Classifieur basé sur un ensemble de profils lorsque les observations ne sont pas identiquement distribuées

1 Introduction

L'apprentissage statistique présenté dans la première partie de cette analyse (voir chapitre III) est élaboré sous l'hypothèse d'indépendance et de distribution identique (i.i.d) des éléments aléatoires $(Y_i, X_i)_{i=1,\dots,n}$ qui ont généré les observations. Dans le présent chapitre, nous cherchons à adapter notre procédure d'apprentissage dans une situation où les données, en plus d'être déséquilibrées, sont réparties entre m clusters (groupes ou blocs) tirés aléatoirement à partir d'une population donnée. On suppose que chaque cluster admet une distribution $[Y, X]_h; h \in \{1, \cdots, m\}$ indépendantes des autres. Etant donné que l'indicateur de performance au tour duquel la procédure d'apprentissage a été élaborée est la valeur prédictive positive, nous proposons un estimateur Bayésien de la valeur prédictive positive de tout profil U conditionnellement à la distribution $[Y, X]_h$ des observations dans un cluster h donné. Cette approche nous permet de tenir en compte l'effet du cluster dans les résultats de l'analyse.

Les méthodes d'analyse classiques permettant de traiter des données groupées (essais multicentriques) introduisent en général la variable d'échantillonnage (groupe, cluster ou centre) comme variables explicatives en autorisant les interactions. Cependant elles ont des limites : (1) Lorsque le nombre de groupes est important, les introduire tous dans le modèle devient problématique. (2) Puisque l'un des groupes est utilisé comme groupe de référence, on ignore les écarts de chaque groupe à la moyenne. (3) Les groupes participant à l'essai constituent un échantillon d'une population plus large de groupes, on peut souhaiter faire des prédictions pour un groupe n'ayant pas participé à l'essai. (4) On peut aussi souhaiter avoir une mesure d'hétérogénéité entre les groupes.

Le modèle Bêta-binomiale figure parmi les méthodes alternatives les plus utilisées dans la littéra-

ture. Ce dernier permet à la fois d'estimer l'espérance de la probabilité de succès conditionnellement à un profil $U(X)$ dans la population et sa variabilité d'un groupe à un autre. De plus, il permet d'inférer sur la probabilité de succès conditionnellement à l'événement $[U(X) = 1]$ dans n'importe quel groupe, pas seulement ceux échantillonnés.

2 Modèle hiérarchique pour le calcul des valeurs prédictives positives

Nous étudions dans ce chapitre un modèle statistique correspondant au cas où les données sont générées par une suite $(Y_i, X_i)_{i=1:n}$ d'éléments aléatoires non identiquement distribués. Il en résulte alors une hétérogénéité des données dont il faudrait tenir compte dans le modèle statistique sur lequel l'analyse du classifieur sera basée.
Nous considérons ici la situation particulière où la suite $(Y_i, X_i)_{i=1:n}$ est structurée suivant une partition de m sous-ensembles $(Y_{ih}, X_{ih})_{\substack{h=1:m \\ i=1:n_h}}$ telles que les éléments de la suite $(Y_{ih}, X_{ih})_{i=1:n_h}$ soient indépendants et de même loi $[Y, X]_h$. Nous supposons que les éléments de la suite $[Y, X]^{\mathcal{L}} = \{[Y, X]_h, h = 1 : m\}$ sont générés de façon indépendante suivant une loi μ sur l'ensemble $Prob(Y, X)$ des lois de probabilités sur $Dom(Y) \times Dom(X)$ muni de la tribu associée à la topologie de la convergence faible. Si on se donne $U(X)$, un profil défini par X, on a alors

- $[Y|\theta_h^U, [Y, X]_h] = \text{Bernoulli}(\theta_h^U)$, où $\theta_h^U = \Pr(Y = 1|U(X) = 1, [Y, X]^{\mathcal{L}} = [Y, X]_h)$
- la suite $\left(\theta_h^U\right)_{h=1:m}$ est un échantillon iid.

On considère désormais que la suite $\theta^U = \left(\theta_h^U\right)_{h=1:m}$ est issue de la loi Bêta de paramètres (α_U, β_U). On désigne par $\left[Y, \theta^U, [Y, X]^{\mathcal{L}}\right]$ et $\left[\theta^U, [Y, X]^{\mathcal{L}}\right]$ les lois de probabilité respectives de $\left(Y, \theta^U, [Y, X]^{\mathcal{L}}\right)$ et $\left(\theta^U, [Y, X]^{\mathcal{L}}\right)$. Le principe de la factorisation permet d'écrire

$$\begin{aligned}
\left[Y, \theta^U, [Y, X]^{\mathcal{L}}\right] &= \left[Y|\theta^U, [Y, X]^{\mathcal{L}}\right] \left[\theta^U, [Y, X]^{\mathcal{L}}\right] \\
\left[Y, \theta^U, [Y, X]^{\mathcal{L}}\right] &= \left[Y|\theta^U, [Y, X]^{\mathcal{L}}\right] \left[\theta^U|[Y, X]^{\mathcal{L}}\right] \left[[Y, X]^{\mathcal{L}}\right] \\
\prod_{h=1}^{m} \left[Y, \theta_h^U, [Y, X]_h\right] &= \prod_{h=1}^{m} \left(\left[Y|\theta_h^U, [Y, X]_h\right] \left[\theta_h^U|[Y, X]_h\right] [[Y, X]_h]\right)
\end{aligned}$$

On peut remplacer la loi $\left[\theta_h^U|[Y, X]_h\right]$ par la loi $\left[\theta_h^U|\alpha_U, \beta_U\right]$ dans l'expression précédente puisqu'il s'agit de la même distribution. Pour réduire la complexité du problème, nous allons nous intéresser pour la suite à la distribution $\left[Y|\theta_h^U, [Y, X]_h\right]$ et à la distribution $\left[\theta_h^U|\alpha_U, \beta_U\right]$. Le modèle hiérarchique à étudier est alors le suivant :

$$\begin{cases}
\left[Y|\theta_h^U, [Y, X]_h\right] &= Bernoulli\left(\theta_h^U\right) \\
\left[\theta_h^U|\alpha_U, \beta_U\right] &= Beta(\alpha_U, \beta_U)
\end{cases}$$

Ce modèle permet d'estimer la probabilité $\Pr(Y=1|U(X),[Y,X]_h)$ qui n'est rien d'autre que la valeur prédictive positive (VPP) du profil $U(X)$ sous la contrainte $[Y,X]_h$.

Proposition 7. Si Y est une variable binaire telle que $Y|\theta_h^U,[Y,X]_h \sim Bernoulli\left(\theta_h^U\right)$ où $\theta_h^U|\alpha_U,\beta_U$ est une variable aléatoire de loi $Beta(\alpha_U,\beta_U)$ alors on a

$$\Pr(Y=1|U(X)) = \frac{\alpha_U}{\alpha_U+\beta_U} \tag{IV.1}$$

$$\left[\theta_h^U|Y=y,\alpha_U,\beta_U\right] = Beta(\alpha_U+y,\beta_U+1-y) \tag{IV.2}$$

Preuve.
On a

$$\begin{aligned}\Pr(Y=1|U(X)) &= \mathbb{E}\left(Y|\theta_h^U\right)\\ &= \mathbb{E}\left(\mathbb{E}\left(Y|\theta_h^U,[Y,X]_h\right)\right)\\ &= \mathbb{E}\left(\theta_h^U\right)\end{aligned}$$

Par ailleurs, on a

$$\left[\theta_h^U|Y=y,\alpha_U,\beta_U\right] = \frac{\left[Y=y|\theta_h^U,\alpha_U,\beta_U\right]\left[\theta_h^U|\alpha_U,\beta_U\right]}{\int_0^1\left[Y=y|\theta_h^U,\alpha_U,\beta_U\right]\left[\theta_h^U|\alpha_U,\beta_U\right]d\theta_h^U}$$

On en déduit alors que

$$\left[\theta_h^U|Y=y,\alpha_U,\beta_U\right] = \frac{\Gamma(\alpha_U+y)\Gamma(\beta_U-y+1)}{\Gamma(\alpha_U+\beta_U+1)}\theta_h^{\alpha_U+y-1}(1-\theta_h)^{\beta_U-y}$$

□

On a $\mathbb{E}(\theta_h^U|\alpha_U,\beta_U) = \dfrac{\alpha_U}{\alpha_U+\beta_U}$ et $\mathbb{V}ar(\theta_h^U|\alpha_U,\beta_U) = \dfrac{\alpha_U}{\alpha_U+\beta_U}\dfrac{\beta_U}{(\alpha_U+\beta_U)(\alpha_U+\beta_U+1)}$

L'application $(\alpha_U,\beta_U) \longrightarrow \begin{pmatrix}\pi_U = \frac{\alpha_U}{\alpha_U+\beta_U}\\ \gamma_U = \frac{1}{\alpha_U+\beta_U+1}\end{pmatrix}$ étant injective, on peut envisager de reparamétrer la famille de loi Bêta par la moyenne π_U et le paramètre γ_U appelé paramètre de dispersion. Pour π_U fixé, le paramètre γ_U détermine la forme de la densité. Nous retiendrons dans la suite du travail cette paramétrisation de la famille des lois Bêta.

3 Lois a posteriori des paramètres relatifs aux clusters : approche Bayésienne empirique

Pour alléger les notations dans cette section, on pose $\tau_U = 1/\gamma_U - 1$. Dans la suite, nous avons choisi d'écrire le modèle en fonction des paramètres $\{\pi_U, \tau_U\}$. Cependant les résultats seront présentés en fonction des paramètres $\{\pi_U, \gamma_U\}$. On pose le modèle suivant :

$$\begin{cases} \left[Y|\theta_h^U, [Y,X]_h\right] = \prod_{k=1}^{m} \left(\theta_h^U\right)^{\mathbb{1}_{[Y=1]}(y)\delta_{\{1,[Y,X]_h\}}(U(X),[Y,X]_h)} \left(1-\theta_h^U\right)^{(1-\mathbb{1}_{[Y=1]}(y))\delta_{\{1,[Y,X]_h\}}(U(X),[Y,X]_k)} \\ \left[\theta_h^U|\pi_U,\tau_U\right] = \frac{\Gamma(\tau_U)}{\Gamma(\pi_U\tau_U)\Gamma((1-\pi_U)\tau_U)} \left(\theta_h^U\right)^{\pi_U\tau_U-1} \left(1-\theta_h^U\right)^{(1-\pi_U)\tau_U-1} \mathbb{1}_{[0,1]}\left(\theta_h^U\right) \end{cases}$$

3.1 Détermination de la loi a posteriori du paramètre θ_h^U par une approche Bayésienne empirique

Dans une approche bayésienne complète, la détermination de la loi a posteriori de θ_h^U nécessite la spécification d'une loi a priori pour le couple (π_U,γ_U). En défaut de la spécification d'une telle loi a priori, on peut adopter une approche empirique pour la détermination a posteriori du vecteur $(\theta_h^U)_{h=1:m}$ et de ses éléments marginaux.

3.2 Loi a posteriori : approche bayésienne empirique

La méthode de Bayes empirique est très souvent utilisée lorsqu'il s'agit d'un problème d'estimation de paramètres multiples où les relations connues (*i.i.d.*) entres les composantes du vecteur de paramètres inconnus $\left(\theta_h^U\right)_{h=1:m}$ suggèrent de partager les informations entre les différentes réalisations similaires du couple (Y,X) pour obtenir une meilleure estimation de chaque paramètre θ_h^U. L'approche de Bayes empirique a été classée en deux catégories par Morris, C.N.[1983][7] dont : le cas non paramétrique (voir [8] pour plus de détails) et le cas paramétrique.
Dans le cas paramétrique, on suppose que la loi a priori du paramètre θ_h^U est dans une classe paramétrique $\left[\theta_h^U|\pi_U,\gamma_U\right]$, où les hyperparamètres π_U et γ_U sont inconnus. L'idée principale consiste à estimer les hyperparamètres d'abord et de les replacer dans la loi a priori avant d'estimer la loi a posteriori (pour plus de détails, consulter [2, 3]).

On considère, $(Y_i, X_i)_{i=1:n_h}$, une suite de n_h réalisations indépendantes de $[Y,X]_h$. On note $n_{hU} = \sum_{i=1}^{n_h} \mathbb{1}(U(X_i)=1)$ le nombre d'observations i telles que $U(X_i)=1$. On suppose que n_{hU} est un entier connu et supérieur strictement à un. On note $S_{hU} = \sum_{i=1}^{n_{hU}} \mathbb{1}(Y_i=1, U(X_i)=1)$ une variable aléatoire qui détermine le nombre d'observations i telles que $U(X_i)=1$ et $Y_i=1$. On suppose que

$(S_{hU}|\theta_h^U)_{h=1:m}$ est une suite de variables aléatoires indépendantes mais pas nécessairement identiquement distribuées. Pour tout cluster h donné, on suppose que

$$\left[S_{hU}|\theta_h^U\right] = \text{Binomiale}(n_{hU}, \theta_h^U)$$

L'objectif est de trouver une estimation ponctuelle pour θ_h^U à partir des observations S_{hU}. On commence par déterminer la loi a posteriori de $\theta_h^U|\pi_U, \gamma_U$ qui dépend des données par S_{hU}. La loi a posteriori est donnée par :

$$\left[\theta_h^U|S_{hU}, \pi_U, \gamma_U\right] = \frac{\left[S_{hU}|\theta_h^U\right]\left[\theta_h^U|\pi_U, \gamma_U\right]}{[S_{hU}|\pi_U, \gamma_U]}$$

En supposant que les hyperparamètres π_U et γ_U sont inconnus, nous les estimerons à partir de la distribution marginale de toutes les données, $[S_{hU}|\pi_U, \gamma_U]$. On obtient la distribution a posteriori estimée :

$$\left[\theta_h^U|S_{hU}, \widehat{\pi}_U, \widehat{\gamma}_U\right]$$

où $\widehat{\pi}_U$ et $\widehat{\gamma}_U$ sont des fonctions de S_{hU} (i.e., $\widehat{\pi}_U(S_{hU})$ et $\widehat{\gamma}_U(S_{hU})$). Ces estimateurs sont habituellement obtenus par la méthode du maximum de vraisemblance (MLE) ou la méthode des moments (MOM) à partir de la distribution marginale $[S_{hU}|\pi_U, \gamma_U]$. Une fois les estimateurs $\{\widehat{\pi}_U, \widehat{\gamma}_U\}$ obtenus, nous pouvons estimer alors $\widehat{\theta}_h^U$ comme étant la moyenne de la distribution a posteriori estimée. Notons que, $\widehat{\theta}_h^U$ dépend de toutes les données par le biais de $\widehat{\pi}_U$ et $\widehat{\gamma}_U$. Dans cette analyse, nous proposons d'estimer les hyperparamètres $\widehat{\pi}_U$ et $\widehat{\gamma}_U$ par la méthode des moments.

4 Estimation des hyperparamètres π_U et γ_U

4.1 Estimation par la méthode des moments

Le principe de la méthode des moments consiste à estimer les paramètres recherchés en égalisant certains moments théoriques (qui dépendent de ces paramètres) avec leurs contreparties empiriques. L'égalisation se justifie par la loi des grands nombres qui implique que l'on peut "approcher" une espérance mathématique par une moyenne empirique. On est donc amené à résoudre un système d'équations.

4.1.1 Moments des variables S_{hU} et θ_h^U

Etant donné que la loi a priori de $\theta_h|\pi_U, \gamma_U$ est connue (la loi Bêta), il est possible de déterminer les expressions explicites de ses moments d'ordre un et deux. Nous commencerons par donner l'expression

des moments d'ordre n. Ensuite nous en déduirons les moments d'ordre un, deux, trois et quatre.

$$\mathbf{E}\left(\left(\theta_h^U\right)^n|\pi_U,\gamma_U\right) = \left[\frac{\Gamma\left[\frac{1-\gamma_U}{\gamma_U}\right]}{\Gamma\left[\frac{\pi_U(1-\gamma_U)}{\gamma_U}\right]}\right]\left[\frac{\Gamma\left[\frac{\pi_U(1-\gamma_U)}{\gamma_U}+n\right]}{\Gamma\left[\frac{1-\gamma_U}{\gamma_U}+n\right]}\right]$$

On obtient alors :

$$\mathbb{E}\left(\theta_h^U|\pi_U,\gamma_U\right) = \pi_U$$

$$\mathbb{E}\left(\left(\theta_h^U\right)^2|\pi_U,\gamma_U\right) = \pi_U^2+\gamma_U\pi_U(1-\pi_U)$$

Nous déduisons des moments de θ_h^U les moments suivants :

$$\begin{aligned}\mathbb{E}\left(S_{hU}\right) &= \mathbb{E}\left[\mathbb{E}\left(S_{hU}|\theta_h^U,\pi_U,\gamma_U\right)\right]\\ &= \mathbb{E}\left(\mathbb{E}\left(S_{hU}|\theta_h^U\right)|\pi_U,\gamma_U\right)\\ &= n_{hU}\pi_U\end{aligned}$$

$$\begin{aligned}\mathbb{V}ar(S_{hU}) &= \mathbb{E}[\mathbb{V}ar(S_{hU}|\pi_U,\gamma_U)]+\mathbb{V}ar[\mathbb{E}(S_{hU}|\pi_U,\gamma_U)]\\ &= n_{hU}\pi_U(1-\pi_U)+\gamma_U\pi_U(1-\pi_U)n_{hU}(n_{hU}-1)\end{aligned}$$

Nous supposons que les observations de n_{hU} sont strictement supérieures à 1 (i.e. $n_{hU}>1$). On obtient alors

$$\begin{cases}\mathbb{E}(S_{hU}) = n_{hU}\pi_U\\ \\ \mathbb{E}[(S_{hU})^2] = n_{hU}\pi_U(1-\pi_U+n_{hU}\pi_U)+\gamma_U\pi_U(1-\pi_U)n_{hU}(n_{hU}-1)\end{cases}$$

En faisant la différence membre à membre des deux égalités ci-dessus, on obtient les égalités suivantes

$$\begin{cases}\mathbb{E}\left(\frac{S_{hU}}{n_{hU}}\right) = \pi_U\\ \\ \mathbb{E}\left(\frac{S_{hU}}{n_{hU}}\frac{S_{hU}-1}{n_{hU}-1}\right) = \pi_U^2+\gamma_U\pi_U(1-\pi_U)\end{cases}$$

4.1.2 Estimation de π_U et γ_U

Dans ses travaux, Griffiths a montré que lorsque les n_{hU} sont inégaux, l'estimation des paramètres π_U et γ_U par des moments empiriques pondérés produit de meilleurs estimateurs que lorsque on utilise des moments empiriques non pondérés [4].

Si on suppose que les variables $\left(\frac{S_{hU}}{n_{hU}}\right)$ sont indépendantes et de variances non nulles de même que les variables $\left(\frac{S_{hU}}{n_{hU}}\frac{S_{hU}-1}{n_{hU}-1}\right)$, il est alors désirable d'utiliser leurs moments empiriques pondérés dans le but d'obtenir de meilleurs estimateurs de π_U et de γ_U.
Soit

$$W_U = \sum_{h=1}^{m} \frac{w_{hU}}{w_U}\left(\frac{S_{hU}}{n_{hU}}\right), \qquad w_U = \sum_{h=1}^{m} w_{hU} \tag{IV.3}$$

et

$$S_U = \sum_{h=1}^{m} \frac{v_{hU}}{v_U}\left(\frac{S_{hU}}{n_{hU}}\frac{S_{hU}-1}{n_{hU}-1}\right), \qquad v_U = \sum_{h=1}^{m} v_{hU} \tag{IV.4}$$

les moments empiriques respectifs de $\left(\frac{S_{hU}}{n_{hU}}\right)$ et $\left(\frac{S_{hU}}{n_{hU}}\frac{S_{hU}-1}{n_{hU}-1}\right)$, où w_{hU} et v_{hU} représente les coefficients de pondération respectifs. Dans la suite, on verra comment ils sont choisis.
En définissant les statistiques des équations (IV.3) et (IV.4) égales à leurs valeurs théoriques et en résolvant les équations qui en résultent par rapport à π_U et γ_U, nous obtenons les estimateurs suivants :

$$\widehat{\pi}_U = \sum_{h=1}^{m} \frac{w_{hU}}{w_U}\frac{S_{hU}}{n_{hU}} \tag{IV.5}$$

$$\widehat{\gamma}_U = \frac{\sum_{h=1}^{m} \frac{v_{hU}}{v_U}\left(\frac{S_{hU}}{n_{hU}}\frac{S_{hU}-1}{n_{hU}-1}\right) - \widehat{\pi}_U^2}{\widehat{\pi}_U(1-\widehat{\pi}_U)} \tag{IV.6}$$

Les estimateurs des moments pondérés dépendent du choix des poids $\{w_{hU},v_{hU}\}$. Il est très connu de la littérature que si $\{w_{hU},v_{hU}\}$ sont choisis proportionnellement aux variances respectives de $\left(\frac{S_{hU}}{n_{hU}}\right)$ et $\left(\frac{S_{hU}}{n_{hU}}\frac{S_{hU}-1}{n_{hU}-1}\right)$ alors W_U et S_U ont les plus petites variances parmi tous les estimateurs linéaires sans biais de π_U et γ_U respectivement. Si nous pondérons chaque variable $\frac{S_{hU}}{n_{hU}}$ et chaque variable $\frac{S_{hU}}{n_{hU}}\frac{S_{hU}-1}{n_{hU}-1}$ par l'inverse de sa variance (supposée être connue) alors W_U et S_U sont les estimateurs linéaires sans biais et de variance minimum de π_U et de $\pi_U^2+\gamma_U\pi_U(1-\pi_U)$ respectivement. Les poids correspondants sont :

$$\begin{aligned} \mathbb{V}ar\left(\frac{S_{hU}}{n_{hU}}\right) &= \frac{\pi_U(1-\pi_U)}{n_{hU}} + \gamma_U\pi_U(1-\pi_U)\left(1-\frac{1}{n_{hU}}\right) \\ \left[\mathbb{V}ar\left(\frac{S_{hU}}{n_{hU}}\right)\right]^{-1} &= \frac{n_{hU}}{\pi_U(1-\pi_U)+\gamma_U\pi_U(1-\pi_U)(n_{hU}-1)} \end{aligned}$$

$$\begin{aligned}\mathbb{V}ar\left(\frac{S_{hU}}{n_{hU}}\frac{S_{hU}-1}{n_{hU}-1}\right) &= \mathbb{E}\left[\left(\frac{S_{hU}}{n_{hU}}\frac{S_{hU}-1}{n_{hU}-1}\right)^2\right] - \left[\mathbb{E}\left(\frac{S_{hU}}{n_{hU}}\frac{S_{hU}-1}{n_{hU}-1}\right)\right]^2 \\ &= \mathbb{E}\left[\left(\frac{S_{hU}}{n_{hU}}\frac{S_{hU}-1}{n_{hU}-1}\right)^2\right] - \left[\pi_U^2+\gamma_U\pi_U(1-\pi_U)\right]^2\end{aligned}$$

avec

$$\mathbb{E}\left[\left(\frac{S_{hU}}{n_{hU}}\frac{S_{hU}-1}{n_{hU}-1}\right)^2\right] = \frac{1}{[n_{hU}(n_{hU}-1)]^2}\left[\mathbb{E}\left(S_{hU}^4\right)-2\mathbb{E}\left(S_{hU}^3\right)+\mathbb{E}\left(S_{hU}^2\right)\right]$$

$$\begin{aligned}\mathbb{E}\left(S_{hU}^2|\theta_h^U\right) &= n_{hU}\theta_h^U + n_{hU}\left(n_{hU}-1\right)\left(\theta_h^U\right)^2 \\ \mathbb{E}\left(S_{hU}^3|\theta_h^U\right) &= n_{hU}\theta_h^U + 2n_{hU}(n_{hU}-1)\left(\theta_h^U\right)^2 + n_{hU}(n_{hU}-1)(n_{hU}-2)\left(\theta_h^U\right)^3 \\ \mathbb{E}\left(S_{hU}^4|\theta_h^U\right) &= n_{hU}\theta_h^U + 4n_{hU}(n_{hU}-1)\left(\theta_h^U\right)^2 + 4n_{hU}(n_{hU}-1)(n_{hU}-2)\left(\theta_h^U\right)^3 \\ &+ n_{hU}(n_{hU}-1)(n_{hU}-2)(n_{hU}-3)\left(\theta_h^U\right)^4\end{aligned}$$

donc

$$\mathbb{E}\left[\left(S_{hU}^2-S_{hU}\right)^2|\theta_h^U\right] = n_{hU}(n_{hU}-1)\left[\left(\theta_h^U\right)^2+2(n_{hU}-2)\left(\theta_h^U\right)^3+(n_{hU}-2)(n_{hU}-3)\left(\theta_h^U\right)^4\right]$$

$$\begin{aligned}\mathbb{E}\left[\left(\frac{S_{hU}}{n_{hU}}\frac{S_{hU}-1}{n_{hU}-1}\right)^2\right] &= \frac{1}{n_{hU}(n_{hU}-1)}\left[\mathbb{E}\left(\left(\theta_h^U\right)^2|\pi_U,\gamma_U\right)+2(n_{hU}-2)\mathbb{E}\left(\left(\theta_h^U\right)^3|\pi_U,\gamma_U\right)\right. \\ &\left.+(n_{hU}-2)(n_{hU}-3)\mathbb{E}\left(\left(\theta_h^U\right)^4|\pi_U,\gamma_U\right)\right]\end{aligned}$$

$$\begin{aligned}\mathbb{E}\left[\left(\frac{S_{hU}}{n_{hU}}\frac{S_{hU}-1}{n_{hU}-1}\right)^2\right] = \frac{\mathbb{E}\left(\left(\theta_h^U\right)^2|\pi_U,\gamma_U\right)}{n_{hU}(n_{hU}-1)}&\left[1+2(n_{hU}-2)\frac{\pi_U(1-\gamma_U)+2\gamma_U}{1+\gamma_U}\right. \\ &\left.+(n_{hU}-2)(n_{hU}-3)\frac{\pi_U(1-\gamma_U)+2\gamma_U}{1+\gamma_U}\frac{\pi_U(1-\gamma_U)+3\gamma_U}{1+2\gamma_U}\right]\end{aligned}$$

Puisque $\pi_U(1-\pi_U)$ est constant (indépendant de h), alors nous considérons pour w_{hU} la valeur suivante :

$$w_{hU} = \frac{n_{hU}}{1+\gamma_U(n_{hU}-1)} \tag{IV.7}$$

et pour v_{hU} la valeur suivante :

$$v_{hU} = \frac{1}{\mathbb{E}\left[\left(\frac{S_{hU}}{n_{hU}}\frac{S_{hU}-1}{n_{hU}-1}\right)^2\right] - \left[\pi_U^2 + \gamma_U \pi_U (1-\pi_U)\right]^2} \tag{IV.8}$$

Cependant l' estimation du paramètre w_{hU} et du paramètre v_{hU} est compliquée par le fait que tous les deux paramètres dépendent des paramètres π_U et γ_U inconnus. Une manière de les estimer consisterait à remplacer π_U et γ_U par leurs estimations respectives $\widehat{\pi}_U$ et $\widehat{\gamma}_U$ dans les équations (IV.7) et (IV.8). Cependant, lorsque m le nombre de cluster n'est pas suffisamment grand, la loi des grands nombres ne s'applique pas et par conséquent, les moments empiriques W_U et S_U n'approchent pas suffisamment bien les moments théoriques. En plus le signe de $\widehat{\gamma}_U$ dépend de la suite (S_{hU}, n_{hU}). Les estimateurs ainsi obtenus peuvent avoir tendance à sortir du support des paramètres (voir annexe C).
Pour parer à cette difficulté, une méthode de pondération empirique a été proposée en premier par Kleinman en [1973][6] puis améliorée par Tchuang-Stein en [1993][1] pour l'estimation de w_{hU}. A partir de cet algorithme, une estimation de $\widehat{\pi}_U$ a été déduite. Nous nous sommes inspirés de cette méthode pour établir l' algorithme d'estimation de π_U et de γ_U décrit ci-dessous.

4.1.3 Algorithme de la méthode de Pondération Empirique

On propose de choisir une valeur initiale $\gamma_0 = 0$ ou $\gamma_0 = 1$ du paramètre γ_U pour obtenir les valeurs initiales w_0 et v_0 de w_{hU} et v_{hU} respectivement. Ensuite on utilise les équations (IV.5) et (IV.6) pour obtenir les estimations de π_U et de γ_U. A partir de cette estimation de γ_U, notée $\widehat{\gamma}_U$, on calcule le couple $\{w_{hU}, v_{hU}\}$ à partir des équations (IV.7) et (IV.8). Et enfin on utilisera ces poids empiriques pour former de nouvelles estimations de W_U et S_U. On répète cette itération jusqu'à ce que les différences entre deux itérations consécutives d'estimations W_U, S_U et $\widehat{\gamma}_U$ soient à la fois plus petites qu'une certaine valeur prédéterminée, par défaut 10^{-6}. Pour des soucis de programmation, nous proposons de réinitialiser à 10^{-6} les estimations négatives de γ_U au lieu de 0 comme proposé par Kleinman.
Pour des raison de programmation, nous avons ajouté la masse de Dirac au point 0 de n_{hU} dans le calcul des statistiques W_U et S_U. Dans la simulation, il n'est pas évident d'avoir toutes les statistiques $(n_{hU})_{k=1}^k$ supérieures strictement à 1. En utilisant cette astuce, nous nous assurons que les dénominateurs de $S_{hU}/[n_{hU} + \delta_0(n_{hU})]$ et $S_{hU}(S_{hU}-1)/[n_{hU}(n_{hU}-1)\delta_0(n_{hU}(n_{hU}-1)) + \delta_0(n_{hU}(n_{hU}-1))]$ soient toujours égaux à 1 si n_{hU} est égale à un ou zéro. Dans le cas où $n_{hU} = 0$, on sait que S_{hU} est presque sûrement nulle. Ceci nous permet de pouvoir faire des estimations de π_U et de γ_U même s'il existe des réalisations $(Y_i, X_i)_{i=1:n_h}$ de $[Y, X]_h$ pour lesquelles le profil $U(X)$ n'a pas été observé $(U(X) = 0)$.

Algorithme : Méthode de pondération empirique

on suppose avoir observé les statistiques suivantes : $(S_{hU})_{h=1:H}$ et $(n_{hU})_{h=1:H}$
on commence par donner une valeur initiale $\gamma_U = 0$ ou $\gamma_U = 1$ et le nombre d'itérations maximum de la procédure : $maxiter = 100$ (par défaut)
on initialise
- $W_U = \frac{1}{K}\sum_{h=1}^{H} \frac{S_{hU}}{n_{hU}+\delta_0(n_{hU})}$
- $S_U = \frac{1}{K}\sum_{h=1}^{H} \frac{S_{hU}(S_{hU}-1)}{n_{hU}(n_{hU}-1)\delta_0(n_{hU}(n_{hU}-1))+\delta_0(n_{hU}(n_{hU}-1))}$

Déclarer une variable booléenne *cond.arret* (condition d'arrêt) initialisée à *vrai* et une variable t initialisée à 0.
Tant que *cond.arret* est toujours vrai **faire** :
initialiser : $t = t + 1$; $\gamma_U^t = \gamma_U$; $\pi_U^t = W_U$ et $S_U^t = S_U$
calculer en fonction de π_U^t et γ_U^t le couple $\{w_{hU}, v_{hU}\}$
En suite calculer les statistiques :
- $W_U = \sum_{h=1}^{m} \frac{w_{hU}}{w_U}\left(\frac{S_{hU}}{n_{hU}+\delta_0(n_{hU})}\right)$
- $S_U = \sum_{h=1}^{m} \frac{v_{hU}}{v_U}\left(\frac{S_{hU}(S_{hU}-1)}{n_{hU}(n_{hU}-1)\delta_0(n_{hU}(n_{hU}-1))+\delta_0(n_{hU}(n_{hU}-1))}\right)$

Puis on associe $\pi_U = W_U$ et $\gamma_U = \frac{S_U - \pi_U^2}{\pi_U(1-\pi_U)}$
- si $\gamma_U < 0 \quad \Rightarrow \quad \gamma_U = 10^{-6}$

cond.arret $= \{|\gamma_U - \gamma_U^t| > 10^{-6}, |\pi_U - \pi_U^t| > 10^{-6}, |S_U - S_U^t| > 10^{-6}, t < maxiter\}$
fin tant que

Tableau IV.1 – Algorithme de la méthode de pondération empirique

4.2 Estimation des hyperparamètres par la méthode du maximum de vraisemblance

Pour simplifier les notations, on a choisi d'omettre l'indice U sur les paramètres π et γ. De plus on considère le changement de paramètre $\tau = \frac{1}{\gamma} - 1$.

4.2.1 Vraisemblance des paramètres

On a

$$\begin{aligned}
[(S_h)_{h=1:m}|\pi,\tau] &= \prod_{h=1}^{m}[S_h|\pi,\tau] \\
&= \prod_{h=1}^{m}\binom{s_h}{n_h}\frac{\Gamma(\tau)}{\Gamma(\tau+n_h)}\frac{\Gamma(\pi\tau+s_h)}{\Gamma(\pi\tau)}\frac{\Gamma((1-\pi)\tau+n_h-s_h)}{\Gamma((1-\pi)\tau)} \\
&= \prod_{h=1}^{m}\left\{\binom{s_h}{n_h}\left[\prod_{j=0}^{n_h-1}\frac{1}{\tau+j}\right]\left[\prod_{k=0}^{s_h-1}(\pi\tau+k)\right]\left[\prod_{l=0}^{n_h-s_h-1}((1-\pi)\tau+l)\right]\right\}
\end{aligned}$$

La vraisemblance des paramètres π et τ est donnée par :

$$L(\pi,\tau) = \sum_{h=1}^{m}\left\{log\left(\binom{s_h}{n_h}\right) - \sum_{j=0}^{n_h-1} log(\tau+j) + \sum_{k=0}^{s_h-1} log(\pi\tau+k) + \sum_{l=0}^{n_h-s_h-1} log((1-\pi)\tau+l)\right\}$$

L'optimisation de la vraisemblance $L(\pi,\tau)$ est très compliquée à implémenter. Il n'est pas possible de trouver une solution analytique. Cependant plusieurs algorithmes itératifs ont été proposés dans la littérature pour venir à bout cette difficulté. Dans cette analyse, nous proposons d'utiliser un algorithme MM.

4.2.2 Présentation du principe et des éléments d'un algorithme MM

Nous allons utiliser l'algorithme MM (Minimisation-Maximisation) pour estimer les paramètres π et τ. Les algorithmes MM ont pour objectif de substituer à un problème d'optimisation numérique d'une fonction f compliquée à implémenter par celui de l'optimisation d'une fonction auxiliaire g dont l'optimum correspond à un optimum local de f. La fonction auxiliaire g est telle que

$$\begin{aligned} f(x) &\geq g(x|x') \qquad x \in \Delta \times \mathbf{D} \\ f(x) &= g(x|x) \end{aligned}$$

On observe que si pour x_0 fixé et $x_1 = \underset{x}{\text{argmax}}\, g(x|x_0)$, alors on a $f(x_1) \geq g(x_1|x_0) \geq g(x_0|x_0) = f(x_0)$ Il en résulte que les algorithmes MM sont des algorithmes monotones. Les algorithmes MM procèdent en deux étapes. La première étape consiste à trouver la fonction g telle que

$$L(\pi,\tau) \geq g(\pi,\tau|\pi',\tau') \tag{IV.9}$$
$$L(\pi,\tau) = g(\pi,\tau|\pi,\tau) \quad \forall\ (\pi,\tau) \tag{IV.10}$$

La deuxième étape consiste à trouver un couple $(\widehat{\pi},\widehat{\tau})$ qui maximise la fonction $g(\pi,\tau|\pi',\tau')$.

$$(\widehat{\pi},\widehat{\tau}) = \underset{(\pi,\tau)}{\text{argmax}}\, g(\pi,\tau|\pi',\tau')$$

4.2.3 Proposition de la fonction auxiliaire et ses propriétés

Proposition 8. Soient $L(\pi,\tau)$ la log-vraisemblance du couple des paramètres (π,τ) et (π',τ') une valeur connue des paramètres (π,τ). La fonction auxiliaire $g(\pi,\tau|\pi',\tau')$ définie par

$$g(\pi,\tau|\pi',\tau') = A(\pi',\tau')[log(\pi)+log(\tau)] + B(\pi',\tau')[log(1-\pi)+log(\tau)] - (\tau-\tau')C(\tau') + D(\pi',\tau')$$

vérifie les conditions (IV.9) et (IV.10).

Preuve. Puisque la fonction $-log(x)$ est convexe, on a :

$$-log(\tau+j) \geq -log(\tau'+j) - \frac{(\tau-\tau')}{\tau'+j}$$

En utilisant la concavité de la fonction $log(x)$, on obtient

$$log(\pi\tau+k) \geq \frac{\pi'\tau'}{\pi'\tau'+k} log\left(\frac{\pi'\tau'+k}{\pi'\tau'}\pi\tau\right) + \frac{k}{\pi'\tau'+k} log\left(\frac{\pi'\tau'+k}{k}k\right)$$

$$log((1-\pi)\tau+l) \geq \frac{(1-\pi')\tau'}{(1-\pi)'\tau'+l} log\left(\frac{(1-\pi')\tau'+l}{(1-\pi')\tau'}(1-\pi)\tau\right) + \frac{l}{(1-\pi')\tau'+l} log\left(\frac{(1-\pi')\tau'+l}{l}l\right)$$

On peut donc poser

$$\begin{aligned}
g(\pi,\tau|\pi',\tau') &= \sum_{h=1}^{m} log\left(\binom{s_h}{n_h}\right) - \sum_{h=1}^{m}\sum_{j=0}^{n_h-1} log(\tau'+j) - (\tau-\tau')\sum_{h=1}^{m}\sum_{j=0}^{n_h-1}\frac{1}{\tau'+j} \\
&+ \sum_{h=1}^{m} \mathbb{1}(s_h \geq 1)\left(\sum_{k=0}^{s_h-1}\frac{\pi'\tau'}{\pi'\tau'+k}\left[log\left(\frac{\pi'\tau'+k}{\pi'\tau'}\pi\tau\right)\right] + \sum_{h=1}^{m}\sum_{k=0}^{s_h-1}\frac{k}{\pi'\tau'+k} log\left(\pi'\tau'+k\right)\right) \\
&+ \sum_{h=1}^{m} \mathbb{1}(n_h > s_h)\left(\sum_{l=0}^{n_h-s_h-1}\frac{(1-\pi')\tau'}{(1-\pi)'\tau'+l}\left[log\left(\frac{(1-\pi')\tau'+l}{(1-\pi')\tau'}(1-\pi)\tau\right)\right]\right) \\
&+ \sum_{h=1}^{m} \mathbb{1}(n_h > s_h)\left(\sum_{l=0}^{n_h-s_h-1}\frac{l}{(1-\pi')\tau'+l} log\left((1-\pi')\tau'+l\right)\right)
\end{aligned}$$

On peut réécrit la fonction $g(\pi,\tau|\pi',\tau')$ de telle sorte que les paramètres π et τ soient séparés. On obtient

$$\begin{aligned}
g(\pi,\tau|\pi',\tau') &= \sum_{h=1}^{m} log\left(\binom{s_h}{n_h}\right) - \sum_{h=1}^{m}\sum_{j=0}^{n_h-1} log(\tau'+j) - (\tau-\tau')\sum_{h=1}^{m}\sum_{j=0}^{n_h-1}\frac{1}{\tau'+j} \\
&+ \sum_{h=1}^{m} \mathbb{1}(s_h \geq 1\left(\sum_{k=0}^{s_h-1}\frac{\pi'\tau'}{\pi'\tau'+k}\left[log\left(\frac{\pi'\tau'+k}{\pi'\tau'}\right) + log(\pi) + log(\tau)\right]\right) \\
&+ \sum_{h=1}^{m} \mathbb{1}(s_h \geq 1)\left(\sum_{k=0}^{s_h-1}\frac{k}{\pi'\tau'+k} log\left(\pi'\tau'+k\right)\right) \\
&+ \sum_{h=1}^{m} \mathbb{1}(n_h > s_h)\left(\sum_{l=0}^{n_h-s_h-1}\frac{(1-\pi')\tau'}{(1-\pi)'\tau'+l}\left[log\left(\frac{(1-\pi')\tau'+l}{(1-\pi')\tau'}\right) + log(1-\pi) + log(\tau)\right]\right) \\
&+ \sum_{h=1}^{m} \mathbb{1}(n_h > s_h)\left(\sum_{l=0}^{n_h-s_h-1}\frac{l}{(1-\pi')\tau'+l} log\left((1-\pi')\tau'+l\right)\right)
\end{aligned}$$

Si on pose

$$\begin{aligned}
A(\pi',\tau') &= \sum_{h=1}^{m} \mathbb{1}\left(s_h \geq 1\right)\left(\sum_{k=0}^{s_h-1}\frac{\pi'\tau'}{\pi'\tau'+k}\right) \\
B(\pi',\tau') &= \sum_{h=1}^{m} \mathbb{1}\left(n_h \geq s_h\right)\left(\sum_{l=0}^{n_h-s_h-1}\frac{(1-\pi')\tau'}{(1-\pi)'\tau'+l}\right) \\
C(\tau') &= \sum_{h=1}^{m}\sum_{j=0}^{n_h-1}\frac{1}{\tau'+j}
\end{aligned}$$

$$
\begin{aligned}
D(\pi',\tau') &= \sum_{h=1}^{m} log\left(\binom{s_h}{n_h}\right) - \sum_{h=1}^{m}\sum_{j=0}^{n_h-1} log(\tau'+j) + \sum_{h=1}^{m}\mathbb{1}(s_h\geq 1)\left(\sum_{k=0}^{s_h-1}\frac{\pi'\tau'}{\pi'\tau'+k} log\left(\frac{\pi'\tau'+k}{\pi'\tau'}\right)\right) \\
&+ \sum_{h=1}^{m}\mathbb{1}(s_h\geq 1)\left(\sum_{k=0}^{s_h-1}\frac{k}{\pi'\tau'+k} log\left(\pi'\tau'+k\right)\right) + \sum_{h=1}^{m}\mathbb{1}(n_h> s_h)\left(\sum_{l=0}^{n_h-s_h-1}\frac{(1-\pi')\tau'}{(1-\pi)'\tau'+l} log\left(\frac{(1-\pi')\tau'+l}{(1-\pi')\tau'}\right)\right) \\
&+ \sum_{h=1}^{m}\mathbb{1}(n_h> s_h)\left(\sum_{l=0}^{n_h-s_h-1}\frac{l}{(1-\pi')\tau'+l} log\left((1-\pi')\tau'+l\right)\right)
\end{aligned}
$$

Il en résulte que

$$g(\pi,\tau|\pi',\tau') = Cste - (\tau-\tau')C(\tau') + A(\pi',\tau')[log(\pi)+log(\tau)] + B(\pi',\tau')[log(1-\pi)+log(\tau)]$$

On a

$$L(\pi,\tau)\geq g(\pi,\tau|\pi',\tau')$$

En plus lorsque on pose $\pi=\pi'$ et $\tau=\tau'$, on obtient

$$L(\pi,\tau) = g(\pi,\tau|\pi,\tau)$$

□

Les couples candidats sont l'ensemble des couples annulant les dérivées partielles de la fonction $g(\pi,\tau|\pi',\tau')$.

$$
\begin{aligned}
\frac{\delta}{\delta\pi}g(\pi,\tau|\pi',\tau') &= \frac{1}{\pi}A(\pi',\tau') - \frac{1}{1-\pi}B(\pi',\tau') \\
\frac{\delta}{\delta\tau}g(\pi,\tau|\pi',\tau') &= -C(\tau') + \frac{1}{\tau}\left[A(\pi',\tau')+B(\pi',\tau')\right]
\end{aligned}
$$

Il en résulte que

$$\widehat{\pi} = \frac{A(\pi',\tau')}{A(\pi',\tau')+B(\pi',\tau')} \tag{IV.11}$$

$$\widehat{\tau} = \frac{A(\pi',\tau')+B(\pi',\tau')}{C(\tau')} \tag{IV.12}$$

En plus on a

$$\frac{\delta^2}{\delta^2\pi}g(\pi,\tau|\pi',\tau') = -\frac{1}{\pi^2}A(\pi',\tau') - \frac{1}{(1-\pi)^2}B(\pi',\tau') \tag{IV.13}$$

$$\frac{\delta^2}{\delta^2\tau}g(\pi,\tau|\pi',\tau') = -\frac{1}{\tau^2}\left[A(\pi',\tau')+B(\pi',\tau')\right] \tag{IV.14}$$

Par conséquent on a

$$\frac{\delta^2}{\delta^2\tau}g(\widehat{\pi},\widehat{\tau}|\pi',\tau') = -C(\tau')\leq 0$$

et

$$\frac{\delta^2}{\delta^2\pi}g(\widehat{\pi},\widehat{\tau}|\pi',\tau') = -\left(\frac{(1-\widehat{\pi}^2)A(\pi',\tau') + \widehat{\pi}^2 B(\pi',\tau')}{\widehat{\pi}^2(1-\widehat{\pi})^2}\right) \leq 0$$

Le couple $(\widehat{\pi},\widehat{\tau})$ donnée par les équations (IV.11) et (IV.12) est donc un maximum local de la fonction $g(\pi,\tau|\pi',\tau')$. En se servant des équations (IV.9) et (IV.10), on obtient $L(\widehat{\pi},\widehat{\tau}) \geq L(\pi',\tau')$. Le couple $(\widehat{\pi},\widehat{\tau})$ maximisant la vraisemblance est atteint lorsque la condition d'arrêt (**??**) est obtenue.

4.2.4 Algorithme

La phase de maximisation consiste à maximiser la fonction $g(\pi,\tau|\pi',\tau')$. Cette dernière partie correspond à l'algorithme numérique itératif de newton pour optimiser la fonction g. Le principe de l'algorithme est le suivant :

Algorithme : MM (Minimisation-Maximisation)

- Entrées : $\mathcal{D} = \{(s_h, n_h); h = 1 : m\}$ un ensemble d'observations ; (π^0, τ^0) valeurs initiales des paramètres à estimer et *maxiter* le nombre d'itération maximum.
- Sortie : le couple $(\widehat{\pi},\widehat{\tau})$

Variables déclarées :

- *cond.arret* : une variable booléenne initialisée à vrai
- t : étape itérative initialisée à 0
- $(\pi^t, \tau^t) \leftarrow (\pi^0, \tau^0)$

Tant que *cond.arret* est vrai **faire** :

On itère $t \leftarrow t+1$ et $\left(\pi^{(t-1)}, \tau^{(t-1)}\right) \leftarrow \left(\pi^{(t)}, \tau^{(t)}\right)$

- $\pi^{(t)} \leftarrow \dfrac{A\left(\pi^{(t-1)}, \tau^{(t-1)}\right)}{A\left(\pi^{(t-1)}, \tau^{(t-1)}\right) + B\left(\pi^{(t-1)}, \tau^{(t-1)}\right)}$
- $\tau^{(t)} \leftarrow \dfrac{A\left(\pi^{(t-1)}, \tau^{(t-1)}\right) + B\left(\pi^{(t-1)}, \tau^{(t-1)}\right)}{C\left(\tau^{(t-1)}\right)}$

$$cond.arret \leftarrow \left\{\left(\left(\pi^{(t)} - \pi^{(t-1)}\right)^2 + \left(\tau^{(t)} - \tau^{(t-1)}\right)^2 + 1 \neq 1\right) \;\&\; (t < maxietr)\right\}$$

fin tant que

résultats : $\left(\pi^{(t)}, \tau^{(t)}\right)$

Tableau IV.2 – Algorithme MM (Minimisation-Maximisation)

5 Eléments pour la formulation d'un classifieur individuel pour les groupes

Soit $U(X)$ un profil donné. Nous observons S_{hU} co-occurrences dans n_{hU} observations pertinentes pour un cluster h donné. Nous modélisons le nombre de co-occurrences par une loi $Binomiale(n_{hU}, \theta_h^U)$ et θ_h^U par une loi $Beta(\pi_U(1-\gamma_U)/\gamma_U, (1-\pi_U)(1-\gamma_U)/\gamma_U)$ de manière hiérarchique pour partager l'information entre les clusters similaires. De manière plus formelle, nous proposons le modèle suivant :

$$\begin{aligned} S_{hU} &\sim \text{Binom}(n_{hU}, \theta_h^U) \\ \theta_h^U &\sim \text{Beta}\left(\frac{\pi_U(1-\gamma_U)}{\gamma_U}, \frac{(1-\pi_U)(1-\gamma_U)}{\gamma_U}\right) \end{aligned}$$

Sous ces hypothèses, on a

$$\begin{aligned} \mathbb{E}\left(\theta_h^U | [Y,X]_h, \pi_U, \gamma_U\right) &= \int_0^1 \Pr\left(Y=1 | U(X)=1, [Y,X]_h\right) \left[\theta_h^U | [Y,X]_h, \pi_U, \gamma_U\right] d\theta_h^U \\ \mathbb{E}\left(\theta_h^U | [Y,X]_h, \pi_U, \gamma_U\right) &= \frac{S_{hU} + \pi_U\left(\frac{1}{\gamma_U} - 1\right)}{n_{hU} + \left(\frac{1}{\gamma_U} - 1\right)} \end{aligned}$$

La valeur prédictive positive a posteriori est donnée par

$$VPP(U,Y,h) = \mathbb{E}\left(\mathbb{E}\left(\theta_h^U | Y, \widehat{\pi}_U, \widehat{\gamma}_U\right)\right)$$

Puisqu'on n'a pas supposé une loi a priori sur les hyperparamètres π_U et γ_U, alors leurs estimations sont faites à partir des données $([Y,X]_h)_{h=1:m}$. Par conséquent la valeur prédictive positive a posteriori obtenue est un estimateur empirique de Bayes de la valeur prédictive positive du classifieur $\phi(X,U)$ généré par le profil $U(X)$.

$$VPP(U,Y,h) = \frac{S_{hU} + \widehat{\pi}_U\left(\frac{1}{\widehat{\gamma}_U} - 1\right)}{n_{hU} + \left(\frac{1}{\widehat{\gamma}_U} - 1\right)}$$

Pour chaque profil $U(X)$ fixé, on a une suite $(VPP(U,Y,h))_{h=1:m}$ dont chaque $VPP(U,Y,h)$ dépend des observations de la loi $[Y,X]_h$. $VPP(U,Y,h)$ est une estimation de la valeur prédictive positive du profil $U(X)$ dans le cluster h en tenant compte de ses fréquences dans les autres clusters. On peut écrire $VPP(U,Y,h)$ sous la forme d'une combinaison linéaire convexe de $\dfrac{S_{hU}}{n_{hU}}$ et de π_U. :

$$VPP(U,Y,h) = \frac{S_{hU}}{n_{hU}}\left(1 - \frac{(1-\widehat{\gamma}_U)/\widehat{\gamma}_U}{n_{hU} + (1-\widehat{\gamma}_U)/\widehat{\gamma}_U}\right) + \widehat{\pi}_U\left(\frac{(1-\widehat{\gamma}_U)/\widehat{\gamma}_U}{n_{hU} + (1-\widehat{\gamma}_U)/\widehat{\gamma}_U}\right)$$

La statistique $\frac{S_{hU}}{n_{hU}}$ représente la valeur prédictive positive du profil $U(X)$ dans le cluster h indépendamment des autres clusters. Tandis que π_U représente la valeur prédictive positive du profil $U(X)$ dans la population.

$$VPP(U,Y) = \widehat{\pi}_U$$

Pour prédire la classe d'une observation dans un cluster h spécifié, on pourra utiliser la statistique $VPP(U,Y,h)$. Par contre, lorsqu'il s'agira de prédire la classe d'une observation dant le cluster n'est pas spécifié ou n'a pas participé à l'estimation des paramètres π_U et γ_U, on pourra se servir de la statistique $VPP(U,Y)$.

Pour adapter la procédure d'apprentissage étudiée dans le chapitre II à une analyse hiérarchique, nous allons construire l'algorithme de la recherche de l'ensemble optimal au tour de la valeur prédictive positive $VPP(U,Y,h)$ du classifieur $U(X)$ pour un cluster h donné.

Si on note par $\phi_h(U,X) = \delta_h(C)U(X)$ le classifieur généré par le profil U pour le cluster h et par $\mathcal{D} = \{(y_i, x_i, c_i); i = 1 : n\}$ l'ensemble des observations du triplet de variables (Y,X,C). On peut interpréter la sensibilité du classifieur $\phi_h(U,X)$ pour le cluster h, $\Pr\{\phi_h(U,X) = 1 \mid Y = 1,\mathcal{D}\}$, comme une fonctionnelle de la loi a posteriori de $\phi_h(U,X)$ conditionnellement aux données $\mathcal{D}$ et à $Y = 1$. Tenant compte que

$$\Pr\{\phi_h(U,X) = 1 \mid Y = 1,\mathcal{D}\} = VPP(U,Y,h)\frac{\Pr\{\phi_h(U,X) = 1 \mid \mathcal{D}\}}{\Pr\{Y = 1 \mid \mathcal{D}\}}$$

on a

$$\frac{\Pr\{\phi_h(U',X) = 1 \mid Y = 1,\mathcal{D}\}}{\Pr\{\phi_h(U,X) = 1 \mid Y = 1,\mathcal{D}\}} = \left[\frac{\Pr\{\phi_h(U',X) = 1 \mid \mathcal{D}\}}{\Pr\{\phi(U,X) = 1 \mid \mathcal{D}\}}\right]\left[\frac{VPP(U',Y,h)}{VPP(U,Y,h)}\right]$$

D'où l'interprétation du quotient $\frac{VPP(U',Y,h)}{VPP(U,Y,h)}$ comme un facteur de Bayes. Comme $U' \prec U$ alors $\frac{\Pr\{\phi_h(U',X) = 1 \mid Y = 1,\mathcal{D}\}}{\Pr\{\phi_h(U,X) = 1 \mid Y = 1,\mathcal{D}\}} \leq 1$. Plus grand est le facteur de Bayes, donc en faveur du classifieur $\phi(U',X)$, plus proche de 1 sera le quotient $\frac{\Pr\{\phi_h(U',X) = 1 \mid Y = 1,\mathcal{D}\}}{\Pr\{\phi_h(U,X) = 1 \mid Y = 1,\mathcal{D}\}}$. Suivant le point de vue exprimé par Kass & Raftery (1995) [5] à savoir, "Le facteur de Bayes est un résumé des preuves fournies par les données en faveur d'une théorie scientifique par un modèle statistique, par opposition aux théories alternatives", on considère la grille ci-dessous pour interpréter le facteur de Bayes en faveur ou non du classifieur associé au profil le plus détaillé $U' \prec U$:

Facteur de Bayes	Interprétation
1-3.2	on ne peut pas soutenir que le profil U' est un meilleur classifieur que U
3.2-10	on peut soutenir que U' est un meilleur classifieur que U
10-100	On peut fortement soutenir que U' est un meilleur classifieur que U
≥ 100	il n'y pas de doute que U' est un meilleur classifieur que U

6 Algorithme de la procédure d'apprentissage

L'adoption de l'algorithme d'apprentissage au cas où les données sont hétérogènes nécessite au préalable un prétraitement des données. En premier lieu, il faut discrétiser les variables numériques, si il en existe, en utilisant l'une des méthodes étudiées au chapitre III. En deuxième lieu, il faut subdiviser les données en trois sous-ensembles : un ensemble d'apprentissage, un ensemble de validation et un ensemble test. La procédure de construction du classifieur peut être résumée en deux grandes étapes. Une fois que nous avons fini de construire le classifieur, il nous reste à évaluer ses performances sur l'ensemble test. Ceci constitue la troisième étape de la procédure d'apprentissage.

1. **Etape 1** : A partir d'un ensemble d'apprentissage
   (a) Générer un ensemble de profils fréquents $\mathcal{U}_\lambda$, en utilisant le paramètre d'apprentissage $\lambda = (s_0, c_0, l_0)$
   (b) Elaguer les profils redondants dans l'ensemble $\mathcal{U}_\lambda$
   (c) Sélectionner les profils qui sont significativement corrélés avec la variable réponse (test fisher)
2. **Etape 2** : A partir d'un ensemble de validation
   (a) Pour chaque profil U : Estimer $\widehat{\pi}_U$ et $\widehat{\gamma}_U$ (par MOM ou MLE)
   (b) Pour chaque cluster h
      i. Estimer la valeur prédictive positive a posteriori de chaque profil U

$$VPP(U, Y, h) = \frac{\sum_{i=1}^{n} Y_i \delta_h(c_i)\, \phi(U, x_i) + \widehat{\pi}_U \left(\frac{1}{\widehat{\gamma}_U} - 1\right)}{\sum_{i=1}^{n} \delta_h(c_i)\, \phi(U, x_i) + \left(\frac{1}{\widehat{\gamma}_U} - 1\right)}$$

      ii. Si il existe deux profils U et U' tels que U' soit emboîté dans U :
         A. Calculer le facteur de Bayes

$$BF(U', U) = \frac{VPP(U', Y, h)}{VPP(U, Y, h)}$$

         B. On supprime le profil U si $BF(U', U) \geq 100$. Sinon on supprime le profil U'.
   (c) fin pour

   Au sortir des étapes 1 et 2, on obtient un ensemble optimal de profils $\mathcal{U}_\lambda^h$.
3. **Etape 3** : A partir d'un ensemble test
   (a) Pour chaque cluster h

i. Définir la règle de classement (classifieur) ϕ d'une observation X par

$$\phi(X,\lambda) = \begin{cases} 1 & \text{si} \quad \sum_{j=1}^{|\mathcal{U}_\lambda^h|} \phi(X,U_j) > 0 \\ \\ 0 & \text{sinon} \end{cases}$$

Le classifieur $\phi(X,\lambda)$ est un cas particulier du classifieur défini au chapitre II à la section 3.2 où on a choisi k égale à zéro. On choisit alors de classer positive une observation X lorsqu'elle vérifie au moins un profil parmi ceux qui sont dans l'ensemble $\mathcal{U}_\lambda^h$.

La première étape consiste à générer $\mathcal{U}_\lambda$ un ensemble de profils à la fois fréquents et significativement corrélés avec la variable réponse, où λ est un paramètre d'apprentissage à spécifier par l'utilisateur. D'ailleurs c'est pour des raisons d'insuffisance de mémoire que le paramètre λ est utilisé. Sinon l'idéal est de générer tous profils existant dans l'ensemble d'apprentissage. Dans la deuxième étape, il est aussi question d'estimation les paramètres π et γ pour chaque profil appartenant à $\mathcal{U}_\lambda$ et de construire un ensemble $\mathcal{U}_\lambda^h$ spécifique à chaque cluster h.
Cette procédure nécessite de subdiviser des données en trois sous-ensembles : apprentissage, validation et test. Il faut subdiviser les données de telle sorte que tous les clusters soient représentés dans chaque sous-ensemble avec la même proportion que dans l'ensemble de départ.

Bibliographie

[1] Chuang-Stein, C. An application of the beta-binomial model to combine and monitor medical event rates in clinical trials. *Drug Information Journal 27*, 2 (1993), 515–523. 113

[2] Efron, B., and Morris, C. Empirical bayes on vector observations : An extension of stein's method. *Biometrika 59*, 2 (1972), 335. 108

[3] Efron, B., and Morris, C. N. Multivariate empirical bayes and estimation of covariance matrices, 1974. 108

[4] Griffiths, D. A. Maximum likelihood estimation for the beta-binomial distribution and an application to the household distribution of the total number of cases of a disease. *Biometrics 29*, 4 (1973), 637. 110

[5] Kass, R., and Raftery, A. Bayes factors. *Journal of the American Statistical Association 90*, 430 (1995), 773–795. 120

[6] Kleinman, J. C. Proportions with extraneous variance : Single and independent sample. *Journal of the American Statistical Association 68*, 341 (1973), 46. 113

[7] Morris, C. N. Parametric empirical bayes inference : Theory and applications. *Journal of the American Statistical Association 78*, 381 (1983), 47. 108

[8] Robbins, H. Asymptotically subminimax solutions of compound statistical decision problems. In *Second Berkeley Symposium on Mathematical Statistics and Probability* (1951), vol. -1, pp. 131–149. 108

Appendices

Annexe C

Annexe Chapitre IV

B.1 Existence de l'estimation des moments des paramètres d'une Bêta-Binomiale

Généralement on pose

$$\frac{S_k}{n_k} = \frac{1}{n_k}\sum_{i=1}^{n_k} Y_{ki}$$

On a

$$\mathbb{V}ar\left(\left[\frac{S_k}{n_k}\right]^2\right) < 1 \Rightarrow \frac{1}{m}\sum_{k=1}^{m}\left(\left[\frac{S_k}{n_k}\right]^2 - \mathbb{E}\left(\left[\frac{S_k}{n_k}\right]^2\right)\right) \xrightarrow{p.s} 0$$

$$\begin{aligned}
\mathbb{E}\left(\left[\frac{S_k}{n_k}\right]^2\right) &= \frac{1}{n_k^2}\mathbb{E}\left(\mathbb{E}\left(S_k^2|\theta_k\right)\right) \\
&= \frac{1}{n_k^2}\mathbb{E}\left(n_k\theta_k(1-\theta_k) + n_k^2\theta_k^2\right) \\
&= \frac{1}{n_k}\pi + \frac{n_k-1}{n_k}\left(\pi(1-\pi)\gamma + \pi^2\right)
\end{aligned}$$

On obtient par la suite

$$\frac{1}{m}\sum_{k=1}^{m}\mathbb{E}\left(\left[\frac{S_k}{n_k}\right]^2\right) = \pi\left(\frac{1}{m}\sum_{k=1}^{m}\frac{1}{n_k}\right) + \left[\pi(1-\pi)\gamma + \pi^2\right]\left(\frac{1}{m}\sum_{k=1}^{m}\left(1-\frac{1}{n_k}\right)\right)$$

Si on remplace le terme à gauche de l'équation par sa valeur empirique, on obtient

$$\widehat{\gamma} = \frac{\frac{1}{m}\sum_{k=1}^{m}\left(\frac{S_k}{n_k}\right)^2 - \widehat{\pi}\left(\frac{1}{m}\sum_{k=1}^{m}\frac{1}{n_k}\right) - \widehat{\pi}^2\left(1-\frac{1}{m}\sum_{k=1}^{m}\frac{1}{n_k}\right)}{\widehat{\pi}(1-\widehat{\pi})\left[\frac{1}{m}\sum_{k=1}^{m}\left(1-\frac{1}{n_k}\right)\right]}$$

Par ailleurs, on a

$$\widehat{\pi}(1-\widehat{\pi})\left[\frac{1}{m}\sum_{k=1}^{m}\left(1-\frac{1}{n_k}\right)\right] \geq 0$$

Donc le signe de $\widehat{\gamma}$ dépend de son numérateur. Or si on pose

$$\begin{aligned} a &= \left(\frac{1}{m}\sum_{k=1}^{m}\frac{1}{n_k}\right) \geq 0 \\ b &= \frac{1}{m}\sum_{k=1}^{m}\left(\frac{S_k}{n_k}\right)^2 \geq 0 \end{aligned}$$

on obtient

$$\widehat{\gamma} = \frac{b-(\widehat{\pi}a+\widehat{\pi}^2(1-a))}{\widehat{\pi}(1-\widehat{\pi})\left[\frac{1}{m}\sum_{k=1}^{m}\left(1-\frac{1}{n_k}\right)\right]}$$

On a $\widehat{\pi}a+\widehat{\pi}^2(1-a) \in [\widehat{\pi}^2, \widehat{\pi}]$ car c'est une combinaison linéaire convexe. A l'aide de l'inégalité de la variance, on a aussi

$$\left(\frac{1}{m}\sum_{k=1}^{m}\frac{S_k}{n_k}\right)^2 \leq b \leq \frac{1}{m}\sum_{k=1}^{m}\frac{S_k}{n_k}$$

Puisque

$$\frac{1}{m}\sum_{k=1}^{m}\frac{S_k}{n_k} = \widehat{\pi}$$

alors $b \in [\widehat{\pi}^2, \widehat{\pi}]$.
Le signe de $\widehat{\gamma}$ dépend donc de la suite (S_k, n_k). Cette équation des moments, comme d'autres proposées dans la littérature, n'admettent pas toujours une solution dans $]0,1[\times]0,1[$; d'où le recourt à une méthode de pondération empirique.

B.2 Estimation par simulation des performances des estimateurs obtenus par la méthode de pondération empirique

B.2.1 Organisation des simulations

Avant d'étudier les propriétés statistiques des estimateurs, nous allons décrire la simulation d'un échantillon Bêta-binomial. Nous simulons un échantillon Bêta-binomial de la manière suivante :

1. On se donne n_U, l'ensemble des observations d'étude vérifiant le profil $U(X)$. Nous supposons avoir disposé de n_U observations constituées à partir de m réalisations de la variable $[Y,X]^{\mathcal{L}}$, où chaque réalisation $[Y,X]_h$ de $[Y,X]^{\mathcal{L}}$ est une suite d'observations indépendantes $(Y_i,X_i)_{i=1:n_h}$ de taille n_h.
2. On génère m réalisations $(\theta_h^U)_{h=1:m}$ d'une loi Bêta de paramètres α_U et β_U donnés. Ensuite on construit une suite $(n_{hU})_{h=1:m}$ telle que $\sum n_{hU} = n_U$.

3. Pour chaque h, on simule n_{hU} observations d'une loi de Bernoulli de probabilité de succès θ_h^U. Ainsi pour chaque couple (α_U, β_U), nous pouvons disposer des statistiques $(S_{hU})_{h=1:m}$ et $(n_{hU})_{h=1:m}$.

On appelle l'échantillon $(S_{hU}, n_{hU})_{h=1:m}$ un échantillon Bêta-Binomial puisqu'il est obtenu à partir d'une combinaison d'une loi Bêta et d'une loi Binomiale.

B.2.2 Présentation et analyse des résultats

Pour étudier des propriétés statistiques des estimations, on suppose avoir $n_U = 100000$ observations constituées à partir de $m = 50$ réalisations de $[Y, X]^{\mathcal{L}}$. On se fixe une valeur de 0.007 pour le paramètre π_U et on fait varier le paramètre γ_U avec les valeurs suivantes :0.005, 0.0075, 0.01, 0.025, 0.05, 0.075, 0.1, 0.25, 0.5, 0.75. Nous avons fait le choix de ces valeurs pour simuler des données semblables à nos données réelles. Par exemple, pour le couple $\pi_U = 0.007$ et $\gamma_U = 0.01$, un aperçu de la forme de la densité de la loi Bêta associée est représentée ci dessous.

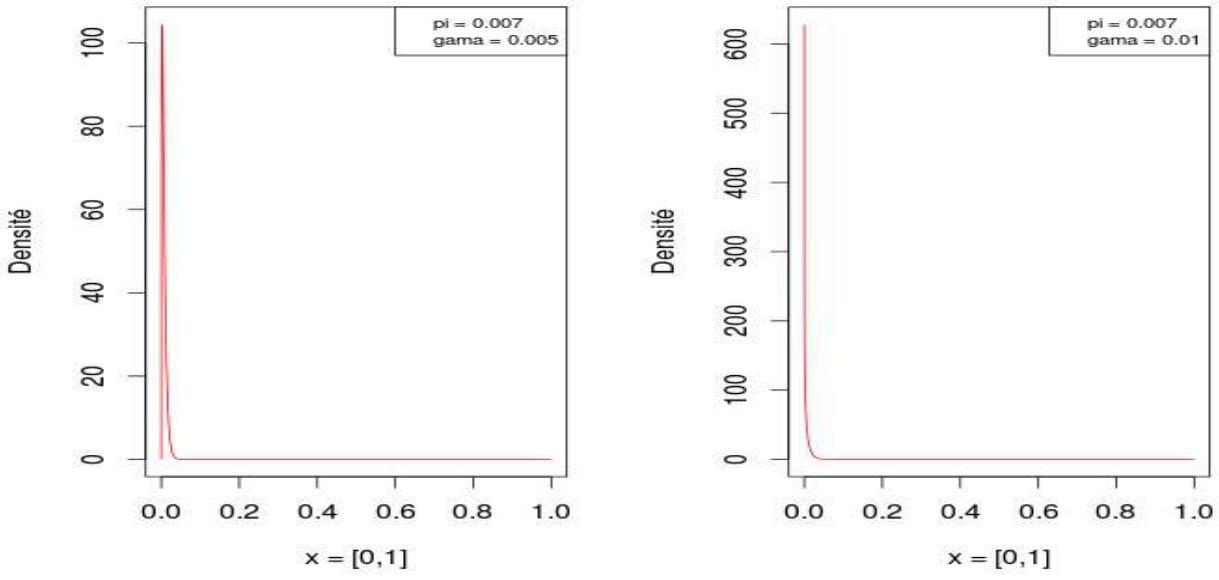

Figure A.1 – Forme de la densité de Bêta

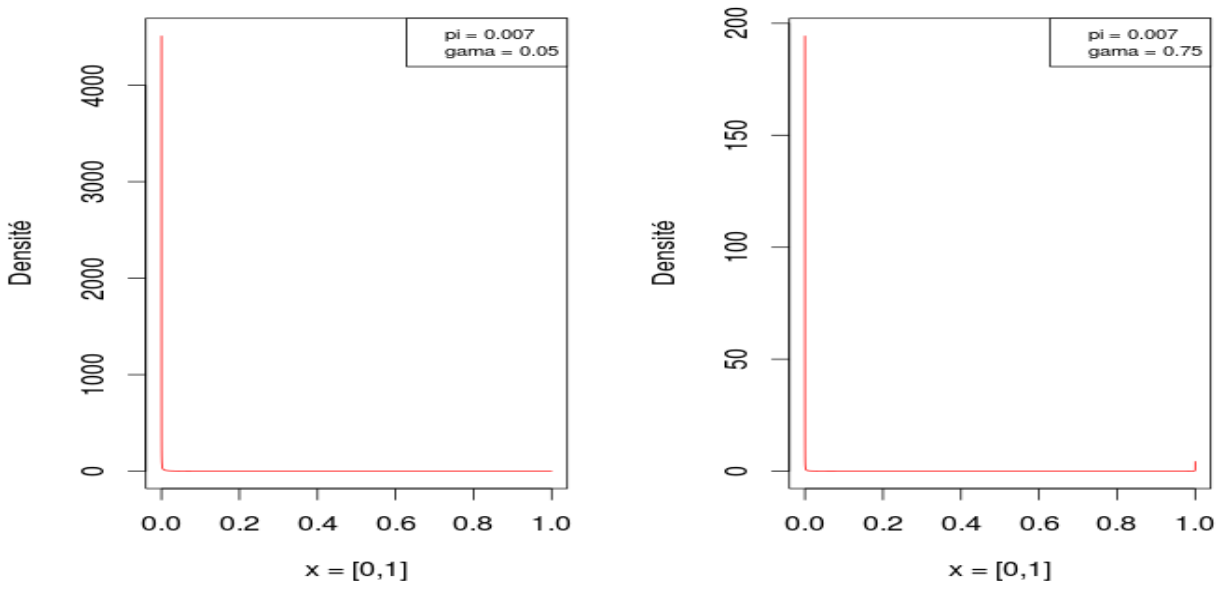

Figure A.2 – Forme de la densité de Bêta

Pour chaque combinaison (π_U, γ_{Uj}) ; $j = 1 : 10$, on en déduit un couple (α_U, β_U) à partir duquel un échantillon Bêta-Binomial $(S_{hU}, n_{hU})_{h=1:m}$ est généré. Ainsi à chaque couple (π_U, γ_U) correspond un échantillon Bêta-Binomial. En combinant les valeurs de π_U et de γ_U, nous simulons 10 échantillons Bêta-Binomial sur lesquels les paramètres π_U et γ_U seront estimés. Dans le tableau A.1, nous présentons les estimations obtenues à partir des équations des moments proposées par Kleinman que nous notons MOMK, les estimations obtenues à partir des équations des moments proposées dans cette analyse que nous notons par MOMG et les estimations obtenues par la méthode du maximum de vraisemblance notées EMV, pour des valeurs de π_U et γ_U fixées.

		MOMK		MOMG		EMV	
π_U	γ_U	$\widehat{\pi}_U$	$\widehat{\gamma}_U$	$\widehat{\pi}_U$	$\widehat{\gamma}_U$	$\widehat{\pi}_U$	$\widehat{\gamma}_U$
0.007	0.0050	0.0062	0.0029	0.0062	0.0029	0.0061	0.0045
0.007	0.0075	0.0088	0.0055	0.0088	0.0054	0.0088	0.0058
0.007	0.0100	0.0080	0.0086	0.0080	0.0090	0.0080	0.0073
0.007	0.0250	0.0079	0.0205	0.0079	0.0202	0.0079	0.0213
0.007	0.0500	0.0091	0.0522	0.0091	0.0511	0.0090	0.0636
0.007	0.0750	0.0045	0.0532	0.0045	0.0526	0.0045	0.0437
0.007	0.1000	0.0077	0.2118	0.0077	0.2072	0.0080	0.1455
0.007	0.2500	0.0070	0.3013	0.0070	0.2946	0.0064	0.3395
0.007	0.5000	0.0018	0.0707	0.0018	0.0697	0.0017	0.1860
0.007	0.7500	0.0197	0.9866	0.0197	0.9666	0.0156	0.8187

Tableau A.1 – Valeurs estimées des paramètres π et γ

A travers ce tableau, on constate que, pour les deux méthodes MOMK et MOMG, nous avons la même estimation de $\widehat{\pi}$ quelque soient les valeurs du couple (π,γ). Ceci est justifié par le fait que nous avons utilisé le même estimateur de π dans les deux méthodes. On constate aussi que la valeur estimée de $\widehat{\pi}$ par la méthode EMV est peu différente de la valeur estimée de $\widehat{\pi}$ par les deux premières méthodes. Cependant on note une différence entre les trois approches aux niveaux des estimations de $\widehat{\gamma}$. Les résultats présentés dans le tableau ci-dessus ne nous permettent pas de départager les trois méthodes. Par contre, on peut comparer les trois approches en calculant les racines carrées des erreurs quadratiques moyennes des estimateurs en procédant par simulation.

Nous considérons les valeurs d'apprentissage suivantes : $\pi_U = 0.007$ et $\gamma_U = (0.005, 0.05)$. Pour chaque couple (π_U, γ_U) fixé, nous porterons nos simulations sur les couples suivants : ($n_U = 20\,000$, $m = 10$), ($n_U = 50\,000$, $m = 50$), ($n_U = 100\,000$, $m = 100$), ($n_U = 200\,000$, $m = 150$), ($n_U = 300\,000$, $m = 200$), ($n_U = 400\,000$, $m = 250$), ($n_U = 500\,000$, $m = 300$), ($n_U = 600\,000$, $m = 350$) et ($n_U = 700\,000$, $m = 400$). Pour chaque couple (n_U, m) fixé, on simule $B = 250$ échantillons Bêta-Binomial sur lesquels on estime π et γ pour chaque échantillon. Et à la fin on calcule la racine carrée de l'erreur quadratique moyenne correspondante de chaque paramètre dans chaque méthode. Les résultats obtenus sont présentés dans les tableaux ci-dessous.

B=250

	n=20.000, m=10			n=50.000, m=50			n=100.000, m=100		
	MOMK	MOMG	EMV	MOMK	MOMG	EMV	MOMK	MOMG	EMV
RMSE ($\widehat{\pi}$)	0.00193	0.00193	0.00193	0.00105	0.00106	0.00105	0.00075	0.00075	0.00075
RMSE ($\widehat{\gamma}$)	0.00407	0.00337	0.00315	0.00238	0.00204	0.00152	0.00155	0.00147	0.00121

B=250

	n=200.000, m=150			n=300.000, m=200			n=400.000, m=250		
	MOMK	MOMG	EMV	MOMK	MOMG	EMV	MOMK	MOMG	EMV
RMSE ($\widehat{\pi}$)	0.00055	0.00055	0.00055	0.00048	0.00048	0.00049	0.00041	0.00041	0.00041
RMSE ($\widehat{\gamma}$)	0.00126	0.00114	0.00087	0.00095	0.00088	0.00067	0.00091	0.00091	0.00064

B=250

	n=500.000, m=300			n=600.000, m=350			n=700.000, m=400		
	MOMK	MOMG	EMV	MOMK	MOMG	EMV	MOMK	MOMG	EMV
RMSE ($\widehat{\pi}$)	0.00037	0.00037	0.00037	0.00040	0.00040	0.00040	0.00037	0.00037	0.00036
RMSE ($\widehat{\gamma}$)	0.00088	0.00075	0.00060	0.00072	0.00070	0.00056	0.00067	0.00065	0.00049

Tableau A.2 – Racines carrées des erreurs quadratiques moyennes des estimateurs de $\pi = 0.007$ et $\gamma = 0.005$

B=250

	n=20.000, m=10			n=50.000, m=50			n=100.000, m=100		
	MOMK	MOMG	EMV	MOMK	MOMG	EMV	MOMK	MOMG	EMV
RMSE ($\widehat{\pi}$)	0.00557	0.00560	0.00542	0.00248	0.00248	0.00244	0.00185	0.00186	0.00184
RMSE ($\widehat{\gamma}$)	0.03650	0.03622	0.04707	0.02239	0.02210	0.02024	0.01952	0.01936	0.01538

B=250

	n=200.000, m=150			n=300.000, m=200			n=400.000, m=250		
	MOMK	MOMG	EMV	MOMK	MOMG	EMV	MOMK	MOMG	EMV
RMSE ($\widehat{\pi}$)	0.00147	0.00147	0.00147	0.00129	0.00130	0.00129	0.00122	0.00122	0.00122
RMSE ($\widehat{\gamma}$)	0.01686	0.01640	0.01134	0.01700	0.01540	0.01077	0.01403	0.01385	0.00971

B=250

	n=500.000, m=300			n=600.000, m=350			n=700.000, m=400		
	MOMK	MOMG	EMV	MOMK	MOMG	EMV	MOMK	MOMG	EMV
RMSE ($\widehat{\pi}$)	0.00115	0.00115	0.00115	0.00104	0.00104	0.00104	0.00091	0.00091	0.00092
RMSE ($\widehat{\gamma}$)	0.01286	0.01255	0.00948	0.01057	0.01056	0.00833	0.01115	0.01112	0.00762

Tableau A.3 – Racines carrées des erreurs quadratiques moyennes des estimateurs de $\pi = 0.007$ et $\gamma = 0.05$

Les résultats présentés dans le tableau A.2 et le tableau A.3 montrent une convergence des erreurs quadratiques moyennes de $\widehat{\pi}_U$ et $\widehat{\gamma}_U$ vers zéro pour toutes les trois méthodes. On peut constater aussi que la méthode d'estimation par le maximum de vraisemblance (EMV) est meilleur que les deux autres méthodes puisqu'elle enregistre la plus petite erreur quadratique moyenne sur les neufs échantillons simulés. Elle est suivie par la méthode MOMG qui a la deuxième plus petite erreur quadratique moyenne. En pratique, on suggère donc d'estimer les hyperparamètres par la méthode du maximum de vraisemblance.

B.3 Loi conditionnelle de θ_h^U

D'après le théorème de Bayes, on peut déterminer la distribution conditionnelle $\left[\left(\theta_h^U\right)_{h=1:m}|Y,\pi_U,\tau_U\right]$ par :

$$\begin{aligned}\left[\left(\theta_h^U\right)_{h=1:H}|Y,\pi_U,\tau_U\right] &= \frac{\left[\left(\theta_h^U\right)_{h=1:H},Y|\pi_U,\tau_U\right]}{[Y|\pi_U,\tau_U]}\\ \left[\left(\theta_h^U\right)_{h=1:m}|Y,\pi_U,\tau_U\right] &= \frac{\left[Y|\left(\theta_h^U\right)_{h=1:m},\pi_U,\tau_U\right]\left[\left(\theta_h^U\right)_{h=1:m}|\pi_U,\tau_U\right]}{[Y|\pi_U,\tau_U]}\end{aligned}$$

On en déduit que

$$\left[\left(\theta_h^U\right)_{h=1:m}|Y,\pi_U,\tau_U\right] = \frac{\left[Y|\left(\theta_h^U\right)_{h=1:m},\pi_U,\tau_U\right]\left[\left(\theta_h^U\right)_{h=1:m}|\pi_U,\tau_U\right]}{\int_{[0,1]^m}\left[Y,|\left(\theta_h^U\right)_{h=1:m},\pi_U,\tau_U\right]\left[\left(\theta_h^U\right)_{h=1:m}|\pi_U,\tau_U\right]\otimes d\theta_h^U}$$

Par ailleurs, on a

$$\left[Y|\left(\theta_h^U\right)_{h=1:m},\pi_U,\tau_U\right] = \left[Y|U(X),[Y,X]^{\mathcal{L}}\right]$$

En plus nous avons $\left[\left(\theta_h^U\right)_{h=1:m}|Y,\pi_U,\tau_U\right] = \prod_{h=1}^{m}\left[\theta_h^U|Y,\pi_U,\tau_U\right]$ puisque la suite $\left(\theta_h^U\right)_{h=1:m}$ est un échantillon iid. Pour simplifier les expressions, nous posons

$$\begin{aligned}\mathbf{a} &= \delta_{(1,[Y,X]_h)}\left(U(X),[Y,X]^{\mathcal{L}}\right)\mathbb{1}_{[Y=1]}(y)\\ \mathbf{b} &= \delta_{(1,[Y,X]_h)}\left(U(X),[Y,X]^{\mathcal{L}}\right)(1-\mathbb{1}_{[Y=1]}(y))\end{aligned}$$

et

$$\Gamma(A) = \frac{\Gamma(\widehat{\tau}_x)}{\Gamma(\widehat{\pi}_x\widehat{\tau}_x)\Gamma((1-\widehat{\pi}_x)\widehat{\tau}_x)}$$

Par la suite, on obtient

$$\begin{aligned}\left[Y|\left(\theta_h^U\right)_{h=1:m},\pi_U,\tau_U\right]\left[\left(\theta_h^U\right)_{h=1:m}|\pi_U,\tau_U\right] &= \left\{\prod_{h=1}^{m}\left[Y|\theta_h^U,[Y,X]_h\right]\right\}\left\{\prod_{h=1}^{m}\left[\theta_h^U|\pi_U,\gamma_U\right]\right\}\\ &= \left\{\prod_{h=1}^{m}\left(\theta_h^U\right)^{\mathbf{a}}\left(1-\theta_h^U\right)^{\mathbf{b}}\right\}\left\{\prod_{h=1}^{m}\Gamma(A)\left(\theta_h^U\right)^{\pi_U\tau_U-1}\left(1-\theta_h^U\right)^{(1-\pi_U)\tau_U-1}\right\}\\ &= [\Gamma(A)]^m\prod_{h=1}^{m}\left\{\left(\theta_h^U\right)^{\mathbf{a}+\pi_U\tau_U-1}\left(1-\theta_h^U\right)^{\mathbf{b}+(1-\pi_U)\tau_U-1}\right\}\end{aligned}$$

Par ailleurs, on a

$$\begin{aligned}&\int_{[0,1]^m}\left[Y,|\left(\theta_h^U\right)_{h=1:m},\pi_U,\tau_U\right]\left[\left(\theta_h^U\right)_{h=1:m}|\pi_U,\tau_U\right]\otimes d\theta_h^U\\ &\qquad= [\Gamma(A)]^m\prod_{h=1}^{m}\left\{\int_0^1\left(\theta_h^U\right)^{\mathbf{a}+\pi_U\tau_U-1}\left(1-\theta_h^U\right)^{\mathbf{b}+(1-\pi_U)\tau_U-1}\right\}\\ &\qquad= [\Gamma(A)]^m\prod_{h=1}^{m}\frac{\Gamma\left[\mathbf{a}+\pi_U\tau_U\right]\Gamma\left[\mathbf{b}+(1-\pi_U)\tau_U\right]}{\Gamma\left[\mathbf{a}+\mathbf{b}+\tau_U\right]}\end{aligned}$$

Puisque les variables $\left(\theta_h^U\right)_{h=1:m}$ sont indépendantes et identiquement distribuées, on obtient

$$\prod_{h=1}^{m}\left[\theta_h^U|Y,\pi_U,\tau_U\right]=\prod_{h=1}^{m}\frac{\Gamma\left[\mathbf{a}+\mathbf{b}+\tau_U\right]\left(\theta_h^U\right)^{\mathbf{a}+\pi_U\tau_U-1}\left(1-\theta_h^U\right)^{\mathbf{b}+(1-\pi_U)\tau_U-1}}{\Gamma\left[\mathbf{a}+\pi_U\tau_U\right]\Gamma\left[\mathbf{b}+(1-\pi_U)\tau_U\right]}$$

donc

$$\prod_{h=1}^{H}\left[\theta_h^U|Y,\pi_U,\tau_U\right] \;=\; \prod_{h=1}^{H}\text{Beta}\left(\mathbf{a},\mathbf{b}\right)$$

On en déduit que la loi conditionnelle de θ_h^U est une loi Bêta définie par :

$$\left[\theta_h^U|Y,\pi_U,\tau_U\right]=\text{Beta}\left(\mathbf{a}+\pi_U\tau_U,\mathbf{b}+(1-\pi_U)\tau_U\right)$$