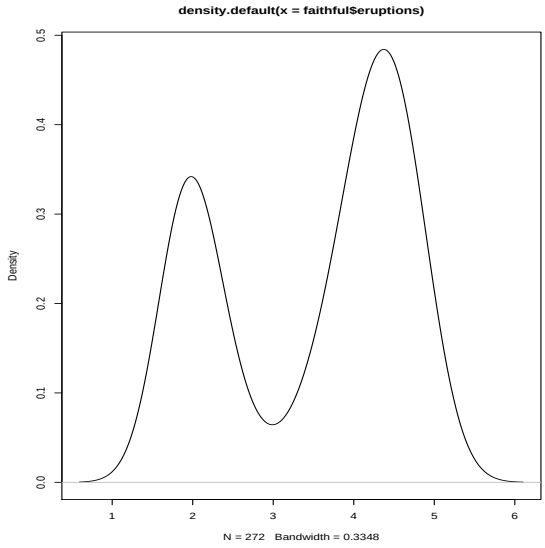
5.6 EM Algorithm

To solve any ML-type problem, we analytically maximize the likelihood function. It works well for the simple distributions, such as Poisson, Binomial, Normal, Gamma distributions. Sometimes, we use numerical methods to find such estimators. In some applications, the distribution of the data may not be well-behaved, or have too many parameters.

Example 5.6.1 (Old Faithful Geyser Data)**.** *In the faithful data set implemented in R, it contains waiting time between eruptions and the duration of the eruption for the Old Faithful geyser in Yellowstone National Park, Wyoming, USA. One would like to model the time of eruptions.*

- *eruptions, Eruption time in mins*
- *waiting, Waiting time to next eruption (in mins)*



All our previous models fail. It seems reasonable to assume the density as a mixture of two Gaussian populations:

$$f(x) = \pi_0 N(\mu_0, \sigma_0^2) + (1 - \pi_0) N(\mu_1, \sigma_1^2).$$

- MOM: need five moments and equations;
- MLE: no explicit formula, Newton-Raphson algorithm is too complicated.

Try the MLE approach.

For each subject, if we know the group of this subject, denoted as $Y_i \in \{0, 1\}$. Then the problem is much easier, we can break the data into two groups. However, $\boldsymbol{Y}$ is missing/hidden/not observed. How to proceed? One of the solutions is to use the Expectation-Maximization algorithm.

The Expectation-Maximization (EM) algorithm has two main applications. The first case occurs when the data has missing values due to limitations or problems with the observation process. The second case occurs when the likelihood function can be obtained and simplified by assuming that there is an additional but missing parameters.

For the EM algorithm, we consider two set of variables:

- Observed variables: directly measurable from the data, e.g.
  - The waveform values of a speech recording;
  - Is it raining today?
  - The eruption time of the Faithful geyser;
- Hidden variables: influence the data, but not trivial to measure, e.g.
  - The phonemes that produce a given speech recording
  - P(rain today—rain yesterday)
  - Activities under the earth;

In this example, we will use the waiting time as the data. Given a statistical model consisting of a set of $\boldsymbol{X}$ as the observations, which is incomplete. Let $\boldsymbol{Z}$ be a set of unobserved latent data or missing values which indicates a long/short waiting, the missing information (short/long waiting). Let $\boldsymbol{\theta}$ be a vector of unknown parameters. Let $\boldsymbol{Y} = (\boldsymbol{X}, \boldsymbol{Z})$ be the completed data. Then the complete likelihood function is

$$L(\boldsymbol{\theta}; \boldsymbol{X}, \boldsymbol{Z}) = f(\boldsymbol{X}, \boldsymbol{Z}|\theta).$$

Note that this function is in fact a random variable since the missing information $\boldsymbol{Z}$ is unknown, random, and presumably governed by an underlying distribution.

The marginal likelihood of the observed data $\boldsymbol{X}$ becomes

$$L(\boldsymbol{\theta}; \boldsymbol{X}) = f(\boldsymbol{X}|\boldsymbol{\theta}) = \sum_{\boldsymbol{Z}} f(\boldsymbol{X}, \boldsymbol{Z}|\boldsymbol{\theta}).$$

This also refers to as the incomplete-data likelihood function. However, this summation contains k^n terms which is intractable when n is large. The exact calculation is extremely difficult.

The EM algorithm first finds the expected value of the complete-data log-likelihood w.r.t. the unknown data $\boldsymbol{Z}$ given the observed data $\boldsymbol{X}$ and the current parameter estimates; and then find the parameter which maximizes the previous expectation.

Usually, the EM algorithm [1] seeks to find the MLE of the marginal likelihood by iteratively applying the following two steps:

- **Expectation Step (E step):** Calculate the expected value of the log likelihood function, with respect to the conditional distribution of $\boldsymbol{X}$ given $\boldsymbol{X}$ under the current estimate of the parameter $\boldsymbol{\theta}^t$:

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^t) = E_{\boldsymbol{Z}|\boldsymbol{X},\boldsymbol{\theta}^t}\left[logL(\boldsymbol{\theta};\boldsymbol{X},\boldsymbol{Z})\right].$$

- **Maximization step (M step):** Find the parameter that maximizes this quantity:

$$\boldsymbol{\theta}^{t+1} = argmax_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^t).$$

Example 5.6.2 (Gaussian Mixture)**.** *Let $\boldsymbol{x} = (x_1, x_2, \cdots, x_n)$ be a sample of n independent observations from a mixture of two normal distributions and $\boldsymbol{z} = (z_1, \cdots, z_n)$ be the latent variables that determine the component from which the observation originates. Model the observations as following:*

$$x_i|z_i = 0 \sim N(\mu_0, \sigma_0^2), x_i|z_i = 1 \sim N(\mu_1, \sigma_1^2),$$

where

$$P(z_i = 0) = \pi_0, P(z_i = 1) = \pi_1 = 1 - \pi_0.$$

[1] http://en.wikipedia.org/wiki/Expectation%E2%80%93maximization_algorithm

Let $\boldsymbol{\theta} = (\pi_0, \mu_0, \sigma_0^2, \mu_1, \sigma_1^2)$.

Then the likelihood function is

$$L(\theta; \boldsymbol{x}, \boldsymbol{z}) = \prod_{i=1}^{n} \sum_{j=0}^{1} 1(z_i = j)\pi_j f(x_i|\mu_j, \sigma_j^2).$$

Given the current values θ^t of the parameters, then

$$T_{j,i}^t = P(Z_i = j|\boldsymbol{x}, \theta^t) = \frac{\pi_j^t f(x_i; \mu_j^t, \sigma_j^{2,t})}{\pi_0^t f(x_i; \mu_0^t, \sigma_0^{2,t}) + \pi_1^t f(x_i; \mu_1^t, \sigma_1^{2,t})};$$

Next, we need to calculate the Q-function,

$$\begin{aligned} & Q(\boldsymbol{\theta}|\boldsymbol{\theta}^t) = E[\log L(\theta; \boldsymbol{x}, \boldsymbol{z})] \\ = \; & E \sum \log L(\theta; x_i, z_i) \\ = \; & \sum_i \sum_j T_{j,i}^t \log(\pi_j f(x_i|z_i = j, \boldsymbol{\theta}^t)) \\ = \; & \sum_i \sum_j T_{j,i}^t (\log \pi_j - \log(\sqrt{2\pi\sigma_j^2}) - \frac{(x_i - \mu_j)^2}{\sigma_j^2}). \end{aligned}$$

M-step:

1. Estimate $\pi_j^{t+1} = \frac{\sum_i T_{j,i}^t}{\sum_i T_{0,i}^t + \sum_i T_{1,i}^t} = \sum_i T_{j,i}^t$;
2. To maximize over μ_0, σ_0^2, only focus on $\sum_i T_{0,i}^t(-\log(\sqrt{2\pi}\sigma_0^2) - \frac{(x_i-\mu_0)^2}{\sigma_0^2})$, then
$$\mu_0^{t+1} = \frac{\sum_i T_{0,i}^t x_i}{\sum_i T_{0,i}^t},$$
and
$$\sigma_0^{2,t+1} = \frac{\sum_i T_{0,i}^t (x_i - \mu_0^{t+1})^2}{\sum_i T_{0,i}}.$$
3. Estimate $\mu_1^{t+1}, \sigma_1^{2,t+1}$ similarly;
4. Replicate steps (1-3) until convergence.

- EM algorithm is frequently used for data clustering in machine learning

and computer vision, such as in the hidden Markov model, the famous Baum-Welch algorithm is derived from the EM algorithm;

- It gains much attention in recent years in the multiple hypothesis testing [Sun and Cai(2009), Liu et al.(2014)Liu, Sarkar, and Zhao].
- With the ability to deal with missing data and observe unidentified variables, EM is a useful tool to price and manage risk of a portfolio;
- The EM is widely used in medical image reconstruction.

5.7 Kernel Density Estimation

In the previous lectures, we have assume that the density function follows a certain parametric form and the likelihood function is known. When assuming the parametric model, if the model is indeed correct, the estimation is very accurate; however, if the model is wrong, then the estimation can go very badly. In a strong parametric model, the class of distributions that is under the consideration is limited.

In the next two sections, we consider the non-parametric estimation. We put little/no assumption on the underlying distribution and use the data to estimate the density function (Section 5.7) or the cumulative distribution function (Section 5.8).

Kernel Density Estimation (KDE) is a non-parametric way to estimate the pdf of a random variable. In signal processing and econometrics, it is also termed the Parzen-Rosenblatt window method. Let $(x_1, x_2, \cdots, x_n)$ be an i.i.d. sample drawn from some distribution with an unknown density $f(x)$. The KDE of $f(x)$ is given as

$$\hat{f}_h(x) = \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{x - x_i}{h}), \tag{5.1}$$

where $K(\cdot)$ is the kernel and $h > 0$ is a smoothing parameter called the bandwidth. A kernel with subscript h is called the scaled kernel and defined as $K_h(x) = \frac{1}{h}K(x/h)$. Intuitively, one wants to choose h as small as the data allow, however, there is always a trade-off between the bias of the estimator and its variance.

A Kernel is a (non-negative) real-valued integrable function K satisfying the following property:

1. $\int K(x)dx = 1$;
2. $K(-u) = K(u)$.

A range of kernel functions are commonly used: uniform, triangular, biweight, triweight, Epanechnikov, normal, and others. See Figure 5.2.

- Uniform: $K(u) = \frac{1}{2}1(|u| \leq 1)$;
- Epanechnikov: $K(u) = \frac{3}{4}(1 - u^2)1(|u| \leq 1)$;
- Normal: $K(u) = \frac{1}{\sqrt{2\pi}}\exp(-\frac{1}{2}u^2)$.

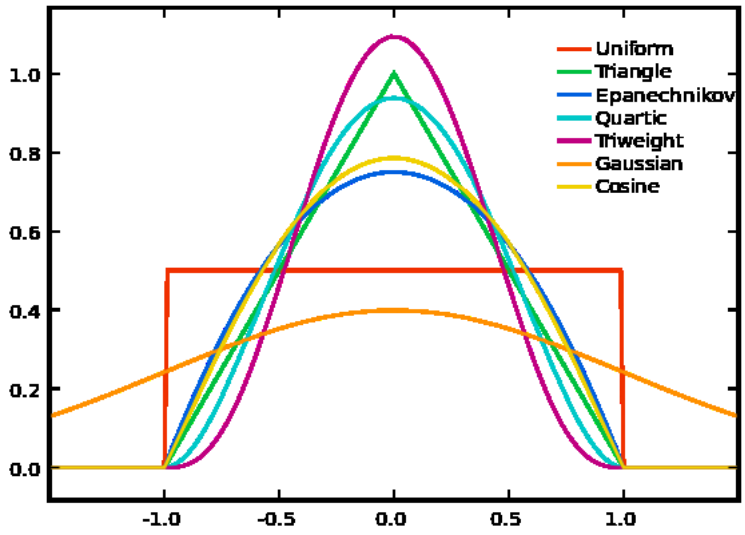


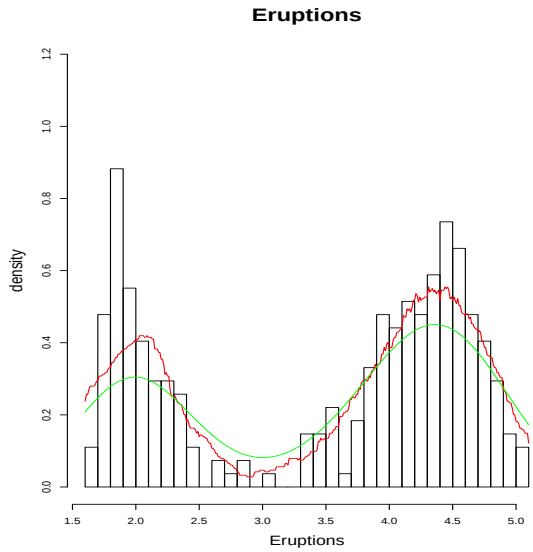
Figure 5.2: Various kernels.

Example 5.7.1 (Geyser)**.**

- Try the uniform kernel;

$$\hat{f}_h(x) = \frac{1}{nh} \sum_i 1(|\frac{x - X_i}{h}| \leq 1).$$

- Try the normal kernel.



How to assess the estimation of pdf? [Wasserman(2006)]
The most common optimality criterion is the expected L_2 risk function, also called the mean integrated squared error:

$$MISE(h) = E \int (\hat{f}_h(x) - f(x))^2 dx.$$

Bias-variance decomposition:

$$Bias(\hat{f}_h(x)) = \frac{h^2}{2} \mu_2(K) f''(x) + o(h^2),$$

and

$$Var(\hat{f}_h(x)) = \frac{1}{nh} R(K) f(x) + o(\frac{1}{nh}),$$

where $R(K) = \int K^2(y) dy$.

When h is large, the variance will decrease, but the biased will increase. If h is too large, this results in the over-smooth. When h is small, the bias will decrease, but the variance will increase. If h is too small, this results in a "choppy" estimator.

Asymptotically, the optimal bandwidth h_{opt} should be choose as $O(\frac{1}{n^{1/5}})$. In practice, how to choose the bandwidth?

1. The natural way for choosing h is to plot out several curves and choose the estimate that best matches one's prior ideas;

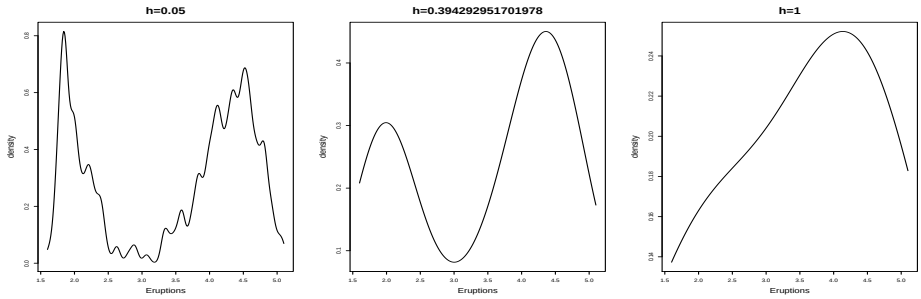


Figure 5.3: h= 0.05, 0.39, and 1

2. Silverman's "Rule of thumb" [Silverman(1986)]

$$h \approx 1.06\hat{\sigma}n^{-1/5},$$

where $\hat{\sigma}$ is the standard deviation of the samples. This is the default option of for density() in R. Some people suggest other estimate of $\hat{\sigma}$ based on the inter quantile range (IQR).

3. Maximum likelihood cross validation. This method was proposed by [Duin(1976)]. They proposed to choose h so that the pseudo-likelihood $\prod_i \hat{f}_h(X_i)$ is maximized. However this has a trivial maximum at $h = 0$. So the cross-validation principle is invoked by replacing $\hat{f}_h(x)$ by the leave-one-out $\hat{f}_{h,i}(x)$, where

$$\hat{f}_{h,i}(X_i) = \frac{1}{(n-1)h}\sum_{j\neq i} K(\frac{X_j - X_i}{h}).$$

Define that h as good which approaches the finite maximum of

$$h_{mlcv} = argmax_{h>0} MLCV(h),$$

where

$$MLCV(h) = \frac{1}{n}\sum_i \log(\sum_{j\neq i} K(\frac{X_j - X_i}{h})) - log((n-1)h).$$

The function h.mlcv from the package "kedd" computes the maximum likelihood CV for bandwidth selection [Guidoum(2014)].

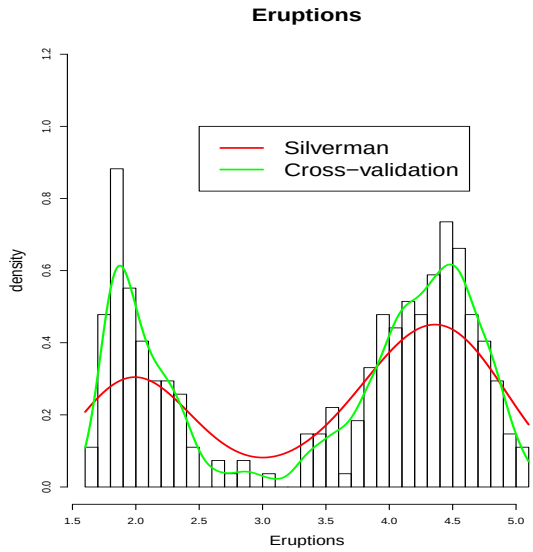


Figure 5.4: Compare h of Silverman and CV.

```
density(x, bw = "nrd0", adjust = 1,
           kernel = c("gaussian", "epanechnikov", "rectangular",
                      "triangular", "biweight",
                      "cosine", "optcosine"),
           weights = NULL, window = kernel, width,
           give.Rkern = FALSE,
           n = 512, from, to, cut = 3, na.rm = FALSE, ...)
```

5.8 Empirical distribution function

So far, we consider the parametric/non-parametric estimation of the density function of the model. The distribution can further be described by its cumulative distribution function. Can we estimate the cdf function directly?

The empirical distribution function, or empirical cdf is the cumulative distribution function associated with the empirical measure of the measure. We can use this function as an estimation of the true cdf.

Definition 5.8.1 (Empirical CDF[Van der Vaart(2000)]). *Let $(x_1, \cdots, x_n)$ be iid real random variables with the common cdf $F(t)$. Then the empirical cdf is defined as*

$$\hat{F}_n(t) = \frac{\textit{number of elements in the sample } \leq t}{n} = \frac{1}{n}\sum_i 1(x_i \leq t).$$

For a given t, $1(X_i \leq t)$ can be viewed as a Bernoulli random variable with parameter $p = F(t)$. Hence, $n\hat{F}_n(t)$ is a binomial r.v. with mean $nF(t)$ and variance $nF(t)(1-F(t))$. Consequently, $\hat{F}_n(t)$ is an unbiased estimator for $F(t)$.

Theorem 5.8.1 (Dvoretzky-Kiefer-Wolfowitz inequality [Dvoretzky et al.(1956)Dvoretzky, Kiefer, and Wolf

$$P(\sqrt{n} \cdot sup_t|\hat{F}_n(t) - F(t)| > z) \leq 2\exp(-2z^2).$$

Example 5.8.1 (Faithful)**.** *Estimate the empirical CDF.*

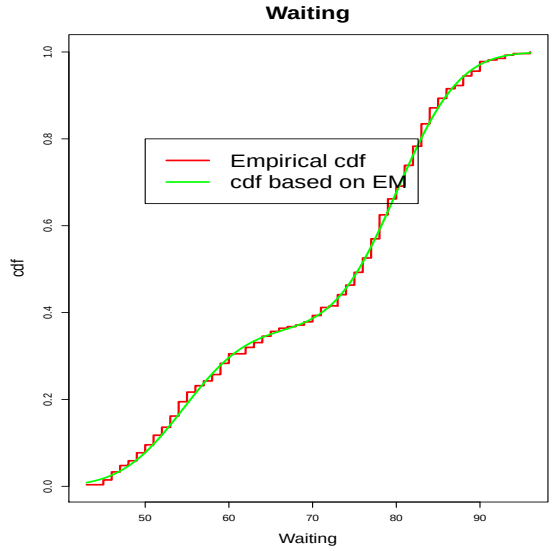


Figure 5.5: Empirical CDF and the cdf based on the EM algorithm.

What is the application of empirical cdf?

- We can use empirical cdf to test whether certain data set follows a given distribution. For instance, in the regression analysis, we usually requires the normality assumption on the data. This assumption can be tested using the Kolomogorov-Smirnov test, which will be introduced in the next chapter.
- The empirical cdf is closely related to the control of false discovery rate in multiple testing. This is the hottest area in the last decades.