

**Le modèle bayésienne de Kaplan Meier sous la présence d'hétérogénéité
Cas : étude d'insertion des chômeurs inscrits à l'Agence Locale de l'Emploi d'Ain El Benian
(janvier 2011-juillet 2013)**

Hamimes Ahmed ^{1,*}, Benamirouche Rachid ²

**Kaplan Meier's Bayesian model under the presence of heterogeneity
Case: integration study of the unemployed registered with the Local Employment Agency of
(Ain El Benian (January 2011-July 2013**

Hamimes Ahmed ^{1,*}, Benamirouche Rachid ²

¹ École nationale de la statistique et d'économie appliquée. (Algérie).

²Professeur, École nationale de la statistique et d'économie appliquée. (Algérie).

Date de réception : 10/11/2020 ; Date de révision : 06/12/2020 ; Date d'acceptation : 31/12/2020

Résumé :

Une approche bayésienne de la survie basée sur le modèle à mélange fini des lois de bêta offre des solutions pratiques, simples et relativement faciles à exploiter numériquement des durées globales pour un ensemble d'individus de nature différente. On s'appuie sur la structure des données manquantes pour réaliser une approximation de la distribution a posteriori basée sur l'échantillonnage de Gibbs et évite le problème de label switching. Dans notre étude, on essaie essentiellement de focaliser l'attention sur l'application des modèles bayésiennes de Kaplan Meier dans l'estimation des durées de chômage des inscrits à l'Agence locale de l'Emploi d'Ain El Benian, dans le but de déterminer le rôle de l'hétérogénéité dans l'amélioration de la qualité d'estimation.

Mots-clés : l'approche bayésienne, le modèle à mélange fini, des durées de chômage, l'Agence locale de l'Emploi d'Ain El Benian.

Codes de classification Jel : C11, J6, J64.

Abstract :

A Bayesian approach to survival based on the finite mixture model of the laws of beta offers practical solutions that are simple and relatively easy to numerically exploit global durations for a set of individuals of different natures. We rely on the missing data structure to approximate the a posteriori distribution based on Gibbs sampling and avoid the label switching problem. In our study, we essentially try to focus attention on the application of Kaplan Meier's Bayesian models in estimating the durations of unemployment of those registered with the Local Employment Agency of Ain El Benian, with the aim of to determine the role of heterogeneity in improving the quality of estimation.

Keywords: Bayesian approach, finite mixture model, unemployment durations, Ain El Benian Local Employment Agency.

Jel classification codes: C11, J6, J64.

-I Introduction :

Un modèle bayésien est constitué d'un modèle statistique paramétré $(\mathcal{X}, \mathcal{F}, P_\theta, \theta \in \Theta)$ avec $f(x/\theta)$ densité de P_θ et d'une loi $\pi(\theta)$ sur le paramètre, où toute l'inférence est basée sur la distribution a posteriori. Face à la complexité des phénomènes observés. Il est possible d'utiliser des hypothèses minimales, dans ce contexte en ayant recours en général aux méthodes non paramétriques où $f(x/\theta)$ est un mélange infini de distributions par contre l'approche paramétrique qui suppose que l'ensemble des paramètres Θ est un sous-ensemble de $\mathbb{R}$. Dans l'approche bayésienne non paramétrique la modélisation repose sur les travaux de Ferguson (1973) et utilise les processus de Dirichlet, dans cette représentation le modèle bayésien est basé sur $g(P_\theta)$ un prior sur la loi des observations et non plus θ ou P_θ . Cette dernière prior est considérée comme une mesure aléatoire. En revanche, l'estimateur de Kaplan-Meier (1958) dans l'approche fréquentiste est une méthode fonctionnelle pour estimer la fonction de survie et prend en compte la présence d'une censure à droite. De plus cet estimateur est convergent, cohérent et asymptotiquement gaussien et il est biaisé positivement. Plusieurs travaux ont été fondés sur l'amélioration de cet estimateur. Khizanov et Maïboroda (2015), ont proposé une modification basée sur un modèle de mélange avec des concentrations variées, Rossa et Zieliński (2006), introduisent un estimateur de Kaplan-Meier basé sur une approximation par la loi de Weibull, Shafiq Mohammad et al (2007), ont présenté une pondération de l'estimateur de Kaplan Meier sous la fonction sinus.

L'utilisation de l'estimateur de Kaplan Meier nécessite le traitement du problème d'hétérogénéité qui constitue une limite dans cette méthode dans l'analyse des durées globales pour un ensemble d'individus de nature différente. Un autre problème dans l'estimateur de Kaplan Meier, parfois même si la différence entre deux points du temps t_1 et t_2 est considérablement large on trouve que $KM(t_1) = KM(t_2)$. Bien qu'il s'agisse d'un problème général d'estimation d'une fonction de distribution lisse et monotone à partir d'échantillons petits ou modérés, dans le contexte de l'estimation des probabilités de survie, l'inconvénient est particulièrement gênant. Dans cet article, nous discutons d'un lissage général de l'estimateur de Kaplan-Meier basé sur l'approche bayésienne. Ces deux problèmes nécessitent un lien avec l'approche paramétrique qui prend en considération la facilité de mise en œuvre. En revanche, une modélisation non paramétrique n'est pas pragmatique par rapport à la modélisation paramétrique parce qu'elle vise à estimer un nombre infini de paramètres par un nombre fini d'observations mais dans la majorité des modèles de durées notre échantillon n'est pas suffisamment grand à cause de la complexité de ces phénomènes observés et la contrainte des ressources (voir Field et Ronchetti, 1990).

On compare dans cet article les différentes structures a priori pour l'estimateur de Kaplan-Meier à travers le critère de déviance d'information dans un exemple réel décrit les durées de sortie du chômage de 1064 individus dans une agence d'emploi à Alger. Cette étude montre clairement l'efficacité d'un mélange bêta basée sur une modification du reparamétrage de Diebolt et Robert (1994).

-I La conception bayésienne de l'estimateur de Kaplan-Meier :

1. L'estimateur de Kaplan-Meier

Soit $(\mathcal{X}, \mathcal{F}, P)$ un espace probabilisé. $T_1, \dots, T_m$ une suite de variables aléatoires $i.i.d$ positives et de fonction de répartition commune F , expriment le temps entre le début de l'étude et l'arrivée d'un événement pour le i ème individu. L'échantillon réellement observé sera donc composé de m couples (Y_i, δ_i) , où δ_i est l'indicateur de censure, qui détermine si T_i a été censuré ou non. Kaplan et Meier (1958) ont introduit un estimateur non-paramétrique écrit de deux manières différentes selon qu'on est en absence d'ex aequo ou en présence d'ex aequo comme suit :

- Cas d'absence d'ex aequo :

L'estimateur de Kaplan-Meier est donné par :

$$\hat{S}(t) = \begin{cases} \prod_{T_{(i)} \leq t} \left(1 - \frac{d_i}{n_i}\right)^{\delta_i} & \text{si } t \geq t_1 \\ 1 & \text{si } t < t_1 \end{cases}$$

avec

$$\delta_i = \begin{cases} 1 & \text{événement réalisé,} \\ 0 & \text{sujet censuré.} \end{cases}$$

- Cas de présence d'ex aequo :

Dans le cas d'application, on est confronté à la présence des événements de natures différentes, on considère que les observations non censurées ont lieu avant les censurées, on a :

$$\hat{S}(t) = \begin{cases} \prod_{T_{(i)} \leq t} \left(1 - \frac{d_i}{n_i}\right) = \prod_{t_i \leq t} (1 - q_i) = \prod_{t_i \leq t} p_i & \text{si } t \geq t_1 \\ 1 & \text{si } t < t_1 \end{cases}$$

L'estimateur de Kaplan-Meier est convergent formé une fonction constante par morceaux, continue à droite et limite à gauche, avec un saut à chaque temps de décès observé.

2. Sur la conception bayésienne du modèle de Kaplan-Meier

On suppose que le nombre de décès dans l'intervalle du temps est une réalisation d'une loi Binomiale s'écrit par

$$d_i \sim \text{Bin}(n_i, q_i),$$

lorsque les sorties dans les intervalles $[t_i, t_{i+1}]$ étant indépendantes les unes des autres, on écrit

$$f(d/q_i) = \prod_{i=1}^m C_{n_i}^{d_i} q_i^{d_i} (1 - q_i)^{n_i - d_i},$$

à travers la fonction de logarithme on trouve

$$\ln L = \sum_{i=1}^m [C_{n_i}^{d_i} + d_i \ln(q_i) + (n_i - d_i) \ln(1 - q_i)],$$

selon la méthode de maximum de vraisemblance, le risque de décès estimé par

$$\hat{q}_i = \frac{d_i}{n_i},$$

cette équation est l'idée de la méthode de Kaplan-Meier, où la probabilité de survivre jusqu'à l'instant t_i est la probabilité de survie t_i sachant qu'on était en vie en t_{i-1} , c'est-à-dire,

$$P(X > t) = P(X > t_{i-1}, X \geq t_i)$$

$$\begin{aligned}
 &= P(X > t_i / X \geq t_{i-1}) P(X > t_{i-1}) \\
 &= P(X > t_i / X \geq t_{i-1}) P(X > t_{i-1} / X \geq t_{i-2}) P(X > t_{i-2}) \dots
 \end{aligned}$$

on peut écrire

$$\begin{aligned}
 S(t_i) &= P(X > t_i / X \geq t_i) * S(t_{i-1}) \\
 &= S(t_{i-1})(1 - q_i) \\
 &= S(t_{i-1}) \frac{n_i - d_i}{n_i},
 \end{aligned}$$

et

$$\frac{S(t_i)}{S(t_{i-1})} = S_{t_i/t_{i-1}} = \frac{n_i - d_i}{n_i},$$

selon cette dernière équation, l'estimateur de Kaplan-Meier est la multiplication des probabilité de survie conditionnelle, l'emploi du terme conditionnel reflète directement l'applications du théorème de Bayes qui apparaît comme un principe « d'actualisation » à partir la fonction de vraisemblance

$$f(d_1, \dots, d_n / q_i)$$

d'un échantillon de n individus, et lorsque on écrit cette fonction dans le bon ordre, on trouve :

$$f(q_i / d_1, \dots, d_n) \propto \prod_{i=1}^m f(d_i / q_i) \pi(q_i)$$

cette distribution a posteriori reflète l'incertitude inhérente aux données et à l'expérimentateur plutôt qu'un aléa physique sur ce paramètre (Robert, 2013).

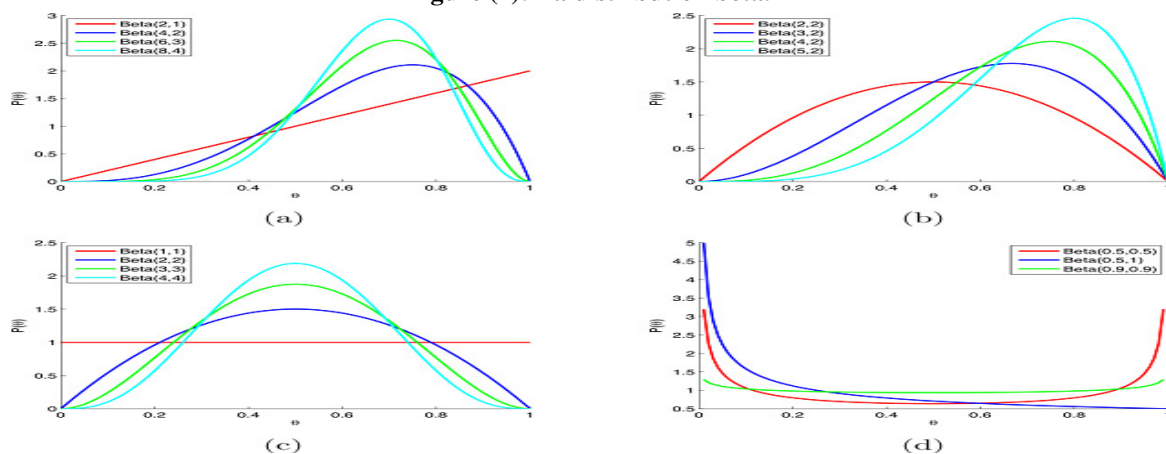
Dans une optique bayésienne on suppose une a priori pour q_i , comme la conjuguée naturelle d'une loi de Binomial est une loi bêta, on pose

$$q_i \sim \text{Be}(\alpha, \beta),$$

d'après Robert (2006), les lois a priori conjuguées peuvent être considérées comme un point de départ pour l'élaboration de distributions a priori fondées sur une information a priori limitée. Aussi la flexibilité de la forme d'une loi de bêta, la facilité de construire la loi a posteriori et le support de la distribution permettent d'une analyse des différents phénomènes et avec une précision élevée.

La figure suivante montre les différentes formes qui peuvent prendre une loi bêta.

Figure (1): La distribution bêta.



La source : réalisée par les chercheurs.

i. L'estimation des hyper-paramètres

Lorsque la loi a priori est le moteur de l'inférence bayésienne, plusieurs méthodes ont été utilisées pour estimer les paramètres de la loi bêta.

Dans le cas non informatif la méthode la plus utilisée est celui de Jeffreys, formée une loi d'arc sinus comme suit

$$\pi(q_i) = \frac{1}{\beta(1/2, 1/2)} q_i^{-1/2} (1 - q_i)^{-1/2}; 0 < q_i < 1,$$

ou, on écrit

$$q_i \sim \mathcal{Be}\left(\frac{1}{2}, \frac{1}{2}\right) \quad (1)$$

cette estimation n'a pas de relation directe avec la loi a priori conjuguée sauf que le résultat est une bêta non informative, cette loi est dépendante de l'intégralité de la loi des nombres de sorties ce qui violer le principe de vraisemblance.

Dans le cas d'une mesure uniforme par rapport à la mesure de Lebesgue, la loi a priori est n'est pas invariante par reparamétrisation, on pose généralement

$$q_i \sim \mathcal{Be}(1, 1),$$

ce choix, il apparait comme biaisée contre les valeurs extrêmes 0 et 1, Novick et Hall, 1995 proposent la loi de Haldane (1931) donnée par

$$\pi^*(q_i) \propto [q_i(1 - q_i)]^{-1}; 0 < q_i < 1,$$

la loi marginale de d_i est donnée par

$$\begin{aligned} f(d_i) &= \int_0^1 [q_i(1 - q_i)]^{-1} C_{n_i}^{d_i} q_i^{d_i} (1 - q_i)^{n_i - d_i} dq_i \\ &= B(d_i, n_i - d_i), \end{aligned}$$

pour $d_i = 0, n_i$ cette dernière fonction est n'est pas définie, et lorsque la constant de normalisation elle n'a pas d'impacte sur la loi a posteriori on utilise une loi a priori π^* d'une limite de lois de bêta dénormalisées lorsque α, β tend vers 0, comme suit

$$\pi(q_i) = q_i^{\alpha-1} (1 - q_i)^{\beta-1}; 0 < q_i < 1,$$

dans la loi a posteriori $\pi(q_i/d_i)$, on trouve

$$\mathcal{Be}(\alpha + d_i, \beta + n_i - d_i)$$

ce qui représente une bonne résolution du problème des valeurs limites. L'inconvénient de cette procédure est que dans les lois impropres comme celui de notre cas et en présence d'un espace paramétrique multidimensionnel, cela peut introduire le paradoxe de marginalisation.

Une autre manière de résoudre le problème de l'incertitude dans la mesure a priori de bêta est le recoure à l'échantillon $(d_1, \dots, d_m)$ et l'utilisation des méthodes statistiques classiques à travers la loi marginale $f_\pi(d_i/\alpha, \beta)$, cette méthode est appelée approche bayésienne empirique. Morris (1983) introduit le concept paramétrique de cette approche. Pour une loi binomiale et une loi apriori conjuguée on pose

$$\begin{aligned} f_\pi(d_i/\alpha, \beta) &= \int_0^1 f(d_i/q_i) \pi(q_i/\alpha, \beta) dq_i \\ &= \int_0^1 [q_i(1 - q_i)]^{-1} C_{n_i}^{d_i} q_i^{d_i} (1 - q_i)^{n_i - d_i} dq_i \end{aligned}$$

$$\begin{aligned}
 &= c_{n_i}^{d_i} \frac{1}{B(\alpha, \beta)} \int_0^1 q_i^{d_i+\alpha-1} (1-q_i)^{n_i-d_i+\beta-1} dq_i \\
 &= c_{n_i}^{d_i} \frac{B(\alpha + d_i, n_i + \beta - d_i)}{B(\alpha, \beta)}
 \end{aligned}$$

Ce qui fournit une loi bêta-binomial pour estimer $\hat{\alpha}, \hat{\beta}$, afin de calculer $\pi(q_i/d_i, \hat{\alpha}, \hat{\beta})$. Les paramètres de bêta binomial peuvent être estimés par plusieurs méthodes paramétriques comme la méthode de maximum de vraisemblance qui nécessite des méthodes itératives pour résoudre la valeur des paramètres de $f_{\pi}(d_i/\alpha, \beta)$ (voir Barsotti et Paroli, 1991), aussi des méthodes basées sur l'estimateur des moments, Tripathi et al (1993) ont proposé plusieurs estimateurs dans ce cadre, l'estimateur basé sur les deux premiers moments s'écrit par

$$\hat{\alpha} = \frac{\bar{\xi}_0(m-1-\bar{\xi}_1)}{\bar{\xi}_0 + m(\bar{\xi}_1 - \bar{\xi}_0)}, \quad \hat{\beta} = \frac{(m-\bar{\xi}_0)(m-1-\bar{\xi}_1)}{\bar{\xi}_0 + m(\bar{\xi}_1 - \bar{\xi}_0)}$$

où

$$\bar{\xi}_j = \frac{\hat{\mu}_{(j+1)}}{\hat{\mu}_{(j)}}, \quad j = 1, 2$$

et

$\hat{\mu}_{(j)}$ est le moment généralisé d'ordre j , calculé par

$$\mu_{(j)} = \frac{(-m)_j (\alpha)_j}{(\alpha + \beta)_j} (-1)^j, j = 1, 2, \dots$$

Dans l'autre côté les méthodes empiriques non paramétriques sont fondées sur les travaux de Robbins (1952, 1955, 1964, 1983). Nous supposons que les q_i sont tous tirés selon la même priori π (de nature inconnue). La loi marginale $f_{\pi}(d_i/\alpha, \beta)$ utilisée pour estimer les paramètres inconnus de $\hat{\pi}_m(q_{m+1})$ où le problème porte sur cette dernière. On écrit donc

$$\hat{\pi}(q_{m+1}/d_{m+1}, \alpha, \beta) \propto f(d_{m+1}/q_{m+1}) \hat{\pi}_m(q_{m+1})$$

l'estimateur bayésien non paramétrique empirique de la survie s'écrit de la manière suivante

$$\begin{aligned}
 \hat{S}_{\text{empirique}} &= \frac{\sum_{i=1}^m n_i \hat{q}_i^{n_i-d_{m+1}} (1-\hat{q}_i)^{d_m}}{\sum_{i=1}^m n_i \hat{q}_i^{n_i-d_m} (1-\hat{q}_i)^{d_m}} \\
 &= \frac{\sum_{i=1}^m n_i \left(1 - \frac{d_i}{n_i}\right)^{n_i-d_{m+1}} \left(\frac{d_i}{n_i}\right)^{d_m}}{\sum_{i=1}^m n_i \left(1 - \frac{d_i}{n_i}\right)^{n_i-d_m} \left(\frac{d_i}{n_i}\right)^{d_m}}
 \end{aligned}$$

où $\hat{q}_i$ est l'estimateur classique de Kaplan-Meier. La variance de cet estimateur s'écrit par

$$\sigma_{\text{empirique}}^2 = \frac{\sum_{i=1}^m n_i \left(1 - \frac{d_i}{n_i}\right)^{n_i-d_{m+1}} \left(\frac{d_i}{n_i}\right)^{d_m}}{\sum_{i=1}^m n_i \left(1 - \frac{d_i}{n_i}\right)^{n_i-d_m} \left(\frac{d_i}{n_i}\right)^{d_m}} - \hat{S}_{\text{empirique}}^2$$

mais dans le cas où $d_i = 0$, l'estimateur $\hat{S}_{\text{empirique}}$ est inexact, pour cela on utilise une loi a priori uniforme pour remplacer l'estimation de Kaplan-Meier dans $\hat{S}_{\text{empirique}}$ par

$$\hat{q}_i = \frac{\sum_{j=0}^{d_i} (-1)^j \frac{d_i!}{(d_i-j)!j! (n_i+j+2)}}{\sum_{j=0}^{d_i} (-1)^j \frac{d_i!}{(d_i-j)!j! (n_i+j+1)}}$$

et à partir de cette équation et pour $d_i = 0$,

$$1 - \frac{d_i}{n_i} = \hat{q}_i = 1 - \frac{1}{n_i - 1}$$

Ces méthodes empiriques souffrent de plusieurs inconvénients :

- Ces techniques ne sont pas bayésiennes parce qu'elles utilisent les données deux fois pour trouver un estimateur sous la loi pseudo a postériori.
- La sous-optimalité des estimateurs empiriques pose des problèmes sur les procédures utilisées.
- l'approximation d'une loi a priori inconnue par une loi marginale n'est pas justifiable sauf dans le cas d'un échantillon de taille important, le problème restera le même dans un grand échantillon car l'inférence basée sur ce paradigme ne correspond qu'à une seule observation.
- Dans l'estimateur empirique non paramétrique et dans le cas où q est un mélange de lois, l'estimation fréquentiste d'une loi marginale est souvent basique¹.

ii. Le prior d'un mélange de distributions

Dans l'estimation d'une densité, l'analyse bayésienne utilise souvent des modèles de mélange de lois de probabilités élémentaires. Les premiers articles dans l'estimation bayésienne des mélanges par la méthode MCMC (Monte Carlo par Chaîne de Markov) ont été les travaux de Diebolt et Robert (1990,1994), Gelman et King (1990), Verdinelli et Wasserman (1991) et Lavine et Wasserman (1992). Cette approche de mélange permet de construire une modélisation flexible contient un ensemble de sous-populations. Les modèles de mélanges se situent à la frontière des modélisations paramétrique et non paramétrique et permettent la description de phénomènes plus complexes (Robert (2006)). L'estimation bayésienne de mélange est possible depuis peu à travers le développement des outils informatiques et les modèles de simulations statistiques avancées comme celui d'MCMC. Pour $i = 1, \dots, n$, on pose un modèle de mélange à k-composantes.

La loi a priori de mélange s'écrit de la manière suivante

$$L(q_i/w, \alpha, \beta) = \sum_{j=1}^k w_j f(q_i/d_i, \alpha_j, \beta_j)$$

et $0 \leq w_j \leq 1, \sum_{j=1}^k w_j = 1$

on considère que les probabilités $q = (q_1, \dots, q_m)$ sont i.i.d dans la vraisemblance est

$$L(q, w/d, \alpha, \beta) = \prod_{i=1}^m \sum_{j=1}^k w_j f(q_i/d_i, \alpha_j, \beta_j) \\ \propto \prod_{i=1}^m \sum_{j=1}^k w_j q_i^{d_i} (1 - q_i)^{m_i - d_i}$$

si on pose le vecteur des paramètres indépendantes $\theta = (\alpha_j, \beta_j)$, dans l'analyse bayésienne où l'inférence repose sur la loi a postériori donnée par

¹ À cause du support fini de la loi marginale.

$$\begin{aligned} \pi(\theta, w/q) &\propto L(q, w/d, \alpha, \beta) \pi(\alpha_j/\xi_j, v_j) \pi(\beta_j/\xi_j^*, v_j^*) \pi(w/\varphi_1, \dots, \varphi_k) \\ &\propto \prod_{i=1}^m \sum_{j=1}^k w_j q_i^{d_i} (1 - q_i)^{n_i - d_i} \pi(\alpha_j/\xi_j, v_j) \pi(\beta_j/\xi_j^*, v_j^*) \pi(w/\varphi_1, \dots, \varphi_k) \end{aligned}$$

la fonction de survie se décompose alors en k^l termes, ce qui rend le calcul difficile à la main ; ces cas nécessitent des méthodes de simulation stochastiques comme la méthode de Monte Carlo par Chaîne de Markov. La complexité de ce modèle est telle n'existe pas de solution autre que d'utiliser l'échantillonnage de Gibbs (Robert, 1996).

On introduit un vecteur d'allocation $z_j (1 \leq j \leq k)$, o'u cette variable annule la structure de mélange, la vraisemblance s'écrit dans la manière suivante :

$$L(q, w/d, \alpha, \beta, z) = \prod_{i=1}^m w_{z_i} f(q_i/d_i, \alpha_{z_i}, \beta_{z_i}),$$

la loi a priori conditionnelle du vecteur d'allocation est

$$z_i/w \sim \mathcal{M}_k(\pi_1, \dots, \pi_k),$$

la loi marginale pour $z_j (1 \leq j \leq k)$ est une famille exponentielle, elle permet une loi a priori conjuguée de Dirichlet $w \sim \mathcal{D}(\varphi_1, \dots, \varphi_k)$, avec une densité

$$\frac{\Gamma(\varphi_1, \dots, \varphi_k)}{\Gamma(\varphi_1) \dots \Gamma(\varphi_k)} w_1^{\varphi_1}, \dots, w_k^{\varphi_k}$$

dans le simplexe de $\mathbb{R}^k$

$$\mathfrak{S} = \left\{ (w_1, \dots, w_k) \in [0,1]^k; \sum_{j=1}^k w_j = 1 \right\}$$

la loi a posteriori des paramètres θ, w est donnée par :

$$\begin{aligned} \pi(\theta, w/q) &\propto L(q, w/d, \alpha_{z_i}, \beta_{z_i}) \pi(\alpha_{z_i}/\xi_{z_i}, v_{z_i}) \pi(\beta_{z_i}/\xi_{z_i}^*, v_{z_i}^*) \pi(w_{z_i}/\varphi_{z_i}, \dots, \varphi_{z_i}) \\ &\propto \prod_{i=1}^m w_{z_i} f(q_i/\alpha_{z_i}, \beta_{z_i}) \pi(\alpha_{z_i}/\xi_{z_i}, v_{z_i}) \pi(\beta_{z_i}/\xi_{z_i}^*, v_{z_i}^*) \pi(w_{z_i}/\varphi_{z_i}, \dots, \varphi_{z_i}) \\ &\propto \prod_{i=1}^m w_i^{n_j} \prod_{i: Z_i = i} q_i^{d_i} (1 - q_i)^{n_i - d_i} \pi(\alpha_j/\xi_j, v_j) \pi(\beta_j/\xi_j^*, v_j^*) \pi(w_{z_i}/\varphi_1, \dots, \varphi_k). \end{aligned}$$

on pose $m_j = \# \{Z_i = j\}$, $\sum_j m_j = m$.

Le problème célèbre des modèles de mélanges est la non-identifiabilité qui causer la présence de $k!$ modes dans la distribution a posteriori, ceci est à l'origine de label-swetching (le changement d'indice) dans l'analyse bayésienne, car la multimodalité de la vraisemblance et dans l'utilisation des lois a priori symétriques elle influe sur la distribution a posteriori à partir de la relation entre eux.

Définition 1.

Soit pour toutes $\in [1, m]$:

$$\mathcal{C} = \left\{ f(d_i/q_i) = \sum_{j=1}^k w_j f_j(d/q_{ij}), w_j > 0 \text{ pour } j = 1 \text{ à } k, \sum_{j=1}^k w_j = 1 \right\}$$

la famille $\mathcal{C}$ est identifiable si, pour tous $f(d_i/q_{ij})$ et $f^*(d_i/q_{ij})$ dans $\mathcal{C}$ tels que :

$$f^*(d/q) = \sum_{j=1}^h w_j f_j(d/q_{ij}) \text{ et } f(d/q) = \sum_{v=1}^c \lambda_v f_v(d/q_{ij})$$

on a

$$f^*(d/q_{ij}) = f(d/q_{ij}) \Leftrightarrow \begin{cases} h = c \\ \text{et} \\ \forall j = 1 \text{ à } m, \exists v : w_j = \lambda_v, f^*(d/q_{ij}) = f(d/q_{ij}) \end{cases}$$

Pour plus de détail sur le problème d'identification voir McLachlan et Peel (2000).

Diebolt et Robert (1994) ont supposées une méthode de reparamétrage pour traiter le problème de changement d'indice dans un modèle gaussien à deux composantes de la façon suivante :

$$x_i \sim \mathcal{N}(\mu_j, \tau^2), \theta \sim U_{[0,1000]}$$

et

$$\begin{aligned} \mu_1 &\sim U_{[-1000,1000]} \\ \mu_2 &= \mu_1 + \theta \end{aligned}$$

Dans notre modèle les hyperparamètres de Béta généralement construisent dans des petits supports, pour cela nous proposons une modification sous la forme suivante :

$$q_{ij} \sim \mathcal{BE}(\alpha_j, \beta_j),$$

et

$$\begin{aligned} \alpha_1 &\sim \text{gamma}(0.1, c) \\ \alpha_2 &= \alpha_1 + \theta_1 \\ \theta_1 &\sim U_{[0,1]} \end{aligned}$$

la même chose pour β_j , on pose :

$$\begin{aligned} \beta_2 &\sim \text{gamma}(0.1, c) \\ \beta_2 &= \beta_1 + \theta_2 \end{aligned}$$

et

$$\theta_2 \sim U_{[0,1000]}$$

la valeur de c détermine la longueur de support, et les censures suivant la distribution vague suivante :

$$c \sim \mathcal{BE}(0.01, 0.01)$$

L'échantillonneur de Gibbs est l'approche la plus utilisée dans l'estimation de mélange bayésien (Diebolt et Robert 1990a, 1994, Lavine et West 1992, Verdinelli et Wasserman 1992, Chib 1995, Escobar et West 1995). Pour simplifier le calcul nous appliquons un cas particulier de l'algorithme de Gibbs en deux étapes qui permet de créer deux Chaînes dépendantes. Cet algorithme est connue aussi sous le nom de l'algorithme d'augmentation des données (Tanner et Wong, 1987). Elle représente le pendant bayésienne de l'algorithme EM en statistique classique. L'algorithme se décompose en les points suivants

1. Initialiser $\theta^0 = \alpha_i^{(0)}, \beta_i^{(0)}$ soit le premier vecteur d'éléments de la chaîne.

2. Poser $t \leftarrow 0$.

Pour passer de t à $t + 1$:

Etape 1 :

Générer z^{t+1} en simulant selon la loi $\pi(z/d_i, w^{(t)}, \alpha_i^{(t+1)}, \beta_i^{(t+1)})$

Etape 2 :

Générer $\alpha_i^{(t+1)}$ en simulant selon la loi $\pi(\alpha_i^{(t+1)} / d_i, z^{(t)})$

Générer $\beta_i^{(t+1)}$ en simulant selon la loi $\pi(\beta_i^{(t+1)} / d_i, \alpha_i^{(t+1)})$

3. Changer la valeur de t à $t \leftarrow t + 1$, et aller en 3.

-II Application :

1. les durées de chômage et le modèle utilisé

Un des outils les plus efficaces de l'analyse des durées de vie est certainement l'estimateur de la fonction de survie ou de séjour. Pour le cas de cette étude, cette fonction est relative à la durée de chômage. Généralement, l'estimateur le plus utilisé pour cette estimation est celui de Kaplan-Meier, qui permet de tenir compte des données censurées à droite. Cet estimateur calcule la probabilité de connaître l'événement dans chaque intervalle de temps, et on obtient ainsi une courbe qui s'interprète simplement comme la proportion de "survivants" pour chaque durée de séjour dans un état donné. Autrement dit, les proportions des individus sortant du chômage pour chaque durée de chômage.

Dans cette application on va analyser les durées de chômage global dans l'agence locale de l'emploi d'Ain El Benian. On travaille sur un échantillon de 1064 individus en chômage observées entre 01/01/2011 et 15/07/2013. Cette application permet à démontrer pratiquement que les procédures bayésienne de mélange bêta constituent un élément efficace pour résoudre de problème de label-swetching et de trouver des estimations de bonne qualité et de bonne précision. En distinguant ceux qui ont trouvés un emploi, le placement des chômeurs pendant cette période donne lieu à 875 observations censurées à droite. Dans ce cas, la variable i représente l'indication que le i ème chômeur a accédé à un emploi après sa période journalière de chômage t_i .

2. Analyse des durées de chômage

La première étape est consacré à l'estimation du nombre des composantes de mélange, pour cette finalité on utilise le critère de déviance d'information (Spiegelhalter et al, 2002) qui prendre en compte l'information a priori.

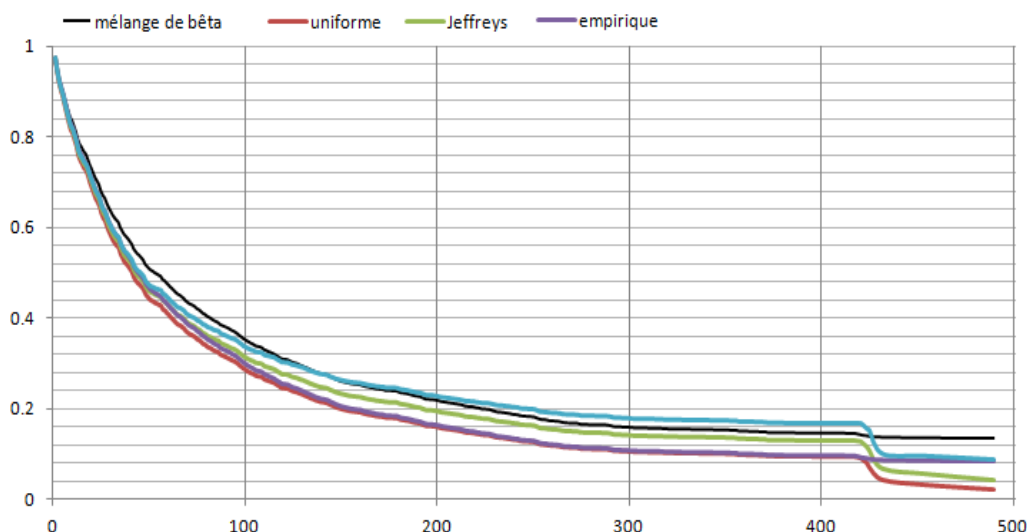
Tableau (1) : La comparaison selon DIC les différentes composantes de mélange.

k	2	3	4	6	7
DIC	964.5	964.4	964.5	965.4	1124.4

La source : réalisée par les chercheurs.

Selon le tableau (1), on remarque que le DIC est significative après 6 composantes de mélange a priori, en effet on choisit le premier modèle qui représente le modèle le moins complexe permet les six premiers.

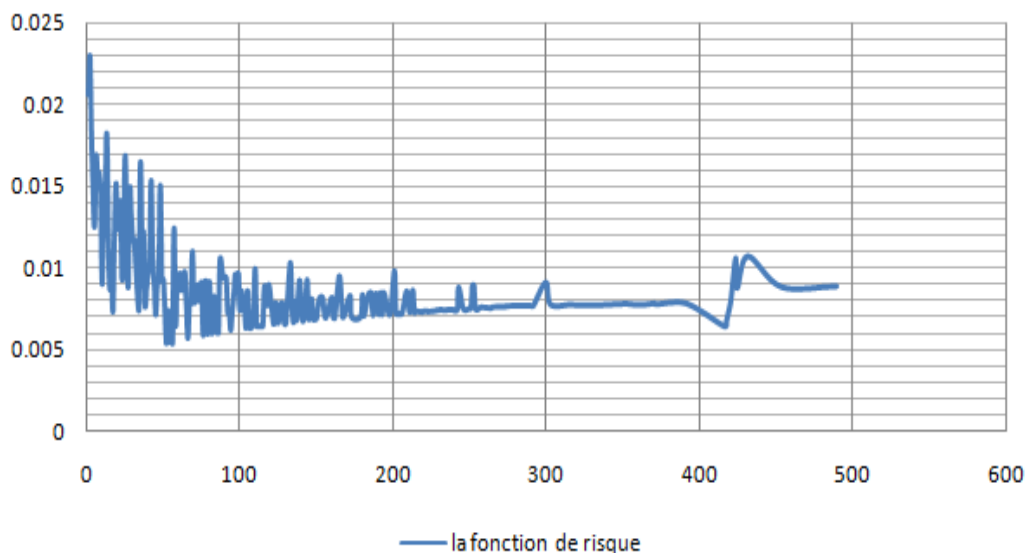
Figure (2) : Les fonctions de survie bayésienne selon plusieurs distributions a priori.



La source : réalisée par les chercheurs.

D'après la figure (2), on remarque qu'au début de la courbe, 100% des individus de l'échantillon sont en chômage, et après 2 mois d'inscription à cette agence, 50% des individus étaient placés dans le marché du travail. Mais, la sortie du chômage pour le reste des individus de l'échantillon s'étale sur une longue durée, pour certain elle dépasse même une année. On remarque aussi que la différence entre les modèles proposés d'estimation se produit graduellement après la durée médiane de sortir de chômage.

Figure 3 : la fonction d risque



La source : réalisée par les chercheurs.

D'après, la figure (3) on constate que la probabilité de sortir du chômage diminue progressivement. Cela, nous conduit à déduire qu'il y a une dépendance de durée négative. Au départ, la probabilité de sortie du chômage est de 0,02 pour les demandeurs d'emploi qui connaissent de courtes durées de chômage. Ensuite, le risque commence à diminuer graduellement jusqu'à atteindre une probabilité de sortie du chômage très faible inférieure à 0,009 pour les chômeurs avec des durées supérieures à 200 jours.

Tableau (2) : La comparaison entre les différentes méthodes bayésiennes.

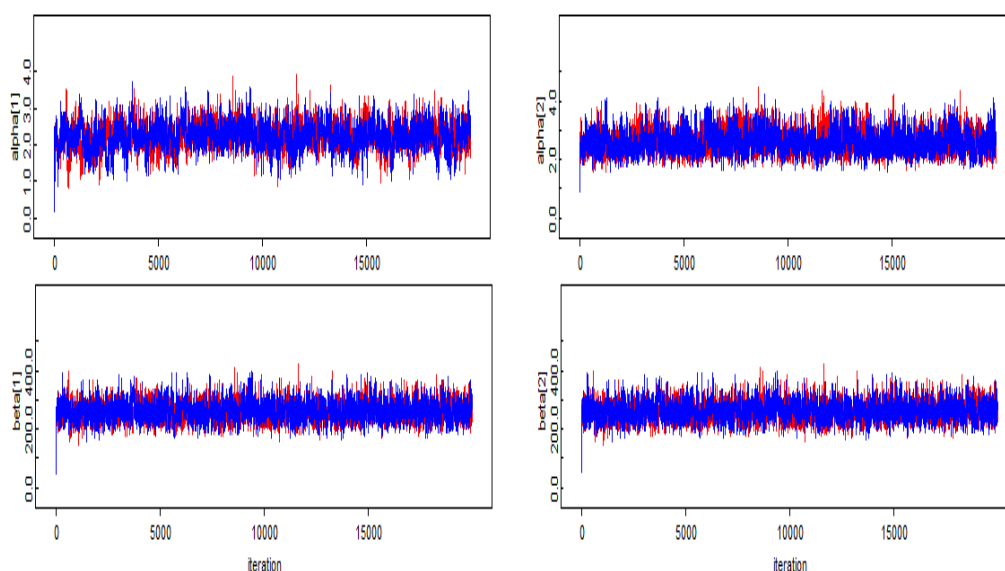
La méthode d'estimation	mélange de bêta	uniforme	Jeffreys	vague	empirique
DIC	925.1	1501	1280	1157	1042

La source : réalisée par les chercheurs.

Le tableau (2) nous indique que le critère de déviance d'information (DIC) de l'estimation bayésienne non paramétrique basée sur un mélange de bêta est inférieur avec un écart important par rapport aux modèles comparatives, ce résultat montre l'efficacité de cette méthode et l'importance de tenir compte de l'hétérogénéité observée dans l'échantillon, il était nécessaire de stratifier l'échantillon selon ses caractéristiques

Sur la figure (4), on peut constater graphiquement une certaine stationnarité des valeurs a posteriori tout au long des 20000 itérations pour le modèle de mélange des lois de bêta. Le modèle utilisé permet de surmonter le problème de label switching. Aussi les deux chaînes se mélangent bien : la convergence est atteinte.

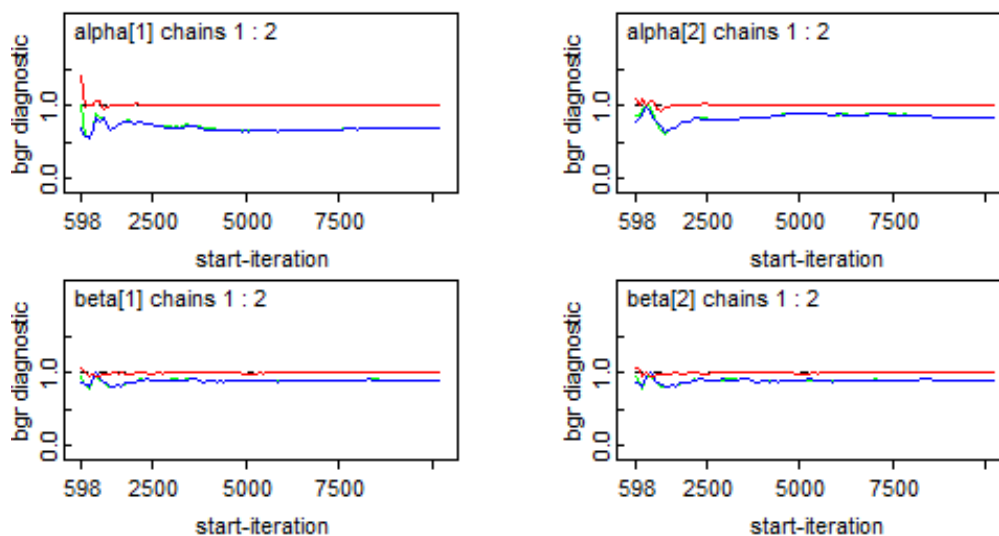
Figure (4) : La trace de la loi a posteriori pour les paramètres (α, β) dans le modèle proposé de mélange bêta.



La source : réalisée par les chercheurs par OpenBUGS.

Brooks et Gelman en 1998 ont proposé une généralisation de la méthode de Gelman et Rubin qui a été introduite dans l'année 1992, c'est une méthode de validation des suites ergodiques des algorithmes MCMC.

Figure (5) : Le graphe de Brooks et Gelman « convergence – diagnostic – graph » après 20000 itérations dans le modèle proposé de mélange bêta.



La source : réalisée par les chercheurs par OpenBUGS.

Dans la figure (5) la courbe verte indique la largeur de l'intervalle de crédibilité inter-chaînes à 80%. La courbe bleue indique la largeur moyenne des intervalles de crédibilité intra-chaîne à 80%. La courbe rouge indique la statistique de Brooks et Gelman (i.e., le ratio des courbes vertes/bleues). La statistique de Brooks et Gelman tend vers le 1, cela signifie qu'il y a convergence.

-III Conclusion :

Lorsque l'utilité intrinsèque de la modélisation bayésienne hiérarchique est de lever la difficulté pour le calcul de la distribution a postérieure et grâce à la flexibilité des modèles de mélange fini il est possible d'associer ces deux magnifiques solutions à partir d'une idée très connue en statistique non paramétrique des modèles de durée il s'agit le modèle de Kaplan Meier. En effet nous pouvons résoudre plusieurs difficultés qu'on trouve dans l'estimation d'une fonction de survie dans certaines conditions. Le présent article donne les résultats suivants :

- On trouve que le modèle à priori de mélange fini basé sur la distribution de Dirichlet porte des résultats prometteurs par rapport aux comparateurs dans l'estimation des probabilités de sortir du chômage sous la présence de censure, nous proposons aussi un code sur l'OpenBUGS qui facilite l'usage de ce modèle dans différentes problématiques.
- D'une manière générale, d'après la courbe de durée de chômage, on déduit que la probabilité de sortir du chômage pour les inscrits à l'Agence Locale de l'Emploi d'Ain el Benian devient très faible pour un chômeur qui dépasse plus d'une année de chômage.
- Après approximativement 2 mois d'inscription à l'Agence locale de l'Emploi d'Ain El Benian, 50% des individus étaient placés dans le marché du travail. Mais, la sortie du chômage pour le reste des individus de l'échantillon s'étale sur une longue durée, pour certains elle dépasse même une année.
- Il y a une dépendance de durée négative entre la probabilité de sortie du chômage et la durée de demande de l'emploi.
- Le critère de déviance d'information (DIC) de l'estimation bayésienne basée sur un mélange de bêta montre l'efficacité de cette méthode et l'importance de tenir compte de l'hétérogénéité observée dans l'échantillon, il était nécessaire de stratifier l'échantillon selon ses caractéristiques.

En perspective, et lorsque l'approche Bayésienne de mélange fini de l'estimateur de Kaplan Meier est bien adaptée aux analyses des données de durées, il serait intéressant de poursuivre la méthodologie suivie dans notre étude pour améliorer ce travail dans l'objectif de détecter les déterminants de l'insertion des chômeurs inscrits à l'agence locale de l'emploi d'Ain el Benian d'une part, et d'autre part, d'examiner la nature de la relation (positive ou négative) existante entre les variables exogènes susceptibles d'influencer leurs placements

-Références:

- Diebolt, J. and Robert, C.P (1994), **Estimation of finite mixture distributions by Bayesian sampling**. Journal of the Royal Statistical Society B, 56, 363-375.
- Haldane, J. (1931). **A note on inverse probability**. Proc. Cambridge Philos.Soc. , 28, 55–61.
- Kaplan, E. L., and Meier, P (1958), **Nonparametric estimation from incomplete observations**. J. Amer. Statist. Assoc, 53(282) : 457 481.
- Khizanov, V. G., Maïboroda, R (2015), **A modified Kaplan-Meier estimator for a model of mixtures with varying concentrations**. Theor. Probability and Math. Statist. 92 (2016), 109-116.
- Novick, M. et Hall, W (1965), **A Bayesian indifference procedure**.J. American Statist. Assoc., 60, 1104–1117.
- Robbins, H (1951), **Asymptotically subminimax solutions to compound statistical decision problems**. In Proc. Second Berkeley Symp. Math. Statist. Probab., volume 1. University of California Press.
- Robbins, H (1955), **An empirical Bayes approach to statistics**. InProc. Third Berkeley Symp. Math. Statist. Probab. , volume 1. University of California Press.
- Robbins, H (1964), **The empirical Bayes approach to statistical decision problems**. Ann. Mathemat. Statist.,35, 1–20.
- Robbins, H (1983), **Some thoughts on empirical Bayes estimation**. Ann. Sta-tist, 11, 713–723.
- Robert, C.P (2006), **Le choix Bayésien : principes et pratiques**. Springer.
- Robert, C.P (2013), **Des spécificités de l'approche bayésienne et de ses justifications en statistique inférentielle**. In **Les approches et méthodes bayésiennes**, sciences et épistémologie (ed. I. Drouet). Éditions Matériologiques (to appear). Available as arxiv:1403.4429.
- Rossa, A and Zieliński, R (2006), **A simple improvement of the kaplan-meier estimator**. Communications in statistics - theory and methods, 31(1), 147-158, doi: 10.1081/sta-120002440.
- Shafiq, Mohammad, Shah, Shuhtrat, Alamgir, M (2007), **Modified Weighted Kaplan-Meier Estimator**. Pakistan Journal of Statistics and Operation Research, 3,39-44.
- Spiegelhalter, D. J., Best, N., Carlin, B.P., Van der Linde, A (2002), **Bayesian measures of model complexity and fit**. Journal of the Royal Statistical Society, Series B, 64, 583–640.