Seconde Étape dans les Modifications de MixtComp

Vincent KUBICKI

10 août 2017

Table des matières

Résumé

Ce document est le troisième document décrivant les évolutions proposées dans MixtComp. Les deux documents précédents présentent une feuille de route [1] et le gros des modifications effectuées [2].

Le présent document revient sur des problèmes évoqués dans [2] et développe les problématiques qui y sont soulevées.

1 Différences dans les méthodes d'initialisation

Dans le précédent document [2], on avait mis en évidence un écart inattendu entre la vraisemblance complétée et la vraisemblance observée. Cet écart n'est visible qu'avec la nouvelle initialisation (utilisant une observation par classe), et pas avec l'ancienne initialisation (utilisant tous les individus par classe), et correspond aux cas où la matrice de confusion est mauvaise. De plus, les visualisations des modèles indiquent une estimation identique dans les deux cas. Ces deux éléments pointent vers des erreurs d'assignation de classe lors de l'échantillonage dans le Gibbs. L'objet de cette section est de rendre compte des essais effectués pour remédier à ces différences.

Le problème est de comprendre pourquoi le problème n'apparaît pas tout le temps. L'initialisation des paramètres, seule différence entre les deux cas, est effectuée avant le SEM. Le Gibbs n'est pas censé être différent, et les paramètres estimés semblent quasi- identiques.

Pour comprendre un peu mieux ce qui se passe, on se place dans le cas de la nouvelle initialisation. On a exporté les classes obtenues en tirant les appartenances aux classes, puis les classes obtenues en calculant les t_{ik} de façon analytique, en prenant le mode par individu dans les deux cas. Cette deuxième méthode a été implémentée de façon exceptionnelle, car on est pas garantis pour tous les modèles de pouvoir calculer analytiquement les probabilités observées de toutes les observations. Les résultats sont résumés dans la table 1.

	1	2
1	22	10
2	68	0

(a) si les t_{ik} sont obtenus à partir de fréquence de tirage

	1	2
1	0	32
2	68	0

(b) si les t_{ik} sont obtenus à partir des probabilités observées

TABLE 1 – Comparaison des matrices de confusion

On voit que si on calcule les t_{ik} analytiquement (ce qui n'est pas complexe pour ce modèle particulier), on obtient une matrice de confusion parfaite avec le nouveau système d'initialisation. Et c'est systématique : sur tous les essais effectués, quelle que soit l'allure de la matrice de confusion obtenue avec les t_{ik} issus de tirage, on a une matrice de confusion parfaite avec les t_{ik} analytique.

On peut en conclure que les appartenances aux classes sont très différentes entre l'apprentissage et le Gibbs. Elles sont "correctes" dans l'apprentissage, et erronées dans le Gibbs, quand on initialise l'apprentissage avec la nouvelle méthode.

On a fait ensuite deux essais, le premier en utilisant l'ancienne initialisation en apprentissage et la nouvelle initialisation en prédiction. En théorie le nouveau code n'impacte que l'initialisation des paramètres, qui n'est pas effectuée en prédiction, donc on s'attend à avoir des prédictions correctes. C'est le cas. On fait ensuite l'essai contraire. On fait un apprentissage en utilisant la nouvelle initialisation, puis une prédiction en utilisant l'ancienne initialisation. On obtient de mauvais résultats. Le problème vient donc des paramètres estimés issus de la nouvelle initialisation, et pas de l'échantilloneur de Gibbs.

Peut-être que l'erreur vient des représentations graphiques ? On regarde donc les paramètres estimés avec les deux initialisations, comme présentés dans la table 2. On ne voit pas non plus de différences frappantes entre les deux jeux de paramètres.

2 Initialisation du Gibbs

De façon orthogonale à cette analyse, on a modifié l'initialisation utilisée pour le Gibbs. Un système compliqué avait été mis en place pour tenir compte des paramètres qui sont censés être connus. Le problème c'est que ça complexifie le code car il fallait séparer les modèles avec chaînes de Markov des modèles sans chaîne de Markov. Et, dans tous les cas, on complète les variables latentes dans un ordre particulier, il n'est donc pas possible d'effectuer une complétion complète en utilisant les modèles.

A partir de maintenant, les initialisations effectuées en début de Gibbs sont donc les mêmes que celles effectuées avant l'apprentissage. Comme il y a une période de burn-in, les variables latentes sont progressivement tirées en utilisant la loi estimée auparavant.

La modification de cette méthode a permis de mettre en évidence une erreur de calcul dans le modèle de fonctionnelles. A un endroit on calcule une grandeur en utilisant le log d'une autre, alors que l'on peut la calculer directement. Le bug ne devait pas se manifester de façon systématique, mais c'est une bonne chose d'avoir effectuer cette correction, pour la robustesse du modèle.

ancienne initialisation	nouvelle initialisation
Composer proportions = 0.68 0.32 Parameters of Functional1 Class : 0 alpha : 0 0 -115566 11667.6 beta : 8.3491 -1.00881 -11.522 0.992039 sigma : 0.741071 0.763944 Class : 1 alpha : 0 0 -219125 22117.9 beta : 1.61716 1.01752 21.5991 -0.99619 sigma : 0.746063 0.751755	Composer proportions = 0.68 0.32 Parameters of Functional1 Class : 0 alpha : 0 0 -237173 24164.2 beta : 8.37728 -1.01723 -11.3817 0.983611 sigma : 0.728408 0.77722 Class : 1 alpha : 0 0 -26556.4 2831.09 beta : 1.43194 1.07491 20.6437 -0.938669 sigma : 0.661943 0.840589

TABLE 2 – Paramètres estimés

3 Conclusion

Pour le moment, on ne trouve pas de différences fondamentales expliquant les différences de comportement entre l'ancienne et la nouvelle initialisation des données fonctionnelles.

Le fait que les prédictions soient correctes dans tous les cas en utilisant les probabilités observées est rassurant est indique que les estimations de paramètres sont correctes, et donc que la correction doit se faire uniquement au niveau de la prédiction.

Références

[1] Vincent KUBICKI. Stabilisation de mixtcomp. 07/2017.

[2] Vincent KUBICKI. Avancement stabilisation mixtcomp. 08/2017.