Journal de la Société Française de Statistique
Vol. 151 No. 2 (2010)

Méthodes bayésiennes variationnelles : concepts et applications en neuroimagerie

Title: Variational Bayesian methods : concepts and neuroimage applications

Christine Keribin[1,2]

Résumé : En estimation bayésienne, les lois a posteriori sont rarement accessibles, même par des méthodes de Monte-Carlo par Chaîne de Markov. Les méthodes bayésiennes variationnelles permettent de calculer directement (et rapidement) une approximation déterministe des lois a posteriori. Cet article décrit le principe des méthodes variationnelles et leur application à l'inférence bayésienne, fait le point sur les principaux résultats théoriques et présente deux exemples d'utilisation en neuroimagerie.

Abstract: Bayesian posterior distributions can be numerically intractable, even by the means of Markov Chains Monte Carlo methods. Bayesian variational methods can then be used to compute directly (and fast) a deterministic approximation of these posterior distributions. This paper describes the principle of variational methods and their applications in the Bayesian inference, surveys the main theoretical results and details two examples in the neuroimage field.

Mots-clés : Méthodes variationnelles, analyse bayésienne, approximation en champ moyen, approximation de Laplace, algorithme EM, données IRMf
Keywords: Variational methods, Bayesian analysis, mean field approximation, Laplace approximation, EM algorithm, IRMf data
Classification AMS 2000 : 49J40, 62F15

[1] Université Paris-Sud 11 - Département de mathématiques - UMR 8628 - 91405 Orsay cedex
E-mail : christine.keribin@math.u-psud.fr
[2] INRIA Saclay Île de France - Projet SELECT

 ISSN: 2102-6238

1. Introduction

Les applications de neuroimagerie utilisent des modèles bayésiens, dont les grandeurs d'intérêt (évidence, lois a posteriori) posent des problèmes de calcul : ceci peut être dû à la trop grande dimension de l'espace des variables latentes, ou à la forme trop complexe des lois a posteriori. Il est alors nécessaire d'avoir recours à des méthodes permettant d'approcher ces grandeurs. La méthode variationnelle en est un exemple, relevant du champ de l'approximation déterministe, et son utilisation s'est récemment répandue dans la communauté de neuroimagerie.

Le succès de l'approximation variationnelle est en effet dû à sa facilité d'utilisation et sa rapidité d'exécution dans des cas d'estimation qu'il peut être difficile de traiter avec les outils classiques (méthodes de Monte Carlo par Chaîne de Markov -MCMC- par exemple). Au lieu de calculer la grandeur exacte d'intérêt, elle maximise une fonctionnelle la minorant, obtenant ainsi une valeur approchée minorant la grandeur exacte à déterminer. Cette méthode n'est pas limitée au cadre bayésien : elle peut également être utilisée dans le cadre fréquentiste, pour approcher la vraisemblance par exemple [11, 22].

L'expression *méthode variationnelle* vient du calcul des variations : il s'agit d'exprimer comment la valeur de la fonctionnelle se modifie en réponse à d'infimes changements de la fonction d'entrée de la fonctionnelle ([5], chap. 10, p.462). Elle fait référence à différents outils mathématiques pour la formulation de problèmes d'optimisation, aussi bien qu'aux techniques associées à leur résolution. L'idée générale est d'exprimer la quantité d'intérêt comme solution d'un problème d'optimisation. Ce problème d'optimisation peut être modifié dans différentes directions, soit en approchant la fonctionnelle à optimiser, soit en approchant l'ensemble sur lequel est optimisé la fonctionnelle. De telles approximations, à leur tour, donnent un moyen d'approcher la quantité d'intérêt initiale : c'est l'*approximation variationnelle* [20].

Même si ces méthodes ont été récemment utilisées dans les applications de neuroimagerie avec des résultats prometteurs, il convient d'être attentif à la qualité de l'approximation obtenue, dont on ne peut quantifier la précision. Ainsi, notre objectif est double : présenter la méthode de l'approximation variationnelle - principalement dans le cadre bayésien - et passer en revue des résultats théoriques connus d'une part, et, d'autre part, montrer son utilisation dans deux exemples d'application de neuroimagerie.

Les principes de l'approximation variationnelle sont présentés en section 2. La section 3 expose l'étude théorique de la qualité de l'approximation dans différents types de modèles, tandis que la section 4 présente son utilisation en neuroimagerie par Friston et al [9] et Woolrich et Behrens [26]. La section 5 propose une discussion générale sur les résultats précédents. Des détails de calculs sont reportés en annexe A.

2. L'approximation variationnelle

Un modèle statistique bayésien est constitué d'un modèle statistique paramétrique $p(y|\theta)$, et d'une *loi a priori* pour le paramètre $\pi(\theta)$, modélisant son incertitude. Le théorème de Bayes

$$p(\theta|y) = \frac{p(y|\theta)\pi(\theta)}{p(y)}$$

 ISSN: 2102-6238

permet d'actualiser l'information sur θ en extrayant l'information contenue dans l'observation y : $p(\theta|y)$ est la *loi a posteriori* du paramètre (consulter par exemple le chapitre 1 du livre de Robert [16] pour une présentation détaillée du cadre bayésien).

La présence de variables cachées x ajoute toujours un degré de complexité : d'une part, par la présence de corrélations possibles entre les variables cachées et les paramètres dans la *loi jointe a posteriori*

$$p(x,\theta|y) = \frac{p(y|x,\theta)p(x|\theta)\pi(\theta)}{p(y)},$$

d'autre part, parce que les *lois marginales a posteriori*,

$$p(\theta|y) = \int p(x,\theta|y)dx \quad ; \quad p(x|y) = \int p(x,\theta|y)d\theta,$$

intègrent l'ensemble des combinaisons des états cachés, rendant les calculs directs souvent inaccessibles. Des outils de calcul numérique comme les méthodes de Monte Carlo par Chaîne de Markov (MCMC) sont bien établis, mais leur utilisation n'est pas toujours simple ; de plus, leur implémentation pratique peut s'avérer impossible à réaliser à cause de l'explosion calculatoire, en particulier si la structure cachée est de grande dimension, ou présente des dépendances non triviales.

Ces problèmes se retrouvent également dans le calcul de l'*évidence*, ou *vraisemblance marginale*

$$p(y) = \int p(x,\theta,y)dxd\theta = \int p(y|x,\theta)p(x|\theta)\pi(\theta)dxd\theta,$$

nécessaire pour l'évaluation de la constante de normalisation d'une loi a posteriori, ou pour le calcul du facteur de Bayes utilisé dans la sélection de modèles. Dans ce cas en effet, une fois définies les probabilités $p(m)$ des différents modèles m, puis les lois a priori $p(\theta|m)$ des paramètres dans chacun des modèles, le choix bayésien est basé sur la loi a posteriori des modèles ([16], chap. 7 ou [5], chap. 3)

$$p(m|y) \propto p(m)p(y|m),$$

où l'on reconnaît dans le terme $p(y|m)$ la vraisemblance marginale

$$p(y|m) = \int p(y|x,\theta,m)p(x|\theta,m)\pi(\theta,m)dxd\theta.$$

La *méthode variationnelle* permet de transformer le calcul de l'intégrale précédente en résolution d'un problème optimisation, en remarquant que l'évidence est un majorant d'une quantité appelée *énergie libre*

$$\mathscr{F}(q_{x,\theta}) = \int q_{x,\theta}(x,\theta)\log\frac{p(x,y,\theta)}{q_{x,\theta}(x,\theta)}d\theta dx,$$

fonction d'une *distribution libre* $q_{x,\theta}$. En effet, l'inégalité de Jensen permet d'écrire :

$$\log p(y) \geq \mathscr{F}(q_{x,\theta})$$

avec égalité pour $q_{x,\theta} = p(x,\theta|y)$.

Ainsi, si l'ensemble des distributions libres $q_{x,\theta}$ n'est pas restreint, la distribution libre qui maximise $\mathscr{F}(.)$ est $p(x,\theta|y)$:

 ISSN: 2102-6238

$$p(x,\theta|y) = \arg\max_{q_{x,\theta}} \mathscr{F}(q_{x,\theta})$$

et la valeur de l'énergie libre au maximum est alors le logarithme de l'évidence

$$\log p(y) = \max_{q_{x,\theta}} \mathscr{F}(q_{x,\theta}) = \mathscr{F}(p(x,\theta|y)).$$

Mais ceci ne simplifie pas le problème, puisque l'évaluation de la loi exacte jointe a posteriori nécessite la connaissance de sa constante de normalisation, l'évidence. Les méthodes d'*approximation variationnelle* vont permettre de chercher la solution d'un problème approché, en modifiant la fonctionnelle à optimiser, ou en approchant l'ensemble des distributions libres sur lequel est optimisé la fonctionnelle : dans ce dernier cas, il s'agit de rechercher une forme approchée $q_{x,\theta}(x,\theta)$ de $p(x,\theta|y)$ dans un ensemble de fonctions dans lequel les calculs sont aisés, et d'en déduire une approximation de la log-évidence comme le majorant de l'énergie libre sur cet ensemble de fonctions.

$$\log p(y) \geq \max_{q_{x,\theta}} \mathscr{F}(q_{x,\theta}). \tag{1}$$

L'erreur d'approximation entre la log-évidence et l'énergie libre est alors la divergence de Kullback entre la distribution libre et la loi jointe a posteriori :

$$\begin{aligned} \log p(y) &= \int q_{x,\theta}(x,\theta) \log \frac{p(x,y,\theta)}{q_{x,\theta}(x,\theta)} d\theta dx + \int q_{x,\theta}(x,\theta) \log \frac{q_{x,\theta}(x,\theta)}{p(\theta,x|y)} d\theta dx \\ &= \mathscr{F}(q_{x,\theta}) + D(q_{x,\theta}||p(\theta,x|y)). \end{aligned} \tag{2}$$

Remarquons que la formulation (1) est présentée sous forme de la maximisation de l'énergie libre, mais elle pourrait l'être sous forme de la minimisation de la divergence de Kullback $D(q_{x,\theta}||p(\theta,x|y))$.

Plusieurs types d'approximation peuvent être définis. Nous détaillerons ci-dessous la factorisation (ou approximation en champ moyen) et l'approximation de Laplace, car elles sont largement utilisées en neuroimagerie. Il est également possible d'utiliser des classes d'approximation paramétrique, ou de factoriser différemment. Minka [15] propose la méthode EP - *Expectation propagation* - qui minimise, non pas la divergence de Kullback $D(q_{x,\theta}||p(\theta,x|y))$, mais une forme 'inverse' $D(p(\theta,x|y)||q_{x,\theta})$, ce qui donne à l'approximation des propriétés différentes, voir Bishop ([5], chap.10, p. 505). Wainwright et Jordan [20] présentent un panorama complet des méthodes variationnelles dans le cas des modèles graphiques, exploitant les propriétés des familles exponentielles.

On peut citer d'autres classes d'approximation bayésienne, hors du cadre des méthodes variationnelles : Tierney et Kadane [18] utilisent directement l'approximation de Laplace (au sens statistique) ; Rue, Martino et Chopin [17] ont développé la méthode INLA - Integrated Nested Laplace Approximation, qui permet de calculer directement des approximations très précises des marginales a posteriori, dans le cadre des modèles gaussiens à données latentes.

2.1. Approximation en champ moyen

La première approximation classiquement utilisée en neuroimagerie est l'*approximation en champ moyen*, qui permet de rechercher une distribution libre factorisée, par exemple en séparant les variables cachées des paramètres

 ISSN: 2102-6238

$$q_{x,\theta}(x,\theta) = q_x(x)q_\theta(\theta). \tag{3}$$

Si les variables cachées sont indépendantes conditionnellement à θ, l'approximation optimale de q_x dans la famille de distributions libres imposées par (3) se factorise alors simplement, d'où :

$$q_{x,\theta}(x,\theta) = q_{x_1}(x_1)\dots q_{x_n}(x_n)q_\theta(\theta).$$

Sinon, il est possible de restreindre l'espace des distributions libres à celles se factorisant suivant les variables cachées et imposer :

$$q_x(x) = \prod_i q_{x_i}(x_i).$$

L'algorithme bayésien variationnel maximise itérativement l'énergie libre $\mathscr{F}(q_x, q_\theta)$ par rapport aux distributions libres q_x (étape VBE, estimation de la loi approchée des variables cachées) et q_θ (étape VBM, maximisation pour obtenir la loi a posteriori des paramètres). Le théorème général suivant (Beal [3]) fournit le cadre général des équations de mise à jour pour l'apprentissage bayésien variationnel (VBEM) :

Théorème 1. *Soit un modèle de paramètre θ, dont on observe un n-échantillon i.i.d, $y = \{y_1,\dots,y_n\}$, avec des variables cachées correspondantes $x = \{x_1,\dots,x_n\}$. Une borne inférieure de la vraisemblance marginale est*

$$\mathscr{F}(q_x, q_\theta) = \int q_x(x)q_\theta(\theta)\log\frac{p(x,y,\theta)}{q_x(x)q_\theta(\theta)}d\theta dx$$

qui peut être optimisée itérativement en effectuant les mises à jour suivantes, l'indice (t) indiquant le numéro d'itération :

$$\textbf{\textit{étape VBE :}}\ q_{x_i}^{(t+1)}(x_i) = \frac{1}{Z_i^{(t+1)}}\exp\left[\int q_\theta^{(t)}(\theta)\log p(x_i,y_i|\theta)d\theta\right]\ \forall i$$

avec

$$q_x^{(t+1)}(x) = \prod_{i=1}^n q_{x_i}^{(t+1)}(x_i),$$

et

$$\textbf{\textit{étape VBM :}}\ q_\theta^{(t+1)}(\theta) = \frac{1}{Z_\theta^{(t+1)}}p(\theta)\exp\left[\int q_x^{(t+1)}(x)\log p(x,y|\theta)dx\right]$$

où $Z_i^{(t+1)} = \int\exp\left[\int q_\theta^{(t)}(\theta)\log p(x_i,y_i|\theta)d\theta\right]dx_i$
et $Z_\theta^{(t+1)} = \int p(\theta)\exp\left[\int q_x^{(t+1)}(x)\log p(x,y|\theta)dx\right]d\theta$ *sont les constantes de normalisation. De plus, l'algorithme converge vers un maximum local de* $\mathscr{F}(q_x, q_\theta)$.

Chaque étape garantit la croissance monotone du minorant de l'évidence : l'évidence elle même ne change pas, c'est son minorant qui s'accroît. D'autre part, puisque la fonction objectif n'est pas convexe, le processus peut ne converger que vers un maximum local ou un col.

Remarque. *Symétrie de l'écriture des deux étapes :*

 ISSN: 2102-6238

Il y a bien symétrie de l'écriture des deux étapes, qui s'observe en réécrivant $q_x^{(t+1)}$ et $q_\theta^{(t+1)}$ sous la forme suivante,

$$\begin{aligned} q_x^{(t+1)}(x) &= \frac{1}{\tilde{Z}_x^{(t+1)}} \exp\left[E_{q_\theta} \log p(x,y,\theta)\right] \\ &= \prod_i \frac{1}{\tilde{Z}_i^{(t+1)}} \exp\left[E_{q_\theta} \log p(x_i,y_i,\theta)\right] \\ q_\theta^{(t+1)}(\theta) &= \frac{1}{\tilde{Z}_\theta^{(t+1)}} \exp\left[E_{q_x} \log p(x,y,\theta)\right] \end{aligned}$$

où E_{q_θ} (et E_{q_x}) désignent l'espérance sous les lois $q_\theta^{(t)}$ (et $q_x^{(t+1)}$ respectivement), et où $\tilde{Z}_x^{(t+1)}$ (et $Z_\theta^{(t+1)}$) sont les constantes de normalisation indépendantes de x (et θ respectivement).

Remarque. *Réécriture de l'étape VBE si q_x n'est pas limitée à une forme factorisée :*

Si q_x n'est pas factorisée, l'étape VBE s'écrit

$$\textbf{étape VBE : } q_{x_i}^{(t+1)}(x_i) = \frac{1}{Z_i^{(t+1)}} \exp\left[\int q_\theta^{(t)}(\theta) \log p(x,y|\theta) d\theta dx_{-i}\right] \forall i,$$

où x_{-i} indique que l'intégration est faite sur toutes les variables cachées, sauf la i-ème.

Remarque. *Utilisation des méthodes variationnelles dans un cadre non bayésien :*

Dans ce cas, il s'agit alors d'approcher la vraisemblance des données $l_0(y;\theta)$ par une fonction approchée $l(y;\theta) = \mathscr{F}(q_x,\theta)$,

$$l_0(y;\theta) \equiv \log \int p(y,x;\theta) dx \geq \int q_x(x) \log\{p(y,x;\theta)/q_x(x)\} dx \equiv l(y;\theta). \tag{4}$$

Si aucune restriction n'est faite sur q_x, l'optimisation itérative de $\mathscr{F}(q_x,\theta)$ suivant q_x et θ permet d'atteindre l'estimateur du maximum de vraisemblance $\widehat{\theta}$ et la loi des variables cachées conditionnellement aux observations : l'algorithme EM [7] possède ainsi une interprétation variationnelle.

Quelle forme prendre pour q_x et q_θ ? Dans certains cas, la simplification induite par l'approximation en champ moyen suffit pour déterminer facilement les formes spécifiques optimales de q_x et q_θ. Quand ce n'est pas le cas, on peut avoir besoin de faire d'autres approximations et attribuer à q_x et q_θ, par exemple, une forme paramétrique n'ayant que quelques statistiques suffisantes.

2.2. Approximation de Laplace

Une méthode concurrente de l'approximation en champ moyen pour le calcul de l'évidence est l'utilisation de l'approximation de Laplace, qui approche l'intégrale en utilisant un développement de Taylor du logarithme de la fonction à intégrer autour de son maximum. En effet, soit $g(\theta)$ une densité non normalisée ayant un mode en θ^* et pour laquelle on cherche à calculer

$$Z_p = \int g(\theta) d\theta.$$

 ISSN: 2102-6238

Le développement de Taylor de $\log g(\theta)$ autour du maximum θ^* permet d'écrire

$$g(\theta) \simeq g(\theta^*) \exp -\frac{1}{2}(\theta - \theta^*)' H(\theta^*)(\theta - \theta^*)$$

où $H(\theta^*)$ est l'opposé du hessien de g en θ^*, d'où

$$\int g(\theta) d\theta \simeq \frac{g(\theta^*)(2\pi)^{d/2}}{|H(\theta^*)|}.$$

Dans le cadre bayésien, $g(\theta) = \pi(\theta)p(y|\theta)$, et l'approximation de Laplace fait une approximation gaussienne locale autour de l'estimateur du Maximum A Posteriori (MAP)

$$\theta^* = \widehat{\theta}_{MAP} = \arg\max_{\theta} p(\theta)p(y|\theta).$$

La validité de cette approximation est basée sur des propriétés de comportement gaussien asymptotique de la loi de l'échantillon et quelques conditions de régularité : la loi a posteriori ainsi approchée peut être mauvaise pour de petits jeux de données ; de même, cette approximation peut être mauvaise pour des paramètres bornés, contraints ou positifs, comme des proportions de mélange ou des précisions ou si le maximum n'est pas proche de la masse principale de la probabilité. Enfin, à cause des problèmes d'identifiabilité, la loi a posteriori peut ne plus être unimodale pour des vraisemblances avec données cachées, et dans ce cas, les conditions de régularité pour la convergence ne sont pas vérifiées [3].

Les deux méthodes -approximation en champ moyen, approximation de Laplace- sont concurrentes, mais elles peuvent aussi être combinées dans certains cas, comme dans l'exemple de neuroimagerie que nous présentons en section 4.1.

3. Etude théorique de la qualité de l'approximation bayésienne variationnelle

L'approximation bayésienne variationnelle a donné cours à de nombreux travaux ces dernières années, et a été appliquée à différents modèles : modèles de mélange (Wang et Titterington [21, 24, 23, 25]), modèles à espace d'états (Wang et Titterington [22]), modèles graphiques (Attias [2], Jordan [13], Beal et Ghahramani [4]), réseaux de neurones (Titterington [19]) par exemple. Nous nous focalisons ici sur les travaux portant sur l'étude de la qualité de l'approximation.

En effet, l'approche variationnelle permet de déterminer un minorant de l'évidence, mais la qualité de ce minorant, définie par l'information de Kullback, ne peut être prédite : si ceci était possible, on pourrait, dès le début, estimer plus précisément la vraisemblance marginale. Cependant, il est possible d'étudier théoriquement la qualité de l'approximation, par exemple dans certains cas où la loi exacte a posteriori est accessible. Nous commencerons par évoquer l'exemple simple d'approximation d'une gaussienne bivariée par un produit de gaussienne, puis nous passerons en revue le modèle de mélange, le modèle probit, le modèle markovien à données manquantes et le modèle à espace d'états.

 ISSN: 2102-6238

3.1. Cas de la gaussienne bivariée

Bishop ([5], chap. 10, p. 466) détaille un exemple simple permettant d'expliquer les limites de l'approximation factorisée, en présentant le cas de l'approximation d'une gaussienne bivariée $p(z) = \mathcal{N}(z|\mu, \Sigma)$ de moyenne et variance

$$\mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \Sigma = \begin{pmatrix} v_1 & v_{12} \\ v_{12} & v_2 \end{pmatrix}$$

sous forme d'un produit factorisé $q(z) = q_1(z_1)q_2(z_2)$. Le facteur optimal q_1^* s'écrit

$$\begin{aligned} \log q_1^*(z_1) &= E_{q_{z_2}}[\log p(z)] + const \\ &= -\frac{1}{2} z_1^2 \Lambda_1 + z_1 \mu_1 \Lambda_1 - z_1 \Lambda_{12}(E[z_2] - \mu_2) + const \end{aligned}$$

avec $\Lambda_1 = \frac{v_2}{v_1 v_2 - v_{12}^2}$, $\Lambda_{12} = \frac{-v_{12}}{v_1 v_2 - v_{12}^2}$, les coefficients de la matrice de précision Σ^{-1}. En remarquant que le terme de droite de l'équation est quadratique en z_1, q_1^* est identifiée à une gaussienne :

$$q_1^*(z_1) = \mathcal{N}(z_1|m_1, \Lambda_1^{-1}), \quad m_1 = \mu_1 - \Lambda_1^{-1}\Lambda_{12}(m_2 - \mu_2)$$

Ce résultat ne provient pas d'une hypothèse a priori sur q_1^*, mais découle de l'optimisation sur toutes les distributions libres possibles q_1. De même, $q_2^*(z_2)$ est également gaussienne :

$$q_2^*(z_2) = \mathcal{N}(z_2|m_2, \Lambda_2^{-1}), \quad m_2 = \mu_2 - \Lambda_2^{-1}\Lambda_{12}(m_1 - \mu_1).$$

Ces équations sont couplées, mais ici sous une forme suffisamment simple pour les résoudre directement et ne pas avoir à utiliser la boucle suivant les différentes variables. On obtient : $m_1^* = \mu_1$, $m_2^* = \mu_2$, les moyennes sont correctement estimées. En revanche, la variance de $q^*(z)$ est contrôlée par la direction de plus petite variance de $p(z)$, et la variance suivant l'axe orthogonal est significativement sous estimée :

$$v_1^* = \Lambda_1^{-1} = v_1 - \frac{v_{12}^2}{v_2} = v_1(1 - c_{12}^2), \quad v_2^* = \Lambda_2^{-1} = v_2 - \frac{v_{12}^2}{v_1} = v_2(1 - c_{12}^2)$$

avec c_{12}, la corrélation entre z_1 et z_2.

3.2. Modèle de mélange

Un modèle de mélange est défini comme une combinaison convexe

$$\sum_{j=1}^{J} p_j f_j(y), \quad \sum_{j=1}^{J} p_j = 1, \quad p_j > 0, \quad J > 1,$$

de lois f_j. Les p_j sont appelés *proportions*, et sont le plus souvent inconnus. Dans la plupart des cas, les f_j sont paramétrés, chacun par un paramètre inconnu θ_j, menant au modèle de mélange paramétrique générique suivant :

$$\sum_{j=1}^{J} p_j f_j(y|\theta_j). \tag{5}$$

 ISSN: 2102-6238

Les modèles de mélanges sont des exemples typiques de modèles à observations cachées (ou variables latentes) : en effet, l'échantillon $y_1, \dots y_n$ peut être vu comme une collection de sous-échantillons provenant chacun d'un groupe j de loi $f_j = f(.|\theta_j)$, quand la taille de chaque échantillon et la provenance de chacune des observations sont à la fois inconnues.
Cette représentation par des observations cachées peut être utilisée comme un moyen technique pour faciliter le calcul numérique. Par marginalisation, il est toujours possible d'associer à une variable aléatoire y_i du mélange (5) une seconde variable aléatoire finie x_i telle que

$$y_i|x_i \sim f(y_i|\theta_{x_i}), \quad p(x_i = j) = p_j.$$

La variable auxiliaire x_i identifie la composante du mélange à laquelle l'observation y_i appartient (Lee et coauteurs [14] pour une étude récente de l'inférence bayésienne des modèles de mélange).

Titterington et différents coauteurs [21, 24, 23, 25] ont investigué les propriétés de l'approximation bayésienne variationnelle dans le cas de modèle de mélange que l'on notera :

$$p(y_i|\Theta) = \sum_{j=1}^{J} p_j(y_i|\Theta)p(x_i = j|\Theta)$$

où Θ regroupe les proportions du mélanges et les paramètres des f_j. Ces auteurs montrent que :

- Dans un modèle de mélange où les f_j appartiennent à la famille exponentielle, la procédure itérative VBEM (avec approximation en champ moyen) définit un $\tilde{\Theta}$ qui converge localement, avec probabilité 1 quand n tend vers l'infini, vers la vraie valeur Θ^* des proportions du mélange. La loi de $\tilde{\Theta}$ est asymptotiquement normale à la vitesse $O(\frac{1}{\sqrt{n}})$ [21]
- Dans un modèle de mélange gaussien, $\tilde{\Theta}$ converge localement vers l'estimateur du maximum de vraisemblance Θ_{MV} à la vitesse $O(\frac{1}{n})$ [24, 25], mais la matrice de covariance de la loi limite de $\tilde{\Theta}$ est inférieure à celle de Θ_{MV} : les intervalles de confiance des paramètres seront en général trop étroits, en particulier quand les composantes du mélange ne sont pas bien séparées [24, 23].
- Mais l'approche échoue à distance finie, dans le sens où la loi variationnelle approchée est non seulement plus simple que la vraie loi a posteriori, mais aussi différente par essence. Par exemple, Humphreys et Titterington [12] étudient un mélange simple à deux composantes connues et à proportions inconnues, et montrent que la loi variationnelle de la proportion du mélange est une simple loi Beta, alors que la vraie loi a posteriori est une mélange compliqué de loi Beta (qui ne peut donc se réduire à une loi Beta).

Ces résultats montrent que dans les modèles de mélange, l'utilisation de la forme factorisée $q_x q_\theta$ de la loi jointe a posteriori ne biaise pas l'estimation pour les grands échantillons, parce que les corrélations viennent des lois a priori sur les paramètres (choix du cadre bayésien), et non du modèle lui-même, comme c'est le cas par exemple dans les modèles de chaînes de Markov.
L'énergie libre pouvant être multimodale dans les modèles de mélanges à paramètres inconnus, l'algorithme VBEM peut converger vers des limites différentes en fonction du point d'initialisation : ainsi, la convergence n'est que locale, comme pour l'algorithme EM.

 ISSN: 2102-6238

3.3. Modèle probit bayésien

Les résultats précédents ont également été constatés par Consonni et Marin [6] dans le cadre du modèle probit à variables latentes. Soient n variables aléatoires continues latentes $x_1,\dots,x_n$ telles que $x_i|\theta \sim \mathcal{N}(v_i'\theta,1)$, où $v_i \in \mathbb{R}^p$ sont des covariables connues et v' désigne la transposée de v. On observe réellement les variables aléatoires binaires $y_1,\dots,y_n$ définies par

$$y_i = \begin{cases} 1 & \text{si } x_i > 0 \\ 0 & \text{sinon} \end{cases}$$

Étant donné θ, les y_i sont des variables indépendantes de loi de Bernoulli

$$\Pr(y_i = 1|\theta) = \Phi(v_i'\theta),$$

où Φ est la fonction de répartition de la loi gaussienne standard : c'est le modèle probit.

Prenant comme loi a priori sur θ la loi gaussienne standard, $p(\theta) \sim \mathcal{N}(0_p, I_p)$, on peut établir que la loi a posteriori conditionnelle aux données complètes s'écrit

$$p(\theta|y,x) \sim \mathcal{N}_p((I_p + V'V)^{-1}V'x, (I_p + V'V)^{-1}),$$

où $V = [v_1|\dots|v_p]$. On en déduit la variance a posteriori de θ

$$\mathrm{var}(\theta|y) = \mathbb{E}(\mathrm{var}(\theta|y,x)) + \mathrm{var}(\mathbb{E}(\theta|y,x)) = (I_p + V'V)^{-1} + \mathrm{var}((I_p + V'V)^{-1}V'x).$$

Le calcul exact de la loi a posteriori est infaisable. Elle peut être estimée par un échantillonneur de Gibbs ou par la méthode variationnelle en champ moyen. Dans ce dernier cas, la valeur courante de la loi a posteriori du paramètre θ est définie par

$$q_\theta^{(t+1)}(\theta) \sim \mathcal{N}((I_p + V'V)^{-1}V'\mathbb{E}_{q_x^{(t)}}(x), (I_p + V'V)^{-1}).$$

L'estimateur bayésien variationnel obtenu $\tilde{\theta}$ converge vers la vraie valeur de θ avec probabilité 1, suivant les résultats de Wang et Titterington [21]. Mais la variance de la loi a posteriori approchée est inférieure à celle de la loi a posteriori exacte, la différence entre les deux étant une matrice définie négative.

Consonni et Marin illustrent ces résultats par des simulations dans le cas d'une covariable ($p = 2$) sur un échantillon de $n = 100$ observations : la convergence de l'estimateur variationnel $\tilde{\theta}$ est très nettement plus rapide que celle de l'estimateur de la loi exacte évalué par un schéma de Gibbs (facteur 200 dans le cas présenté), mais les approximations de $\tilde{\theta}$ peuvent être mauvaises et la précision très nettement mal estimée (facteur 10 à 50). Toutefois, on observe bien une amélioration de la qualité de l'estimation avec n.

3.4. Modèle markovien à données manquantes

Dans un cadre non bayésien (cf remarque 2.1), Hall, Humphreys et Titterington [11] comparent théoriquement et numériquement la forme de la surface de la vraisemblance exacte $l_0(y|\theta)$ à

 ISSN: 2102-6238

celle obtenue après approximation variationnelle $l(y|\theta)$ dans le cas des modèles markoviens comportant des observations manquantes. Il faut choisir la distribution libre q_x dans une classe de fonctions pour laquelle les calculs sont aisés : typiquement, la forme de $p(y,x)$ implique que le calcul de $l(y|\theta)$ est relativement simple si on applique l'approximation en champ moyen pour q_x, en l'écrivant sous forme d'un produit de termes $q_x(x) = \prod_i q_{x_i}(x_i)$ où x_i est la composante de x au site i. Mais il faut aussi que la borne inférieure $l(y|\theta)$ produite soit utilisable : ce n'est pas tant la distance entre $l_0(y|\theta)$ et $l(y|\theta)$ qui est intéressante, que la comparaison de la forme de ces deux surfaces : plus elles sont similaires (en mode, et en courbure), meilleure sera l'approximation du lieu du maximum de l'une par celui du l'autre.

Si les indices des observations manquantes sont suffisamment éloignés les uns des autres, alors les événements arrivant à ces sites sont conditionnellement indépendants, et

$$p_{x|y}(x|y) = p(x|y^x) = \prod_{i \in \mathscr{H}} p(x_i|y^i).$$

où $\mathscr{H}$ est l'ensemble des indices des observations cachées, et y^i le voisinage markovien du site i. Alors, on peut prendre $q_{x_i} = p(.|y^i)$ pour tout $i \in \mathscr{H}$ et $l_0(y|\theta) = l(y|\theta)$.
Sinon, si deux observations se trouvent dans un voisinage markovien, elles ne sont plus indépendantes. La densité jointe approchée est alors calculée en itérant les deux étapes suivantes :

1. estimer, pour chaque observation manquante au site i, les observations manquantes de son voisinage markovien (par moyenne sur les voisins en initialisation, en calculant la moyenne de la loi approchée a posteriori de l'étape 2 pour les itérations suivantes) ; on note $\hat{y}^{0i}$, l'ensemble des valeurs au voisinage de i, qu'elles soient observées ou estimées.
2. calculer la densité approchée :

$$\widehat{p}(x|y) = \prod_{i \in \mathscr{H}} p(x_i|\hat{y}^{0i})$$

En notant $\bar{p}(x|y)$ l'approximation finale de $\widehat{p}(x|y)$, et en prenant $q_x(x) = \bar{p}(x|y)$, on obtient une approximation en champ moyen de la vraisemblance, qui peut améliorer de façon substantielle le temps d'exécution en comparaison d'approches exactes comme l'algorithme EM .

D'un point de vue théorique, Hall, Humphreys et Titterington [11] montrent que pour un champ de Markov gaussien dont les sites sont les sommets d'une treillis régulier de $\mathbb{R}^d$, et pour lesquels les sites manquants forment t grappes séparées de s sites chacune et de même configuration spatiale, si s est fixe, mais $n,t \to \infty$, avec $t/n \to 0$, et sous quelques conditions de régularité, alors l'estimateur $\tilde{\theta}$ de θ maximisant la vraisemblance approchée par champ moyen a un biais additionnel à distance finie par rapport à l'estimateur du maximum de vraisemblance exacte θ_{MV} :

$$\theta_{MV} - \tilde{\theta} = O(\frac{t}{n}),$$

et ces deux estimateurs ont la même variance asymptotique.

Ces résultats sont illustrés par des simulations avec un processus AR(1) à observations manquantes, pour lequel la vraisemblance exacte est calculable : dans ce cas, la surface approchée est très similaire en forme à la surface exacte, même dans une situation non asymptotique.

 ISSN: 2102-6238

3.5. Modèles à espace d'états

Dans le cas du modèle précédent, une hypothèse forte pour l'obtention de la consistance asymptotique est que la proportion de sites non observés tende vers zéro quand n tend vers l'infini. Cependant, cette condition suffisante n'est pas satisfaite dans de nombreux cas.
Par exemple, Wang et Titterington [22] ont montré, pour les modèles à espace d'états, que l'approximation en champ moyen est asymptotiquement efficace quand les variances des variables de bruit sont suffisamment petites, mais que l'estimateur de la vraisemblance approchée en champ moyen ou l'estimateur bayésien variationnel ne sont pas toujours asymptotiquement consistants.

En effet, considérons le modèle à espace d'états suivant

$$\begin{aligned} X_{i+1} &= \theta X_i + \sigma W_i,\ X_1 \sim \mathcal{N}(\mu_0, \sigma^2) \\ Y_i &= \alpha X_i + \sigma V_i \end{aligned}$$

pour $i = 1, \ldots, n$, où $X_i \in \mathbb{R}$, $Y_i \in \mathbb{R}$ et où $\{V_i\}_n^{i=1}$ et $\{W_i\}_{i=1}^n$ sont des suites de variables aléatoires indépendantes gaussiennes centrées réduites. Les paramètres α, σ, μ_0 sont connus et θ est le paramètre à estimer, $|\theta| < 1$. Dans un premier temps, l'approximation en champ moyen est utilisée pour calculer l'estimateur du maximum de vraisemblance θ_{MV} : celui est approché par la valeur $\tilde{\theta}$ qui maximise, en θ, l'énergie libre

$$\mathcal{F}(q_x, \theta) = \int q_x(x) \log \frac{p(x, y|\theta)}{q_x(x)} dx.$$

L'énergie libre est calculée avec une loi factorisée $q_x(x) = \prod_{k=1}^n q_i(x_i)$. En supposant de plus que $q_i(x_i) \sim \mathcal{N}(\mu_i, \sigma_i)$ pour $i = 1, \ldots, n$, Wang et Titterington définissent les deux étapes de l'algorithme itératif. Comme la loi a posteriori exacte $p(x|y, \theta)$ est connue analytiquement, ils calculent alors la distance de Kullback $D(q_x(x)||p(x|y, \theta))$ et montrent que

- La distance de Kullback ne tend pas vers zéro, même si n tend vers l'infini
- La distance de Kullback est indépendante de σ : la loi a posteriori variationnelle ne devient pas la loi exacte, quelle que que soit la variance du bruit
- La distance de Kullback tend vers zéro si θ tend vers 0 : l'approximation en champ moyen donne alors une solution proche de la vraie loi si θ est petit
- La distance de Kullback ne tend pas vers zéro, quel que soit α.

Ces résultats montrent l'impossibilité d'estimer la vraisemblance par l'énergie libre, sauf si θ est petit. Cependant, si la variance du bruit σ tend vers zéro, alors Wang et Titterington prouvent que l'estimateur $\tilde{\theta}$ est consistant et asymptotiquement normal, même si la distance de Kullback ne tend pas vers zéro.
Cette propriété ne s'étend pas au cas où la variance du bruit ne tend pas vers zéro : dans cette situation, en effet, $\tilde{\theta}$ n'est pas consistant si $\theta \neq 0$.

De même, l'estimateur bayésien variationnel de θ n'est pas consistant.

 ISSN: 2102-6238

4. Deux exemples en neuroimagerie

La localisation des zones réactives du cerveau dépend du stimulus auquel est soumis un individu. En imagerie par résonance magnétique fonctionnelle (IRMf), cette réaction est mesurée par l'augmentation de l'oxygénation du sang dans la zone activée (signal BOLD), apportant le "carburant" nécessaire à la réaction. Il est ainsi possible de dresser des cartes d'activation cérébrale, enregistrant pour chaque volume élémentaire (*voxel*) l'intensité de la réponse BOLD constatée lors du stimulus. Le but de l'analyse des cartes d'activation est de détecter les zones activées lors d'un stimulus, de les comparer avec les zones anatomiques, et de confronter ces résultats pour une meilleure connaissance du cerveau. À terme, il est envisageable d'utiliser cette méthode pour prédire une pathologie donnée (voir le livre de Frackowiak et coauteurs [8] pour une introduction générale).

Le modèle linéaire est largement utilisé dans le traitement des données IRMf (par exemple, Worsley et coauteurs [28]). Mais le problème est délicat, et entre dans le cadre de la sélection de variables en grande dimension. En effet, on observe en général quelques dizaines d'images 3D (*scans*), alors qu'il y a des dizaines, voire des centaines de milliers de voxels. De nombreuses études font une sélection monovariée (voxel par voxel, en fonction d'un score d'Anova) des voxels significatifs. Les deux études présentées ici proposent des modèles de sélection alternatifs, nécessitant l'utilisation de l'approximation bayésienne variationnelle.

- Friston et coauteurs [9] utilisent un modèle bayésien hiérarchique dans lequel les voxels sont regroupés dans des sous-ensembles (motifs) successifs (peu nombreux) de poids croissants. Les loi a posteriori sont impossibles à calculer explicitement et l'estimateur bayésien variationnel est calculé en utilisant l'approximation en champ moyen et l'approximation de Laplace.
- Woolrich et Behrens [26] définissent un modèle de mélange spatial (mélange de lois représentant l'activation, la désactivation et la non activation), et approchent les labels discrets en chaque voxel par un vecteur de poids continus, ce qui permet de poser un a priori sous forme d'un champ de Markov gaussien. Ils montrent qu'il est alors possible de calculer adaptativement le taux de régularisation spatiale. Des simulations comparent les algorithmes MCMC et l'approximation variationnelle.

4.1. Modèle bayésien hiérarchique de définition de motifs

A partir de s images 3D (*scans*) de résonance magnétique fonctionnelle (IRMf) comportant n volumes élémentaires de mesure (*voxels*), Friston and al [9] proposent le schéma suivant de décodage des états du cerveau :

$$WX = RY\beta + \zeta$$

où X est un vecteur de dimension s représentant la mesure de l'état comportemental, perceptif ou cognitif pour chacun des s scans de l'expérience, Y est une matrice de taille $s \times n$ représentant la mesure du flux de la réponse hémodynamique en chaque voxel, W et R sont des matrices données et connues de dimension $s \times s$, β est un vecteur de dimension n représentant l'influence de chaque voxel dans la mesure de l'état comportemental, et ζ est un vecteur aléatoire corrélé de dimension

 ISSN: 2102-6238

s de covariance $\text{cov}(\zeta) = \Sigma^{\zeta}(\lambda) = \exp(\lambda^{\zeta})Q_0$.
Les paramètres à estimer sont ici β et λ^{ζ}. En général, il y a peu de scans (de l'ordre de la centaine), par rapport au nombre de voxels (plusieurs dizaines de milliers) et le problème est mal posé. Ainsi, l'estimation de β nécessite des contraintes ou des a priori, ce qui est traité en invoquant un second niveau dans le modèle. Soit U la matrice d'une fonction de base spatiale (U peut éventuellement être l'identité), de dimension $\mathbb{R}^{n\times u}$, et $\eta \in \mathbb{R}^u$ les poids (inconnus) des vecteurs de la base dans la définition de β :

$$\beta = U\eta.$$

La contrainte de parcimonie est définie sur la covariance de η, qui est posée diagonale et dont les coefficients de la diagonale ne peuvent prendre que m valeurs distinctes :

$$\text{cov}(\eta) = \Sigma^{\eta}(\lambda) = \exp(\lambda_1^{\eta})I^{(1)} + \ldots + \exp(\lambda_m^{\eta})I^{(m)}.$$

Les matrices $I^{(i)}$ sont des matrices diagonales permettant de coder l'appartenance du poids d'une colonne de U à un ensemble de poids de même variance, formant une suite emboîtée de sous-ensembles de poids $s^{(1)} \supset s^{(2)} \supset \ldots \supset s^{(m)}$ de plus en plus petits regroupant des éléments de variance de plus en plus grande. La variance du poids d'une colonne dans un sous-ensemble $s^{(i)}$ est toujours supérieure à celle du poids d'une colonne dans sur-ensemble $s^{(i-1)}$.

Cette modélisation permet de séparer la spécification de l'a priori spatial, codé par la base spatiale U, des variances codées comme un mélange de composantes de covariance dans $\Sigma^{\eta}(\lambda)$. Le modèle est donc défini par :

$$WX = L\eta + \zeta$$

où

- ζ est un bruit gaussien inconnu corrélé, de loi normale $p(\zeta/\lambda) \sim \mathcal{N}_s(0, \Sigma^{\zeta}(\lambda))$
- η est un effet aléatoire de loi normale $p(\eta/\lambda) \sim \mathcal{N}_u(0, \Sigma^{\eta}(\lambda))$
- $\lambda = (\lambda^{\zeta}, \lambda_1^{\eta}, \ldots, \lambda_m^{\eta})$ est le paramètre à estimer de dimension $m+1$, largement inférieure à s

Sous cette forme, les seules quantités inconnues sont les paramètres λ, contrôlant la covariance $\Sigma(\lambda)$:

$$\text{cov}(WX) = \Sigma(\lambda) = \Sigma^{\zeta}(\lambda) + L\Sigma^{\eta}(\lambda)L' = \exp(\lambda^{\zeta})Q_0 + \sum_{i=1}^{m} \exp(\lambda_i^{\eta})Q_i.$$

La loi a priori sur λ est choisie gaussienne $\mathcal{N}_{m+1}(\pi, \Pi^{-1})$, ses hyperparamètres sont pris tels que $\pi_1 = \ldots = \pi_{m+1} = -32$, et $\Pi = \frac{1}{256} I_{m+1}$, ce qui permet une loi a priori relativement peu informative (très grande variance) et de faible espérance pour $\exp(\lambda)$.

Ni le calcul de la loi a posteriori $p(\lambda|WX) = \frac{p(WX|\lambda)p(\lambda)}{p(WX)}$, ni le calcul de l'évidence $p(WX) = \int_{\lambda} p(WX, \lambda)d\lambda$ ne sont explicites à cause de la forme de $\Sigma(\lambda)$. De plus, on souhaite calculer une estimation des données cachées η afin de déterminer progressivement la suite des sous-ensembles. Pour une suite donnée de sous-ensembles $(I^{(1)}, \ldots, I^{(m)})$, on procède en deux étapes : d'abord l'estimation des paramètres λ, puis l'estimation les données manquantes η. Les résultats sont alors utilisés pour créer un nouveau sous-ensemble emboîté de plus grande variance, et le processus est itéré jusqu'à ce que l'énergie libre ne s'accroisse plus.

 ISSN: 2102-6238

4.1.1. Calcul de l'évidence

Le logarithme de l'évidence est la somme de l'énergie libre $\mathscr{F}(q_\lambda)$ et de la distance de Kullback entre la loi a posteriori $p(\lambda|WX)$ du paramètre et son approximation q_λ, d'après un calcul analogue à celui qui a conduit à (1) :

$$\begin{aligned}\log p(WX) &= \int \log\left(\frac{p(WX,\lambda)}{q_\lambda(\lambda)}\right) q_\lambda(\lambda)d\lambda + \int \log\left(\frac{q_\lambda(\lambda)}{p(\lambda|WX)}\right) q_\lambda(\lambda)d\lambda \\ &= \mathscr{F}(q_\lambda) + D(q_\lambda(\lambda)||p(\lambda|WX)).\end{aligned}$$

La log évidence est approchée par l'énergie libre maximisée fonctionnellement sur un ensemble des lois q_λ. Cependant, l'énergie libre ne peut être calculée ici quelle que soit la loi q_λ. L'utilisation de l'approximation de Laplace permet de résoudre cette difficulté, en approchant quadratiquement $\log p(WX,\lambda)$. Il est alors naturel de considérer q_λ gaussienne $q_\lambda \sim \mathscr{N}(\mu^\lambda, \Sigma^\lambda)$, amenant à l'expression suivante pour l'énergie libre (voir détails de calcul en annexe A.1) :

$$\begin{aligned}\mathscr{F}(q_\lambda) &= -\frac{1}{2}[(WX)'\Sigma(\mu^\lambda)^{-1}WX + \log|\Sigma(\mu^\lambda)| + w\log(2\pi) \\ &\quad - \log|\Sigma^\lambda\Pi| + (\mu^\lambda - \pi)'\Pi(\mu^\lambda - \pi)].\end{aligned}$$

L'expression écrite dans l'article de Friston comporte des coquilles de signe ; c'est pourquoi nous reprenons en annexe les détails du calcul, qui n'étaient pas explicités. Les deux premiers termes reflètent la précision du modèle, les deux derniers sa complexité, $w\log(2\pi)$ étant un terme constant. Cette expression dépend des paramètres (donnés) de la loi a priori π, Π, et de ceux la loi a posteriori μ^λ, Σ^λ, calculés par l'itération d'un schéma de Newton (*M-Step* de Friston). Soit L_λ et $L_{\lambda\lambda}$ le gradient et le Hessien de l'énergie libre, les étapes suivantes sont répétées jusqu'à la convergence (cf annexe A.2) :

$$\begin{aligned}P_i &= -\exp(\mu_i^\lambda)\Sigma(\mu^\lambda)^{-1}Q_i\Sigma(\mu^\lambda)^{-1} \\ L_{\lambda_i} &= -\frac{1}{2}[\mathrm{trace}(P_i(WXX'W' - \Sigma(\mu^\lambda))] - \Pi_{i.}(\mu^\lambda - \pi) \\ L_{\lambda\lambda ij} &= -\frac{1}{2}\mathrm{trace}(P_i\Sigma(\mu^\lambda)P_j\Sigma(\mu^\lambda)) - \Pi_{ij} \\ \Delta\mu^\lambda &= -L_{\lambda\lambda}^{-1}L_\lambda \\ \Sigma^\lambda &= -L_{\lambda\lambda}^{-1}\end{aligned}$$

où $\Pi_{i.}$ est la i-ème ligne de Π.

4.1.2. Estimation de η

Étant données les espérances conditionnelles des hyper-paramètres obtenus à l'étape précédente, il est maintenant possible d'obtenir analytiquement les moyennes a posteriori des états cachés η (*E-step* de Friston). L'énergie libre est maximum pour la loi jointe a posteriori $P(\eta,\lambda|WX)$. L'*hypothèse de champ moyen* permet de chercher une approximation factorisée en η et λ de la loi a posteriori :

$$P(\eta,\lambda|WX) \simeq q_\lambda(\lambda)q_\eta(\eta)$$

 ISSN: 2102-6238

et la solution optimum en q_η est

$$q_\eta(\eta) = \frac{1}{K_\eta} \exp \int d\lambda q_\lambda(\lambda) \log P(WX, \eta|\lambda).$$

En utilisant l'*approximation de Laplace*, on cherche q_η sous forme d'une gaussienne $\mathcal{N}(\mu^\eta, \Sigma^\eta)$ dont on calcule explicitement l'espérance et la variance (cf. annexe A.3) :

$$\begin{aligned} \Sigma^\eta &= \Sigma^\eta(\mu^\lambda) - \Sigma^\eta(\mu^\lambda) L' \Sigma^\eta(\mu^\lambda)^{-1} L \Sigma^\eta(\mu^\lambda) \\ \mu^\eta &= \Sigma^\eta L' \Sigma^\zeta(\mu^\lambda)^{-1} WX \end{aligned}$$

S'en déduisent immédiatement l'estimation des paramètres de l'approximation de la loi a posteriori de $\beta = U\eta$

$$\begin{aligned} \Sigma^\beta &= U \Sigma^\eta U' \\ \mu^\beta &= U \mu^\eta \end{aligned}$$

On résout ainsi facilement le problème d'inférence dans un modèle mal conditionné. Il reste cependant une difficulté à lever : le choix de la partition codée par les matrices diagonales $I^{(i)}$

4.1.3. Recherche des sous-ensembles de poids

La solution proposée est de faire une recherche progressive ascendante. Partant de l'hypothèse d'égalité de variance $I^1 = I$ pour tous les poids, les espérances conditionnelles des poids sont utilisées pour créer un sous-ensemble avec les plus hauts poids (en prenant les poids du sous-espace précédent de $|\mu^\eta|$ supérieurs à la médiane par exemple). Puis l'EM est ré-exécuté, et le sous-ensemble de plus haut poids est à nouveau scindé. La procédure est répétée jusqu'à ce que la log-évidence cesse de croître. L'algorithme peut ne converger que vers un maximum local.

4.1.4. Résultats

Friston et al illustrent d'abord leur méthode sur un jeu de données simulées. Ils montrent qu'elle permet de distinguer un modèle nul (c'est à dire sans motif ni influence quelconque de voxels), d'un modèle défini avec un codage parcimonieux. Cependant, seuls les voxels ayant un poids important sont bien récupérés. Et il faudrait sans doute étudier l'influence du signal sur bruit sur ces résultats.

Dans le cas d'un jeu de données réelles (étude de l'attention dans un mouvement visuel), la distribution des poids pour le modèle optimum affirme bien leur parcimonie. Cependant, le graphique de l'évidence en fonction du nombre de partitions présente une croissance jusqu'à 4 partitions, puis une décroissance, ce qui est en contradiction avec le fait que l'énergie libre ne peut qu'augmenter ou se stabiliser avec des composants supplémentaires (cf 4.1.3). Ceci est vraisemblablement du au fait que l'algorithme a convergé vers un minimum local quand le nombre de partitions devient important. Enfin, la méthode est utilisée pour comparer des modèles avec des codages spatiaux différents ou des régions d'intérêt différentes. Les résultats sont positifs, mais effectués sur des régions très différentes et demanderaient à être précisés.

 ISSN: 2102-6238

4.2. Modèle de mélange avec régularisation spatiale

Les modèles de mélanges spatiaux sont utilisés dans le traitement d'image pour effectuer une régularisation spatiale des observations. Ceci permet de prendre en compte la conviction que des voxels voisins dans une image sont susceptibles de provenir de la même classe. Dans le cas de l'imagerie par résonance magnétique fonctionnelle, on peut penser que des voxels voisins réagissent de façon identique aux stimuli de l'expérience (ie, sont dans le même état : activé, non-activé, désactivé par exemple).

Un a priori sous forme de champ de Markov discret pénalise simplement les voxels voisins de classes différentes. Mais le niveau de pénalisation dépend d'un paramètre de contrôle de granularité du champ, qu'il n'est pas possible de calculer analytiquement. Woolrich, Behrens, Beckmann et Smith [27] proposent d'approcher les labels discrets par un vecteur de poids continus. Les propriétés imposées sur le vecteur de poids assurent que la nouvelle loi a posteriori est la même que celle calculée avec les labels discrets. La loi a priori sous forme de champ de Markov discret sur les labels est remplacée par un champ de Markov gaussien continu sur les paramètres afférents aux poids continus, pour lequel il est possible de calculer la constante de normalisation.

On considère des observations Y sur une grille spatiale régulière, où y_i est l'observation au voxel i, et $i = 1,\dots,N$. Il y a K classes dans le mélange (par exemple, $K = 3$ pour les états activé, non-activé, désactivé). Les paramètres des lois des classes sont représentés par le vecteur $\theta = \{\theta_k : k = 1,\dots K\}$. Les labels des classes sont codés par X, où x_i est le label au lieu i. En supposant l'indépendance conditionnelle de la vraisemblance sachant les labels, la loi a posteriori complète conditionnellement aux observations est

$$p(x = \kappa, \theta, \phi_x | y) \propto \{\prod_i^n p(y_i | x_i = \kappa_i, \theta_{k_i})\} \times p(x = \kappa | \phi_x) p(\phi_x) p(\theta)$$

où κ est une configuration spécifique de la classe de labels x et la loi a priori sur x est un champ de Markov discret :

$$p(x = \kappa | \phi_x) \propto f(\phi_x) \exp(-\frac{\phi_x}{4} \sum_i \sum_{j \in \mathcal{N}_i} \mathbb{I}[x_i \neq x_j])$$

où $\mathcal{N}_i$ est le voisinage spatial de i, ϕ_x le paramètre de contrôle et $f(\phi_x)$ est la fonction (de partition) que l'on ne sait pas calculer.

La loi a posteriori est approchée en replaçant les labels discrets x_i par des vecteurs de poids continus $w_i = \{w_{ik} : k = 1,\dots,K\}$, $w = \{w_i : i = 1,\dots,n\}$, de loi

$$p(w, \theta, \phi_{\tilde{w}} | y) \propto \prod_i^n \{\sum_{k=1}^K w_{ik} p(y_i | x_i = k, \theta_k)\} \times p(w | \phi_{\tilde{w}}) p(\phi_{\tilde{w}}) p(\theta). \quad (6)$$

L'utilisation d'une transformation déterministe *logit* permet d'assurer les contraintes sur w_{ik} :

$$w_{ik} = \frac{\exp(\tilde{w}_{ik}/\gamma)}{\sum_{l=1}^K \exp(\tilde{w}_{il}/\gamma)}.$$

γ est choisi pour permettre l'approximation sous forme de Dirac ($\gamma = 0.5$). La loi a priori sur $\tilde{w}$, est un processus auto-régressif gaussien

$$p(\tilde{w} | \phi_{\tilde{w}}) = \prod_k p(\tilde{w}_k | \phi_{\tilde{w}}) \sim \mathcal{N}_n(0, (I - C)^{-1} M),$$

 ISSN: 2102-6238

où $M = (1/\phi_{\tilde{w}})I_n$, et C définit la dépendance spatiale.

Le bénéfice de cette modélisation est le calcul analytique de la constante de normalisation de

$$f(\phi_{\tilde{w}}) = (2\pi)^{-nK/2}\phi_{\tilde{w}}^{nK/2}|I - C|^{K/2} \propto \phi_{\tilde{w}}^{nK/2},$$

permettant la détermination adaptative du paramètre de contrôle $\phi_{\tilde{w}}$. La loi a priori sur $\phi_{\tilde{w}}$ est une loi gamma conjuguée

$$\phi_{\tilde{w}}|b_{\tilde{w}_0}, c_{\tilde{w}_0} \sim Ga(b_{\tilde{w}_0}, c_{\tilde{w}_0})$$

dont on choisit les paramètres $b_{\tilde{w}_0} = 1000$, $c_{\tilde{w}_0} = 0.001$, pour rendre la loi faiblement informative (espérance 1, variance 1000). Dans le cas de trois classes d'activation, les vraisemblances $p(y_i|x_i = k, \theta_k)$ sont choisies de la façon suivante :

- loi gaussienne pour une classe non-activée : $y_i|x_i = k_n, \theta_{k_n} \sim \mathcal{N}(\mu_{k_n}, \sigma^2_{k_n})$,
- loi gamma pour une classe activée : $y_i|x_i = k_a, \theta_{k_a} \sim Ga(a_{k_a}, b_{k_a})$,
- lois gamma "négative" pour une classe désactivée : $y_i|x_i = k_d, \theta_{k_d} \sim Ga((-y_i), a_{k_d}, b_{k_d})$.

Notons que le modèle initial à labels discrets a été remplacé par un modèle à labels continus, plus facile à approcher. Ce choix de modélisation est indépendant de l'approximation variationnelle, ou de l'approximation de Laplace. Il va permettre en revanche un bon compromis entre la qualité de l'analyse et la facilité de calcul.

4.2.1. Inférence

Le modèle ci-dessus peut être estimé en utilisant un algorithme MCMC, mais les temps d'exécution sont parfois prohibitifs. Woolrich et Behrens [26] en ont adapté une version bayésienne variationnelle.

Si θ est considéré comme connu, la loi a posteriori (6) s'écrit

$$p(\tilde{w}, \phi_{\tilde{w}}|y) \propto p(\tilde{w}|\phi_{\tilde{w}})p(\phi_{\tilde{w}}) \prod_i \left\{ \sum_{k=1}^{K} \frac{\exp(\tilde{w}_{ik})}{1 + \exp(\tilde{w}_{ik})} p(y_i|x_i = k, \theta_k) \right\}.$$

On en cherche une approximation (en champ moyen) de la forme

$$q(\phi_{\tilde{w}}, \tilde{w}) = q_{\phi_{\tilde{w}}}(\phi_{\tilde{w}}) \prod_i q_{\tilde{w}_i}(\tilde{w}_i),$$

avec $q_{\phi_{\tilde{w}}}(\phi_{\tilde{w}}|y) \sim Ga(b_{\tilde{w}}, c_{\tilde{w}})$ et $q(\tilde{w}_i|y) \sim \mathcal{N}_k(m_i, \Lambda_i^{-1})$. L'approximation en champ moyen a permis la factorisation par voxel, alors que la loi exacte est un mélange compliqué. De plus, considérer l'indépendance de la structure de covariance de la loi posteriori permet de remplacer l'estimation de K matrices de covariance de dimension $n \times n$ par n matrices de covariance de dimension $K \times K$. Mais cette invariance de la structure de covariance de la loi a posteriori n'empêche pas les espérances a posteriori m_i des poids de dépendre de l'information spatiale du champ de Markov.

La loi maximisant l'énergie libre est calculée itérativement en effectuant successivement les deux étapes de mise à jour, par application du théorème 1 et de la remarque 1 :

$$\begin{aligned} q_{\phi_{\tilde{w}}}(\phi_{\tilde{w}}|y) &\propto \exp(\mathbb{E}_{q_{\phi_{\tilde{w}}}}[\log p(\tilde{w}, \phi_{\tilde{w}}|y)]) \\ q_w(\tilde{w}_i|y) &\propto \exp(\mathbb{E}_{q_{\tilde{w}-i}q_{\phi_{\tilde{w}}}}[\log p(\tilde{w}, \phi_{\tilde{w}}|y)]) \end{aligned}$$

 ISSN: 2102-6238

où $\mathbb{E}_{q_{\tilde{w}-i}q_{\phi_{\tilde{w}}}}$ indique que l'intégration est faite en $\phi_{\tilde{w}}$ et sur tous les $\tilde{w}_j$ sauf le i-ème. D'habitude, les espérances sont calculables en choisissant des lois a priori conjuguées de la vraisemblance ; mais ici, le modèle a des composantes non linéaires complexes contenues dans $p(y|\tilde{w})$, et il n'y a pas d'a priori conjugué disponible. On procède alors au développement de Taylor au second ordre de ce terme autour du mode courant des poids de classification m_i (approximation de Laplace) :

$$\begin{aligned} -\log(p(y_i|\tilde{w}_i)) &= f(\tilde{w}_i) \\ &\simeq f(m_i) + (\tilde{w}_i - mi)' h_i + \frac{1}{2}(\tilde{w}_i - mi)' \Sigma_i (\tilde{w}_i - mi)' \end{aligned}$$

où h_i et Σ_i sont les dérivées premières et secondes de f par rapport à $\tilde{w}_i$ prises en m_i. Tous les calculs sont accessibles après cette approximation.

Il est possible d'inclure la loi des paramètres θ dans la mise à jour en incorporant un facteur $q_\theta(\theta)$ dans l'approximation de la loi a posteriori. Mais il n'est pas besoin de marginaliser pour trouver le MAP de θ, comme c'était le cas pour $\phi_{\tilde{w}}$: une simple estimation ponctuelle $\widehat{\theta}$ de θ est suffisante, par exemple en choisissant $\widehat{\theta}$ maximisant

$$p(\theta) \prod_i \sum_{k=1}^{K} \frac{\exp(m_{ik})}{1+\exp(m_{ik})} p(y_i|x_i = k, \theta_k).$$

4.2.2. Résultats

Woolrich et Behrens construisent des cartes d'activation spatiales à partir des moyennes a posteriori de w_{ik}, ie $\mathbb{E}_{w_{ik}}(w_{ik})$ pour les classes activées et désactivées, seuillées à $0,5$. Leurs résultats sur des données simulées montrent clairement le besoin d'estimer adaptativement le paramètre de régularisation spatiale $\phi_{\tilde{w}}$. Les différences entre la méthode variationnelle bayésienne et l'inférence par MCMC sont faibles, avec un léger avantage pour les méthodes MCMC en ce qui concerne l'erreur de classement, et une réelle amélioration de temps de calcul pour la méthode VB (d'un rapport 15 pour un volume de 10 000 voxels). Ces résultats s'observent également sur un jeu de données réelles qu'ils ont testé.

5. Discussion

L'approximation variationnelle est une méthode séduisante permettant de remplacer une intégration coûteuse et souvent impossible à mener par l'optimisation d'une fonctionnelle sur un ensemble de fonctions pour lesquelles les calculs sont aisés : c'est le cas en particulier pour les modèles comportant des variables cachées. En fonction de l'étude, l'intérêt de la qualité de l'approximation se porte sur le lieu du maximum ou sur la valeur du maximum :

1. Si la quantité d'intérêt est le lieu du maximum (par exemple, recherche de l'estimateur du Maximum A Posteriori), l'approximation est de bonne qualité si les deux surfaces exacte et approchée ont la même forme (mode, courbure) : dans le cas de modèle de mélange de lois de la famille exponentielle, l'estimateur variationnel est asymptotiquement consistant, mais la variance de la loi limite est inférieure à celle attendue : les intervalles

 ISSN: 2102-6238

de confiance calculés avec des lois variationnelles a posteriori seront trop étroits. Pour le modèle markovien avec peu de données manquantes, l'estimateur est consistant et de même variance que l'estimateur du maximum de vraisemblance ; en revanche, il n'y pas consistance dans le modèle à espace d'états.

2. Si la quantité d'intérêt est la valeur du maximum (au titre de la comparaison de modèle par exemple), il faut pouvoir quantifier la différence entre l'énergie libre et l'évidence sous peine de biaiser le résultat. Par exemple, le logarithme du facteur de Bayes s'écrit

$$\log \frac{p(y|m)}{p(y|m')} = F - F' + d - d'$$

où $d = D(q_{x,\theta}||p(\theta|y,m))$ (resp. $d' = D(q'_{x,\theta}||p(\theta|y,m'))$) sont les distances de Kullback résiduelles dans le premier (resp. second) modèle après optimisation. On ne peut utiliser la différence des énergies libres $F - F'$ en sélection de modèle que si on suppose que la différence des distances de Kullback $d - d'$ ne change pas le signe de l'expression $F - F' + d - d'$. À distance finie, même pour des exemples très simples (modèle de mélange de deux lois connues), la loi approchée est différente de la loi exacte. Dans le modèle à espace d'états, même asymptotiquement, la distance de Kullback ne tend pas vers zéro. Ainsi, baser la comparaison de modèle sur la différence des énergies libres est risqué, et dépend du comportement de $d - d'$, qu'on ne sait en général pas calculer.

L'approximation en champ moyen n'est souvent pas suffisante pour simplifier le problème en neuroimagerie. La résolution cumule dans ce cas l'approximation en champ moyen (pour factoriser la loi a posteriori) et l'approximation de Laplace (pour en calculer les facteurs). Ces deux approximations peuvent bien sûr affecter l'estimation de l'évidence, et il faudrait étudier théoriquement leur influence.

Les simulations de Woolrich et Behrens sur un modèle de mélange de régularisation spatiale montrent peu de différence entre la méthode MCMC et la méthode variationnelle, ce qui est encourageant. De plus la méthode variationnelle est beaucoup plus rapide d'exécution ; mais il est possible que l'estimation en résultant souffre des inconvénients énoncés ci-dessus (en particulier, une variance trop petite).

Friston et coauteurs [10] utilisent l'énergie libre pour sélectionner un nombre de composantes de variance, la recherche s'arrêtant quand l'énergie libre n'augmente plus à l'ajout d'une composante : la solution est donc théoriquement sensible à la variation de la proximité entre l'énergie libre et l'évidence suivant les modèles. Pour valider leur approche, ils estiment l'évidence par MCMC (échantillonnage de Metropolis-Hasting et identité de la moyenne harmonique) d'une part, et la calculent par approximation variationnelle d'autre part, dans des modèles ayant de plus en plus de composantes de variance. Ils montrent alors que l'évidence estimée par MCMC et celle calculée par approximation variationnelle varient dans les mêmes proportions en fonction du nombre de composantes, donnant, pour cet exemple, l'avantage à l'approximation variationnelle, qui est beaucoup plus rapide d'exécution.

Malgré des résultats théoriques parfois peu engageants, les simulations montrent des résultats intéressants pour les méthodes variationnelles. Cette étude renforce la conclusion qu'il faut pousser plus loin les investigations théoriques et les tests afin de pouvoir préciser le cadre d'utilisation des méthodes variationnelles et de les déployer comme un outil de calcul fiable de la statistique bayésienne.

 ISSN: 2102-6238

Annexe A: Détails de calcul

Les calculs n'étant pas détaillés dans Friston et al[9], nous les reprenons ici, d'autant plus que la formule de leur article comporte des erreurs de signes. On rappelle le modèle

$$\begin{aligned} WX &= RY\beta + \zeta \\ &= L\eta + \zeta \\ \beta &= U\eta \end{aligned}$$

Les lois a priori sur η et ζ sont gaussiennes, centrées et de variance respective $\Sigma^{\eta}(\lambda)$ et $\Sigma^{\zeta}(\lambda)$. $\Sigma^{\eta}(\lambda)$ est la somme de m matrices diagonales :

$$\Sigma^{\eta}(\lambda) = \sum_{i=1}^{m} \exp(\lambda_i^{\eta}) I^{(i)}.$$

La loi de $WX|\lambda$ est gaussienne, centrée, et de covariance

$$\begin{aligned} \operatorname{cov}(WX) = \Sigma(\lambda) &= \Sigma^{\zeta}(\lambda) + L\Sigma^{\eta}(\lambda)L' \\ &= \exp(\lambda^{\zeta})Q_0 + \sum_{i=1}^{m} \exp(\lambda_i^{\eta})Q_i \\ &= \sum_{i=0}^{m} \exp(\lambda_i)Q_i \end{aligned} \tag{7}$$

A.1. Calcul de l'énergie libre $\mathscr{F}(q_\lambda)$

Notons $\mathbb{E}_{q_\lambda}$ l'espérance sous la loi q_λ. L'énergie libre à maximiser est

$$\mathscr{F}(q_\lambda) = \mathbb{E}_{q_\lambda}(\log p(WX,\lambda)) - \mathbb{E}_{q_\lambda}(q_\lambda).$$

L'approximation de Laplace permet d'approcher quadratiquement la log-vraisemblance complète $\log p(WX,\lambda)$ autour de son maximum μ^{λ} :

$$\log P(WX,\lambda) \simeq \log p(WX,\mu^{\lambda}) - \frac{1}{2}(\lambda-\mu^{\lambda})' H(\mu^{\lambda})(\lambda-\mu^{\lambda})$$

où $H(\mu^{\lambda}) = -\partial^2_{\lambda=\mu^{\lambda}} P(WX,\lambda)$ est l'opposé du hessien évalué en μ^{λ}.
Comme $p(\lambda|WX) \propto p(WX,\lambda)$, il est naturel de maximiser $\mathscr{F}$ en q_λ sur l'ensemble des gaussiennes $\mathscr{N}(\mu^{\lambda},\Sigma^{\lambda})$. Alors, si q_λ, est gaussienne, on a

$$E_{q_\lambda}(\log q_\lambda) = -\frac{1}{2}\left((m+1)\log(2\pi) + \log|\Sigma^{\lambda}| + m + 1\right).$$

D'où

$$E_{q_\lambda}((\lambda-\mu^{\lambda})' H(\mu^{\lambda})(\lambda-\mu^{\lambda})) = \operatorname{trace}(H(\mu^{\lambda})\Sigma^{\lambda}),$$

 ISSN: 2102-6238

et

$$\begin{aligned}\mathscr{F}(q_\lambda) &= \log P(WX,\mu^\lambda) - \frac{1}{2}\operatorname{trace}(H(\mu^\lambda)\Sigma^\lambda) \\ &\quad + \frac{1}{2}\left((m+1)\log(2\pi) + \log|\Sigma^\lambda| + m + 1\right).\end{aligned}$$

En dérivant l'énergie libre par rapport à Σ^λ, on obtient la forme de la covariance conditionnelle approchée

$$\frac{\partial \mathscr{F}(q_\lambda)}{\partial \Sigma^\lambda} = -(\Sigma^\lambda)^{-1} + H(\mu^\lambda) = 0 \Leftrightarrow \Sigma^\lambda = H(\mu^\lambda)^{-1}$$

d'où $\operatorname{trace}(\Sigma^\lambda H(\mu^\lambda)) = m+1$. L'énergie libre se simplifie :

$$\mathscr{F}(q_\lambda) = \log P(WX,\mu^\lambda) + \frac{1}{2}\left((m+1)\log(2\pi) + \log|\Sigma^\lambda|\right).$$

Enfin, soit w le rang de W, la loi de $WX|\lambda$ est gaussienne centrée :

$$\log p(WX|\mu^\lambda) = -\frac{1}{2}\left(\log|\Sigma(\mu^\lambda)| + w\log(2\pi) + (WX)'\Sigma(\mu^\lambda)^{-1}WX\right),$$

tandis que la loi a priori de λ est gaussienne $\mathscr{N}(\pi,\Pi)$:

$$\log p(\mu^\lambda) = -\frac{1}{2}\left(\log|\Pi^{-1})| + (m+1)\log(2\pi) + (\mu^\lambda - \pi)'\Pi(\mu^\lambda - \pi)\right).$$

D'où l'approximation de la log-évidence $\log p(WX)$ par l'énergie libre calculée de la façon suivante :

$$\begin{aligned}\mathscr{F}(q_\lambda) &= -\frac{1}{2}[(WX)'\Sigma(\mu^\lambda)^{-1}WX + \log|\Sigma(\mu^\lambda)| + w\log(2\pi) \\ &\quad - \log|\Sigma^\lambda\Pi| + (\mu^\lambda - \pi)'\Pi(\mu^\lambda - \pi)].\end{aligned}$$

A.2. Éléments de calcul des étapes du M-step

Grâce à la forme particulière (7) de la covariance $\Sigma(\lambda)$, on a :

$$\frac{\partial\Sigma(\lambda)}{\partial\lambda_i} = \exp(\lambda_i)Q_i, \tag{8}$$

d'où, la forme de la dérivée partielle de la matrice de précision $\Sigma(\lambda)^{-1}$

$$\begin{aligned}P_i = \frac{\partial}{\partial\lambda_i}\Sigma(\lambda)^{-1} &= -\Sigma(\lambda)^{-1}\frac{\partial\Sigma(\lambda)}{\partial\lambda_i}\Sigma(\lambda)^{-1} \\ &= -\exp(\lambda_i)\Sigma(\lambda)^{-1}Q_i\Sigma(\lambda)^{-1},\end{aligned} \tag{9}$$

et

$$\begin{aligned}\frac{\partial}{\partial\lambda_i}((WX)'\Sigma(\lambda)^{-1}WX) &= (WX)'P_iWX \\ &= \operatorname{trace}[(WX)'P_iWX] \\ &= \operatorname{trace}[P_iWX(WX)'].\end{aligned}$$

 ISSN: 2102-6238

On en déduit l'expression suivante de la dérivée de l'énergie libre $\mathscr{F}(q_\lambda)$, $\Pi_{i.}$ étant la i-ème ligne de Π.

$$L_{\lambda_i} = -\frac{1}{2}[\text{trace}(P_i(WXX'W' - \Sigma(\lambda))] - \Pi_{i.}(\lambda - \pi) :$$

La dérivée seconde s'obtient grâce à quelques manipulations élémentaires de matrices :

$$\begin{aligned} L_{\lambda\lambda ij} &= -\frac{1}{2}\,\text{trace}\,(P_i(-\exp(\lambda_j)Q_j)) - \Pi_{ij} \\ &= -\frac{1}{2}\,\text{trace}\,\big(P_i\Sigma(\lambda)(-\exp(\lambda_j))\Sigma(\lambda)^{-1}Q_j\Sigma(\lambda)^{-1}\Sigma(\lambda)\big) - \Pi_{ij} \\ &= -\frac{1}{2}\,\text{trace}(P_i\Sigma(\lambda)P_j\Sigma(\lambda)) - \Pi_{ij}. \end{aligned}$$

A.3. Éléments de calcul de l'étape E-step

L'utilisation de l'approximation de Laplace permet d'écrire

$$\log P(WX,\eta,\lambda) = \log P(WX,\mu^\eta,\mu^\lambda) - \frac{1}{2}\begin{pmatrix} (\lambda-\mu^\lambda)' & (\eta-\mu^\eta)' \end{pmatrix} H \begin{pmatrix} \lambda-\mu^\lambda \\ \eta-\mu^\eta \end{pmatrix},$$

où $H = H(\mu^\eta,\mu^\lambda)$ est l'opposée de la matrice des dérivées secondes de $\log P(WX,\eta,\lambda)$ calculée au maximum (μ^η,μ^λ). Or,

$$q_\eta(\eta) \propto \exp \int d\lambda q_\lambda(\lambda) \log P(WX,\eta,\lambda),$$

ce qui revient à prendre pour q_η une gaussienne $\mathscr{N}(\mu^\eta,\Sigma^\eta)$

$$q_\eta(\eta) \propto \exp -\frac{1}{2}(\eta-\mu^\eta)'(\Sigma^\eta)^{-1}(\eta-\mu^\eta),$$

avec

$$\Sigma^\eta = -\left(\frac{\partial^2}{\partial\eta^2}\log(WX,\eta,\mu^\lambda)_{|\eta=\mu^\eta}\right)^{-1}.$$

La dérivée du logarithme de la vraisemblance complète amène directement à

$$\mu^\eta = \Sigma^\eta L'\Sigma^\zeta(\mu^\lambda)^{-1}WX$$

et la variance a posteriori Σ^η s'écrit, en utilisant le lemme d'inversion matricielle de Woodbury (annexe A d'Anderson p. 638 [1]) pour supprimer les matrices de grande taille :

$$\begin{aligned} \Sigma^\eta &= \left(L'\Sigma^\zeta(\mu^\lambda)^{-1}L + \Sigma^\eta(\mu^\lambda)^{-1}\right)^{-1} \\ &= \Sigma^\eta(\mu^\lambda) - \Sigma^\eta(\mu^\lambda)L'(\Sigma^\zeta(\mu^\lambda) + L\Sigma^\eta(\mu^\lambda)L')^{-1}L\Sigma^\eta(\mu^\lambda) \\ &= \Sigma^\eta(\mu^\lambda) - \Sigma^\eta(\mu^\lambda)L'\Sigma(\mu^\lambda)^{-1}L\Sigma^\eta(\mu^\lambda). \end{aligned}$$

 ISSN: 2102-6238

Références

[1] T.W. ANDERSON : *An introduction to multivariate Statistical Analysis*. John Wiley and Sons, 2003.

[2] H. ATIAS : Inferring parameters ans structure of latent variable models by variational bayes. *In Proc. 15th Conference of Uncertainty in Artificial Intelligence*, pages 21–30, 1999.

[3] M. J. BEAL : *Variational algorithms for approximate bayesian inference*. Thèse de doctorat, University College London, 2003.

[4] M. J. BEAL et Z. GHAHRAMANI : Variational bayesian learning of directed graphical models with hidden variables. *Bayesian Analysis*, 1(4):793–832, 2004.

[5] C. M. BISHOP : *Pattern Recognition and Machine Learning*. Springer, 2008.

[6] G. CONSONNI et J-M. MARIN : Mean-field variational approximate bayesian inference for latent variable models. *Computational Statistics & Data Analysis*, 52:790–798, 2007.

[7] A.P. DEMPSTER, N.M. LAIRD et D.B. RUBIN : Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[8] R. S. J. FRACKOWIAK, K.J. FRISTON, C. FRITH, R. DOLAN, C. PRICE, J. ASHBURNER, W. PENNY et S. ZEKI : *Human Brain Function*. Academic Press, second édition, 2003.

[9] K. FRISTON, C. CHU, J Mourão MIRANDA, O. HULME, G. REES, W. PENNY et J. ASHBURNER : Bayesian decoding of brain images. *NeuroImage*, 39:181–205, 2008.

[10] K. FRISTON, J. MATTOUT, N. Trujillo BARRETO, J. ASHBURNER et Penny W. : Variational free energy and the laplace approximation. *NeuroImage*, 34:220–234, 2006.

[11] P. HALL, K. HUMPHREYS et D. M. TITTERINGTON : On the adequacy of variational lower bound functions for likelihood-based inference in markovian models with missing values. *J. R. Statist. Soc B*, 64(3):549–564, 2002.

[12] K. HUMPHREYS et D. M. TITTERINGTON : Approximate bayesian inference for simple mixtures. *In Proc. Computational Satistics*, pages 331–336. Physica-Verlag, 2000.

[13] M. I. JORDAN : Graphical models. *Statistical Science*, 19(1):140–155, 2004.

[14] K. LEE, K.L. MENGERSEN, J.-M. MARIN et C.P ROBERT : Bayesian inference on mixtures of distributions. *In Proceedings of the Platinum Jubilee of the Indian Statistical Institute*, 2008.

[15] T. MINKA : Expectation propagation and approximate bayesian inference. *In Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence*, pages 362–369, San Francisco, CA, 2001. Morgan Kaufmann.

[16] C.P. ROBERT : *Le choix bayésien : Principes et pratiques*. Springer, 2006.

[17] H. RUE, S. MARTINO et N. CHOPIN : Approximate bayesian inference for latent gaussian models by using intergated nested laplace approximations. *J. R. Statist. Soc. B*, 71(2):1–35, 2009.

[18] L. TIERNEY et J. B. KADANE : Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, 81:82–86, 1986.

[19] D. M. TITTERINGTON : Bayesian methods for neural networks and related methods. *Statistical Science*, 19(1):128–139, 2004.

[20] M. J. WAINWRIGHT et M. I. JORDAN : Graphical models, exponential families, and variational inference. *Foundations and trends in Machine Learning*, 1(1-2):1–305, 2008.

[21] B. WANG et D. M. TITTERINGTON : Convergence and asymptotique normality of vatiational bayesian approximations for exponential familty models with missing values. *In Proceeding of the twentieth conference on uncertainty and artificial interlligence*, pages 577–584, Banff, Canada, 2004. AUAJ Press.

[22] B. WANG et D. M. TITTERINGTON : Lack of consistency of mean field and variational bayes approximations for state space models. *Neural Proceesing Letters*, 20(3):151–170, 2004.

[23] B. WANG et D. M. TITTERINGTON : Inadequacy of interval estimates corresponding to variational bayesian approximations. *In Proc. 10th. Int. Workshop Artific. Intell. Statist.*, pages 373–380, 2005.

[24] B. WANG et D. M. TITTERINGTON : Variational bayes estimation of mixing coefficients. *In Deterministic and statistical methods in machine learning*, volume 3635, pages 281–295, 2005.

[25] B. WANG et D. M. TITTERINGTON : Convergence properties of a general algorithm for calculating variational bayesian estimates for a normal mixture model. *Bayesian Analysis*, 1(3):625–650, 2006.

[26] M. W. WOOLRICH et BEHRENS : Variational bayes inference for spatial mixture models for segmentation. *In IEEE Trans. Med. Imag.*, 2006.

 ISSN: 2102-6238

[27] M. W. WOOLRICH, T. E. BEHRENS, C. F. BECKMANN et S. M. SMITH : Mixture models with adaptative spatial regularisation for segmentation with application to fmri data. *In IEEE Trans. Med. Imag.*, 2005.

[28] K.J. WORSLEY, C.H. LIAO, J. ASTON, V. PETRE, G.H. DUNCAN, F. MORALES et A.C. EVANS : A general statistical analysis for fmri data. *Neuroimage*, 15(1):1–15, 2002.

 ISSN: 2102-6238