Maximum d'entropie, une seule ligne de bus

FB-RL-GG

UNIL

Notations et formalisation du problème

Une ligne orientée allant de l'arrêt $i = 1$ jusqu'à $i = l$. Soit x_i le nombre de passagers montant à l'arrêt i, et y_i le nombre de passagers descendant à l'arrêt i. On a $y_1 = 0$ et $x_l = 0$.

Soit N_{ij} (avec $i < j$; sinon $N_{ij} = 0$) le nombre de personnes montant en i et descendant en j.

Soit $n_{i,i+1}$ le nombre de personnes transportées dans le tronçon i, $i+1$. Par construction:

$$n_{i,i+1} = n_{i-1,i} + x_i - y_i \qquad\qquad n_{01} = 0 \tag{1}$$

On veut estimer les données N_{ij}. Par construction,

$$N_{i\bullet} = x_i \qquad\qquad N_{\bullet j} = y_j \qquad\qquad N_{\bullet\bullet} = x_\bullet \stackrel{!}{=} y_\bullet \tag{2}$$

Soit $f^D_{ij} := \frac{N_{ij}}{N_{\bullet\bullet}}$ la distribution empirique à estimer. Soit $g_i := \frac{x_i}{x_\bullet}$ et $h_j := \frac{y_j}{y_\bullet}$ les distributions marginales correspondantes. En fait f^D_{ij} est une matrice $(l-1) \times (l-1)$, où $i = 1, \ldots, l-1$ et $j = 2, \ldots, l$

Pour la distribution théorique f^M, donnée par une matrice $(l-1) \times (l-1)$, on peut imaginer un prior

$$f^M_{ij} = a_i b_j 1(i < j) \qquad\qquad \sum_{i=1}^{l-1} a_i B_i \stackrel{!}{=} 1 \qquad\qquad B_i := \sum_{j=i+1}^{l} b_j \tag{3}$$

dépendant des $2l-3$ paramètres libres $a_1, \ldots, a_{l-1}$ et $b_2, \ldots, b_l$ (contraints par la normalisation).

Les contraintes se réécrivent

$$f^D_{i\bullet} = g_i \quad i = 1, \ldots, l-1 \qquad\qquad f^D_{\bullet j} = h_j \quad j = 2, \ldots, l \tag{4}$$

Il y en a aussi $2l - 3$ (car $\sum_i f^D_{i\bullet} = \sum_j f^D_{\bullet j}$). Cela permet d'espérer de déterminer a et b de telle sorte que $f^D = f^M \equiv f$, qui donnerait un minimum absolu de $K(f^D||f^M)$.

Les premiers termes non nuls sont $f_{12} = a_1 b_2$, $f_{13} = a_1 b_3$.... $f_{1,l} = a_1 b_l$, dont la somme $f_{1\bullet} = a_1 B_1$ doit être g_1.

Puis $f_{23} = a_2 b_3$, $f_{24} = a_2 b_4$.... $f_{2,l} = a_2 b_l$, dont la somme $f_{2\bullet} = a_2 B_2$ doit être g_2.

En général, on a $f_{i\bullet} = a_i B_i \stackrel{!}{=} g_i$ pour $i = 1, \ldots, l-1$. De même, la normalisation (3) peut aussi s'écrire

$$\sum_{j=2}^{l} A_j b_j \stackrel{!}{=} 1 \qquad\qquad A_j = \sum_{a=1}^{j-1} a_i$$

d'où l'on tire que $f_{\bullet j} = A_j b_j \stackrel{!}{=} h_j$ pour $j = 2, \ldots, l$.

Pas sûr que $f^D = f^M$ puisse être réalisé, peut-être faut il changer de prior f^M : piste à ne pas abandonner. Mais...

Approche: "estimer une table de contingence N dont les marges sont fixées"

... le problème d'estimer une table de contingence N dont les marges sont fixées (équation (2)) a fait l'objet d'une énorme littérature... A étudier et poursuivre.

Modèle de Guillaume

(avec quelques notations utilisées ici).

- Probabilité de monter en i: $p_i^{\text{in}} = x_i / x_\bullet$.
- Probabilité de descendre en i: $p_i^{\text{out}} = y_i / n_{i-1,i}$.
- Probabilité de continuer de i à $i+1$: $c_i = 1 - p_i^{\text{out}}$.

- Probabilité P_{ij} de trajet de i à $j > i$:

$$P_{ij} = p_i^{\text{in}} c_{i+1} \dots c_{j-1} p_j^{\text{out}} \quad \text{pour} \quad j \geq i+2 \qquad\qquad P_{i,i+1} = p_i^{\text{in}} p_{i+1}^{\text{out}} \quad (5)$$

  Le produit commence par c_{i+1}, car, si l'on est monté en i, la probabilité d'effectuer le tronçon $i \to i+1$ vaut 1.

Soit $X_i := \sum_{k=1}^{i} x_k$ le nombre cumulé de montées, et $Y_i := \sum_{k=1}^{i} y_k$ le nombre cumulé de descentes. On a

$$X_i \geq Y_i \quad i = 1, \dots, l \qquad\qquad X_l = Y_l \qquad\qquad n_{i,i+1} = X_i - Y_i \qquad (6)$$

Il est pratique de définir le "transit d'avant" $t_i := X_{i-1} - Y_{i-1} = n_{i-1,i}$, en posant $t_1 = 1$ (au lieu de $t_0 = 0$) afin que $p_1^{\text{out}} = y_1/t_1 = 0/1 = 0$. On a alors $p_i^{\text{out}} = y_i/t_i$ pour tout $i = 1, \dots, l$, avec

$$p_l^{\text{out}} = \frac{y_l}{t_l} = \frac{y_l}{X_{l-1} - Y_{l-1}} = \frac{y_l}{y_l} = 1$$

comme il se doit, où on a utilisé $X_{l-1} - Y_{l-1} = X_{l-1} + 0 - Y_{l-1} - y_l + y_l = X_l - Y_l + y_l = y_l$.

On observe que $P_{\bullet\bullet} = 1$. On va redéfinir comme avant $f_{ij} := P_{ij}$. Le nombre attendu de trajets N_{ij} est alors $N_{ij} = x_\bullet f_{ij}$. On observe que $N_{i\bullet} = x_i$ et $N_{j\bullet} = y_j$.

On observe aussi que les trajets attendus sont (pour les cas étudiés) de la forme (cf. (3))

$$N_{ij} = N_{\bullet\bullet} a_i b_j I(j > i) \qquad (7)$$

et qu'ainsi la forme des histogrammes du nombre de sorties j depuis un départ i variable reste la même:

$$N_{j|i} := \frac{N_{ij}}{N_{i\bullet}} = \frac{N_{\bullet\bullet} a_i b_j I(j > i)}{N_{\bullet\bullet} a_i \sum_{k>i} b_k} = \frac{b_j I(j > i)}{B_i} \qquad (8)$$

avec $B_i := \sum_{k>i} b_k$.

Par construction, a_l et b_1 sont indéfinis dans (7); on peut les poser égaux à zéro.

Estimation des paramètres a, b dans $N_{ij} = a_i b_j I(i < j)$

On a $N_{i\bullet} = x_i$ et $N_{i\bullet} = y_j$. Soit

$$s := \min_i \{i | x_i > 0\} \qquad\qquad t := \max_j \{j | y_j > 0\} \qquad (9)$$

Par construction, $1 \leq s < t \leq l$. On suppose que n_{ij} est irréductible (pas de tronçon à vide); sinon il faut considérer chaque tronçon plein séparément. De fait, l'expression $N_{ij} = a_i b_j I(i < j)$ présuppose explicitement l'irréductibilité.

En particulier, $a_i = 0$ pour $i < s$ et $i = l$; $b_j = 0$ pour $j > t$ et $j = 1$; $a_s > 0$ et $b_t >$. Alors, pour $s \le i < j \le t$,

$$N_{sj} = a_s b_j \qquad N_{it} = a_i b_t \qquad N_{st} = a_s b_t \tag{10}$$

et donc

$$\boxed{a_i b_j = \tfrac{N_{it} N_{sj}}{N_{st}}} \tag{11}$$

qui exprime une sorte "d'indépendance asymétrique". On peut alors poser, dans le cas irréductible

$$a_i = \begin{cases} \frac{N_{it}}{\sqrt{N_{st}}} & \text{if } i \ge s \ , \\ 0 & \text{otherwise.} \end{cases} \qquad b_j = \begin{cases} \frac{N_{sj}}{\sqrt{N_{st}}} & \text{if } j \le t \ , \\ 0 & \text{otherwise.} \end{cases} \tag{12}$$

En particulier, par sommation,

$$a_\bullet = \frac{N_{\bullet t}}{\sqrt{N_{st}}} = \frac{y_t}{\sqrt{N_{st}}} \qquad b_\bullet = \frac{N_{s\bullet}}{\sqrt{N_{st}}} = \frac{x_s}{\sqrt{N_{st}}} \qquad x_s y_t = a_\bullet b_\bullet N_{st} \tag{13}$$

Simulations numériques

Voir test_markov_Guillaume_Francois.R : tout semble jouer avec l'exemple 1, pour lequel le transit $X_i - Y_i$ n'est jamais nul (sauf en $i = l$). Mais difficultés avec l'exemple 2 ($l = 10$), pour lequel le bus est vide entre les arrêts 8 et 9 ($X_8 - Y_8 = 0$), et donc p_9^{out} *n'est pas défini.*

Clairement, l'absence de voyageurs entre les arrêts 8 et 9 "simplifie" le problème, qui doit être résolu comme deux problèmes "disjoints": de la station 1 à la station 8 d'une part, et de la station 9 à la station 10 d'autre part: il faut commencer par déterminer les tronçons vides, puis résoudre les sous-problèmes délimités par les tronçons vides.