Příklad 4: (Intervalové) cenzorování

Chci pozorovat $T_i \in \underbrace{(L_i, U_i]}_{\text{data}}$, $i = 1, \dots, n$

$0 \leq L_i < U_i \leq \infty$, $U_i = \infty \equiv$ cenzor. zprava

model pro T_i:

$F_i(\cdot\,;\theta) \equiv$ distrib fce, kterou předpoll. pro T_i

- nezávislé, neinformativní cenzorování

$$L_{obs}(\theta) = \prod_{i=1}^{n} P(L_i < T_i \leq U_i; \theta) =$$

$$= \prod_{i=1}^{n} \{F_i(U_i; \theta) - F_i(L_i; \theta)\} =$$

$$= \prod_{i=1}^{n} \int_{L_i}^{U_i} f_i(t; \theta)\, dt$$

$f_i \equiv$ hustota pro F_i

VIDÍME:

$L_{obs}(\theta) = p(y|\theta)$ ne vždy „hezká"

řešení: uvažuj DOPLNĚNÁ (AUGMENTED) data ≡ latentní data

motivace – hierarchický model

$$p(y|\theta) = \int p(y|\psi,\theta)\, p(\psi|\theta)\, d\psi$$

2. úroveň „dat"
(např. náhodné efekty ve smíš. modelu)

pro bayes. počítání (na základě MCMC)

(ψ, θ) = parametry

„věrohodnost" pro Bayes. větu byla

$$p(y|\psi,\theta) =: L_{augm}(\psi,\theta)$$

≡ doplněná věrohodnost

BYLO: $p(\psi,\theta) = p(\psi|\theta) \cdot p(\theta)$

muselo odpovídat původnímu modelu

$\Rightarrow p(\theta|y)$ bylo marginálním rozdělením z $p(\theta,\psi|y)$

$$p(\theta,\psi|y) \propto \underbrace{p(y|\psi,\theta)}_{L_{augm}(\psi,\theta)}\,\underbrace{p(\psi,\theta)}_{p(\psi|\theta)p(\theta)}$$

BDA (Bayesian data augmentation)

obecně

$p(y|\theta)$ daná a je „nehezká" (věrohodnost, model)

<u>BDA</u>: navrhneme ψ /doplněná/ augmented) data ~~data~~, ~~aby~~ a k nim

$p(y|\psi,\theta)$, $p(\psi|\theta)$ /2 další modely/

potřeba udělat tak, aby

$$p(y|\theta) = \int p(y|\psi,\theta)\,p(\psi|\theta)\,d\psi$$

(Potom) stejně jako výše bude

$$p(\theta|y) = \int p(\theta,\psi|y)\, d\psi$$ v

Bayes. modelu s apriorním

$$p(\theta,\psi) = \cancel{p(\theta)}\, p(\underline{\psi}|\theta)\, p(\underline{\theta})$$

aby bylo prakticky užitečné, je potřeba,

aby $p(y|\psi,\theta)$ i $p(\psi|\theta)$

bylo „hezké"

<u>Příklady:</u>

1) LMM, mohl by začít s tímže

model:

$$p(y_i|\theta) = N\Big(X_i\underline{\beta} + Z_i\underline{\mu},\ \underbrace{Z_i \underline{D} Z_i^T + \underline{\sigma}^2 I_{n_i}}_{\text{musí být } \geq 0}\Big)$$

2) Logistická regrese s náh. efekty

$$P(y_i \mid b_i, \theta) = \text{dopln. věrohodnost}$$

$$= \prod_{j=1}^{n_i} \underbrace{p(y_{ij} \mid b_i, \theta)}_{\sim \mathrm{Alt}\left(\frac{e^{b_i}}{1+e^{b_i}}\right)}$$

---

3) směs

$$L_{obs}(\theta) = p(y \mid \theta) = \prod_i \left(\sum_k w_k \, f_k(y_i; \theta) \right)$$

my: $f_k = \mathcal{N}(\mu_k, \Sigma_k)$

cíl: navrhni ψ tak, aby

$P(y_i \mid \psi_i, \theta)$ bylo bez součtu

$P(\psi_i \mid \theta)$ bylo „hezké" (též)

a aby platilo

$$P(y_i \mid \theta) = \int P(y_i \mid \psi_i, \theta) \, p(\psi_i \mid \theta) \, d\psi_i$$

$$= \sum_k \underbrace{P(y_i \mid \psi_i = k, \theta)}_{f_k(y_i; \theta)} \underbrace{P(\psi_i = k \mid \theta)}_{w_k}$$

BDA: $\psi = (\psi_1, \dots, \psi_n)^T$

$\psi_i \equiv Z_i$, kde $\underbrace{P(Z_i = k \mid \theta) = w_k}_{= p(\psi_i \mid \theta)}$

$\in \{1, \dots, K\}$

$$p(y_i \mid \psi_i = k, \theta) \equiv f_k(y_i; \theta) \quad (\equiv N(\mu_k, \Sigma_k))$$

$$p(\theta, \psi \mid y) \propto p(y \mid \theta, \psi)\, p(\psi \mid \theta)\, p(\theta) =$$

$$= \prod_{i=1}^{n} \Big\{ \underbrace{p(y_i \mid \psi_i, \theta)}_{f_{\psi_i}(y_i; \theta)} \; \underbrace{p(\psi_i \mid \theta)}_{w_{\psi_i}} \Big\}\, p(\theta)$$

Bayes. doplnění $\longrightarrow$ hierar. model

$$\begin{tikzcd}
w_k, k = 1, \dots, K \arrow[d] & & \\
Z_i = \psi_i \arrow[d] & \mu_k, \Sigma_k, k = 1, \dots, K \arrow[dl] & \\
Y_i & &
\end{tikzcd}$$

potřeba apriorní

$$\theta = (w_k, \mu_k, \Sigma_k, k = 1, \dots, K)$$

<u>Gibbs:</u> $p(\theta \mid \ldots) \propto$

$$\propto \prod_{i=1}^{n} p(y_i \mid z_i, \underbrace{\mu_k, \Sigma_k}_{k=1..K}) \, p(z_i \mid \underbrace{w_k}_{k=1..K}) \cdot p(\theta)$$

$$p(\mu_k, \Sigma_k \mid \ldots)_{k=1..K} \propto \underbrace{p(y_i \mid \mu_k, \Sigma_k, z_i)}_{\substack{k=1..K \\ = N(\mu_{z_i}, \Sigma_{z_i})}} \cdot p(\mu_k, \Sigma_k)$$

BUDE SNADNÉ, pokud $p(\mu_k, \Sigma_k)$ je (semi) konjugované k normálnímu rozdělení

$$p(w_k \mid \ldots)_{k=1..K} \propto \prod_{i=1}^{n} \underbrace{p(z_i \mid w_k)}_{k=1..K} \, p(w) =$$

$$\boxed{z_i \in \{1, K\}} = \prod_{i=1}^{n} w_{z_i} \, p(w)$$

konjugovaná, zde je Dirichletovo rozděl.

a) pouze 2 skupiny

$$w = (w_1, w_2)$$
$$= (w, 1-w)$$

$$p(w) = \text{Beta}(a, b)$$

$$p(w) \propto w^{a-1}(1-w)^{b-1}$$

b) $\underline{K > 2}$

$$p(w) = \text{Dirichlet}(a_1, \dots, a_K)$$

$$p(w) \propto w_1^{a_1 - 1} \, w_2^{a_2 - 1} \cdots w_K^{a_K - 1}$$

$$p(w \mid \dots) \propto \prod_{i=1}^{n} w_{z_i} \, p(w) =$$

$$= w_1^{N_1} \cdots w_K^{N_K}$$

$$N_k = \sum_{i=1}^{n} \mathbb{1}(z_i = k)$$

$$\underline{\propto} \text{ Dirichlet}(a_1 + N_1, a_2 + N_2, \dots, a_K + N_K)$$

$$p(z_i \mid \dots) = ?$$

$$\underline{P(z_i = k \mid \dots)} \propto P(z_i = k \mid w) \cdot p(y_i \mid z_i = k, \theta) =$$

$$= \underline{w_k} \cdot f_k(y_i; \mu_k, \Sigma_k)$$

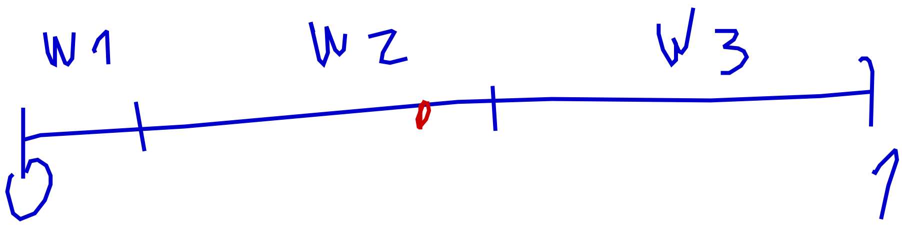
$$P(z_i = k \mid \dots) = \frac{w_k f_k(y_i; \mu_k, \Sigma_k)}{\sum_l w_l f_l(y_i; \mu_l, \Sigma_l)}$$

lze použít ke klasifikaci

$z_i \mid \dots$ má diskrétní rozdělení na $\{1, \dots, K\}$

$P(z_i = k \mid \dots) =$

Jak generovat z diskr. rozdělení když mám generátor z $U(0,1)$.



Celá směs jde rozšířit

DATA $\equiv$ jednotky z K populací

v každé populaci platí obecně jiný model, nevíme, kdo kam patří

$y_i, \; i = 1, \dots, n$

celkový model = směs modelů

$$P(y_i \mid \Theta) = \sum_{k=1}^{K} w_k \underline{P_k(y_i \mid \Theta)}$$

model pro y_i, pokud patří do k-té populace