Consider the following two data generating processes:

1. Generate X_i as follows:

$$Y_{i1} \sim \mathsf{N}(-1, \sigma^2), \quad Y_{i2} \sim \mathsf{N}(-1, \sigma^2), \quad X_i = \frac{1}{2}\big(Y_{i1} + Y_{i2}\big).$$

2. Generate X_i as follows:

$$Z_i = \begin{cases} 1, & \text{w/p } 1/2; \\ 2, & \text{w/p } 1/2. \end{cases} \qquad X_i \sim \begin{cases} \mathsf{N}(-1, \sigma^2), & \text{if } Z_i = 1; \\ \mathsf{N}(1, \sigma^2), & \text{if } Z_i = 2. \end{cases}$$

1. Generate X_i as follows:

$$Y_{i1} \sim \mathsf{N}(-1, \sigma^2), \quad Y_{i2} \sim \mathsf{N}(-1, \sigma^2), \quad X_i = \frac{1}{2}\big(Y_{i1} + Y_{i2}\big).$$

$X_i \sim \mathsf{N}(0, \sigma^2/2)$.

2. Generate X_i as follows:

$$Z_i = \begin{cases} 1, & \text{w/p } 1/2; \\ 2, & \text{w/p } 1/2. \end{cases} \qquad X_i \sim \begin{cases} \mathsf{N}(-1, \sigma^2), & \text{if } Z_i = 1; \\ \mathsf{N}(1, \sigma^2), & \text{if } Z_i = 2. \end{cases}$$

X_i follows a mixture distribution with pdf

$$f_X(x) = \frac{1}{2}\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x+1)^2}{2\sigma^2}\right\} + \frac{1}{2}\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-1)^2}{2\sigma^2}\right\}.$$

Mixture Distributions

- $F_1(x), \ldots, F_k(x)$: k different CDFs
- $w_1, \ldots, w_k$: k positive numbers with $\sum_{j=1}^{k} w_j = 1$
- Define a new distribution with CDF given by

$$G(x) = w_1 F_1(x) + w_2 F_2(x) + \cdots + w_k F_k(x).$$

  Easy to check that the function $G(x)$ is a valid CDF. The corresponding distribution is called a **mixture distribution** since it mixes k distributions according to weights w_1, $w_2, \ldots, w_k$.

- If the k distributions are continuous with pdfs $f_1(x), \ldots, f_k(x)$, then the pdf for the mixture distribution G is given by

$$g(x) = w_1 f_1(x) + w_2 f_2(x) + \cdots + w_k f_k(x).$$

We can generate the mixture random variable by the following two-stage process:

- Generate $Z \sim$ Multinomial$(w_1, \ldots, w_k)$, i.e.,

$$Z = j, \quad \text{w/p } w_j, \quad j = 1, 2, \ldots, k.$$

- Conditioning on Z, we can generate X,

$$X \mid Z = j \quad \sim F_j.$$

Z: Latent Variable. Latent variables refer to rvs we do not observe directly.

Inference on the Latent Variable

Recall that each observation from the two-normal mixture model is either from $\mathsf{N}(-1, \sigma^2)$ or from $\mathsf{N}(1, \sigma^2)$.

Suppose we collect a sample X from this mixture model.

- If $X = 3$, which component do you think it's from?
- If $X = -2$, which component do you think it's from?
- If $X = 0.5$, which component do you think it's from?

Consider the complete data (X, Z) although we do not observe Z. Based on two-stage data generating process, we know that their joint distribution can be factorized as

$$p(x, z) = p(z)p(x \mid z), \qquad \text{e.g., } w_j f_j(x) \text{ if } z = j.$$

To answer the questions asked before, we need to evaluate the condition distribution of Z given X.

Using the Bayes' Theorem, we have

$$p(Z = j \mid X = x) = \frac{p(x, j)}{p_X(x)} = \frac{w_j f_j(x)}{w_1 f_1(x) + w_2 f_2(x) + \cdots + w_k f_k(x)}.$$

Applications of Mixture Models

- Email spam filter,
- Market segmentation,
- Automatic news categorization,
- and many more