Now let us discretise the problem using a standard Euler approximation, rewriting the system as

$$\delta X_i = X_{i+1} - X_i = \mu(X_i;\theta)\delta t + \sigma(X_i)\sqrt{\delta t}u_i,$$
$$\implies \quad \frac{\delta X_i}{\sigma(X_i)\sqrt{\delta t}} = \frac{\mu(X_i;\theta)}{\sigma(X_i)}\sqrt{\delta t} + u_i,$$

where $u_i \sim \mathcal{N}(0,1)$ are independent Gaussian increments. The corresponding log-likelihood function is then

$$l(\theta;X) = \sum_i -\frac{1}{2}\ln(2\pi) - \frac{1}{2}\left(\frac{\delta X_i}{\sigma(X_i)\sqrt{\delta t}} - \frac{\mu(X_i;\theta)}{\sigma(X_i)}\sqrt{\delta t}\right)^2.$$

After discarding terms independent of θ and let $\delta t \to 0$, we arrive at the *infill log-likelihood* function,

$$l_{IF}(\theta;X) = \int_0^T \frac{\mu(X_t;\theta)}{\sigma^2(X_t)}\,dX_t - \frac{1}{2}\int_0^T \frac{\mu^2(X_t;\theta)}{\sigma^2(X_t)}\,dt. \tag{1.27}$$

Example 1.15. Consider the following OU process

$$dX_t = \kappa X_t dt + \sigma dW_t.$$

The volatility is estimated from the quadratic variation

$$\hat{\sigma} = \frac{1}{T}[X]_t.$$

The infill log-likelihood is

$$l_{IF}(\kappa) = \int_0^T \frac{\kappa X_t}{\hat{\sigma}^2}\,dX_t - \frac{1}{2}\int_0^T \frac{\kappa^2 X_t^2}{\hat{\sigma}^2}\,dt,$$

which reaches maximum at

$$\hat{\kappa} = \left(\int_0^T X_t^2\,dt\right)^{-1}\int_0^T X_t\,dX_t.$$

Infill likelihood via the Girsanov theorem

Again starting with

$$dX_t = \mu(X_t;\theta)dt + \sigma(X_t)dW_t^{\mathbb{P}}.$$

The likelihood of a sample path $X_t(\omega)$ can be written as $d\mathbb{P}(\omega)$. Now define

$$W_t^{\mathbb{Q}} := W_t^{\mathbb{P}} + \int_0^t \frac{\mu(X_s;\theta)}{\sigma(X_s)}\,ds.$$

We can construct the measure $\mathbb{Q}$ via the Radon-Nikodym derivative

$$\eta_\cdot = \frac{d\mathbb{P}}{d\mathbb{Q}} = \exp\left(\int_0^\cdot \frac{\mu(X_t;\theta)}{\sigma(X_t)}\, dW_t^{\mathbb{Q}} - \frac{1}{2}\int_0^\cdot \frac{\mu^2(X_t;\theta)}{\sigma^2(X_t)}\, dt \right).$$

By Girsanov theorem, we have

$$dX_t = \sigma(X_t) dW_t^{\mathbb{Q}},$$

where W_t^Q is a $\mathbb{Q}$-Brownian motion. Furthermore

$$d\mathbb{P}(\omega) = \eta_T d\mathbb{Q}(\omega).$$

But since X_t is independent of θ under $\mathbb{Q}$, so is its sample path likelihood $d\mathbb{Q}(\omega)$. So it suffices to maximise the infill log-likelihood

$$\begin{aligned} l_{IF}(\theta) = \ln \eta_T &= \int_0^T \frac{\mu(X_t;\theta)}{\sigma(X_t)}\, dW_t^{\mathbb{Q}} - \frac{1}{2}\int_0^T \frac{\mu^2(X_t;\theta)}{\sigma^2(X_t)}\, dt \\ &= \int_0^T \frac{\mu(X_t;\theta)}{\sigma^2(X_t)}\, dX_t - \frac{1}{2}\int_0^T \frac{\mu^2(X_t;\theta)}{\sigma^2(X_t)}\, dt. \end{aligned}$$

1.5 EM algorithm

Suppose we have some observed data X whose distribution depends on unknown parameters θ. The log-likelihood function is $l(\theta; x) = \ln f_X(x \mid \theta)$. In some cases, it might be easier to maximise $l(\theta; x)$ if we also had access to some other latent (unobserved) data Y, i.e., it would be easier to maximise $l(\theta; x, y) = \ln f_{X,Y}(x, y \mid \theta)$. Here, (X, Y) is known as the *complete data*, with X being the observed data and Y being the missing data. Note that the density functions relate to each other via

$$f_X(x \mid \theta) = \int_y f_{X,Y}(x, y \mid \theta)\, dy.$$

The *expectation-maximisation (EM) algorithm* is an iterative algorithm which maximises $l(\theta; x)$ by maximising $l(\theta; x, y)$. Each iteration has two parts.

E-step: Suppose $\hat{\theta}^{(t)}$ was the output of the previous iteration. Define

$$Q(\theta \mid \hat{\theta}^{(t)}) := \mathbb{E}(l(\theta; X, Y) \mid X, \hat{\theta}^{(t)}). \tag{1.28}$$

Note the expectation is taken with respect to the conditional density $f(Y \mid X, \hat{\theta}^{(t)})$. The variable $\hat{\theta}^{(t)}$ only affects the density of Y. It does not replace θ in $l(\theta; X, Y)$. In cases where the expectation is difficult to evaluate analytically, a Monte Carlo approach can be used by sampling from $f(Y \mid X, \hat{\theta}^{(t)})$ and then computing the average of $l(\theta; X, Y)$.

M-step: Compute the value $\hat{\theta}^{(t+1)}$ by maximising $Q(\theta \mid \hat{\theta}^{(t)})$,

$$\hat{\theta}^{(t+1)} := \arg\max_\theta Q(\theta \mid \hat{\theta}^{(t)}). \tag{1.29}$$

The algorithm iterates until a desired level of accuracy is obtained.

Monotonicity and stationarity

It is not immediately clear that the output $\hat{\theta}^{(t+1)}$ would actually improve $l(\theta; x)$, since we are have been maximising $Q(\theta \,|\, \hat{\theta}^{(t)})$ instead. The following proposition shows that the EM algorithm indeed improves the required likelihood function.

Proposition 1.16.
$$l(\hat{\theta}^{(t+1)}; X) \geq l(\hat{\theta}^{(t)}; X).$$

Proof. Consider the function $g(\theta \,|\, \hat{\theta}^{(t)}) = l(\theta; X) - Q(\theta \,|\, \hat{\theta}^{(t)})$, we claim $g(\theta \,|\, \hat{\theta}^{(t)})$ is minimised at $\hat{\theta}^{(t)}$. Begin by rewriting it as follows,

$$\begin{aligned}
g(\theta \,|\, \hat{\theta}^{(t)}) &= l(\theta; X) - Q(\theta \,|\, \hat{\theta}^{(t)}) \\
&= \mathbb{E}(l(\theta; X) - l(\theta; X, Y) \,|\, X, \hat{\theta}^{(t)}) \\
&= \mathbb{E}\left(\ln\left(\frac{f(X \,|\, \theta)}{f(X, Y \,|\, \theta)} \right) \,\middle|\, X, \hat{\theta}^{(t)} \right) \\
&= -\mathbb{E}(\ln(f(Y \,|\, X, \theta)) \,|\, X, \hat{\theta}^{(t)}).
\end{aligned}$$

Then, by Jensen's inequality,

$$\begin{aligned}
g(\theta \,|\, \hat{\theta}^{(t)}) - g(\hat{\theta}^{(t)} \,|\, \hat{\theta}^{(t)}) &= -\mathbb{E}\left(\ln\left(\frac{f(Y \,|\, X, \theta)}{f(Y \,|\, X, \hat{\theta}^{(t)})} \right) \,\middle|\, X, \hat{\theta}^{(t)} \right) \\
&\geq -\ln \mathbb{E}\left(\frac{f(Y \,|\, X, \theta)}{f(Y \,|\, X, \hat{\theta}^{(t)})} \,\middle|\, X, \hat{\theta}^{(t)} \right) \\
&= -\ln \int_y \frac{f(y \,|\, X, \theta)}{f(y \,|\, X, \hat{\theta}^{(t)})} f(y \,|\, X, \hat{\theta}^{(t)}) \, dy \\
&= -\ln \int_y f(y \,|\, X, \theta) \, dy \\
&= -\ln(1) = 0.
\end{aligned}$$

So $g(\theta \,|\, \hat{\theta}^{(t)})$ is indeed minimised at $\hat{\theta}^{(t)}$. Back to the problem at hand, using the fact that $\hat{\theta}^{(t+1)}$ maximises $Q(\theta \,|\, \hat{\theta}^{(t)})$, we have

$$\begin{aligned}
l(\hat{\theta}^{(t+1)}; X) &= g(\hat{\theta}^{(t+1)} \,|\, \hat{\theta}^{(t)}) + Q(\hat{\theta}^{(t+1)} \,|\, \hat{\theta}^{(t)}) \\
&\geq g(\hat{\theta}^{(t)} \,|\, \hat{\theta}^{(t)}) + Q(\hat{\theta}^{(t)} \,|\, \hat{\theta}^{(t)}) \\
&= l(\hat{\theta}^{(t)}; X)
\end{aligned}$$

as required. □

Remark 1.17. The key to the proof of Proposition 1.16 is that the following inequality,

$$g(\theta \,|\, \hat{\theta}^{(t)}) - g(\hat{\theta}^{(t)} \,|\, \hat{\theta}^{(t)}) = \mathbb{E}\left(\ln\left(\frac{f(Y \,|\, X, \hat{\theta}^{(t)})}{f(Y \,|\, X, \theta)} \right) \,\middle|\, X, \hat{\theta}^{(t)} \right) \geq 0$$

This difference is also known as the *Kullback-Leibler divergence*, denoted by

$$D_{KL}(f(Y \mid X, \hat{\theta}^{(t)}) \parallel f(Y \mid X, \theta)) = \mathbb{E}\left(\ln \left(\frac{f(Y \mid X, \hat{\theta}^{(t)})}{f(Y \mid X, \theta)} \right) \,\middle|\, X, \hat{\theta}^{(t)} \right).$$

It measures the relative entropy of one probability measure with respect to another. It is known to be non-negative and in this case it reaches the minimum of 0 when $\theta = \hat{\theta}^{(t)}$.

If both $l(\theta; X)$ and $Q(\theta \mid \hat{\theta}^{(t)})$ satisfies certain smoothness conditions, it can be shown that $\hat{\theta}^{(\infty)} = \lim_{t\to\infty} \hat{\theta}^{(t)}$ converges to a (possibly local) maximum. At the limit we have

$$\hat{\theta}^{(\infty)} := \arg\max_{\theta} Q(\theta \mid \hat{\theta}^{(\infty)}).$$

The first order condition satisfies

$$\frac{\partial}{\partial \theta} l(\hat{\theta}^{(\infty)}; X) = \frac{\partial}{\partial \theta} g(\hat{\theta}^{(\infty)} \mid \hat{\theta}^{(\infty)}) + \frac{\partial}{\partial \theta} Q(\hat{\theta}^{(\infty)} \mid \hat{\theta}^{(\infty)}) = 0,$$

so $l(\hat{\theta}^{(\infty)}; X)$ is indeed a stationary point and hence maximum. Note that we have again use the fact that $g(\theta \mid \hat{\theta}^{(\infty)})$ is minimised at $\hat{\theta}^{(\infty)}$.

Example 1.18. Let $X = (X_1, \dots, X_n)$ be a sample of n independent observations draw from a mixture of two Gaussian distributions,

$$X_i|(Z_i = 1) \sim \mathcal{N}(\mu_1, \sigma_1^2), \quad X_i|(Z_i = 2) \sim \mathcal{N}(\mu_2, \sigma_2^2),$$

where $Z = (Z_1, \dots, Z_n)$ are latent (unobserved) variables with

$$\mathbb{P}(Z_i = 1) = p, \quad \mathbb{P}(Z_i = 2) = 1 - p.$$

The unknown parameters to be estimated are

$$\theta = (p, \mu_1, \mu_2, \sigma_1^2, \sigma_2^2).$$

The incomplete data likelihood function is

$$L(\theta; X) = \prod_{i=1}^{n} (p f(X_i \mid \mu_1, \sigma_1^2) + (1 - p) f(X_i \mid \mu_2, \sigma_2^2)),$$

whereas the complete data likelihood is

$$\begin{aligned} L(\theta; X, Z) &= \prod_{i=1}^{n} (\mathbb{1}(Z_i = 1) p f(X_i \mid \mu_1, \sigma_1^2) + \mathbb{1}(Z_i = 2)(1 - p) f(X_i \mid \mu_2, \sigma_2^2)), \\ l(\theta; X, Z) &= \sum_{i=1}^{n} (\mathbb{1}(Z_i = 1)(\ln p + \ln f(X_i \mid \mu_1, \sigma_1^2)) \\ &\qquad + \mathbb{1}(Z_i = 2)(\ln(1 - p) + \ln f(X_i \mid \mu_1, \sigma_1^2))). \end{aligned} \tag{1.30}$$

E-step: In the E-step, we need to take conditional expectation over the latent variable Z. The conditional density of Z_i is given by

$$\begin{aligned}
P_i^{(t)} = \mathbb{P}(Z_i = 1 \mid X, \hat{\theta}^{(t)}) = \mathbb{P}(Z_i = 1 \mid X_i, \hat{\theta}^{(t)}) &= \frac{f(X_i, Z_i = 1 \mid \hat{\theta}^{(t)})}{f(X_i \mid \hat{\theta}^{(t)})} \\
&= \frac{\hat{p}^{(t)} f(X_i \mid \hat{\mu}_1^{(t)}, (\hat{\sigma}_1^2)^{(t)})}{\hat{p}^{(t)} f(X_i \mid \hat{\mu}_1^{(t)}, (\hat{\sigma}_1^2)^{(t)}) + (1-\hat{p}^{(t)}) f(X_i \mid \hat{\mu}_2^{(t)}, (\hat{\sigma}_2^2)^{(t)})}. \qquad (1.31)
\end{aligned}$$

Using the log-likelihood function from (1.30), the function $Q(\theta \mid \hat{\theta}^{(t)})$ is given by

$$\begin{aligned}
Q(\theta \mid \hat{\theta}^{(t)}) &= \mathbb{E}(l(\theta; X, Z) \mid X, \hat{\theta}^{(t)}) \\
&= \sum_{i=1}^{n} (P_i^{(t)} (\ln p + \ln f(X_i \mid \mu_1, \sigma_1^2)) \\
&\qquad + (1 - P_i^{(t)})(\ln(1-p) + \ln f(X_i \mid \mu_2, \sigma_2^2))). \qquad (1.32)
\end{aligned}$$

M-step: Now we have to find $\hat{\theta}^{(t+1)} = \arg\max_\theta Q(\theta \mid \hat{\theta}^{(t)})$. This is reasonably straight forward since $Q(\theta \mid \hat{\theta}^{(t)})$ is fairly easy to deal with. First we look at $\hat{p}^{(t+1)}$, whose maximum likelihood estimate is similar to the binomial distribution,

$$\hat{p}^{(t+1)} = \arg\max_p \sum_{i=1}^{n} (P_i^{(t)} \ln p + (1 - P_i^{(t)}) \ln(1-p)) = \frac{1}{n} \sum_{i=1}^{n} P_i^{(t)}. \qquad (1.33)$$

For $j = 1, 2$, the function $f(X_i \mid \mu_j, \sigma_j^2)$ is the density of $\mathcal{N}(\mu_j, \sigma_j^2)$, so the maximum likelihood estimate is similar to the Gaussian distribution,

$$\begin{aligned}
(\hat{\mu}_j^{(t+1)}, (\hat{\sigma}_j^2)^{(t+1)}) &= \arg\max_{(\mu_j, \sigma_j^2)} \sum_{i=1}^{n} P_{i,j}^{(t)} \ln f(X_i \mid \mu_j, \sigma_j^2) \\
&= \left(\frac{\sum_{i=1}^n P_{i,j}^{(t)} X_i}{\sum_{i=1}^n P_{i,j}^{(t)}}, \frac{\sum_{i=1}^n P_{i,j}^{(t)} (X_i^2 - \hat{\mu}_j^{(t+1)})^2}{\sum_{i=1}^n P_{i,j}^{(t)}} \right), \qquad (1.34)
\end{aligned}$$

where we have introduced the notations $P_{i,1}^{(t)} = P_i^{(t)}$ and $P_{i,2}^{(t)} = 1 - P_i^{(t)}$.

Mixture models

Example 1.18 is a simple example of a mixture model where the EM algorithm is used to identify the underlying distributions. The principal works in general. Suppose we have independent observations $X = (X_1, \ldots, X_n)$ taken from a mixture of K (possibly different) distributions $D_1, \ldots, D_k$ involving unknown parameters θ. Suppose further that the probabilities of an observation being in any mixture are $\alpha = (\alpha_1, \ldots, \alpha_K)$ with $\sum_{i=1}^K \alpha_i = 1$. We can once again define latent variables $Z = (Z_1, \ldots, Z_n)$, with each Z_i taking values from $\{1, \ldots, K\}$, indicating the distribution of X_i.

In the **E-step**, the expectation of the complete log-likelihood is given by

$$
\begin{aligned}
Q(\alpha, \theta \mid \hat{\alpha}^{(t)}, \hat{\theta}^{(t)}) &= \mathbb{E}\left(\sum_{i=1}^{n} \sum_{j=1}^{K} \mathbb{1}(Z_i = j) \ln(\alpha_j f_{D_j}(X_i \mid \theta)) \,\middle|\, \hat{\alpha}^{(t)}, \hat{\theta}^{(t)} \right) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{K} \mathbb{E}(\mathbb{1}(Z_i = j) \mid \hat{\alpha}^{(t)}, \hat{\theta}^{(t)}) \ln(\alpha_j f_{D_j}(X_i \mid \theta)) \\
&= \sum_{i=1}^{n} \sum_{j=1}^{K} P_{i,j}^{(t)} \ln(\alpha_j f_{D_j}(X_i \mid \theta)),
\end{aligned}
$$

where

$$
P_{i,j}^{(t)} := \mathbb{P}(Z_i = j \mid \hat{\alpha}^{(t)}, \hat{\theta}^{(t)}) = \frac{\hat{\alpha}_j^{(t)} f_{D_j}(X_i \mid \hat{\theta}^{(t)})}{\sum_{j=1}^{K} \hat{\alpha}_j^{(t)} f_{D_j}(X_i \mid \hat{\theta}^{(t)})}.
$$

In the **M-step**, we can maximise α using Lagrange multipliers,

$$
\begin{aligned}
\hat{\alpha}^{(t+1)} &= \underset{(\alpha_1, \ldots, \alpha_K)}{\arg\max} \left\{ \sum_{i=1}^{n} \sum_{j=1}^{K} P_{i,j}^{(t)} \ln(\alpha_j) : \sum_{i=1}^{K} \alpha_i = 1 \right\} \\
&= \left(\frac{1}{n} \sum_{i=1}^{n} P_{i,1}^{(t)}, \ldots, \frac{1}{n} \sum_{i=1}^{n} P_{i,K}^{(t)} \right).
\end{aligned}
$$

On the other hand, maximising θ is reduced to

$$
\hat{\theta}^{(t+1)} = \underset{\theta}{\arg\max} \sum_{i=1}^{n} \sum_{j=1}^{K} P_{i,j}^{(t)} \ln(f_{D_j}(X_i \mid \theta)).
$$

In the multivariate-Gaussian case, where $D_j = \mathcal{N}(\mu_j, Q_j), j = 1, \ldots, K$ and $\theta = (\mu_1, Q_1, \ldots, \mu_K, Q_k)$, we find

$$
\hat{\mu}_j^{(t+1)} = \frac{\sum_{i=1}^{n} X_i P_{i,j}^{(t)}}{\sum_{i=1}^{n} P_{i,j}^{(t)}}, \quad \hat{Q}_j^{(t+1)} = \frac{\sum_{i=1}^{n} (X_i - \hat{\mu}_j^{(t+1)})(X_i - \hat{\mu}_j^{(t+1)})^T P_{i,j}^{(t)}}{\sum_{i=1}^{n} P_{i,j}^{(t)}}.
$$

1.6 Maximum a posteriori (MAP)

Maximum likelihood estimates do not make prior assumptions about the parameters θ. So if we flip a coin and obtain 7 heads and 3 tails, the maximum likelihood estimate suggests that the probability of the coin showing heads is 0.7. This is perhaps a little unreasonable given our prior knowledge of a typical coin and previous observations of coin flips. The method of maximum a posteriori (MAP) estimate incorporates our existing knowledge and beliefs of the parameters θ in the form of a prior distribution

$p(\theta)$. Then, instead of maximising the likelihood $p(X \mid \theta)$, we maximise the posterior distribution

$$p(\theta \mid X) = \frac{p(X \mid \theta)p(\theta)}{p(X)} \propto p(X \mid \theta)p(\theta).$$

Formally speaking, the MAP estimate is defined to be

$$\hat{\theta}_{MAP} = \arg\max_{\theta} p(\theta \mid X) = \arg\max_{\theta} p(X \mid \theta)p(\theta).$$

Example 1.19. Let the probability of a coin toss resulting in heads be θ with a Beta distribution as a prior $\theta \sim Beta(\alpha, \beta)$. In particular larger values of α and β enforces a stronger prior belief on the value of θ. For example we can let $\alpha - 1$ and $\beta - 1$ be the previously observed amounts of heads and tails. If $\alpha = \beta = 1$ then θ has a uniform distribution and the MAP estimate is equivalent to the ML estimate.

Now suppose 7 heads were obtained from 10 coin tosses, then the posterior is given by

$$p(\theta \mid X) \propto p(X \mid \theta)p(\theta) = \binom{10}{7}\theta^7(1-\theta)^3\frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha,\beta)} \propto \theta^{7+\alpha-1}(1-\theta)^{3+\beta-1}$$

Maximising the log-likelihood, we obtain

$$\hat{\theta}_{MAP} = \frac{7+\alpha-1}{10+\alpha+\beta-2}.$$

So if α and β are both equal to the same large number (i.e., we have a strong belief that the coin is fair), then $\hat{\theta}_{MAP}$ will still be very close to 0.5.

The EM algorithm can also be used to find MAP estimates. In order to maximise $p(\theta \mid X) \propto p(X \mid \theta)p(\theta)$, we simply modify the definition of $Q(\theta \mid \hat{\theta}^{(t)})$ in the E-step to

$$Q(\theta \mid \hat{\theta}^{(t)}) := \mathbb{E}(\ln p(\theta \mid X, Y) \mid X, \hat{\theta}^{(t)}) \propto \mathbb{E}(\ln p(X, Y \mid \theta) \mid X, \hat{\theta}^{(t)}) + \ln p(\theta),$$

and then proceed with the M-step as per usual. Using arguments similar to the ML estimate case, monotonicity and stationarity results can also be established for the MAP estimate.