Clustering Multivariate Ordinal Data

Thomas Michel, Théo Rudkiewicz and Ali Ramlaoui

February 6, 2024

0.1 BOS Model

$$
\begin{aligned}
\Pr(x|x \in e_j, \mu, \pi) &= \sum_{e_{j+1} \subset e_j} \Pr(x, e_{j+1}|x \in e_j, \mu, \pi) && (1)\\
&= \sum_{e_{j+1} \subset e_j} \Pr(x|e_{j+1}, x \in e_j, \mu, \pi) \Pr(e_{j+1}|e_j, \mu, \pi) && (2)\\
&= \sum_{e_{j+1} \subset e_j ; x \in e_{j+1}} \Pr(x|x \in e_{j+1}, \mu, \pi) \Pr(e_{j+1}|e_j, \mu, \pi) && (3)
\end{aligned}
$$

We now suppose $e_j = [\![l, h-1]\!]$:

$$
\begin{aligned}
&\Pr(x|x \in [\![l, h-1]\!], \mu, \pi) = && (4)\\
&\sum_{y=x+1}^{h-1} \Pr([\![l, y-1]\!] \,|\, [\![l, h-1]\!], \mu, \pi) \Pr(x|x \in [\![l, y-1]\!], \mu, \pi)\\
&+ \Pr(\{x\} \,|\, [\![l, h-1]\!], \mu, \pi) \Pr(x|x \in \{x\}, \mu, \pi) && (5)\\
&+ \sum_{y=l}^{x-1} \Pr([\![y+1, h-1]\!] \,|\, [\![l, h-1]\!], \mu, \pi) \Pr(x|x \in [\![y+1, h-1]\!], \mu, \pi)
\end{aligned}
$$

$$
= \begin{aligned}
&\frac{1}{h-l} \sum_{y=x+1}^{h-1} \left[\pi \mathbb{1}\left\{\mu < y\right\} + (1-\pi)\frac{y-l}{h-l}\right] \Pr(x|x \in [\![l, y-1]\!], \mu, \pi)\\
&+ \frac{1}{h-l} \left[\pi \mathbb{1}\left\{\mu = x \vee (x = l \wedge \mu \leqslant x) \vee (x = h-1 \wedge \mu \geqslant x)\right\} + (1-\pi)\frac{1}{h-l}\right]\\
&\Pr(x|x \in \{x\}, \mu, \pi)\\
&+ \frac{1}{h-l} \sum_{y=l}^{x-1} \left[\pi \mathbb{1}\left\{\mu > y\right\} + (1-\pi)\frac{h-y-1}{h-l}\right] \Pr(x|x \in [\![y+1, h-1]\!], \mu, \pi)
\end{aligned} \tag{6}
$$

As $\Pr(x|x \in \{x\}, \mu, \pi) = 1$ this allows to compute the probability of x being in the interval $\llbracket l, h-1 \rrbracket$ recursively.

As:

$$\Pr(x|x \in \llbracket l, y-1 \rrbracket, \mu, \pi) = \Pr(x-l|x-l \in \llbracket 0, y-l-1 \rrbracket, \max(0, \mu - l), \pi) \tag{7}$$

We can rewrite the previous equation as:

$$h \Pr(x|x \in \llbracket 0, h-1 \rrbracket) = \sum_{y=x+1}^{h-1} \left[\pi \mathbb{1}\{\mu < y\} + (1-\pi)\frac{y}{h} \right] \Pr(x|x \in \llbracket 0, y-1 \rrbracket, \mu, \pi) \tag{8}$$

$$+ \pi \mathbb{1}\{\mu = x \vee (x = 0 \wedge \mu \leqslant x) \vee (x = h-1 \wedge \mu \geqslant x)\} + (1-\pi)\frac{1}{h} \tag{9}$$

$$+ \sum_{y=0}^{x-1} \left[\pi \mathbb{1}\{\mu > y\} + (1-\pi)\frac{h-y-1}{h} \right] \Pr(x-y-1|x-y-1 \in \llbracket 0, h-y \tag{10}$$

We can now prove that $\forall x \in \llbracket 0, h-1 \rrbracket, \forall \mu \in \llbracket 0, h-1 \rrbracket, \pi \mapsto \Pr(x|x \in \llbracket 0, h-1 \rrbracket, \mu, \pi)$ is concave on $[0, 1]$

Lemma 1 (Log concavity affine times polynomial). *Let P a* log-*concave polynomial postive polynomail (for all x considered) and $a, b \in \mathbb{R}$ with $ax+b \geqslant 0$. Then $f : x \mapsto (ax+b)P(x)$ is* log-*concave.*

Proof. Using the lemma **??** we have that $P'(x)^2 - P(x)P''(x) \geqslant 0$.

As

$$f'(x)^2 - f(x)f''(x) = a^2 P(x)^2 + (ax+b)\left[P'(x)^2 - P(x)P''(x)\right]$$

we have that $f'(x)^2 - f(x)f''(x) \geqslant 0$ hence using the lemma **??** we have that f is log-concave. □

Theorem 1 (Log concavity of the BOS model). *$\forall x \in \llbracket 0, h-1 \rrbracket, \forall \mu \in \llbracket 0, h-1 \rrbracket, f : \pi \mapsto \Pr(x|x \in \llbracket 0, h-1 \rrbracket, \mu, \pi)$ is* log-*concave on $[0, 1]$ and a postive (for $\pi \in [0, 1]$) polynomial of degree less than $h-1$.*

Proof. We proceed by induction on h:

Initialization: $h = 1$:

$$\forall x \in \llbracket 0, h-1 \rrbracket, \forall \mu \in \llbracket 0, h-1 \rrbracket, \Pr(x|x \in \llbracket 0, h-1 \rrbracket, \mu, \pi) = 1$$

which is log-concave and a positive polynomial of degree 0.

Induction: Suppose the theorem holds for $h-1$ and let us prove it for h.

Using the previous formula we have that f is a sum of psotive affine function in π times $\Pr(x|x \in [\![0, y-1]\!], \mu, \pi)$ which is log-concave by induction hypothesis and a positive polynomial of degree $y-1$. We immediately deduce that f is postive and polynomial of degree less than $h-1$. Moreover using the previous lemma 1 we have that f is log-concave.

Hence the theorem holds for h. □

0.1.1 Sum of Z

$$\sum_{j=k}^{m-1} \Pr(z_j = 1|x \in e_k, \mu, \pi) = \Pr(z_k = 1|x \in e_k, \mu, \pi) + \sum_{j=k+1}^{m-1} \Pr(z_j = 1|x \in e_k, \mu, \pi) \tag{11}$$

$$\sum_{j=k+1}^{m-1} \Pr(z_j = 1|x \in e_k, \mu, \pi) = \sum_{e_{k+1} \subset e_k} \sum_{j=k+1}^{m-1} \Pr(z_j = 1, e_{k+1}|x \in e_k, \mu, \pi) \tag{12}$$

$$= \sum_{e_{k+1} \subset e_k} \sum_{j=k+1}^{m-1} \Pr(z_j = 1|e_{k+1}, x \in e_k, \mu, \pi) \Pr(e_{k+1}|x \in e_k, \mu, \pi) \tag{13}$$

$$= \sum_{e_{k+1} \subset e_k; x \in e_{k+1}} \Pr(e_{k+1}|x \in e_k, \mu, \pi) \sum_{j=k+1}^{m-1} \Pr(z_j = 1|e_{k+1}, x \in e_k, \mu, \pi) \tag{14}$$

$$\Pr(z_k = 1|x \in e_k, \mu, \pi) = \sum_{e_{k+1} \subset e_k} \Pr(z_k = 1, e_{k+1}|x \in e_k, \mu, \pi) \tag{15}$$

$$= \sum_{e_{k+1} \subset e_k; \mu, x \in e_{k+1}} \Pr(z_k = 1, e_{k+1}|x \in e_k, \mu, \pi) \tag{16}$$

$$= \sum_{e_{k+1} \subset e_k; \mu, x \in e_{k+1}} \Pr(z_k = 1|x \in e_{k+1}, \mu, \pi) \Pr(e_{k+1}|x \in e_k, \mu, \pi) \tag{17}$$

1 GOD Model proofs