# Density estimation

Objective is to find

$f(\underset{\sim}{x})$ - density of feature vector

## Approaches

(i) non-parametric approach

(ii) parametric approach

## Non-parametric density estimation methods

(A) Histogram method

For one-dimensional data:

$$\hat{p}(x) = \frac{n_j}{\left(\sum_{j=1}^{N} n_j\right) dx}$$

Where, $n_j$ : # of samples in the histogram cell of width $dx$ the contains the pt $x$

$N$ : # of cells in the histogram

$dx$ : width of the cell

For multidimensional feature vector:

$$\hat{p}(\underset{\sim}{x}) = \frac{n_j}{\left(\sum_{j=1}^{N} n_j\right) dV}$$

$dV$ : volume of the $j^{th}$ bin

(B) K-nearest neighbor method

Note that if $X$ is a r.v. (cont type),

$$P(x \leq X \leq x+\Delta x) = F(x+\Delta x) - F(x)$$

$$\lim_{\Delta x \to 0} \frac{F(x+\Delta x) - F(x)}{\Delta x} = p(x) \quad \text{p.d.f}$$

i.e. $F(x+\Delta x) - F(x) = P(x \leq X \leq x+\Delta x)$

$\approx p(x)\, \Delta x$ for small $\Delta x$

For multivariate set up $\underset{\sim}{X}$ with p.d.f $p(\underset{\sim}{x})$

$P(\underset{\sim}{X}$ will fall in a given region $C$, centered at say $\underset{\sim}{x})$

$$= \int_C p(\underset{\sim}{x})\, d\underset{\sim}{x} \approx V(C)\, p(\underset{\sim}{x})$$ , if we assume $C \ni$ volume of $V(C)$ is small and $p(\underset{\sim}{x})$ does not vary appreciably within region $C$

Let $\theta = V(C)\, p(\underset{\sim}{x})$

Realize that $\theta$ can also be approximated by the proportion of samples falling within $C$

i.e $\theta \approx \dfrac{K}{n}$ ;

where ; $K$, the number of samples falling within $C$ out of total $n$ samples

i.e. $\dfrac{K}{n} \approx p(\underset{\sim}{x})\, V$

$$\Rightarrow \hat{p}(\underset{\sim}{x}) = \frac{K}{n\,V}$$

K-nearest neighbor approach fixes $K$ and then determines the volume $V$ which contains $K$ samples centered at the point $\underset{\sim}{x}$.

If $\underset{\sim}{x}_K$ is the $K^{th}$ nearest neighbor pt to $\underset{\sim}{x}$, then $C$ may be taken to be sphere centered at $\underset{\sim}{x}$ with radius $\|\underset{\sim}{x} - \underset{\sim}{x}_K\|$. The volume of such a sphere in $p$ dimension is $2\, r^p \, \pi^{p/2} / p\, \Gamma^{p}/_{2}$

**Remark**: This approach is different from the histogram approach wherein bin size is fixed.

(C) Kernel methods (Parzen methods)

Consider a 1-dimensional sample, $x_1, \ldots x_n$

An estimate of cumulative dist$^n$ f$^n$ at $x$ is

$$\hat{F}(x) = \frac{\#\text{ of observations} \le x}{n}$$

estimate of p.d.f at $x$:

$$\hat{p}(x) = \frac{\hat{F}(x+h) - \hat{F}(x-h)}{2h}$$

proportion of obsns falling within the interval $[x-h, x+h]$ $/2h$

i.e. using a Kernel $f^n$ (rectangular Kernel)

$$K(z) = \begin{cases} \frac{1}{2}, & |z| \le 1 \\ 0, & o/w \end{cases}$$

$$\hat{p}(x) = \frac{1}{nh} \underbrace{\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)}_{= \frac{1}{2}(\#\text{ of observations with } h \text{ distance from } x)}$$

i.e. pts within h distance from x contintribute $\frac{1}{2nh}$ to the density and pts outside this distance contribute 0.

Remark: 'h' is referred to as spread or smoothing parameter (or bandwidth)

Remark: Examples of popular univariate ~~a~~ Kernel $f^n$s.

(i) ~~Rect~~ Rectangular: $K(z) = \begin{cases} 1/2, & |z| \le 1 \\ 0 & o/w \end{cases}$

(ii) Triangular: $K(z) = \begin{cases} 1 - |z|, & \text{for } |z| \le 1 \\ 0, & o/w \end{cases}$

(iii) Gaussian: $K(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \quad \forall z$

(iv) Bi-weight/quartic: $K(z) = \begin{cases} \frac{15}{16}(1 - z^2)^2, & |z| \le 1 \\ 0, & o/w. \end{cases}$

(v) Bartlett - Epanechnikov :

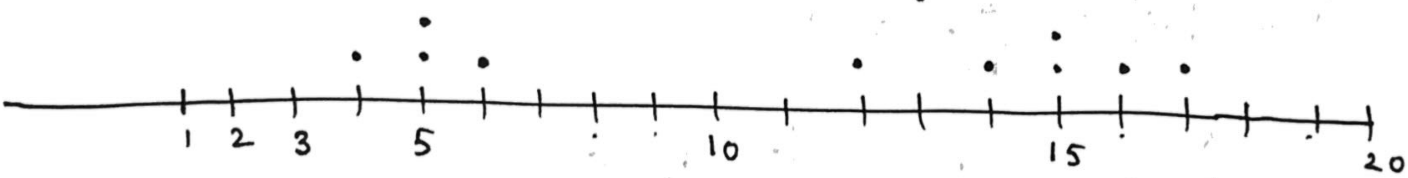
$$K(z) = \begin{cases} \frac{3}{4}\left(1 - z^2/5\right)/\sqrt{5}, & |z| \le \sqrt{5} \\ 0, & o/w \end{cases}$$

Example : $\mathcal{X} = \{4, 5, 5, 6, 12, 14, 15, 15, 16, 17\}$

$n = 10$ samples

(a) Knn density estimate with $K = 4$

$$\hat{p}(3) = \frac{4}{10}\left(V_4(3)\right)^{-1} \qquad \left( \hat{p}(x) = \frac{K}{n} V^{-1} \right)$$



For $V_4(3)$, $r = 3 \Rightarrow V_4(3) = 2r = 6$

$$\hat{p}(3) = \frac{4}{10 \times 6} = \frac{1}{15}$$

$$\hat{p}(10) = \frac{4}{10}\left(V_4(10)\right)^{-1}$$

For $V_4(10)$, $r = 5 \Rightarrow V_4(10) = 10$

$$\Rightarrow \hat{p}(10) = \frac{4}{10 \times 10} = \frac{1}{25}$$

Sly $\hat{p}(15) = \frac{4}{10 \times 2} = \frac{1}{5}$ $(r = 1)$

. . . .

Kernel density estimate with rectangular Kernel

with $\underline{h = 4 \text{ bandwidth}}$

$$\hat{p}(x) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) ; \quad K(z) = \begin{cases} 1/2, & |z| \leq 1 \\ 0, & o/w \end{cases}$$

i.e $$\hat{p}(x) = \frac{1}{nh} \sum_{i=1}^{n} \left(\frac{1}{2} I_{(|x - x_i| \leq h}\right)$$

e.g.

$$\hat{p}(3) = \frac{1}{10 \times 4} \left( \sum_{i=1}^{10} \frac{1}{2} I_{(|3 - x_i| \leq 4)} \right)$$

i.e. $$\hat{p}(3) = \frac{1}{40} \left( \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + \frac{1}{2} + 0 + 0 + \cdots + 0 \right)$$

i.e $$\hat{p}(3) = \frac{2}{40} = \frac{1}{20}$$

$$\hat{p}(10) = \frac{1}{10 \times 4} \left( \sum_{i=1}^{10} \frac{1}{2} I_{(|10 - x_i| \leq 4)} \right)$$

i.e $$\hat{p}(10) = \frac{1}{40} \times \frac{3}{2} = \frac{3}{80}.$$

$$\hat{p}(15) = \frac{1}{40} \times \left(6 \times \frac{1}{2}\right) = \frac{3}{40}$$

- - -

# Multivariate Kernel density estimate

**Approach I** : Assume independence of the component variables and estimate univariate Kernel density estimates. for the components and get

$$\hat{p}(\underset{\sim}{x}) = \prod_{i=1}^{p} \hat{p}_i(x_i)$$

**Approach II** : Generalization of univariate approach for multivariate case.

$$\hat{p}(\underset{\sim}{x}) = \frac{1}{n \cancel{p} h^p} \sum_{i=1}^{n} K\left(\frac{\underset{\sim}{x} - \underset{\sim}{x}_i}{h}\right)$$

i.e. $$\hat{p}(\underset{\sim}{x}) = \frac{1}{n h^p} \sum_{i=1}^{n} K\left(\frac{x_1 - x_{i_1}}{h}, \ldots, \frac{x_p - x_{i_p}}{h}\right)$$

**Remark** : A more general form is using different bandwidth

$$\hat{p}(\underset{\sim}{x}) = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\prod_{j=1}^{p} h_j} K\left(\frac{x_1 - x_{i_1}}{h_1}, \ldots, \frac{x_p - x_{i_p}}{h_p}\right)$$

**Remark** : A simple approach is to use a product Kernel

$$\hat{p}(\underset{\sim}{x}) = \frac{1}{n h^p} \sum_{i=1}^{n} \left( \prod_{j=1}^{p} \tilde{K}\left(\frac{x_j - x_{i_j}}{h}\right)\right)$$

or $$\frac{1}{n} \sum_{i=1}^{n} \left( \prod_{j=1}^{p} \frac{1}{h_j} \tilde{K}\left(\frac{x_j - x_{i_j}}{h_j}\right)\right).$$

using diff bandwidth

or $\frac{1}{n}\sum_{i=1}^{n}\left(\prod_{j=1}^{p}\frac{1}{h_j}\tilde{K}_j\left(\frac{x_j - x_{ij}}{h_j}\right)\right)$

using diff bandwidth & diff Kernel $f^n$

$\tilde{K}$ (or $\tilde{K}_j$) is taken as one of the univariate Kernels discussed earlier

<u>Remark</u>: Alternatively, one can use a genuine multivariable kernel

e.g. multivariate Gaussian Kernel

$$K(\underset{\sim}{y}) = \frac{1}{(2\pi)^{p/2}}\exp\left(-\frac{1}{2}\underset{\sim}{y}'\underset{\sim}{y}\right)$$

multivariate Epanechnikov Kernel, multivariate quartic kernel are other choices of mult Kernel.

Epanechnikov Kernel

$$K(\underset{\sim}{y}) = \begin{cases}(1-\underset{\sim}{y}'\underset{\sim}{y})(p+2)/2c_p & \text{for } |\underset{\sim}{x}| \leq 1\\ 0, & o/w\end{cases}$$

$$c_p = \pi^{p/2}\Big/\Gamma(p/2+1) = 2\pi^{p/2}\Big/p\,\Gamma p/2$$

## Parametric density estimation

Most commonly used assumption : multivariate Gaussian or a mixture of mult. Gaussian

Multivariate Gaussian :

$\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n$ realizations of $N_p(\underline{\mu}, \Sigma)$ ; $\Sigma > 0$

Use $\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n$ to find $\hat{f}(\underset{\sim}{x})$ $\underset{\sim}{x} \in \mathbb{R}^p$.

$$f(\underline{x}) = (2\pi)^{-p/2} |\Sigma|^{-1/2} \exp\left(-\tfrac{1}{2}(\underline{x} - \underline{\mu})' \Sigma^{-1} (\underline{x} - \underline{\mu})\right)$$

$\theta = (\underset{\sim}{\mu}, \Sigma)$ set of unknown parameters

Likelihood $f^n$

$$L(\theta) = \prod_{j=1}^{n} f(\underline{x}_j)$$

$$L(\theta) = (2\pi)^{-np/2} |\Sigma|^{-n/2} \exp\left(-\tfrac{1}{2}\sum_{j=1}^{n}(\underline{x}_j - \underline{\mu})' \Sigma^{-1} (\underline{x}_j - \underline{\mu})\right)$$

Note that

$$\sum_{j=1}^{n} (\underline{x}_j - \underline{\mu})' \Sigma^{-1} (\underline{x}_j - \underline{\mu})$$

$$= \sum_{j=1}^{n} (\underline{x}_j - \bar{\underset{\sim}{x}} + \bar{\underset{\sim}{x}} - \underline{\mu})' \Sigma^{-1} (\underline{x}_j - \bar{\underset{\sim}{x}} + \bar{\underset{\sim}{x}} - \underline{\mu})$$

$$= \sum_{j=1}^{n} (\underline{x}_j - \bar{\underset{\sim}{x}})' \Sigma^{-1} (\underline{x}_j - \bar{\underset{\sim}{x}}) + n(\bar{\underset{\sim}{x}} - \underline{\mu})' \Sigma^{-1} (\bar{\underset{\sim}{x}} - \underline{\mu})$$

$$+ 2 \underbrace{\cancel{\sum_{j=1}^{n} (\underline{x}_j - \bar{\underset{\sim}{x}})' \Sigma^{-1} (\bar{\underset{\sim}{x}} - \underline{\mu})}}_{0}$$

$$= \sum_{j=1}^{n} (\underline{x}_j - \bar{\underset{\sim}{x}})' \Sigma^{-1} (\underline{x}_j - \bar{\underset{\sim}{x}}) + n(\bar{\underline{x}} - \underline{\mu})' \Sigma^{-1} (\bar{\underline{x}} - \underline{\mu})$$

$$= tr\left( \sum_{j=1}^{n} (\underline{x}_j - \bar{\underline{x}})' \Sigma^{-1} (\underline{x}_j - \bar{\underline{x}}) \right) + n(\bar{\underset{\sim}{x}} - \underline{\mu})' \Sigma^{-1} (\bar{\underset{\sim}{x}} - \underline{\mu})$$

$$= \sum_{j=1}^{n} tr (\underline{x}_j - \bar{\underline{x}})' \Sigma^{-1} (\underline{x}_j - \bar{\underset{\sim}{x}}) + n(\bar{\underline{x}} - \underline{\mu})' \Sigma^{-1} (\bar{\underset{\sim}{x}} - \underline{\mu})$$

$$= \sum_{j=1}^{n} tr\, \Sigma^{-1} (\underline{x}_j - \bar{\underset{\sim}{x}})(\underline{x}_j - \bar{\underset{\sim}{x}})' + n(\bar{\underset{\sim}{x}} - \underline{\mu})' \Sigma^{-1} (\bar{\underset{\sim}{x}} - \underline{\mu})$$

$$= tr\, \Sigma^{-1} \sum_{j=1}^{n} (\underline{x}_j - \bar{\underline{x}})(\underline{x}_j - \bar{\underline{x}})' + n(\bar{\underset{\sim}{x}} - \underline{\mu})' \Sigma^{-1} (\bar{\underset{\sim}{x}} - \underline{\mu})$$

$$= tr\, \Sigma^{-1} (n-1)S + n(\bar{\underline{x}} - \underline{\mu})' \Sigma^{-1} (\bar{\underline{x}} - \underline{\mu})$$

$$= tr\, \Sigma^{-1} A + n(\bar{\underset{\sim}{x}} - \underline{\mu})' \Sigma^{-1} (\bar{\underset{\sim}{x}} - \underline{\mu})$$

$$\Rightarrow L(\theta) = (2\pi)^{-np/2} |\Sigma|^{-n/2} \exp\left( -\frac{1}{2} tr\, \Sigma^{-1} A - \frac{n}{2} (\bar{\underset{\sim}{x}} - \underline{\mu})' \Sigma^{-1} (\bar{\underset{\sim}{x}} - \underline{\mu}) \right) \quad \text{---} (*)$$

Note that for a fixed $\Sigma > 0$

$$(\bar{\underset{\sim}{x}} - \underset{\sim}{\mu})' \Sigma^{-1} (\bar{\underset{\sim}{x}} - \underline{\mu}) \geq 0 \text{ and } 0 \text{ only for } \bar{\underset{\sim}{x}} = \underset{\sim}{\mu}$$

$L(\underline{\mu}, \Sigma)$ for a fixed $\Sigma (>0)$ is max if exponent is max w.r.t $\underset{\sim}{\mu}$

i.e. if $(\bar{\underset{\sim}{x}} - \underline{\mu})' \Sigma^{-1} (\bar{\underline{x}} - \underline{\mu})$ is min w.r.t $\underline{\mu}$

i.e. if $\underset{\sim}{\mu} = \bar{\underset{\sim}{x}}$ ← indep of the fixed level of $\Sigma$

$$\Rightarrow \hat{\underset{\sim}{\mu}}_{MLE} = \bar{\underset{\sim}{X}}$$

The log-likelihood at $\underset{\sim}{\mu} = \underset{\sim}{\bar{x}}$

$$l(\hat{\underset{\sim}{\mu}}, \Sigma) = \log L(\hat{\underset{\sim}{\mu}}, \Sigma)$$

$$= -\frac{np}{2} \log 2\pi + \frac{n}{2} \log |\Sigma^{-1}| - \frac{1}{2} \operatorname{tr} \Sigma^{-1} A \quad — (*)$$

maximization of $(*)$ w.r.t. $\Sigma$ is equiv to maximization

of $\frac{n}{2} \log |\Sigma^{-1}| - \frac{1}{2} \operatorname{tr} \Sigma^{-1} A$

$$= \frac{n}{2} \log |\Sigma^{-1} A| - \frac{1}{2} \operatorname{tr} \Sigma^{-1} A - \frac{n}{2} \log |A|$$

i.e. maximization of

$$\frac{n}{2} \log |\Sigma^{-1} A| - \frac{1}{2} \operatorname{tr} \Sigma^{-1} A \quad — (**)$$

Let $\lambda_1, \ldots \lambda_p$ be eigenvalues of $\Sigma^{-1} A$

$$(**) = \frac{n}{2} \log \prod_1^p \lambda_j - \frac{1}{2} \sum_1^p \lambda_j$$

$$= \frac{n}{2} \sum_1^p \log \lambda_j - \frac{1}{2} \sum \lambda_j$$

$$= \frac{1}{2} \sum_{j=1}^p \left( n \log \lambda_j - \lambda_j \right) \quad — (***)$$

$(***)$ is maximized w.r.t $\lambda_j$ at $\lambda_j = n \; \forall n$

i.e. $\Sigma^{-1} A = P \, n I_p \, P' = n I_p$.

$\Rightarrow \Sigma^{-1} = n A^{-1}$ i.e. $\Sigma = \frac{1}{n} A$ maximizes likelihood w.r.t $\Sigma$

$$\Rightarrow \hat{\Sigma}_{MLE} = \frac{1}{n} \sum_{j=1}^n (\underset{\sim}{X}_j - \hat{\underset{\sim}{X}})(\underset{\sim}{X}_j - \hat{\underset{\sim}{X}})'$$

Based $\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n$ obtain

$$\hat{\underset{\sim}{\mu}} = \bar{\underset{\sim}{x}} \text{ and } \hat{\Sigma} = \frac{1}{n} A = S_n$$

Estimate density as

$$\hat{f}(\underset{\sim}{x}) = (2\pi)^{-p/2} |S_n|^{-1/2} \exp\left(-\frac{1}{2}(\underset{\sim}{x} - \bar{\underset{\sim}{x}})' S_n^{-1} (\underset{\sim}{x} - \bar{\underset{\sim}{x}})\right)$$

<u>Mixture Normal setup.</u>

One of the most widely used assumption

$$f(\underset{\sim}{x}) = \sum_{j=1}^{g} \pi_j \, f(\underset{\sim}{x} | \theta_j)$$

$g$ : # of mixing densities

$\pi_j$ : mixing proportion for $j$th group / component

$f(\underset{\sim}{x} | \theta_j)$ : density for $j$th component in the mixture

$j$th component $N_p(\underset{\sim}{\mu}_j, \Sigma_j)$; $j = 1(1)g$, $\Sigma_j > 0$

$$\theta_j = (\underset{\sim}{\mu}_j, \Sigma_j)$$

$$\Phi = (\pi_1, \ldots, \pi_g, \underset{\sim}{\mu}_1, \Sigma_1, \ldots, \underset{\sim}{\mu}_g, \Sigma_g)$$

unknown parameters

Likelihood $f^n$

$$L(\Phi) = \prod_{i=1}^{n} \left( \sum_{j=1}^{g} \pi_j \, f(\underset{\sim}{x}_i | \theta_j) \right)$$

<u>Remark</u> : $(\underset{\sim}{x}_1, \ldots, \underset{\sim}{x}_n)$ is incomplete data

Use E-M algorithm

$\underset{\sim}{x}$ : incomplete data without class labels

$\underset{\sim}{y}' = (\underset{\sim}{x}', \underset{\sim}{z}')$ complete data

where, $\underset{\sim}{z}$ = indicator vector of length $g$ with 1 at the $k^{th}$ position

$$g(\underset{\sim}{y} \mid \underline{\Phi}) = g(\underset{\sim}{x}, \underset{\sim}{z} \mid \underline{\Phi}) = \frac{p(\underset{\sim}{x}, \underset{\sim}{z}, \underline{\Phi})}{p(\underset{\sim}{z}, \underline{\Phi})} \cdot \frac{p(\underset{\sim}{z}, \underline{\Phi})}{p(\underline{\Phi})}$$

$$= p(\underset{\sim}{x} \mid \underset{\sim}{z}, \underline{\Phi})\, p(\underset{\sim}{z} \mid \underline{\Phi})$$

$$= p(\underset{\sim}{x} \mid \theta_K)\, \pi_K$$

i.e. $g(\underset{\sim}{y} \mid \underline{\Phi}) = \left(p(\underset{\sim}{x} \mid \theta_1)\pi_1\right)^0 \cdots \left(p(\underset{\sim}{x} \mid \theta_K)\pi_K\right)^1 \cdots \left(p(\underset{\sim}{x} \mid \theta_g)\pi_g\right)^0$

Let $z_j = \begin{cases} 1, & \text{if } j = K \\ 0, & \text{o/w} \end{cases}$

$$g(\underset{\sim}{y} \mid \underline{\Phi}) = \prod_{j=1}^{g} \left(p(\underset{\sim}{x} \mid \theta_j)\pi_j\right)^{z_j}$$

Consider for simplicity $g = 2$

$\underset{\sim}{z} = (z_1, z_2)$ $\quad z_1 = 1$ if $\underset{\sim}{x}$ corresp to component 1

$\quad z_2 = 1$ if $\underset{\sim}{x}$ corresp to component 2

Let $\pi_2 = \pi$, $\pi_1 = 1 - \pi$

$$g(\underset{\sim}{y} \mid \underline{\Phi}) = \left(p(\underset{\sim}{x} \mid \theta_1)\pi_1\right)^{z_1} \left(p(\underset{\sim}{x} \mid \theta_2)\pi_2\right)^{z_2}$$

i.e. $g(\underset{\sim}{y} \mid \underline{\Phi}) = \left(p(\underset{\sim}{x} \mid \theta_1)(1-\pi)\right)^{z_1} \left(p(\underset{\sim}{x} \mid \theta_2)\pi\right)^{z_2}$

$$g(\underset{\sim}{y}_1, \ldots, \underset{\sim}{y}_n \mid \underline{\Phi}) = \prod_{i=1}^{n} \prod_{j=1}^{2} \left(p(\underset{\sim}{x}_i \mid \theta_j)\pi_j\right)^{z_{ji}}$$

$$= \prod_{i=1}^{n} \left( \left\{ p(\underset{\sim}{x}_i \mid \theta_1)(1-\pi) \right\}^{z_{1i}} \left\{ p(\underset{\sim}{x}_i \mid \theta_2)\pi \right\}^{z_{2i}} \right.$$

Note: For a general '$g$',

$$g(\underset{\sim}{y}_1, \ldots, \underset{\sim}{y}_n \mid \Phi) = \prod_{i=1}^{n} \left( \prod_{j=1}^{g} \left( p(\underset{\sim}{x}_i \mid \theta_j) \pi_j \right)^{z_{ji}} \right).$$

log likelihood $f^n$

$$\ell(\Phi) = \log g(\underset{\sim}{y}_1, \ldots, \underset{\sim}{y}_n \mid \Phi)$$

$$= \sum_{i=1}^{n} \log \left( \left\{ p(\underset{\sim}{x}_i \mid \theta_1)(1-\pi) \right\}^{z_{1i}} \left\{ p(\underset{\sim}{x}_i \mid \theta_2)\pi \right\}^{z_{2i}} \right)$$

i.e $$\ell(\Phi) = \sum_{i=1}^{n} \Big( z_{1i} \log \left( p(\underset{\sim}{x}_i \mid \theta_1)(1-\pi) \right) + z_{2i} \log \left( p(\underset{\sim}{x}_i \mid \theta_2)\pi \right) \Big).$$

$$= \sum_{i=1}^{n} \Big( z_{1i} \log \left( p(\underset{\sim}{x}_i \mid \theta_1) \right) + z_{2i} \log \left( p(\underset{\sim}{x}_i \mid \theta_2) \right) \Big) + \sum_{i=1}^{n} \left( z_{1i} \log(1-\pi) + z_{2i} \log \pi \right)$$

Note: For a general '$g$'

$$\ell(\Phi) = \sum_{i=1}^{n} \left( \sum_{j=1}^{g} z_{ji} \log \left( p(\underset{\sim}{x}_i \mid \theta_j) \right) \right) + \sum_{i=1}^{n} \left( \sum_{j=1}^{g} z_{ji} \log \pi_j \right).$$

Remark:

Note that if $(z_{1i}, z_{2i})$ is known $\forall i$, the MLE is simple

$$\left\{ \begin{array}{l} \underset{\sim}{\mu}_1 \to \bar{X}_1 \\ \Sigma_1 \to S_1 \end{array} \right\} \text{ from all } \underset{\sim}{x}_i \ni z_{1i} = 1$$

$$\left\{ \begin{array}{l} \underset{\sim}{\mu}_2 \to \bar{\underset{\sim}{X}}_2 \\ \Sigma_2 \to S_2 \end{array} \right\} \text{ from all } \underset{\sim}{x}_i \ni z_{2i} = 1$$

But $\underset{\sim}{z}_i$ s are unknown!

## E-M algorithm steps

E - step :

$$E\left(z_{ji} \mid \underset{\sim}{x}_i, \underline{\Phi}^{(m)}\right) = P\left(z_{ji} = 1 \mid \underset{\sim}{x}_i, \underline{\Phi}^{(m)}\right) = \omega_{ji}$$

$\omega_{ji}$ : prob that $\underset{\sim}{x}_i \in$ group $j$ given current estimates $\underline{\Phi}^{(m)}$

$$\omega_{ji} = \frac{\pi_j^{(m)} \; p(\underset{\sim}{x}_i \mid \theta_j^{(m)})}{\pi_1^{(m)} p(\underset{\sim}{x}_i \mid \theta_1) + \pi_2^{(m)} p(\underset{\sim}{x}_i \mid \theta_2^{(m)})}$$

Form the $f^n$

$$Q(\underline{\Phi}, \underline{\Phi}^{(m)}) = \sum_{i=1}^{n} \left( \omega_{1i} \log\left(p(\underset{\sim}{x}_i \mid \theta_1)\right) + \omega_{2i} \log\left(p(\underset{\sim}{x}_i \mid \theta_2)\right)\right)$$

$$+ \sum_{i=1}^{n} \sum_j \omega_{ji} \log \pi_j$$

Note that $Q(\underline{\Phi}, \underline{\Phi}^{(m)}) = E\left(\log g(\underset{\sim}{y}_1, \ldots \underset{\sim}{y}_n) \mid \underset{\sim}{x}_1, \ldots \underset{\sim}{x}_n, \underline{\Phi}\right)$

M - step : Maximise $Q$ w.r.t. $\pi_i$ & $\theta_i$

Maximisation of $Q$ w.r.t. $\pi_i$ subject to $\sum \pi_j = 1$

$$\tilde{Q} = Q - \lambda\left(\sum \pi_j - 1\right)$$

$$\frac{\partial \tilde{Q}}{\partial \pi_j} = \frac{\partial}{\partial \pi_j}\left(\sum_{i=1}^{n}\sum_j \omega_{ji} \log \pi_j\right) - \frac{\partial}{\partial \pi_j}\left(\lambda\left(\sum \pi_j - 1\right)\right)$$

$$= \sum_{i=1}^{n} \frac{\omega_{ji}}{\pi_j} - \lambda = 0$$

$$\Rightarrow \lambda = \sum_{i=1}^{n} \omega_{ji} \Big/ \pi_j \quad \text{i.e. } \lambda \pi_j = \sum_{i=1}^{n} \omega_{ji}$$

$$\lambda \pi_j = \sum_i \omega_{ji}$$

$$\lambda \sum_j \pi_j = \sum_j \sum_i \omega_{ji} = \sum_{i=1}^{n} \left( \sum_j \omega_{ji} \right)$$

i.e. $\lambda = n$

$$\& \quad \hat{\pi}_j = \frac{1}{n} \sum_{i=1}^{n} \omega_{ji}$$

Also for $\theta_j$

$$\hat{\mu}_j = \frac{\sum_{i=1}^{n} \omega_{ji} \, \underset{\sim}{x}_i}{\sum_{i=1}^{n} \omega_{ji}} = \frac{1}{n \hat{\pi}_j} \sum_{i=1}^{n} \omega_{ji} \underset{\sim}{x}_i$$

$$\& \quad \hat{\Sigma}_j = \frac{\sum_{i=1}^{n} \omega_{ji} (\underline{x}_i - \underline{\hat{\mu}}_j)(\underline{x}_i - \underline{\hat{\mu}}_j)'}{\sum_{i=1}^{n} \omega_{ji}}$$

$$\text{i.e. } \hat{\Sigma}_j = \frac{1}{n \hat{\pi}_j} \sum_{i=1}^{n} \omega_{ji} (\underline{x}_i - \underline{\hat{\mu}}_j)(\underline{x}_i - \underline{\hat{\mu}}_j)'$$

The E-M algorithm alternates bet$^n$ E-step of estimating $\omega_{ji}$ and M-step of calculating $\hat{\pi}_j$, $\underline{\hat{\mu}}_j$ and $\hat{\Sigma}_j$, given $\omega_{ji}$.

The iteration continues till convergence of likelihood.

<u>Example</u> : $p = 1$ ; $g = 2$

Comp 1 : $N(\mu_1, \sigma_1^2)$ ; $\theta = (\mu_1, \sigma_1^2)$

Comp 2 : $N(\mu_2, \sigma_2^2)$ ; $\theta = (\mu_2, \sigma_2^2)$

$$\Phi = (\pi, \mu_1, \sigma_1^2, \mu_2, \sigma_2^2)$$

$$Q(\Phi, \Phi^{(m)}) = \sum_{i=1}^{n} \left( \omega_{1i} \log p(x_i|\theta_1) + \omega_{2i} \log p(x_i|\theta_2) \right)$$

<u>E-step</u> :

$$+ \sum_{i=1}^{n} \left( \sum_{j=1}^{2} \omega_{ji} \log \pi_j \right) \quad — (*)$$

$$\omega_{1i} = \frac{\pi_1^{(m)} p(x_i|\theta_1^{(m)})}{\pi_1^{(m)} p(x_i|\theta_1^{(m)}) + (1-\pi_1^{(m)}) p(x_i|\theta_2^{(m)})}$$

Starting r.h.s. can be obtained from cluster analysis output.

<u>M-step</u> : $$\hat{\pi}_j = \frac{\sum_{i=1}^{n} \omega_{ji}}{n} \quad ; \quad j = 1, 2$$

and
$$\frac{\partial Q}{\partial \mu_j} = \sum_{i=1}^{n} \omega_{ji} \frac{\partial}{\partial \mu_j} \left( -\frac{1}{2} \log 2\pi \sigma_j^2 - \frac{1}{2\sigma_j^2} (x_i - \mu_j)^2 \right)$$

$$= \sum_{i=1}^{n} \omega_{ji} \left( \frac{1}{2\sigma_j^2} (x_i - \mu_j) \right) = 0$$

i.e. $$\sum_{i=1}^{n} \omega_{ji} x_i = \mu_j \sum_{i=1}^{n} \omega_{ji}$$

$$\Rightarrow \hat{\mu}_j = \frac{\sum_i \omega_{ji} x_i}{\sum_i \omega_{ji}} = \frac{\sum_i \omega_{ji} x_i}{n \hat{\pi}_j}$$

$$\frac{\partial Q}{\partial \sigma_j^2} = \sum_{i=1}^{n} w_{ji} \frac{\partial}{\partial \sigma_j^2}\left(-\frac{1}{2}\log 2\pi - \frac{1}{2}\log \sigma_j^2 - \frac{1}{2\sigma_j^2}(x_i - \mu_j)^2\right)$$

$$= \sum_{i=1}^{n} w_{ji}\left(-\frac{1}{2\sigma_j^2} + \frac{1}{2(\sigma_j^2)^2}(x_i - \mu_j)^2\right)$$

$$\left.\begin{aligned} \frac{\partial Q}{\partial \mu_j} &= 0 \\ \frac{\partial Q}{\partial \sigma_j^2} &= 0 \end{aligned}\right\} \Rightarrow \quad \begin{aligned} \hat{\mu}_j &= \frac{1}{\sum_i w_{ji}} \sum_i w_{ji} x_i \\ \hat{\sigma}_j^2 &= \frac{1}{\sum_i w_{ji}} \sum_i w_{ji}(x_i - \hat{\mu}_j)^2 \end{aligned}$$

Start with initial $(\pi_1^{(0)}, \theta_1^{(0)}, \theta_2^{(0)}) \rightarrow$ obtain $(w_{1i}, w_{2i})$ $i = 1(1)n \longrightarrow$ obtain $(\hat{\pi}_1, \hat{\mu}_1, \hat{\sigma}_1^2, \hat{\mu}_2, \hat{\sigma}_2^2) = (\pi_1^{(1)}, \theta_1^{(1)}, \theta_2^{(1)}) \rightarrow$ alternate betn E-step & M-step till convergence of the likelihood.