Mixture of Multinomial PCA

Nicolas Jouvin [1] & Pierre Latouche [2] & Charles Bouveyron [3]

[1] *Laboratoire SAMM EA 4543, 90 rue de tolbiac, 75013 PARIS*
nicolas.jouvin@univ-paris1.fr
[2] *Laboratoire MAP5 UMR 8145, 45 rue des Saints-Pères, 75006 PARIS*
pierre.latouche@maths.cnrs.fr
[3] *Laboratoire J.A. Dieudonné UMR 7351, 28 avenue de Valrose, 06108 NICE Cedex 02*
charles.bouveyron@math.cnrs.fr

Résumé. Nous proposons le modèle MMPCA (pour mixture of multinomial PCA) pour la classification de données de comptages. Basé sur le modèle Latent Dirichlet allocation, il permet d'allier la flexibilité des *topic models*, tout en identifiant une partition explicite des données. L'inférence des paramètres et le clustering sont réalisés via un algorithme C-VEM (classification variational expectation-maximisation) qui optimise une vraisemblance complétée, couplé avec une stratégie de type *branch & bound* sur la borne variationnelle. Un critère ICL est proposé pour la sélection de modèles, et la performance de la méthodologie proposée est évaluée sur données simulées.

Mots-clés. Classification non-supervisée, modèles de mélange, données de comptages, Latent Dirichlet Allocation.

Abstract. We introduce the mixture of multinomial PCA (MMPCA), a mixture model for the clustering of count data. Relying on Latent Dirichlet Allocation, it offers the flexibility of *topic modeling* while being able to assign each observation to a unique cluster. Inference and clustering are done jointly by mixing a classification variational expectation maximisation algorithm (C-VEM), with a branch & bound like strategy on the variational lower bound. An integrated classification likelihood (ICL) criterion is derived, and experiments on simulated data illutrate the performance of the model.

Keywords. Clustering, mixture models, count data, Latent Dirichlet Allocation.

1 Introduction

La classification non supervisée, ou *clustering*, est un problème classique consistant à partitionner N observations en Q groupes distincts. La littérature est fournie sur ce sujet, et plusieurs méthodes sont proposées comme les approches hiérarchiques ou se basant sur une mesure de similarité. Le modèle de mélange est une des approches statistiques les plus plébiscitée en analyse exploratoire, dont l'instance la plus célèbre est sans doute le mélange de lois gaussiennes. Le modèle de mélange offre une grande flexibilité dans la modélisation,

ainsi que des outil statistiques pour l'inférence et la sélection de modèles. Nous nous intéressons ici au clustering de données de comptages. C'est un problème transversal et qui nécessite de sortir du cadre classique de mélanges gaussiens de part la nature discrète des données. Il apparaît par exemple en analyse textuelle pour le clustering de documents représentés en *sac-de-mots*, ou bien en génomique pour l'analyse d'expression de gènes via des techniques de séquençage de nouvelle génération.

Les modèles thématiques, ou *topic models*, représentent une classe de modèles génératifs dits à appartenance mixte, issus de l'analyse textuelle, où les corpus apparaissent sous la forme de comptages de mots. L'exemple le plus connu est le modèle Latent Dirichlet Allocation (LDA) [Blei et al., 2003], qui reste très utilisé et a connu de multiple extensions. Comme il est montré dans [Buntine, 2002], le modèle LDA peut être reformulé comme une version discrète de l'analyse en composantes principales probabiliste appelée Multinomial PCA (MPCA). Le modèle LDA n'est pas adapté lorsque l'objectif premier est d'obtenir une partition des observations En réalité, il fait l'hypothèse que chaque observation appartient, dans des proportions différentes, à tous les clusters. Plusieurs travaux proposent des extensions de LDA ou MPCA pour la classification. Le *cluster based topic model* (CTM) [Wallach, 2008] propose par exemple un modèle de mélange de MPCA où les variables latentes sont tirées selon un mélange de lois de Dirichlet, l'inférence est faite via des méthodes de type Monte-Carlo (MCMC), qui ont du mal à s'étendre en grande dimension. Dans [Xie and Xing, 2013], les auteurs proposent une extension du modèle précédent en considérant qu'un document est issu d'un mélange entre une MPCA global et une MPCA dont les paramètre dépendent de son groupe. L'inférence est faite par méthode variationnelle, le nombre de paramètres à estimer dans un tel modèle augmente grandement avec le nombre de groupes.

Dans un premier temps, nous présentons le modèle MPCA et proposons un nouveau modèle génératif pour la classification de données entières, où chaque observation suit un mélange de MPCA. Ensuite, nous proposons un algorithme pour estimer les paramètres et la partition de manière conjointe en maximisant une borne inférieure d'une vraisemblance complète. Enfin, nous proposons un critère de type vraisemblance classifiante intégrée [Biernacki et al., 2000] pour la sélection de modèle, et illustrons les performances sur un jeu de données simulées.

2 Le modèle mixture of multinomial PCA

Dans la suite, nous noterons $X = \{x_i\}_{i=1,\dots,N}$ un ensemble d'observations, où $x_i \in \mathbb{N}^V$. Le comptage total pour l'observation i sera noté $L_i = \sum_{v=1}^{V} x_{iv}$. En analyse textuelle, dont est issu notre modèle, V représente la taille du *vocabulaire*, tandis que les observations x_i sont appelées *documents*, modélisés en sac de *mots*.

Comme son nom l'indique, ce modèle proposé dans [Buntine, 2002] s'inspire de la version probabiliste de l'analyse en composante principale (pPCA) [Tipping and Bishop,

1999]. L'hypothèse générale des modèles statistiques pour la réduction de dimension est que chaque observation peut être reliée à une variable latente, appelons la θ_i, dans un espace de dimension K inférieure à V, via une transformation linéaire β et une *fonction de lien* probabiliste. Dans la pPCA classique θ_i suit une loi normale standardisée $\mathcal{N}(0_K, I_K)$ et la loi conditionnelle des observations s'écrit :

$$x_i \mid \theta_i \sim \mathcal{N}(\beta\theta_i + \mu, \sigma^2 I_V)\,.$$

Les matrices β, μ, ainsi que la variance σ^2 sont alors estimées par maximum de vraisemblance.

Pour modéliser des observations discrètes, il est souhaitable de s'affranchir du cadre gaussien. Le modèle multinomial PCA (MPCA) s'inspire de l'approche ci-dessus en supposant que θ_i est un vecteur du simplexe de dimension $K-1$, représentant une mesure de probabilité discrète sur $\{1,\dots,K\}$ (*i.e.* $\theta_{ik} \in \Delta_K := \{p \in \mathbb{R}^K \ : \ p \succcurlyeq 0 \text{ and } \sum_k p_k = 1\}$). La fonction de lien est alors une loi multinomiale et les observations sont supposées être générées comme suit :

1. $\theta_i \sim \mathcal{D}(\cdot;\alpha)$, où $\mathcal{D}$ est la loi de Dirichlet sur Δ_K : $\mathcal{D}(\theta_i;\,\alpha) = \frac{1}{Z(\alpha)} \prod_{k=1}^K \theta_{ik}^{\alpha_k - 1} \mathbb{1}_{\Delta_K}(\theta_i)$,
2. $x_i \mid \theta_i \sim \mathcal{M}_V(\cdot;\, L_i,\, \beta\theta_i)$,
   avec $K \in \mathbb{N}^\star$, et $\alpha = (\alpha_1,\dots,\alpha_K) \succ 0$.

Ici, la matrice β contient K lois discrètes sur $\{1,\dots,V\}$ dans ses colonnes ($\beta_{\cdot,k} \in \Delta_V$), appelées *topics* et caractérisant de manière globale l'espace des observations. Pour chaque observation, le paramètre de la fonction de lien est un mélange de ces distributions, dont les proportions sont déterminées par θ_i. Une observation est donc caractérisée par un mélange particulier des différents topics.

La vraisemblance des observations x_i n'est pas calculable explicitement à cause du couplage entre β et θ_i dans la multinomiale qui rend difficile la marginalisation sur θ_i. L'inférence est alors réalisée dans une autre formulation du modèle, qui est en fait le modèle LDA.

Bien que le modèle MPCA soit spécifiquement construit pour la réduction de dimension, il n'est pas conçu pour classer les observations elles-mêmes. Considérons le problème de trouver une partition des données en Q groupes. Pour le résoudre, tout en tirant parti de la flexibilité de MPCA, nous proposons un modèle où, conditionnellement à leur groupe, les observations sont tirées selon un modèle multinomial PCA dont la variable latente dépend de leur groupe. Le modèle génératif est détaillé Figure 1 et s'écrit :

Données : $(L_i)_i$, Q (nombre de clusters) et K (dimension de l'espace latent).

Paramètres : β, π

1. Tirer Q fois de manière indépendante $\theta_q \sim \mathcal{D}(\cdot;\,\alpha)$.
2. De manière indépendante, pour tout i,
   (a) tirer $Y_i \sim \mathcal{M}_K(\cdot;\,1,\,\pi)$ puis,
   (b) tirer $x_i \mid \{Y_{iq} = 1\}$, $(\theta_q)_q \sim \mathcal{M}_V(\cdot;\, L_i,\, \beta\theta_q)$.

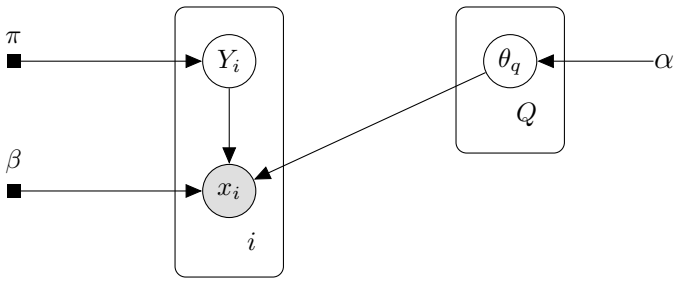

FIGURE 1 – Modèle graphique de MMPCA.

Contrairement à MPCA il n'y a pas une variable latente par observation mais une par cluster. Conditionnellement aux $(\theta_q)_q$, et en marginalisant sur Y_i, nous obtenons un mélange de Q lois MPCA :

$$\mathrm{p}(x_i \mid \theta;\, \beta, \pi) = \sum_{Y_i} \mathrm{p}(x_i, Y_i \mid \theta;\, \beta,\, \pi) = \sum_{q=1}^{Q} \pi_q \mathcal{M}_V(x_i;\, L_i,\, \beta\theta_q)\,.$$

De plus, si l'on considère une vraisemblance conditionnellement à Y, celle-ci s'écrit comme celle d'un modèle MPCA à Q observations :

$$\begin{aligned}
\mathrm{p}(X, \theta \mid Y;\, \beta) &= \prod_q \mathrm{p}(\theta_q) \prod_i \prod_q \mathcal{M}_V(x_i;\, L_i,\, \beta\theta_q)^{Y_{iq}}\,, \\
&\propto \prod_q \mathrm{p}(\theta_q) \prod_i \prod_v (\beta_{v,\cdot}^\top\, \theta_q)^{Y_{iq} x_{iv}}\,, \\
&\propto \prod_q \mathrm{p}(\theta_q) \prod_v (\beta_{v,\cdot}^\top\, \theta_q)^{\sum_i Y_{dq} x_{iv}}\,.
\end{aligned}$$

Il apparaît clairement que, à partition Y fixée, notre modèle est équivalent à une MPCA où les observations ont été agrégées en Q *méta-observations* obtenues en sommant les comptages dans chaque groupes $X_q := \sum_{i=1}^{N} Y_{iq} x_i$. Donc faire de l'inférence dans notre modèle en connaissant Y, revient à faire de l'inférence dans un modèle MPCA à Q observations agrégées selon Y. Dans la section suivante nous proposons un algorithme s'appuyant sur cette propriété, en alternant entre inférence des paramètres à Y fixé, via les méthodes déjà connues pour MPCA, et optimisation en Y.

3 Inférence

Dans une optique de classification, nous nous appuyons ici sur une log-vraisemblance *complétée* intégrée pour réaliser l'inférence :

$$\log \mathrm{p}(x, Y \mid \beta, \pi) = \log \int_\Theta \mathrm{p}(x, Y, \theta \mid \beta, \pi) = \log \int_\Theta \mathrm{p}(x, \theta \mid Y, \beta) + \log \mathrm{p}(Y \mid \pi)\,. \qquad (1)$$

Le premier terme de l'équation (1) fait apparaître la vraisemblance intégrée d'une MPCA sur les méta-observations, et n'est donc pas explicitement calculable. Nous prenons appui sur la borne variationnelle établie dans [Buntine, 2002], que nous n'explicitons pas ici par souci de place et notons $\mathcal{L}_{X|Y}$. Notre modèle vérifie la propriété suivante :

Proposition 1. *Pour toute partition Y, et pour toute loi sur θ, $\mathrm{q}(\theta) = \prod_q \mathrm{q}(\theta_q)$, l'inégalité suivante est vérifiée :*

$$\log \mathrm{p}(X, Y \mid \beta, \pi) \geq \mathcal{L}_{X|Y}(Y; q(\cdot), \beta) + \log \mathrm{p}(Y \mid \pi) =: \mathcal{L}(Y; \mathrm{q}(\cdot), \beta, \pi)\,. \tag{2}$$

Supposons Y fixé, la borne de l'équation (2) peut-être maximisée en (q, β) et π en utilisant respectivement les équations de [Buntine, 2002], et l'estimateur $\hat{\pi} = \sum_i Y_i/N$. Reste à déterminer comment trouver la partition Y inconnue. L'optimisation de la borne en Y nécessite Q^N partitions à explorer ce qui est irréalisable. Nous proposons de combiner inférence et clustering en nous appuyant sur une méthode de type *branch & bound*. La stratégie consiste à partir d'une partition aléatoire Y, de faire l'inférence variationnelle sur les Q méta-observations induites, et ensuite de parcourir les observations une à une. Pour chaque observation x_i, tous les $Q-1$ échanges possibles de groupes sont testés, les méta-observations mises à jour et une procédure d'inférence variationnelle est effectuée pour actualiser les paramètres (β, q, π). L'échange induisant la plus grande variation positive de la borne est ensuite retenu (s'il existe). La procédure s'arrête lorsqu'aucun échange maximisant la borne n'est plus possible, ou bien lorsqu'un nombre maximum d'itérations fixé à l'avance est atteint. L'étape d'inférence variationnelle pour chaque échange étant peu couteuse, cette algorithme peut-être appliqué sur des jeux de données contenant plusieurs milliers d'observations en grande dimension. Bien entendu, comme pour l'inférence variationnelle, il n'y a pas de garantie de trouver un maximum *global* de la log-vraisemblance et plusieurs initialisations doivent-être testées, celle apportant la plus grand borne étant retenue. La Figure 2 montre les résultat sur des données simulées selon notre modèle avec $Q = 6$ groupes et $K = 4$ *topics*, $N = 100, V = 1000, L_i = 100\,, \forall i$. Deux scénarios correspondants à deux vecteurs π différents sont investigués, et les performances sont évaluées en terme d'*Adjusted Rand Index*. La comparaison est faite avec des benchmarks classiques en clustering de données discrètes : K-means dans l'espace latent d'une MPCA avec différents K, argmax sur une MPCA à $K = Q$, factorisation de matrices non négatives (NMF), et le clustering de données de comptages issues de la génomiques (HTSClust).

4 Sélection de modèle

Les valeurs de Q et K n'étant pas connues *à priori*, nous proposons un critère de type vraisemblance classifiante intégrée pour les sélectionner. En suivant l'approche de [Biernacki et al., 2000], nous introduisons des distributions a priori sur les paramètre β et

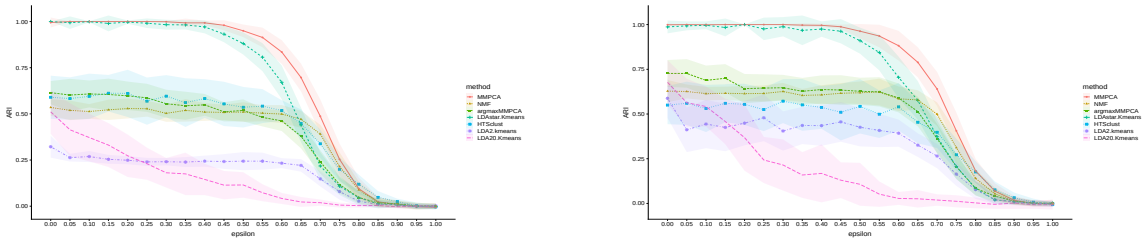

FIGURE 2 – ARI de la partition trouvée par différentes méthodes sur données simulée avec groupes proportionné (gauche) et groupe de tailles différentes (droite).

π, et travaillons avec la quantité :

$$\log \mathrm{p}(X, Y \mid Q, K) = \log \int_{\beta} \mathrm{p}(w \mid Y, \beta; Q, K)\, \mathrm{p}(\beta \mid K)\, \mathrm{d}\beta + \log \int_{\Pi} \mathrm{p}(Y \mid \pi; Q)\, \mathrm{p}(\pi \mid Q)\, \mathrm{d}\pi\,.$$

Les détails sont spécifiés dans le papier original : l'utilisation d'une approximation de Laplace sur le premier terme, ainsi que d'un a priori de Dirichlet sur π combiné à une approximation de Stirling permet de trouver une approximation de la quantité précédente :

$$ICL_{MMPCA}(Q, K) := \mathcal{L}_{X|Y}(\hat{Y}; \hat{q}, \hat{\beta}) - \frac{K(V-1)}{2} \log(Q) + \log \mathrm{p}(Y \mid \hat{\pi}) - \frac{Q-1}{2} \log(D)\,,$$

où le maximum de la log-vraisemblance conditionnelle dans l'approximation de Laplace est remplacé par sa borne variationnelle.

Références

[Biernacki et al., 2000] Biernacki, C., Celeux, G., and Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE transactions on pattern analysis and machine intelligence*, 22(7) :719–725.

[Blei et al., 2003] Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan) :993–1022.

[Buntine, 2002] Buntine, W. (2002). Variational extensions to em and multinomial pca. In *European Conference on Machine Learning*, pages 23–34. Springer.

[Tipping and Bishop, 1999] Tipping, M. E. and Bishop, C. M. (1999). Probabilistic principal component analysis. *Journal of the Royal Statistical Society : Series B (Statistical Methodology)*, 61(3) :611–622.

[Wallach, 2008] Wallach, H. M. (2008). *Structured topic models for language*. PhD thesis, University of Cambridge.

[Xie and Xing, 2013] Xie, P. and Xing, E. P. (2013). Integrating document clustering and topic modeling. *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence*.