Projet de statistique bayésienne
Marginal Likelihood from the Gibbs Output

Rémi Boutin, Alexis Gerbeaux

January 2020

1 Présentation du problème

L'article de Siddhartha Chib dans le *Journal of the American Statistical Association*, intitulé *Marginal Likelihood From the Gibbs Output (1995)*, permet de calculer la distribution d'une observation étant donnée la vraisemblance et d'autres éléments (que l'on précisera plus bas). Pour comprendre comment arriver à ce résultat, on définit d'abord les éléments nécessaires à notre étude. Soit $y = (y_i)_{i=1}^n$ nos données de loi $\mathbb{P}$, notre modèle $(P_\theta, \theta \in \Theta)$, la log vraisemblance de notre modèle, $\forall \theta \in \Theta, l(\theta) = \log(f(y \mid \theta))$ où $\frac{dP_\theta}{d\lambda}(x) = f(x \mid \theta)$.

Calcul de $l(y) = \log(f(y))$ L'objectif principal est d'estimer la loi $f(y)$. Pour ce faire, on utilise la relation suivante :

$$f(y) = \frac{f(y \mid \theta)\pi(\theta)}{\pi(\theta \mid y)} = \frac{f(y \mid \theta)\pi(\theta)}{\int \pi(\theta \mid y, z)\pi(z \mid y)dz} \tag{1}$$

où z est une variable latente. Une variable z est dite latente si les observations sont indépendantes conditionnellement à z, mais z n'est pas directement observée. Par exemple, supposons que les données soient distribuées selon un modèle de mélange. Si le groupe auquel appartient chaque observation n'est pas observé, il est considéré comme une variable latente (c.f l'exemple des galaxies dans la région de la couronne boréale). Dans la relation précédente, le numérateur est connu puisqu'il correspond à la valeur de la vraisemblance et du prior mis sur le paramètre du modèle.
Ainsi, on obtient :

$$\ln \hat{f}(y) = \ln f(y \mid \theta^*) + \ln \pi(\theta^*) - \ln \hat{\pi}(\theta^* \mid y) \tag{2}$$

Et le reste du travail repose sur l'estimation du terme $\pi(\theta \mid y)$.

Estimation de $\pi(\theta \mid y)$ en 3 étapes

1.1 Estimation du maximum de vraisemblance

L'algorithme proposé ne peut être appliqué que dans un cadre relativement restreint. On choisit de le présenter selon pour un paramètre θ décomposé en 2 blocs. Soit $\theta = (\theta_1, \theta_2) \in \mathbb{R}^d$ avec $\theta_i \in \mathbb{R}^{d_i}$, $i = 1, 2$ et $d_1 + d_2 = d$. Sachant que $\pi(\theta \mid y) = \pi(\theta_1 \mid y)\pi(\theta_2 \mid \theta_1, y)$, on va

estimer les deux quantités successivement. On suppose que l'on connaît les densités des distributions suivantes:

$$\pi(\theta_1 \mid y, z, \theta_2), \pi(\theta_2 \mid y, z, \theta_1),\ \pi(z \mid y, \theta)$$

Elles permettent la mise à jour dans l'échantillonneur de Gibbs. On note $(\theta_1^{(g)}, \theta_2^{(g)}, z^{(g)})_{g=1}^{G}$, l'échantillon issu de cet échantillonneur. On peut donc choisir $(\theta_1^*, \theta_2^*, z^*)$ qui maximise la vraisemblance. Remarquons que la formule 2 est vraie pour tout θ, prendre le maximum de vraisemblance n'a un intérêt que purement numérique, afin d'assurer plus de stabilité (car il est plus facile d'échantillonner dans cette zone). On obtient donc le premier estimateur naturel :

$$\hat{\pi}(\theta_1^* \mid y) = G^{-1} \sum_{g=1}^{G} \pi(\theta_1^* \mid y, z^{(g)}, \theta_2^{(g)}) \tag{3}$$

Pour estimer $\pi(\theta_2^* \mid y)$, on a besoin que les échantillons $(z^{(g)})$ soient tirés selon la loi $\pi(z \mid y, \theta_1^*)$ (on le voit dans l'équation 4, lors de l'estimation de l'intégrale). On rééchantillonne selon $\pi(\theta_2 \mid y, z, \theta_1^*), \pi(z \mid y, \theta_2, \theta_1^*)$ et on obtient l'estimateur suivant à partir du nouvel échantillon.

$$\hat{\pi}(\theta_2^* \mid y, \theta_1^*) = G^{-1} \sum_{g=1}^{G} \pi(\theta_2^* \mid y, z^{(g)}, \theta_1^*) \tag{4}$$

Il ne alors qu'à prendre le produit des deux quantités comme estimation de $\pi(\theta \mid y)$ recherchée.

2 Exemple dans un modèle de classification probit

Dans cette section, on cherche à reproduire les résultats de la section 4.1 de l'article. L'auteur souhaite modéliser la propagation du cancer de la prostate aux ganglions lymphatiques environnants. Les données concernent 53 patients atteints d'un cancer de la prostate pour lesquels on collecte 5 variables explicatives. Si une propagation du cancer a lieu, y, la variable que l'on cherche à prédire vaut 1 et 0 sinon.

L'auteur souhaite calculer la loi marginale, notée $f_i(y)$ (l'indice i fait référence au modèle considéré le cas échéant), pour plusieurs modèles probit afin d'évaluer le facteur de Bayes pour chacun des couples de modèle et de trouver le meilleur modèle.

On rappelle qu'une fois que l'on trouve le logarithme des vraisemblances marginales des différents modèle il suffit de calculer l'exponentielle de la différence pour trouver le facteur de Bayes. Autrement dit, le facteur de Bayes entre le modèle M_i et le modèle M_j est :

$$B_i j(y) = \frac{f_i(y)}{f_j(y)} = \exp\left\{\ln f_i(y) - \ln f_j(y)\right\} \tag{5}$$

où les $\ln f_j(y)$ sont calculés pour les 9 modèles testés dans l'article dans l'avant dernière colonne de la table 2.

Le problème est modélisé de la façon suivante pour un modèle probit donné. Soit β le vecteur des paramètres de la régression. β est de taille $p * 1$ où p est le nombre de features considérées pour le modèle en question. X, la matrice des variables explicatives, est de taille

$n*p$ avec $n = 53$ le nombre d'individus pour lesquels on évalue le modèle. Et x_i est le vecteur de taille $p*1$ des caractéristiques de l'individu i. On définit alors la vraisemblance du modèle de la façon suivante, avec Φ la fonction de répartition de la loi Normale $\mathcal{N}(0,1)$:

$$f(y \mid \beta) = \prod_{i=1}^{n} \Phi(x_i'\beta)^{y_i}(1 - \Phi(x_i'\beta))^{1-y_i} \tag{6}$$

Les hypothèses du modèle sont que le vecteur β suit une loi Normale $\mathcal{N}(a, A^{-1})$ de moyenne le vecteur a où tous les valeurs valent $0,75$ et de matrice de variance covariance A^{-1} où seuls les termes diagonaux sont non nuls et valent 25.

Pour faciliter la simulation à l'aide de l'algorithme de Gibbs, on introduit la variable latente $z \in \mathcal{M}_{n1}$, telle que $y_i = 1_{\{z_i>0\}}$. On a alors, $z_i \sim \mathcal{N}(x_i'\beta, 1)$.

L'algorithme de Gibbs est défini par :

$$\begin{aligned}(\beta \mid y, z) &\sim \mathcal{N}(\hat{\beta}_z, B)\\ (z_i \mid y, \beta) &\sim \mathcal{N}_+(x_i'\beta, 1), y_i = 1\\ &\sim \mathcal{N}_-(x_i'\beta, 1), y_i = 0\end{aligned}$$

avec $\hat{\beta}_z = (A+X'X)^{-1}(Aa+X'z)$, $B = (A+X'X)^{-1}$ et $\mathcal{N}_+$ et $\mathcal{N}_-$, les lois normales tronquées sur $\mathcal{R}_+$ ou $\mathcal{R}_-$.

L'algorithme de Gibbs nous permet d'obtenir pour un modèle de gibbs $G = 5000$ simulations pour $\beta^{(g)}, \hat{\beta}_z^{(g)}$. Cela nous permet de calculer β^*, puis d'évaluer le logarithme de la loi marginale de la façon suivante :

$$\beta^* = \frac{1}{G}\sum_{g=1}^{G} \beta^{(g)} \tag{7}$$

$$\ln \hat{f}(y) = \ln f(y \mid \theta^*) + \ln \pi(\theta^*) - \ln \hat{\pi}(\theta^* \mid y) \tag{8}$$

Nous avons dans un premier temps retrouvé le maximum de vraisemblance des différents modèles étudiés. Les maximums de vraisemblance sont retrouvés sauf pour les modèles avec le log de la variable x_2 où un petit écart est observé de l'ordre de 10^{-2}.

Ensuite, nous avons cherché à évaluer le logarithme de la loi marginale comme le fait l'auteur dans la table 2. Nous obtenons des résultats très proches de ceux obtenus dans la table 2. Pour le modèle 8 par exemple, nous obtenons $-34,53$ contre $-34,55$ dans l'article. Les différences sont de l'ordre de 10^{-2} pour l'évaluation de la log marginal likelihood des autres modèles également. Ces calculs permettent de calculer, comme dans l'article, le facteur de Bayes entre deux modèles et d'en déduire quel est le meilleur modèle.

3 Exemple dans un modèle de mélange gaussien

Dans cette section, on reproduit les résultats de la section 4.2 de l'article. Le but est d'estimer la loi marginale $p(y)$ dans un modèle de mélange, sans connaissance à priori sur les différents groupes composant les données, représentées figure 1.

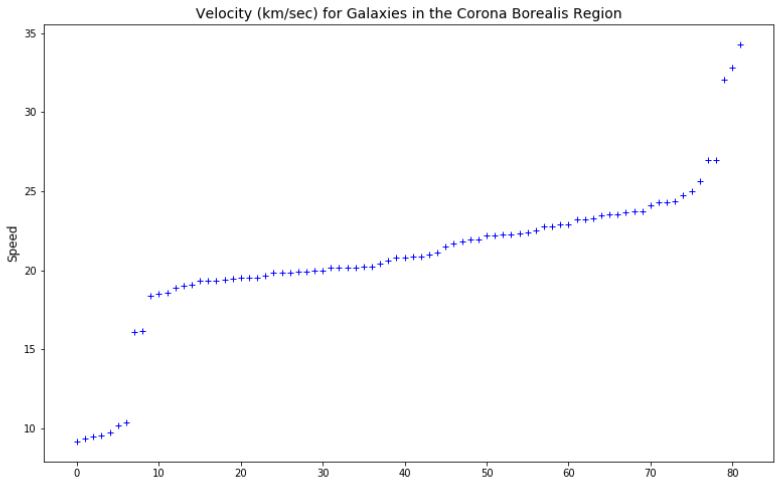

Figure 1: Vitesse des galaxies

On se retrouve donc avec la vraisemblance suivante

$$f(y \mid \theta) = \int f(y \mid \theta, z) p(z \mid \theta) dz = \prod_{i=1}^{n} \sum_{k=1}^{K} \phi(y_i \mid \mu_k, \sigma_k^2, z = k) q_k$$

où ϕ correspond à la densité de la loi normale, de moyenne μ et de variance σ^2. En prenant des à priori, gaussiens sur μ, inverse-gamma sur σ et une loi de dirichlet pour q, on obtient le modèle conjugué avec les à prioris suivant:

$$\mu_j \sim \mathcal{N}(\mu_0, phi) \quad \sigma_j \sim \mathcal{IG}(\frac{v_0}{2}, \frac{\delta_0}{2}) \quad q \sim Dirichlet(\alpha_1, ..., \alpha_d)$$

Les lois à posteriori sont données par :

$$\begin{aligned} \mu \mid y, \sigma^2, z &\sim \mathcal{N}(\hat{\mu}, \hat{\Sigma}) \\ \sigma^2 \mid y, z, \mu &\sim \otimes_{k=1}^{K} IG((v_0 + n_k)/2, (\delta_0 + \delta_k)/2) \\ q \mid y, z &\sim Dirich(\alpha + n) \\ \mathbb{P}(z = j \mid y, \mu, \sigma, q) &\propto q_j \times \phi(\mu_j, \sigma_j^2) \end{aligned}$$

On obtient les résultats suivants pour des variances intra clusters différentes.

K	$l(y \mid \theta^*)$	$l(\theta^*)$	$\hat{l}(\mu^* \mid y)$	$\hat{l}(\sigma^* \mid y, \mu^*)$	$\hat{l}(q^* \mid y, \mu^*, \sigma^*)$	$l(\pi(y))$	Chib's value
3	-209.39	-22.78	2.15	-6.83	1.66	-229.15 (0.181)	-224.138 (0.086)
2	-220.41	-17.69	1.58	-5.32	0.29	-234.66	None

Figure 2: Représentation des moyennes des clusters échantillonnées par l'algorithme de Gibbs

Il est intéressant de comparer la figure 2 avec la figure 3 en annexe qui montre comment l'algorithme se comporte lorsque l'on considère la variance intra-classe comme non constante, avec 2 clusters. Le groupe en vert sera toujours bien identifié (du fait qu'il comporte près de 70 observations sur 82) et les deux groupes extrêmes (les faibles vitesses et les hautes) seront mélangés, au détriment d'une variance très forte dans ce groupe et très faible dans le groupe du milieu. Nos résultats ne sont pas conformes à ceux trouvés par Chib dans son papier, comme en atteste l'estimation de l'écart type, bien trop grande (3 fois supérieur à l'estimation donnée par Chib) pour être exacte, des erreurs d'arrondis pourraient s'être glissées dans des approximations lors des calculs, bien que des passages au log aient été utilisés.

References

[1] Siddhartha Chib *Marginal Likelihood from the Gibbs Output*. Journal of the American Statistical Association (1995)

Appendix

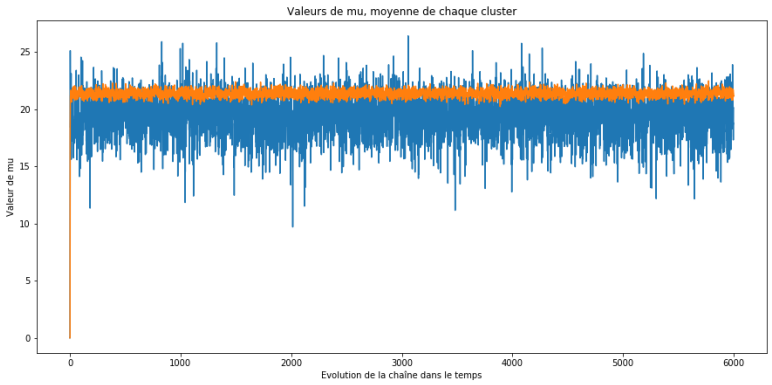

Figure 3: Modèle de mélange avec 2 clusters, et variance intra classe non constante