Recall Example 5.2 (pp. 72–73)

$X_1, X_2, ..., X_n$ independent, each $\sim$ Poisson(θv_i)

$$L(\theta) = \prod_{i=1}^{n} \frac{e^{-\theta v_i}(\theta v_i)^{x_i}}{x_i!} \quad \Rightarrow \quad \ell(\theta) = \sum_{i=1}^{n} -\theta v_i + x_i \log(\theta v_i) - \log x_i!$$

$$\ell'(\theta) = 0 \quad \Rightarrow \quad \widehat{\theta} = \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} v_i}$$

interesting "twist" (Example 5.5, p. 79–80)

$X_1, X_2, ..., X_n$ not fully observable

instead, observe only whether $X_i > 0$ or $X_i = 0$

© 2017-23 by M. Zhu, PhD

MLE with Missing Data

Components

$\boldsymbol{D}_{obs}$:	observed data
$\boldsymbol{D}_{mis}$:	unobserved data (e.g., missing, latent, ...)
$\boldsymbol{\Theta}$:	model parameters

MLE Want to

$$\max_{\boldsymbol{\Theta}} \quad \ell(\boldsymbol{\Theta}; \boldsymbol{D}_{obs}, \boldsymbol{D}_{mis})$$

but "stuck" because $\boldsymbol{D}_{mis}$ not available.

© 2017-23 by M. Zhu, PhD

EM Algorithm

E-Step Compute

$$Q(\boldsymbol{\Theta}; \boldsymbol{D}_{obs}, \widehat{\boldsymbol{\Theta}}_{t-1}) \quad = \quad \mathbb{E}_{\boldsymbol{D}_{mis}|\boldsymbol{D}_{obs};\widehat{\boldsymbol{\Theta}}_{t-1}}[\ell(\boldsymbol{\Theta}; \boldsymbol{D}_{obs}, \boldsymbol{D}_{mis})],$$

i.e., take best guess.

M-Step Compute

$$\widehat{\boldsymbol{\Theta}}_t \quad = \quad \arg\max_{\boldsymbol{\Theta}} \quad Q(\boldsymbol{\Theta}; \boldsymbol{D}_{obs}, \widehat{\boldsymbol{\Theta}}_{t-1}),$$

i.e., update parameter estimate.

Remarks E = expectation. M = maximization. Will skip theory about why it works, convergence, etc.

© 2017-23 by M. Zhu, PhD

Continuation: Example 5.6 (pp. 84–85)

$$\begin{aligned}
\ell(\theta) &= \sum_{i=1}^{n} -\theta v_i + X_i \log(\theta v_i) \\
\mathbb{E}_{\boldsymbol{D}_{mis}|\boldsymbol{D}_{obs};\widehat{\theta}_{t-1}}[\ell(\theta)] &= \sum_{i=1}^{n} -\theta v_i + [\mathbb{E}_{\boldsymbol{D}_{mis}|\boldsymbol{D}_{obs};\widehat{\theta}_{t-1}}(X_i)] \log(\theta v_i)
\end{aligned}$$

$$\mathbb{E}_{\boldsymbol{D}_{mis}|\boldsymbol{D}_{obs};\widehat{\theta}_{t-1}}(X_i) = \begin{cases} \mathbb{E}(\overset{\boldsymbol{D}_{mis}}{\overset{\downarrow}{X_i}} \mid \underbrace{X_i > 0}_{\boldsymbol{D}_{obs}}; \widehat{\theta}_{t-1}), & X_i > 0; \\ 0, & X_i = 0 \end{cases}$$

Exercise 5.11 (p. 85) Show that, if $X \sim \text{Poisson}(\lambda)$, then $\mathbb{E}(X|X > 0; \lambda) = \lambda/(1 - e^{-\lambda})$.

© 2017-23 by M. Zhu, PhD

Two-Stage Generating Mechanism

- $Z_i \in \{1, 2, \ldots, K\} \sim$ multinomial$(1; \pi_1, \pi_2, \ldots, \pi_K)$

- $X_i | (Z_i = k) \sim f(x; \theta_k)$
  - here (and often), same family f, different parameters θ_k
  - in principle, can also do f_k, but need to be careful

- but label $Z_i \in \{1, 2, ..., K\}$ unobserved

© 2017-23 by M. Zhu, PhD

Components

- parameters $\boldsymbol{\Theta}$

$$\boldsymbol{\Theta} = \{\theta_1, \ldots, \theta_K, \pi_1, \ldots, \pi_K\}$$

- observed data $\boldsymbol{D}_{obs}$

$$\boldsymbol{X} = \{X_1, \ldots, X_n\} \quad \text{or} \quad \{x_1, \ldots, x_n\}$$

- unobserved "data" $\boldsymbol{D}_{mis}$

$$\boldsymbol{Z} = \{Z_1, \ldots, Z_n\}$$

© 2017-23 by M. Zhu, PhD

MLE with Missing Data

Likelihood (pretend z_i available)

$$L(\boldsymbol{\Theta}; \boldsymbol{X}, \boldsymbol{Z}) = \prod_i f(x_i, \overset{\text{NA}}{\downarrow} z_i) = \prod_i \left\{ \prod_k \left[\underbrace{f(x_i; \theta_k)}_{f(x_i|z_i)} \overset{f(z_i)}{\downarrow} \pi_k \right]^{I(z_i=k)} \right\}$$

Log-Likelihood (remember z_i not available)

$$\ell(\boldsymbol{\Theta}; \boldsymbol{X}, \boldsymbol{Z}) = \sum_i \sum_k [I(Z_i = k)] \log [f(x_i; \theta_k)\pi_k]$$

MLE Would like to maximize $\ell(\boldsymbol{\Theta}; \boldsymbol{X}, \boldsymbol{Z})$ over $\boldsymbol{\Theta}$ but $\boldsymbol{Z}$ missing.

© 2017-23 by M. Zhu, PhD

EM Algorithm

E-Step Take a guess — suffices to compute

$$\mathbb{E}_{\boldsymbol{Z}|\boldsymbol{X};\widehat{\boldsymbol{\Theta}}_{t-1}}[I(Z_i = k)]$$

as log-likelihood is linear in $I(Z_i = k)$.

M-Step Update parameter estimate

$$\begin{aligned}
\widehat{\boldsymbol{\Theta}}_t &= \arg\max_{\boldsymbol{\Theta}} \sum_i \sum_k \underbrace{\mathbb{E}_{\boldsymbol{Z}|\boldsymbol{X};\widehat{\boldsymbol{\Theta}}_{t-1}}[I(Z_i = k)]}_{w_{ik}} \log\left[f(x_i;\theta_k)\pi_k\right] \\
&= \arg\max_{\boldsymbol{\Theta}} \sum_i \sum_k w_{ik} \log\left[f(x_i;\theta_k)\right] + \sum_i \sum_k w_{ik} \log(\pi_k).
\end{aligned}$$

© 2017-23 by M. Zhu, PhD

M-Step

- for the "θ_k part", just weighted maximum likelihood

$$\max_{\theta_k} \quad \sum_i w_{ik} \log\left[f(x_i;\theta_k)\right] \quad \forall \quad k \tag{1}$$

- for the "π_k part", usual multinomial likelihood with w_{ik}

$$\max_{\pi_1,\ldots,\pi_K} \quad \sum_i \sum_k w_{ik} \log(\pi_k) \quad \text{s.t.} \quad \sum_k \pi_k = 1 \tag{2}$$

Remark A "prototypical" model here (for $\boldsymbol{x}_i \in \mathbb{R}^d$) is to take $f(\boldsymbol{x}_i;\boldsymbol{\theta}_k)$ as $\mathrm{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$. [Fraley & Raftery (2002; *JASA*) discuss why the temptation $\mathrm{N}(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ is "problematic" and what can be done to compromise.]

© 2017-23 by M. Zhu, PhD

E-Step

Simply writing $\widehat{\boldsymbol{\Theta}}$ rather than $\widehat{\boldsymbol{\Theta}}_{t-1}$,

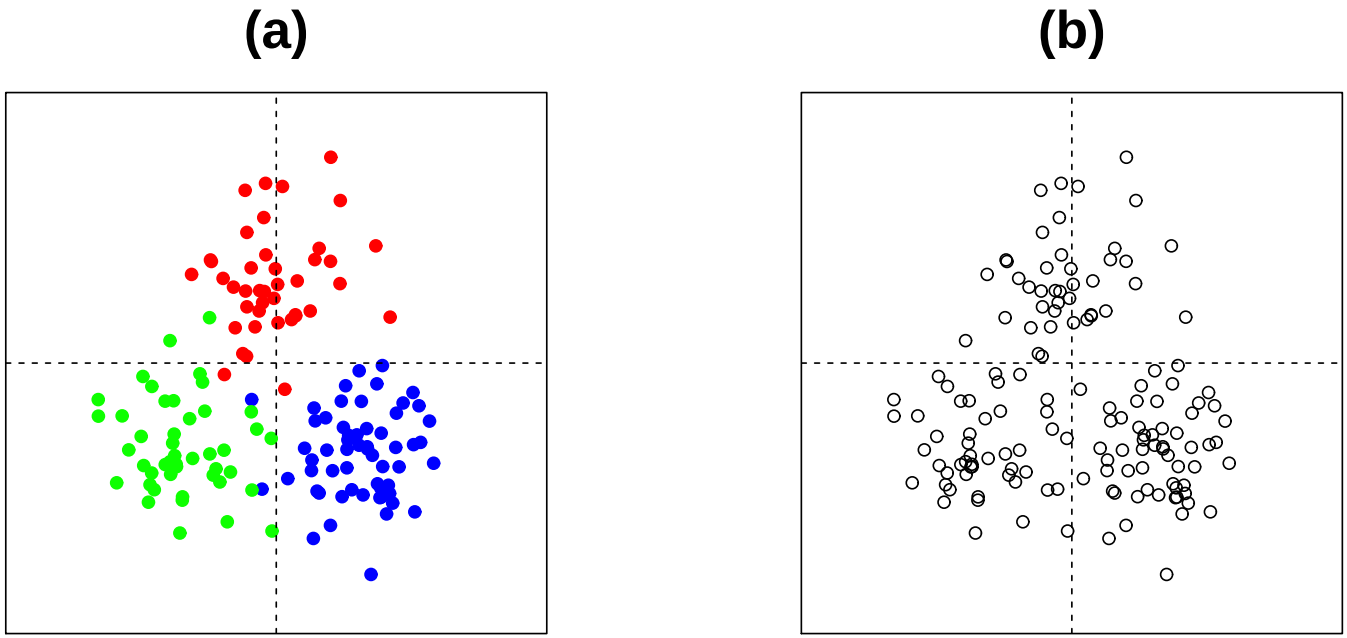
$$\begin{aligned} w_{ik} &\equiv \mathbb{E}_{\boldsymbol{Z}|\boldsymbol{X};\widehat{\boldsymbol{\Theta}}}[I(Z_i = k)] \\ &= \mathbb{P}(Z_i = k | X_i = x_i; \widehat{\boldsymbol{\Theta}}) \\ &= \frac{\widehat{\pi}_k f(x_i; \widehat{\theta}_k)}{\widehat{\pi}_1 f(x_i; \widehat{\theta}_1) + ... + \widehat{\pi}_K f(x_i; \widehat{\theta}_K)}. \end{aligned}$$

The denominator shows the marginal distribution of each X_i is a finite mixture distribution

$$f(x_i; \boldsymbol{\Theta}) = \sum_{k=1}^{K} \pi_k f(x_i; \theta_k).$$

© 2017-23 by M. Zhu, PhD

Finite Mixture Model



(a) hidden generating mechanism; (b) observed data

© 2017-23 by M. Zhu, PhD

Variation for Text Data

- each $\boldsymbol{x}_i = (x_{i1}, ..., x_{id})^\top$ a text document
- x_{ij} = count of word j in document i (bag of words)
- can use the same framework, except choose

$$f(\boldsymbol{x}_i; \boldsymbol{\theta}_k) \quad \text{as} \quad \text{multinomial}(m_i; \boldsymbol{p}_k),$$

  where $\boldsymbol{p}_k = (p_{1k}, ..., p_{dk})^\top$, and m_i = total number of words in document i (treated as fixed constants)

- many sophisticated variations/extensions, e.g., latent Dirichlet allocation (Blei, Ng & Jordon, 2003; *JMLR*), mostly because vanilla MLE does not work well — most $\widehat{p}_{jk}$ would be zero

© 2017-23 by M. Zhu, PhD

Back to Basketball Passes

- recall: given $\mathbf{Z} = \{z_1, ..., z_n\}$,

$$\mathcal{M}(\mathbf{T}|\mathbf{Z}) = \prod_{i,j} \left\{ J(\rho_{z_i z_j}) \times \prod_{h=1}^{m_{ij}} \rho_{z_i z_j}(t_{ijh}) \right\}$$

- likelihood with latent variables Z_i, Z_j (not observed)

$$\prod_{i,j} \left\{ \prod_{k,\ell} \left[J(\rho_{k\ell}) \times \prod_{h=1}^{m_{ij}} \rho_{k\ell}(t_{ijh}) \right]^{I(Z_i=k, Z_j=\ell)} \right\} \times \prod_i \prod_k \pi_k^{I(Z_i=k)}$$

- requires $\mathbb{P}(Z_i = k, Z_j = \ell|...)$ in addition to $\mathbb{P}(Z_i = k|...)$, but Z_i, Z_j conditionally dependent here

© 2017-23 by M. Zhu, PhD