Description formelle des modifications de l'été 2017 de MixtComp

Vincent KUBICKI

20 octobre 2017

Résumé

Le présent document est un complément formel aux précédents documents décrivant les dernières modifications effectuées sur MixtComp. Il ne reprend pas l'ensemble des informations de ces derniers. Comme il n'y a pour le moment aucune documentation scientifique dans MixtComp, il pourrait en servir de base.

1 Notations

Les variables aléatoires sont notées en majuscules et leurs réalisations en minuscules. Les vecteurs aléatoires sont notés en gras.

L'échantillon étudié est composé de N observations chacune décrite par P variables. Cet échantillon est considéré comme une réalisation d'un vecteur aléatoire à N éléments $\boldsymbol{X} = \{\boldsymbol{X}_1, ..., \boldsymbol{X}_N\}$. Chaque observation est elle-même un vecteur aléatoire à P composantes $\boldsymbol{X}_i = \{X_i^1, ..., X_i^P\}$. Le modèle de mélange de cet ensemble de variables est composé de K classes et inclue l'hypothèse de l'indépendance conditionnelle des P variables aux K classes.

Pour la i-ème observation, certaines variables peuvent être observées, et d'autres sont censurées voir manquantes. On note l'ensemble des indices de variables $\bar{P} = \{1, 2, ..., P\}$. Pour la i-ème observation, on introduit les deux sous-ensembles suivants :

- $O_i \subseteq \bar{P}$ le sous-ensemble des indices des variables observées ;
- $L_i \subseteq \bar{P}$ le sous-ensemble des indices des variables partiellement observées ou manquantes, complément de O_i dans $\bar{P}$.

Certains modèles introduisent des variables latentes internes (par exemple les ordres de présentation pour les données de rang). Leur nombre varie en fonction du modèle. On note $\boldsymbol{Y}_i^j$ le vecteur aléatoire composé des variables latentes contenu dans le modèle décrivant X_i^j. L'ensemble des variables latentes pour une observation i est noté $\boldsymbol{Y}_i$.

L'appartenance aux classes de la i-ème observation est une variable aléatoire latente Z_i.

Une réalisation des variables latentes du modèle de la j-ème variable est $\boldsymbol{y}_i^j$. $\tilde{\boldsymbol{y}}_i^j$ désigne une complétion soit par tirage, soit par le calcul des valeurs médianes (pour les variables continues) ou des modes (pour les variables discrètes) des lois estimées pour $\boldsymbol{Y}_i^j$ (la nature de la complétion sera précisée au besoin).

Pour les variables $\boldsymbol{X}_i$, on note pour simplifier $\boldsymbol{x}_i$ la réalisation observée (uniquement composée des réalisations x_i^j pour $j \in O_i$). On note $\tilde{\boldsymbol{x}}_i$ une réalisation complétée contenant :

— pour X_i^j avec $j \in O_i$: la valeur observée fournie par l'utilisateur ;
— pour X_i^j avec $j \in L_i$: une valeur complétée.

L'ensemble des paramètres du modèle est noté $\boldsymbol{\Theta}$. Il s'agit de la collection de paramètres $\{\boldsymbol{\pi}, \boldsymbol{\theta}_1, ..., \boldsymbol{\theta}_P\}$, où $\boldsymbol{\pi} = \{\pi_1, ..., \pi_K\}$ représente les proportions des appartenances aux classes et $\boldsymbol{\theta}_j$ les paramètres de X_i^j. $\boldsymbol{\theta}_j$ est la collection des paramètres $\{\theta_1^j, ..., \theta_K^j\}$, avec θ_k^j les paramètres pour la k-ième composante de la variable X_i^j. Les paramètres estimés par le SEM sont notés $\hat{\boldsymbol{\Theta}}$, et les paramètres particuliers estimés pour l'initialisation du SEM $\hat{\boldsymbol{\Theta}}^0$.

On dénote par f_j la densité de probabilité (pour les variables continues) ou la fonction masse (pour les variables discrètes) de la j-ième variable. Pour une valeur censurée, on dénote par A_{ij} le sous-ensemble du domaine autorisé. Par exemple pour une variable à valeurs réelles dont les bornes sont observées entre a et b, on aura $A_{ij} = [a, b]$ et pour une variable catégorielle où on n'autorise que les modalités 1, 3 et 5 on aura un sous-ensemble autorisé de la forme $A_{ij} = \{m_1, m_3, m_5\}$. Si la variable n'est pas observée pour une observation donné, alors A_{ij} est égal au domaine.

2 Expressions des densités de probabilité

La densité de probabilité [1] observée d'une observation peut être exprimée en détaillant le modèle de mélange :

$$f_o(\boldsymbol{x}_i; \Theta) = \sum_{k=1}^{K} \pi_k \prod_{j \in O_i} f_j(x_i^j; \theta_k^j) \prod_{j \in L_i} \int_{A_{ij}} f_j(t; \theta_k^j)\, dt\,. \tag{1}$$

La densité de probabilité complétée d'une observation est utilisée dans les étapes "S" (d'échantillonnage) du SEM et du Gibbs. Son expression est :

$$f_c(\tilde{\boldsymbol{x}}_i, \tilde{\boldsymbol{y}}_i, z_i = k; \Theta) = \pi_k \prod_{j \in O_i} f_j(x_i^j, \tilde{\boldsymbol{y}}_i^j; \theta_k^j) \prod_{j \in L_i} f_j(x_i^j, \tilde{\boldsymbol{y}}_i^j; \theta_k^j). \tag{2}$$

La densité de probabilité complétée pour les appartenances aux classes mais marginalisée sur toutes les autres variables latentes est utilisée dans le calcul du critère ICL. Son expression est :

$$f_{oc}(\boldsymbol{x}_i, z_i = k; \Theta) = \pi_k \prod_{j \in O_i} f_j(x_i^j; \theta_k^j) \prod_{j \in L_i} \int_{A_{ij}} f_j(t; \theta_k^j)\, dt\,. \tag{3}$$

3 Utilisation des probabilités observées lors des initialisations

Les initialisations des variables latentes étaient auparavant obtenues par des tirages suivant des lois uniformes. Comme on effectue à présent une initialisation des paramètres présentant une variance importance (en utilisant un individu représentant par classe, vois les

1. Cette expression mélange densités de probabilités et probabilités via les intégrales, donc le terme de densité est peut-être impropre.

documents précédents), il devient intéressant de tirer les variables latentes suivant le modèle $\hat{\boldsymbol{\Theta}}^0$.

Le problème est que les lois de tirages sont conditionnelles. On connaît les expressions des probabilités $p(\boldsymbol{Y}_i|\boldsymbol{X}_i, Z_i; \hat{\boldsymbol{\Theta}}^0)$ et $p(Z_i|\boldsymbol{X}_i, \boldsymbol{Y}_i; \hat{\boldsymbol{\Theta}}^0)$ car elles sont utilisées à chaque itération du calcul SEM. Or, en début de calcul, on ne connaît les valeurs prises ni par $\boldsymbol{Y}_i$ ni par Z_i. L'idée de la modification présentée dans ce chapitre est d'initialiser le calcul en utilisant la distribution marginale $p(Z_i|\boldsymbol{X}_i; \hat{\boldsymbol{\Theta}}^0)$, et donc d'intégrer la distribution sur toutes les valeurs possibles de $\boldsymbol{Y}_i$. Le calcul de probabilité observée est plus long que celui de probabilité complétée, et avant on ne l'effectuait qu'en fin de calcul pour exporter les vraisemblances observées ainsi que les critères de choix de modèle comme le BIC ou l'ICL.

Les $t_{ik,c}$ complétés (utilisés à chaque étape du SEM et du Gibbs) ont pour expression :

$$t_{ik,c} = \frac{\pi_k \prod_{j \in O_i} f_j(x_i^j, \tilde{\boldsymbol{y}}_i^j; \theta_k^j) \prod_{j \in L_i} f_j(x_i^j, \tilde{\boldsymbol{y}}_i^j; \theta_k^j)}{\sum_{k'=1}^{K} \pi_{k'} \prod_{j \in O_i} f_j(x_i^j, \tilde{\boldsymbol{y}}_i^j; \theta_{k'}^j) \prod_{j \in L_i} f_j(x_i^j, \tilde{\boldsymbol{y}}_i^j; \theta_{k'}^j)}. \tag{4}$$

On a donc rajouté, lors de l'initialisation, un tirage de Z_i utilisant $t_{ik,o}$ qui a pour expression :

$$t_{ik,o} = \frac{\pi_k \prod_{j \in O_i} f_j(x_i^j; \theta_k^j) \prod_{j \in L_i} \int_{A_{ij}} f_j(t; \theta_k^j)\, dt}{\sum_{k'=1}^{K} \pi_{k'} \prod_{j \in O_i} f_j(x_i^j; \theta_{k'}^j) \prod_{j \in L_i} \int_{A_{ij}} f_j(t; \theta_{k'}^j)\, dt}. \tag{5}$$