Methods of Estimation

2.1 The plug-in principles

Framework: $\mathbf{X} \sim P \in \mathcal{P}$, usually $\mathcal{P} = \{\mathbf{P}_\theta : \theta \in \Theta\}$ for parametric models.

More specifically, if $X_1, \cdots, X_n \sim i.i.d. P_\theta$, then $\mathbf{P}_\theta = P_\theta \times \cdots \times P_\theta$.

Unknown parameters: A certain aspects of population. $\nu(P)$ or $q(\theta) = \nu(P_\theta)$.

Empirical Dist.: $\widehat{P}[X \in A] = \frac{1}{n}\sum_{i=1}^n I(X_i \in A)$ or $\widehat{F}(x) = \frac{1}{n}\sum_{i=1}^n I(X_i \leqslant x)$

Substitution principle: Estimate $\nu(P)$ by $\nu(\widehat{P})$.

Note: As to be seen later, most methods of estimation can be regarded as using

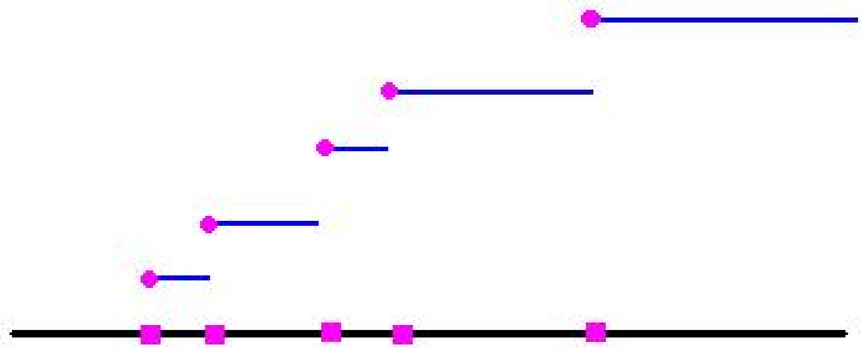


Figure 2.1: Empirical distribution of observed data

"substitution principle", since the functional form ν is not unique.

Example 1. Suppose that $X_1, \cdots, X_n \sim N(\mu, \sigma^2)$. Then

$$\mu = EX = \int x\, dF(x) = \mu(F) \qquad \text{and} \qquad \sigma^2 = \int x^2\, dF(x) - \mu^2.$$

Hence,

$$\widehat{\mu} = \mu(\widehat{F}) = \int x\, d\widehat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

and

$$\widehat{\sigma}^2 = \int x^2 \, d\widehat{F}(x) - \widehat{\mu}^2 = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2.$$

This is also a non-parametric estimator, as the normality assumption has not been explicitly used.

Example 2. Let $X_1, \cdots, X_n$ be a random sample from the following box:

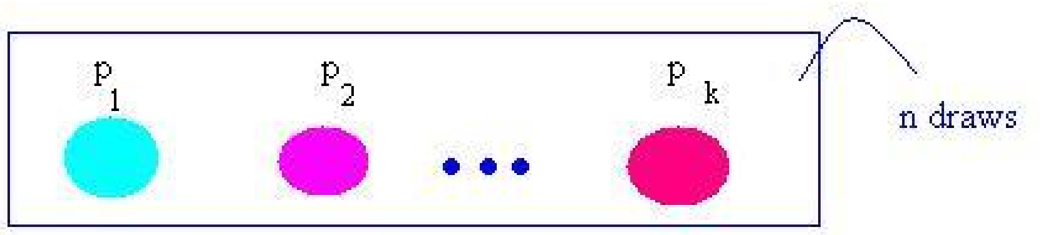


Figure 2.2: Illustration of multinomial distribution

Interested in parameters: $p_1, \cdots, p_k$ and $q(p_1, \cdots, p_k)$.

e.g. dividing the job in the population by 5 categories, interested in p_5 and $(p_4 + p_5 - p_1 - p_2)$.

The empirical distribution:

$$p_j = P(X = j) = F(j) - F(j-) \equiv P_j(F)$$

Hence,

$$\widehat{p}_j = P_j(\widehat{F}) = \widehat{F}(j) - \widehat{F}(j-) = \frac{1}{n}\sum_{i=1}^{n} I(X_i = j),$$

namely, the empirical frequency of getting j. Hence,

$$q(p_1, \cdots, p_k) = q(P_1(F), \cdots, P_k(F))$$

is estimated as

$$\widehat{q} = q(\widehat{p}_1, \cdots, \widehat{p}_k) \text{ — frequency substitution.}$$

Example 3. In population genetics, sampling from a equilibrium population with respective to a gene with two alleles

$$\begin{cases} A & \text{with prob. } \theta \\ a & \text{with prob. } 1-\theta \end{cases},$$

three genotypes can be observed with proportions (Hardy-Weinberg formula).

AA	Aa	aa
$p_1 = \theta^2$	$p_2 = 2\theta(1-\theta)$	$p_3 = (1-\theta)^2$

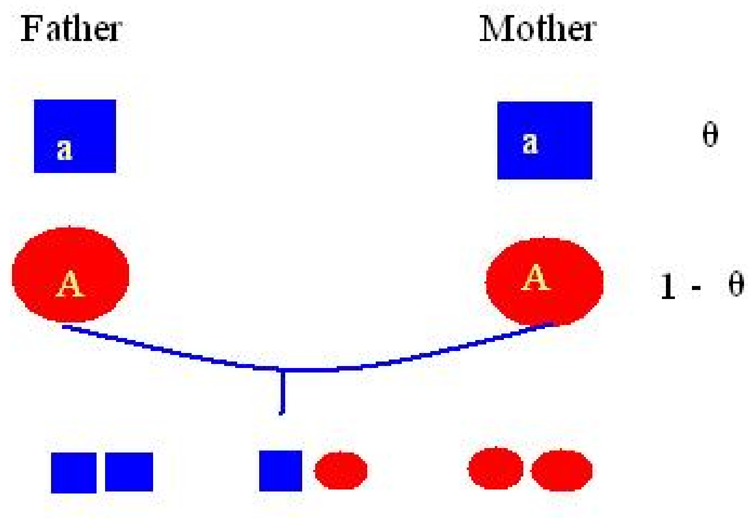


Figure 2.3: Illustration of Hardy-Weinberg formula.

One can estimate θ by $\sqrt{\widehat{p}_1}$ or $1-\sqrt{\widehat{p}_3}$, etc.

Thus, the representation

$$q(\theta) = h(p_1(\theta), \cdots, p_k(\theta))$$

is not necessarily unique, resulting in many different procedures.

Method of Moments: Let

$$m_j(\theta) = E_\theta X^j \quad \text{— theoretical moment}$$

and

$$\widehat{m}_j = \int x^j \, d\widehat{F}(x) = \frac{1}{n}\sum_{i=1}^{n} X_i^j \quad \text{— emprirical moment}$$

By the law of average, the empirical moments are close to theoretical ones. The method of moments is to solve the following estimating equations:

$$m_j(\theta) = \widehat{m}_j, \;\; j = 1, \cdots, r,$$

— smallest r to make enough equations. Why smallest?

$$\begin{cases} \text{inaccurate estimate of high order moment} \\ \text{inaccuracy of modeling of high order moment} \end{cases}$$

Consequently, the method of moment estimator for

$$q(\theta) = g(m_1(\theta), \cdots, m_r(\theta))$$

is $\widehat{q}(\mathbf{x}) = g(\widehat{m}_1, \cdots, \widehat{m}_r)$.

Exampel 4. Let $X_1, \cdots, X_n \sim i.i.d. N(\mu, \sigma^2)$. Then

$$EX = \mu \qquad and \qquad EX^2 = \mu^2 + \sigma^2.$$

Thus,

$$\widehat{\mu} = \bar{X} = \widehat{m}_1$$

$$\widehat{\mu}^2 + \widehat{\sigma}^2 = \widehat{m}_2$$

$$\Longrightarrow \quad \widehat{\sigma}^2 = \widehat{m}_2 - \widehat{\mu}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

Example 5. Let $X_1, \cdots, X_n \sim i.i.d.\, Poisson(\lambda)$. Then,

$$EX = \lambda \qquad and \qquad Var(X) = \lambda.$$

So

$$\lambda = m_1 = m_2 - m_1^2.$$

Thus,

$$\widehat{\lambda}_1 = \bar{X} \;\text{ and }\; \widehat{\lambda}_2 = \widehat{m}_2 - \widehat{m}_1^2 = \text{sample variance}.$$

The method of moments is

$$\begin{cases} \text{not necessarily unique} \\ \text{usually crude, serving a preluminary estimator} \end{cases}$$

Generalized method of moment (GMM):

Let $g_1(X), \cdots, g_r(X)$ be given functions. Write

$$\mu_j(\theta) = E_\theta\{g_j(X)\},$$

which are generalized moments. The GMM solves the equations

$$\widehat{\mu}_j = n^{-1}\sum_{i=1}^n g_j(X_i) = \mu_j(\theta), \;\; j = 1, \cdots, r.$$

If r > the number of parameters, find θ to minimize

$$\sum_{j=1}^r (\widehat{\mu}_j - \mu_j(\theta))^2$$

(this has a scale problem) or more generally

$$(\widehat{\mu} - \mu(\theta))^T \Sigma^{-1} (\widehat{\mu} - \mu(\theta)).$$

Σ can be found to optimize the performance of the estimator (EMM).

Example 6. For any random sample $\{(\mathbf{X}_i, Y_i), \;\; i = 1, \cdots, n\}$, define the coefficient of the best linear prediction under the loss function $d(\cdot)$ by

$$\beta(P) = \arg\min_{\beta} E_P d(|Y - \beta^T \mathbf{X}|).$$

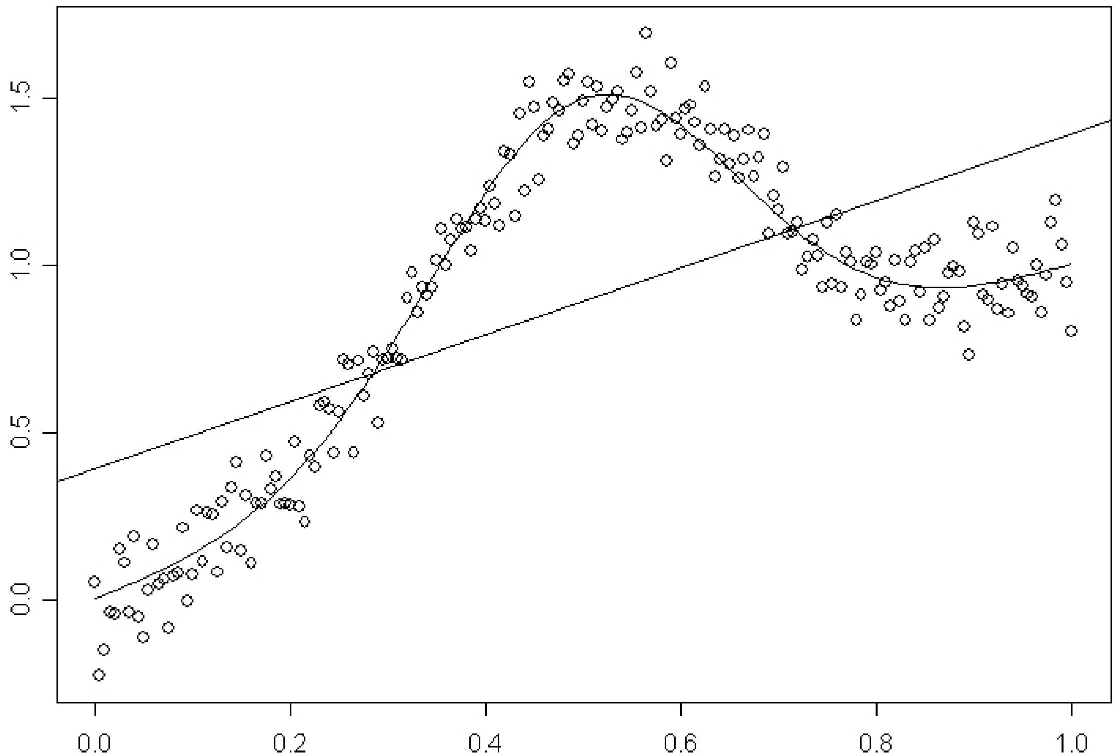


Figure 2.4: Illustration of best linear and nonlinear fittings.

Thus, its substitution estimator is

$$\beta(\widehat{P}) = \arg\min_{\beta} \frac{1}{n} \sum_{i=1}^{n} d(|Y_i - \beta^T \mathbf{X}_i|).$$

Thus, $\beta(\widehat{P})$ is always a consistent estimator of $\beta(P)$, whether the linear $Y = \beta^T \mathbf{X} + \varepsilon$ holds or not. In this view, the least-squares estimator is a substitution estimator.

2.2 Minimum Contrast Estimator and Estimating Equations

Let $\rho(X, \theta)$ be a contrast (discrepancy) function. Define

$$D(\theta_0, \theta) = E_{\theta_0} \rho(X, \theta), \text{ where } \theta_0 \text{ is the ture parameter.}$$

Suppose that $D(\theta_0, \theta)$ has a unique mimimum θ_0. Then, the minimum contrast estimator for a random sample is defined as the minimizer of

$$\widehat{D}(\theta) = \frac{1}{n} \sum_{i=1}^{n} \rho(X_i, \theta).$$

Under some regularity conditions, the estimator satisfies the estimating equations

$$\widehat{D}'(\theta) = \frac{1}{n}\sum_{i=1}^{n} \rho'(X_i, \theta) = 0.$$

Minimum contrast estimator. In general, the method applies to general situation:

$$\widehat{\theta} = \arg\min_{\theta} \rho(\mathbf{X}, \theta).$$

as long as θ_0 minimizes

$$D(\theta, \theta_0) = E_{\theta_0} \rho(\mathbf{X}, \theta).$$

Usually, $\rho(\mathbf{X}, \theta) = \frac{1}{n}\sum_{i=1}^{n} \rho(X_i, \theta) \longrightarrow D(\theta, \theta_0)$ (as $n \longrightarrow \infty$).

.

Similarly, **estimating equation method** solves the equations

$$\psi_j(\mathbf{X}, \theta) = 0, \; j = 1, \cdots, r$$

as long as

$$E_{\theta_0}\psi_j(X, \theta_0) = 0, \; j = 1, \cdots, r$$

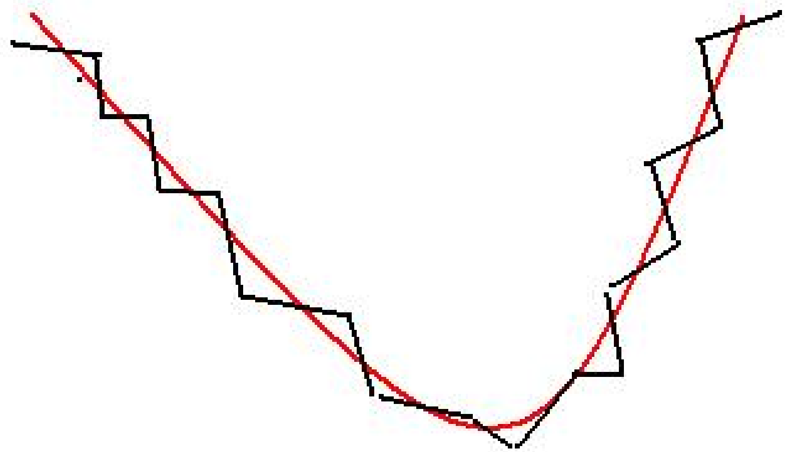


Figure 2.5: Minimum contrast estimator

Apparently, these two approaches are closely related.

Example 7 (Least-squares). Let $(\mathbf{X}_i, Y_i)$ be i.i.d. from

$$\begin{aligned} Y_i &= g(\mathbf{X}_i, \beta) + \varepsilon_i \\ &= \mathbf{X}_i^T \beta + \varepsilon_i, \qquad \text{if linear model} \end{aligned}$$

Then, by letting

$$\rho(\mathbf{X}, \beta) = \sum_{i=1}^{n} [Y_i - g(\mathbf{X}_i, \beta)]^2$$

be a contract function, we have

$$\begin{aligned} D(\beta_0, \beta) &= E_{\beta_0} \rho(\mathbf{X}, \beta) \\ &= n E_{\beta_0} [Y - g(\mathbf{X}, \beta)]^2 \\ &= n E\{g(\mathbf{X}, \beta_0) - g(\mathbf{X}, \beta)\}^2 + n\sigma^2, \end{aligned}$$

which is indeed minimized at $\beta = \beta_0$. Hence, the minimum contrast estimator is

$$\widehat{\beta} = \arg\min_{\beta} \sum_{i=1}^{n} [Y_i - g(\mathbf{X}_i, \beta)]^2 \;\; \text{— least-squares.}$$

It satisfies the system of equations

$$\sum_{i=1}^{n} (Y_i - g(\mathbf{X}_i, \widehat{\beta})) \frac{\partial g(\mathbf{X}_i, \beta)}{\partial \beta_j} = 0, \;\; j = 1, \cdots, d,$$

under some mild regularity conditions. One can easily check that $\psi_j(\beta) = (Y_i -$

$g(\mathbf{X}_i, \beta))\frac{\partial g(\mathbf{X}_i,\beta)}{\partial \beta_j}$ satisfies

$$E_{\beta_0}\psi_j(\beta)|_{\beta=\beta_0} = E\{g(\mathbf{X}_i, \beta_0) - g(\mathbf{X}_i, \beta_0)\}\frac{\partial g(\mathbf{X}_i, \beta_0)}{\partial \beta_j} = 0.$$

Thus, it is also an estimator based on the estimating equations.

Weighted least-squares: Suppose that $\text{var}(\varepsilon_i) = \omega_i \sigma^2$. The OLS continues to apply. However, it is not efficient. Through the transform

$$\frac{Y_i}{\sqrt{\omega_i}} = \frac{g(X_i, \beta)}{\sqrt{\omega_i}} + \frac{\varepsilon_i}{\sqrt{\omega_i}}$$

or

$$\tilde{Y}_i = \tilde{g}(X_i, \beta) + \tilde{\varepsilon}_i, \quad \tilde{\varepsilon}_i \sim N(0, \sigma^2),$$

we apply the OLS

$$\sum_{i=1}^{n}(\tilde{Y}_i - \tilde{g}(X_i, \beta))^2 = \sum_{i=1}^{n}(Y_i - g(\mathbf{X}_i, \beta))^2/\omega_i = \rho(\mathbf{X}, \beta).$$

Obviously,

$$E_{\beta_0}\rho(\mathbf{X},\beta) = \sum_{i=1}^{n} \omega_i \sigma^2/\omega_i + \sum_{i=1}^{n} \omega_i^{-1} E(g(\mathbf{X},\beta) - g(\mathbf{X},\beta_0))^2$$

is minimized at $\beta = \beta_0$. Thus,WLS is a minimum contrast estimator.

Example 8 (L_1-regression) Let $Y = \mathbf{X}^T\beta_0 + \varepsilon$, X and ε. Consider

$$\rho(\mathbf{X}, Y, \beta) = |Y - \mathbf{X}^T\beta|.$$

Then,

$$D(\beta_0, \beta) = E_{\beta_0}|Y - \mathbf{X}^T\beta| = E|\mathbf{X}^T(\beta - \beta_0) + \varepsilon|.$$

For any a, define

$$f(a) = E|\varepsilon + a|.$$

Then,

$$
\begin{aligned}
f'(a) &= E\text{sgn}(\varepsilon + a) \\
&= P(\varepsilon + a > 0) - P(\varepsilon + a < 0) \\
&= 2P(\varepsilon + a > 0) - 1.
\end{aligned}
$$

If $med(\varepsilon) = 0$, then $f'(0) = 0$. In other words, $f(a)$ is minimized at $a = 0$, or $D(\beta_0, \beta)$ is minimized at $\beta = \beta_0$! Thus, if $med(\varepsilon) = 0$, then

$$\frac{1}{n}\sum_{i=1}^{n} |Y_i - \mathbf{X}_i^T \beta|.$$

is a minimum contrast estimator.

2.3 The maximum likelihood estimator

Suppose that $\mathbf{X}$ has joint density $p(\mathbf{x}, \theta)$. This shows the "probability" of observing "$\mathbf{X} = \mathbf{x}$" under the parameter θ. Given $\mathbf{X} = \mathbf{x}$, there are many $\theta's$ that can have observed value $\mathbf{X} = \mathbf{x}$. We pick the one that is most probable to produce the observed $\mathbf{x}$:

$$\widehat{\theta} = \max_{\theta} L(\theta),$$

where $L(\theta) = p(\mathbf{x}, \theta)$= "likelihood of observing $\mathbf{x}$ under θ". This corresponds to the minimum contrast estimator with

$$\rho(\mathbf{x}, \theta) = -\log p(\mathbf{x}, \theta).$$

In particular, if $X_1, \cdots, X_n \sim i.i.d.\ f(\cdot\,, \theta)$, then

$$\rho(\mathbf{X}, \theta) = -\sum_{i=1}^{n} \log f(X_i, \theta).$$

To justify this, observe that

$$\begin{aligned} D(\theta_0, \theta) &= -E_{\theta_0} \log f(X, \theta) \\ &= D(\theta_0, \theta_0) - E_{\theta_0} \log \frac{f(X, \theta)}{f(X, \theta_0)} \\ &\geqslant D(\theta_0, \theta_0) - \log E_{\theta_0} \frac{f(X, \theta)}{f(X, \theta_0)} \\ &= D(\theta_0, \theta_0). \end{aligned}$$

Thus, θ_0 minimizes $D(\theta_0, \theta)$ or equivalently

$$D(\theta_0, \theta) - D(\theta_0, \theta_0) = -E_{\theta_0} \log \frac{f(X, \theta)}{f(X, \theta_0)}$$

— Kullback-Leibler information divergence. Thus, the MLE is a minimum contrast estimator.

The MLE is usually found by solving the **likelihood equations**:

$$\frac{\partial \log L(\theta)}{\partial \theta_j} = 0, \ \ j = 1, \cdots, d,$$

or

$$\ell'(\theta) = 0, \qquad \ell(\theta) = \log L(\theta).$$

For a given θ_0 that is close to $\widehat{\theta}$, then

$$0 = \ell'(\widehat{\theta}) \approx \ell'(\theta_0) + \ell''(\theta_0)(\widehat{\theta} - \theta_0)$$

or

$$\widehat{\theta} = \theta_0 - \ell''(\theta_0)^{-1}\ell'(\theta_0).$$

Newton-Raphson algorithm:

$$\widehat{\theta}_{new} = \widehat{\theta}_{old} - \ell''(\widehat{\theta}_{old})^{-1}\ell'(\widehat{\theta}_{old}).$$

One-step estimator: With a good initial estimator $\widehat{\theta}_0$,

$$\widehat{\theta}_{os} = \widehat{\theta}_0 - \ell''(\widehat{\theta}_0)^{-1}\ell'(\widehat{\theta}_0).$$

Example 9. (Hardy-Weinberg formula)

AA	Aa	aa
1	2	3
$p_1 = \theta^2$	$p_2 = 2\theta(1-\theta)$	$p_3 = (1-\theta)^2$

$$P_\theta(X_i = j) = \begin{cases} \theta^2, & j = 1 \\ 2\theta(1-\theta), & j = 2 \\ (1-\theta)^2, & j = 3 \end{cases}$$

$$L(\theta) = \prod_{i=1}^n P_\theta\{X_i = x_i\} = [\theta^2]^{n_1}[2\theta(1-\theta)]^{n_2}[(1-\theta)^2]^{n_3}$$

Thus,

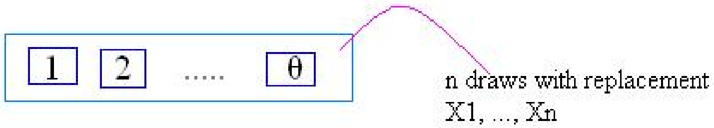
$$\ell(\theta) = (2n_1 + n_2)\log\theta + (n_2 + 2n_3)\log(1-\theta) + n_2 \log 2$$

$$\ell'(\theta) = (2n_1 + n_2)/\theta - (n_2 + 2n_3)/(1-\theta) = 0$$

$$\Longrightarrow \widehat{\theta} = \frac{2n_1 + n_2}{2(n_1 + n_2 + n_3)} = \frac{2n_1 + n_2}{n} = 2\widehat{p}_1 + \widehat{p}_2.$$

Obviously, $\ell''(\theta) < 0$. Hence, $\widehat{\theta}$ is the maxima.

Example 10. Estimating Population Size:



$$L(\theta) = \prod_{i=1}^{n} \frac{1}{\theta} I\{X_i \leqslant \theta\} = \theta^{-n} I\{\theta \geqslant \max_i X_i\}$$

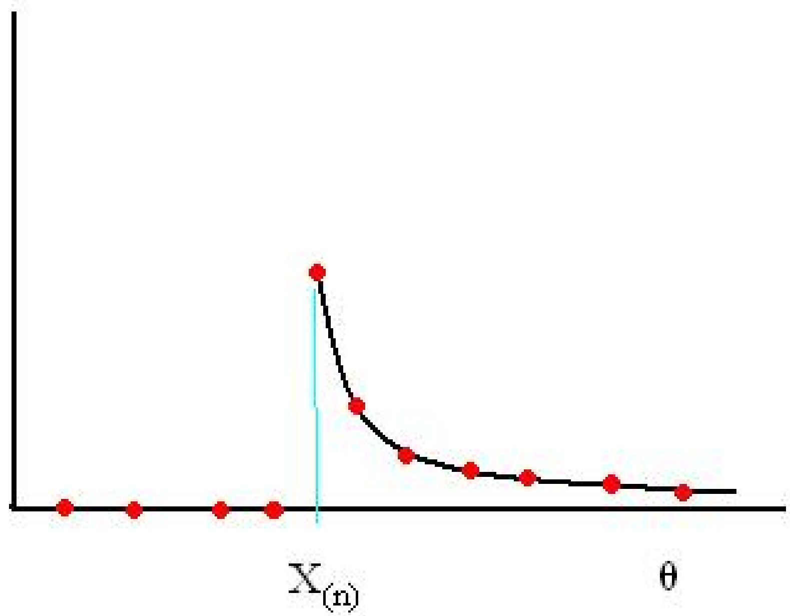


Figure 2.6: Likelihood function

Thus, $\widehat{\theta} = \max_i X_i = X_{(n)}$ is the MLE.

Example 11. Let $Y_i = g(X_i, \beta) + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma^2)$

Then,

$$\ell(\sigma, \beta) = \log(\frac{1}{\sqrt{2\pi}\sigma})^n - \frac{1}{2\sigma^2}\sum_{i=1}^n [Y_i - g(X_i, \beta)]^2.$$

Thus, the MLE for β is equivalent to minimize

$$\sum_{i=1}^n [Y_i - g(X_i, \beta)]^2 \quad \text{— least-squares.}$$

Let $\widehat{\beta}$ be the minimizer. Define

$$\text{RSS} = \sum_{i=1}^n [Y_i - g(X_i, \widehat{\beta})]^2.$$

Then, after dropping a constant

$$\ell(\sigma, \widehat{\beta}) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\text{RSS}.$$

which is maximized at $\widehat{\sigma}^2 = \frac{\text{RSS}}{n}$. In particular, if $g(X_i, \beta) = \mu$, then $\widehat{\mu} = \bar{Y}$ and

RSS $= \frac{1}{n}\sum_{i=1}^{n}[Y_i - \bar{Y}]^2$. Hence, the MLE is

$$\widehat{\mu} = \bar{Y} \qquad \textit{and} \qquad \widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \bar{Y})^2.$$

Remark:

MLE — use full likelihood function $\Longrightarrow$ more efficient, less robust.

MM — use the first few moments $\Longrightarrow$ less efficient, more robust.

2.4 The EM algorithm

(Reading assignment — read the whole section 2.4.)

Objective: Used to deal with missing data. [Dempster,Laird and Rubin(1977) and Baum, Petrie, Soules, and Weiss(1970).]

Problem: Suppose that we have a situation in which the full likelihood $\mathbf{X} \sim p(\mathbf{x}, \theta)$ is easy to compute and to maximize. Unfortunately, we only observe the partial information $S = S(\mathbf{X}) \sim q(s,\theta)$. $q(s,\theta)$ itself is hard to compute and to maximize. The algorithm is to maximize $q(s,\theta)$.

Example 12 (Lumped Hardy-Weinberg data)

The full information is $X_1, \cdots, X_n$ with

$$\log p(\mathbf{x},\theta) = n_1 \log \theta^2 + n_2 \log 2\theta(1-\theta) + n_3 \log(1-\theta)^2$$

Partial information:

$$\begin{aligned} &\text{complete cases} \quad S_i = (X_{i1}, X_{i2}, X_{i3}), \ \ i = 1, \cdots, m \\ &\text{incomplete cases} \ \ S_i = (X_{i1}+X_{i2}, X_{i3}), \ \ i = m+1, \cdots, n. \end{aligned}$$

The likelihood of the available data is

$$\begin{aligned} &\log q(s,\theta) = m_1 \log \theta^2 + m_2 \log 2\theta(1-\theta) + m_3 \log(1-\theta)^2 \\ &+ n_{12}^* \log(1-(1-\theta)^2) + n_3^* \log(1-\theta)^2 \\ &n_{12}^* = \sum_{i=m+1}^{n} (X_{i1}+X_{i2}), \ \ n_3^* = \sum_{i=m}^{n} X_{i3}. \end{aligned}$$

The maximum likelihood can be found by maximizing the above expression. For many other problems, this log-likelihood can be hard to compute.

Intuition for E-M algorithm: Guess the full likelihood using the available and maximum the conjectured likelihood.

E-M algorithm: Given an initial value θ_0,

E-step: Compute $\ell(\theta, \theta_0) = E_{\theta_0}(\ell(\mathbf{X}, \theta)|S(\mathbf{X}) = s)$,

M-step: $\widehat{\theta} = \arg\max \ell(\theta, \theta_0)$,

and iterate.

Example 2.12. (continued) Full likelihood:

$$\log p(\mathbf{x}, \theta) = n_1 \log \theta^2 + n_2 \log 2\theta(1-\theta) + n_3 \log(1-\theta)^2$$

E-step:

$$\ell(\theta, \theta_0) = E_{\theta_0}(n_1|S) \log \theta^2 + E_{\theta_0}(n_2|S) \log 2\theta(1-\theta) + n_3 \log(1-\theta)^2.$$

$$E_{\theta_0}(n_1|S) = m_1 + n_{12}^* \frac{\theta_0^2}{\theta_0^2 + 2\theta_0(1-\theta_0)}$$

and

$$E_{\theta_0}(n_2|S) = m_2 + n_{12}^* \frac{2\theta_0(1-\theta_0)}{\theta_0^2 + 2\theta_0(1-\theta_0)}$$

M-step:

$$\widehat{\theta} = \frac{2E_{\theta_0}(n_1|S) + E_{\theta_0}(n_2|S)}{2(E_{\theta_0}(n_1|S) + E_{\theta_0}(n_2|S) + n_3)} = \frac{n_{12} + E_{\theta_0}(n_1|S)}{2n},$$

where n_{12} is the number of data points for genotypes $\boxed{1}$ and $\boxed{2}$. When the algorithm converges, it solves the following equation:

$$2n\theta = n_{12} + m_1 + n_{12}^* \theta/(2-\theta).$$

This is indeed the maximum likelihood estimator based on the available (partial) data, which we now justify.

Rationale of the EM algorithm: $p(\mathbf{x}, \theta) = q(s, \theta) P_\theta(\mathbf{X} = \mathbf{x} | S = s) I(S(\mathbf{X}) = s)$.

Let $r(x|s, \theta) = P_\theta(\mathbf{X} = \mathbf{x}|S = s)I(S(\mathbf{x}) = s)$. Then,

$$\ell(\theta, \theta_0) = \log q(s, \theta) + E_{\theta_0}\{\log r(\mathbf{X}|s, \theta)|S(\mathbf{X}) = s\}.$$

Hence,

$$0 = \ell(\widehat{\theta}, \theta_0)' = (\log q(s, \theta))'|_{\theta=\widehat{\theta}} + E_{\theta_0}\{(\log r(\mathbf{X}|s, \theta))'|_{\theta=\widehat{\theta}}|S = s\}.$$

If the algorithm converges to θ_1, then

$$(\log q(s, \theta_1))' + E_{\theta_1}\{(\log r(\mathbf{X}|s, \theta_1))'|S = s\} = 0.$$

The second term vanishes by noticing that for any regular function f,

$$E_\theta(\log f(\mathbf{X}, \theta))' = \int \frac{f'(\mathbf{x}, \theta)}{f(\mathbf{x}, \theta)} f(\mathbf{x}, \theta) d\mathbf{x} = 0.$$

Hence

$$\{\log q(s, \theta_1)\}' = 0,$$

which solves the likelihood equation based on the (partial) data. In other words, the EM algorithm converges to the true likelihood.

Theorem 1

$$\log q(s, \theta_{new}) \geqslant \log q(s, \theta_{old}),$$

namely, each iteration always increases the likelihood.

Proof. Note that

$$\begin{aligned} \ell(\theta_n, \theta_0) &= \log q(s, \theta_n) + E_{\theta_0}\{\log r(X|s, \theta_n)|S(\mathbf{X}) = s\} \\ &\geqslant \log q(s, \theta_0) + E_{\theta_0}\{\log r(X|s, \theta_0)|S(\mathbf{X}) = s\} \end{aligned}$$

$\Longrightarrow$

$$\begin{aligned} \log q(s, \theta_n) &\geq \log q(s, \theta_0) + E_{\theta_0}\{\log \frac{r(X|s, \theta_0)}{r(X|s, \theta_n)}|S(\mathbf{X}) = s\} \\ &\geqslant \log q(s, \theta_0). \end{aligned}$$

Example 2.13. Let $X_1, \cdots, X_{n+4}$ be i.i.d. $N(\mu, 1/2)$. Suppose that we observe

$S_1 = X_1, \cdots, S_n = X_n$, $S_{n+1} = X_{n+1} + 2X_{n+2}$, and $S_{n+2} = X_{n+3} + X_{n+4}$. Use the EM algorithm to find the maximum likelihood estimator based on the observed data.

Note that the full likelihood is

$$\begin{aligned}\log p(\mathbf{X}, \mu) &= -\sum_{i=1}^{n+4}(X_i - \mu)^2 \\ &= -\sum_{i=1}^{n+4} X_i^2 + 2\mu \sum_{i=1}^{n+4} X_i - (n+4)\mu^2.\end{aligned}$$

At the E-step, we compute

$$\begin{aligned}E_{\mu_0}\{\log p(\mathbf{X}, \mu)|\boldsymbol{S}\} &= a(\mu_0) - 2\mu\{\sum_{i=1}^{n} X_i + E_{\mu_0}\{X_{n+1} + X_{n+2}|S_{n+1}\} \\ &\quad + S_{n+2}\} - (n+4)\mu^2,\end{aligned}$$

where $a(\mu_0) = (\mu_0^2 + 1/2)$. To compute $E_{\mu_0}\{X_{n+1}|S_{n+1}\}$, we note that $2X_{n+1} - X_{n+2}$

is uncorrelated with S_{n+1}. Hence, we have

$$E_{\mu_0}\{2X_{n+1} - X_{n+2}|S_{n+1}\} = \mu_0$$

$$E_{\mu_0}\{X_{n+1} + 2X_{n+2}|S_{n+1}\} = S_{n+1}.$$

Solving the above two equations gives

$$E_{\mu_0}(X_{n+1}|S_{n+1}) = (S_{n+1} + 2\mu_0)/5, \quad E_{\mu_0}(X_{n+2}|S_{n+1}) = (2S_{n+1} - \mu_0)/5$$

and that

$$E_{\mu_0}\{X_{n+1} + X_{n+2}|S_{n+1}\} = (3S_{n+1} + \mu_0)/5.$$

Hence, the conditional likelihood is given by

$$\ell(\mu, \mu_0) = a(\mu_0) - 2\mu\{\sum_{i=1}^{n} X_i + 0.6S_{n+1} + 0.2\mu_0 + S_{n+2}\} - (n+4)\mu^2.$$

At the M-step, we maximize $\ell(\mu, \mu_0)$ with respect to μ, resulting in

$$\widehat{\mu} = (n+4)^{-1}\{\sum_{i=1}^{n} X_i + 0.6S_{n+1} + 0.2\mu_0 + S_{n+2}\}.$$

The EM algorithm is to iterate the above step. When the algorithm converges, the estimate solves

$$\widehat{\mu} = (n+4)^{-1}\{\sum_{i=1}^{n} X_i + 0.6S_{n+1} + 0.2\widehat{\mu} + S_{n+2}\}.$$

or

$$\widehat{\mu} = (n+3.8)^{-1}\{\sum_{i=1}^{n} X_i + 0.6S_{n+1} + S_{n+2}\}.$$

This is the maximum likelihood estimator for the "missing" data.

Example 2.14. Mixture normal distribution:

$$S_1, \cdots, S_n \sim_{i.i.d.} \lambda N(\mu_1, \sigma_1^2) + (1-\lambda) N(\mu_2, \sigma_2^2)$$

Challenge: The likelihood of $S_1, \cdots, S_n$ is easy to write down, but hard to compute.

EM Algorithm: Thinking of the full information as $X_i = (\triangle_i, S_i)$, in which $\triangle_i$ tells the population under which it is drawn from, but missing.

$$P(\triangle_i = 1) = \lambda.$$

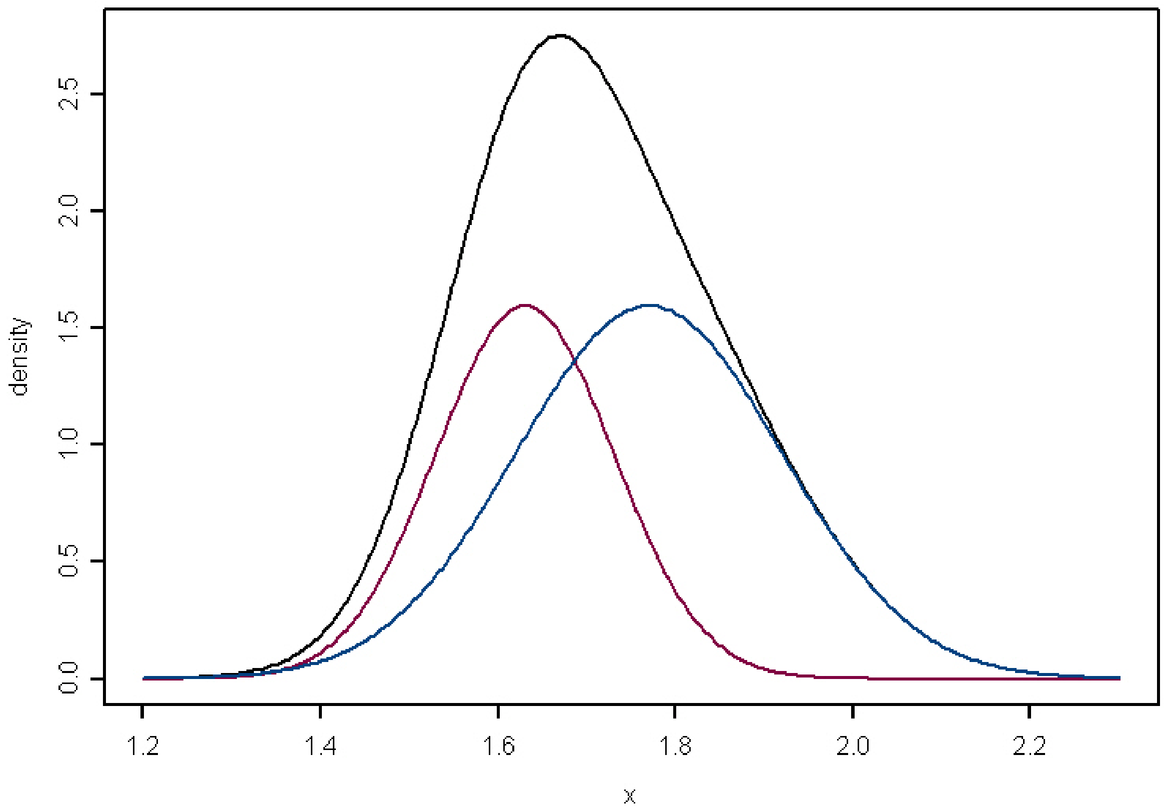


Figure 2.7: Mixture of two normal distributions

$$P(S_i|\triangle_i) \sim \begin{cases} N(\mu_1, \sigma_1^2), & \text{if } \triangle_i = 1 \\ N(\mu_2, \sigma_2^2), & \text{if } \triangle_i = 0 \end{cases}$$

Then, the full likelihood is

$$
\begin{aligned}
p(\mathbf{x}, \theta) &= \lambda^{\sum \triangle_i}(1-\lambda)^{n-\sum \triangle_i}\Pi_{\triangle_i=1}(\frac{1}{\sqrt{2\pi}\sigma_1})\exp(-\frac{(S_i-\mu_1)^2}{2\sigma_1^2}) \\
&\quad \times \Pi_{\triangle_i=0}(\frac{1}{\sqrt{2\pi}\sigma_2})\exp(-\frac{(S_i-\mu_2)^2}{2\sigma_2^2}).
\end{aligned}
$$

It follows that

$$
\begin{aligned}
\log p(\mathbf{x}, \theta) &= \sum \triangle_i \log\lambda + \sum (1-\triangle_i)\log(1-\lambda) \\
&\quad + \sum_{\triangle_i=1}\{-\log\sigma_1 - \frac{(S_i-\mu_1)^2}{2\sigma_1^2}\} \\
&\quad + \sum_{\triangle_i=0}\{-\log\sigma_2 - \frac{(S_i-\mu_2)^2}{2\sigma_2^2}\}
\end{aligned}
$$

To find the E-step, we need to find the conditional distribution of $\triangle_i|\mathbf{S}$.

Note that

$$\begin{aligned} P_{\theta_0}\{\triangle_i = 1 | \mathbf{S} = \mathbf{s}\} &= P\{\triangle_i = 1 | S_i \in s_i \pm \varepsilon\} \\ &= \frac{P\{\triangle_i = 1, S_i \in s_i \pm \varepsilon\}}{P(S_i \in s_i \pm \varepsilon)} \\ &= \frac{\lambda_0 \sigma_{10}^{-1} \phi\left(\frac{s_i - \mu_{10}}{\sigma_{10}}\right)}{\lambda_0 \sigma_{10}^{-1} \phi\left(\frac{s_i - \mu_{10}}{\sigma_{10}}\right) + (1 - \lambda_0) \sigma_{20}^{-1} \phi\left(\frac{s_i - \mu_{20}}{\sigma_{20}}\right)} \\ &\equiv p_i \end{aligned}$$

Then,

$$\begin{aligned} \ell(\theta, \theta_0) &= \sum_{i=1}^n p_i \log \lambda + \sum_{i=1}^n (1 - p_i) \log(1 - \lambda) \\ &\quad + \sum_{i=1}^n p_i \{-\log \sigma_1 - \frac{(s_i - \mu_1)^2}{2\sigma_1^2}\} \\ &\quad + \sum_{i=1}^n (1 - p_i) \{-\log \sigma_2 - \frac{(s_i - \mu_2)^2}{2\sigma_2^2}\}. \end{aligned}$$

The M-step is to maximize the above quantity with respective to $\lambda, \sigma_1, \mu_1, \sigma_2, \mu_2$, which can be explicitly found. e.g.

$$\frac{\sum_{i=1}^n p_i}{\lambda} - \frac{\sum_{i=1}^n (1-p_i)}{1-\lambda} = 0 \;\Rightarrow\; \lambda = \frac{\sum_{i=1}^n p_i}{n}$$

$$\sum_{i=1}^n p_i(s_i - \mu_1) = 0 \;\Rightarrow\; \widehat{\mu}_1 = \frac{\sum_{i=1}^n p_i s_i}{\sum_{i=1}^n p_i},$$

The EM algorithm is to iterate these two steps.