How to leverage Bayesian mixtures for dynamic clustering and classification

Beatrice Franzolini[1]

Bocconi Institute for Data Science and Analytics, Bocconi University, IT
beatrice.franzolini@unibocconi.it

Abstract. Bayesian mixtures are well-established models for density estimation and probabilistic clustering of cross-sectional data. In the last decades, they have also been successfully extended to regression settings. However, there is still currently no consensus on whether and how Bayesian mixture models can handle the analysis of longitudinal data and conduct classification in general frameworks. This work presents some recent advances in longitudinal clustering and classification via Bayesian mixture models, showing novel promising results for the applicability of such models in these settings. The contents of these pages summarize some of the results derived in [12] and [11].

Keywords: Bayes classifiers, Bayesian nonparametrics, conditional partial exchangeability, longitudinal data, model-based clustering

1 BMM for cross-sectional density estimation, clustering, and regression

Bayesian mixture models (BMM) [10, 20, 18, 13, 5, 9, 24, 21, 7, 29] are a popular and well-established tool for density estimation and clustering of cross-sectional data. BMM assume the sequence of data $(X_i)_{i\geq 1}$ being generated by

$$X_i \overset{iid}{\sim} \int k(x_i,\theta)\mathrm{d}P(\theta) \overset{a.s.}{=} \sum_{h=1}^{\infty} w_h k(x_i,\theta_h) \qquad \text{for } i=1,2,\dots \tag{1}$$

where k is a deterministic probability mass or density function, typically Gaussian and referred to as *kernel*. The Bayesian approach comes into play once a prior is established for the *mixing measure* P or, equivalently, for the sequences of weights $(w_h)_{h\geq 1} \in \Delta_\infty$ and atoms $(\theta)_{h\geq 1}$, with $\Delta_\infty = \{(w_h)_{h\geq 1},: w_h \geq 0 \text{ and } \sum_{h=1}^{\infty} w_h = 1\}$ denoting the infinite-dimensional probability simplex. Note that (1) is general and does not require that the number of mixture components be almost surely infinite. In particular, finite mixture models with $H < \infty$ number of components can be recovered imposing that $w_h \sim \delta_0$ for any $h \geq H$, mixtures of finite mixtures [25, 24, 2] requiring $(w_h)_{h\geq 1}$ to be a.s. an eventually-zero sequence, and infinite mixtures [10, 20] imposing $\mathrm{pr}(w_h > 0) = 1$, $\forall h \geq 1$. When BMM are used for density estimation, they may achieve high flexibility (usually provided by the full weak support of the mixing measure P) while

also guaranteeing good frequentist consistency properties for most popular kernel and prior choices [see 14, 16, 19, 28, 15, 31, 32, 33]. When used for clustering, some advantages of BMM compared to algorithmic techniques are that they are based on explicit definitions of clusters via the kernel k, allow for an automatic data-driven selection of the number of clusters, and provide posterior-based uncertainty quantification. Differently than for density estimation, frequentist consistency properties of the clustering arrangement are puzzling. This is due not only to the posterior behavior of the number of clusters of most popular priors [22, 23, 3, 1, 8] but also to an intrinsic identifiability issue of model-based clustering problems anytime *true clusters* correspond to overlapping distributions.

Building upon models as in (1), in the last two decades several proposals have emerged to extend BMM to data grouped into J distinct populations. In this case, denoting via $((X_{ji})_{i\geq 1})_{j=1}^{J}$ an infinite partially exchangeable sequence of data, BMM takes the form

$$X_{ji} \overset{ind}{\sim} \int k_j(x_{ji},\theta)\mathrm{d}P_j(\theta) \overset{a.s.}{=} \sum_{h=1}^{\infty} w_{jh}k_j(x_{ji},\theta_h) \qquad \text{for } i=1,2,\dots \tag{2}$$

for $j=1,\dots,J$. No restrictions are imposed on the weights except that $(w_{jh})_{h\geq 1} \in \Delta_\infty$, and thus the complete sharing of atoms across populations is possibly just apparent. Different priors can be used for (2) based on which type of dependence and borrowing of information has to be considered across populations, allowing for modeling either ordered or unordered populations, with or without sharing of mixture components. Models of the type as in (2) appear to inherit frequentist posterior consistency for density estimation under mild additional assumptions [see, for instance, 26, 4, 6], while still guaranteeing flexibility, explicit definition of clusters, automatic selection of the number of clusters, and uncertainty quantification. However, contrary to BMM in (1) they accommodate different levels of heterogeneity across observations, based on the population assignment j, while allowing for borrowing across populations. Finally, both versions of BMM in (1) and (2) can be extended to be nonparametric regression models for a dependent variable Y_i on a vector of covariates Z_i (possibly continuous). This can be done either by setting $X_i = (Y_i, Z_i)$ or modeling weights and atoms in a mixture model for Y_i as functions of Z_i [see, 27, 30, for recent reviews].

While the settings mentioned so far, i.e., cross-sectional clustering and regression via BMM, are nowadays well-studied and mostly well-established, it is not yet unanimous how and if the core construction of BMM can be extended to effectively model longitudinal data and perform classification. The next two sections present some answers and ideas for these two settings, respectively.

2 BBM for dynamic clustering

Recently, in [12], a general extension to dynamic clustering based on BMM has been established, via the notion of conditional partial exchangeability (CPE) and a Markovian assumption. Let $X_{t,i}$, be data observed at time-point t for the

$i-$th item to be clustered and let $i = 1, \ldots, n$. The goal of dynamic clustering is to estimate a sequence of clustering configurations $(\rho_t)_{t=1}^T$, each corresponding to a distinct time-point. The partition ρ_t, can be represented by the vector $\boldsymbol{c}_t = (c_{t1}, \ldots, c_{tn})$ of subject-specific allocation variables, whose elements take value in the set $[n] := \{1, 2, \ldots, n\}$ and are such that $c_{ti} = c_{tj}$ if and only if subjects i and j belong to the same cluster according to ρ_t.

Exchangeability of the multivariate observations $(X_{1,i}, \ldots, X_{T,i})$ across i does not imply *conditional exchangeability* at different time points, given the past. To better understand this point, let's see what happens if we do assume exchangeability of $(X_{t,i})_{i\geq 1}$ conditionally on ρ_{t-1}. This implies, for instance, that for any set of three subjects i, j and k,

$$\text{pr}(X_{t,i}, X_{t,j} \mid \rho_{t-1} = \{\{i,j\},\{k\}\}) = \text{pr}(X_{t,i}, X_{t,k} \mid \rho_{t-1} = \{\{i,j\},\{k\}\}),$$

meaning that knowing that subjects i and j belong to the same cluster at time $t-1$ does not provide any information specific to those same two subjects at time t. This odd behavior of the learning mechanism under *conditional exchangeability* is because such an assumption prevents subject-level information (such as which subjects belong to the same cluster) from being carried from one time-point to the next, and allows only population-level information (such as knowledge about the number of clusters or the clusters' frequencies) to be transferred across time.

To avoid this issue, [12] propose CPE, that in the context of longitudinal data requires that for any $\sigma \in \mathcal{P}(n; \rho_{t-1,n})$, where $\mathcal{P}(n; \rho_{t-1,n})$ denotes the space of permutations of n elements that preserve $\rho_{t-1,n}$, i.e, $c_{t-1,\sigma(i)} = c_{t-1,i}$, for any i, the following two conditions hold

$$\text{pr}\left[(X_{t,1}, \ldots, X_{t,n}) \in A \mid \rho_{t-1,n}\right] = \text{pr}\left[(X_{t,\sigma(1)}, \ldots, X_{t,\sigma(n)}) \in A \mid \rho_{t-1,n}\right]$$

for any measurable set A, $n \geq 1$, and

$$\begin{aligned}&\text{pr}\left[(X_{t,i_1}, \ldots, X_{t,i_\ell}) \mid c_{t-1,i_1} = \ldots = c_{t-1,i_\ell}\right] = \\ &= \text{pr}\left[(X_{t,j_1}, \ldots, X_{t,j_\ell}) \mid c_{t-1,j_1} = \ldots = c_{t-1,j_\ell}\right]\end{aligned}$$

for any $(i_1, \ldots, i_\ell), (j_1, \ldots, j_\ell) \subset [n]$, allowing to obtain

$$\begin{aligned}&\text{pr}(X_{t,i}, X_{t,j} \in A^2 \mid \rho_{t-1} = \{\{i,j\},\{k\}\}) \geq \\ &\geq \text{pr}(X_{t,i}, X_{t,k} \in A^2 \mid \rho_{t-1} = \{\{i,j\},\{k\}\}).\end{aligned}$$

with a strict inequality on non-trivial sets A guaranteeing the preservation of subjects' identities across time. In [12], we show how models based on this assumption overperformed models that disregard subjects' identity in longitudinal settings.

3 BBM for classification

Classification tasks involve the process of categorizing input data into predefined classes or categories based on their features or attributes. In this setting, the

primary objective is to build a predictive model that can accurately assign new instances to the appropriate class labels. This task typically entails estimating a model (training a classifier) using a train set, where each data point is associated with a known category, and evaluating its performance on a test set. The most popular models and techniques for classification are logistic regression, support vector machines, random forests, naïve Bayes classifiers, and neural networks. All have advantages and disadvantages based on the specific applied problem and the amount of available data. Denoting with $(C_i, X_i)_{i=1}^n$ the train set, where $X_i = (X_{i,1}, \ldots, X_{i,p})$ is a vector of covariates, and $C_i \in \mathcal{C} = \{C_1^*, \ldots, C_K^*\}$ is the category, a naïve Bayes classifier associates to each new item with $X_i = x_i$, for $i \geq n+1$, the category $C(x_i)$ such that

$$C(x_i) = \underset{C \in \mathcal{C}}{\operatorname{argmax}} \operatorname{pr}(C|x_i) = \underset{C \in \mathcal{C}}{\operatorname{argmax}} \operatorname{pr}(C) \prod_{j=1}^{p} \operatorname{pr}(x_{ij}|C) \tag{3}$$

Equation (3) is derived employing Bayes theorem and under the conditional global independence assumption $\operatorname{pr}(x_i|C_k) = \prod_{j=1}^p \operatorname{pr}(x_{ij}|C_k)$ which, while reducing the number of parameters to feasible levels even in large p settings, also constitute the main limitation of naïve Bayes classifiers. Nonetheless, [11], in a spirit close to that of [17], explore the use of BMM to estimate the conditional multivariate distribution $\operatorname{pr}(x_i|C_k)$ assuming local (but not global) independence, meaning that the classification rule is obtained as

$$C(x_i) = \operatorname{pr}(x_i|C_k) = \underset{C \in \mathcal{C}}{\operatorname{argmax}} \operatorname{pr}(C) \sum_{h=1}^{\infty} w_h \prod_{j=1}^{p} \operatorname{pr}(x_{ij}|\theta_h, C_k)$$

The table below shows the out-of-sample predictive accuracy of a local independent model based on BMM obtained via category-specific mixtures of Dirichlet process mixtures of normal kernels, (multinomial) logistic regression and random forests on the `iris` dataset available in `R` for different proportion of train and test sets, code to replicate the results is available at https://github.com/beatricefranzolini/BMMclassifier.

Table 1. Out-of-sample accuracy on iris data-set

	DPM	Logistic	Random Forest
`iris`: train 80%, test: 20%	**100.00 %**	**100.00 %**	**100.00 %**
`iris`: train 50%, test: 50%	**96.00 %**	93.33 %	94.66 %
`iris`: train 30%, test: 70%	**97.14 %**	96.19 %	94.28 %
`iris`: train 10%, test: 90%	**96.29 %**	95.55 %	94.81 %

Acknowledgements

This work is supported by the National Recovery and Resilience Plan of Italy (PE1 FAIR - CUP B43C22000800006).

Bibliography

[1] Alamichel, L., Bystrova, D., Arbel, J., King, G.K.K.: Bayesian mixture models (in) consistency for the number of clusters. arXiv preprint arXiv:2210.14201 (2022)

[2] Argiento, R., De Iorio, M.: Is infinity that far? a Bayesian nonparametric perspective of finite mixture models. The Annals of Statistics **50**(5), 2641–2663 (2022)

[3] Ascolani, F., Lijoi, A., Rebaudo, G., Zanella, G.: Clustering consistency with Dirichlet process mixtures. Biometrika **110**(2), 551–558 (2023)

[4] Barrientos, A.F., Jara, A., Quintana, F.A.: Fully nonparametric regression for bounded data using dependent Bernstein polynomials. Journal of the American Statistical Association **112**(518), 806–825 (2017)

[5] Barrios, E., Lijoi, A., Nieto-Barajas, L.E., Prünster, I.: Modeling with normalized random measure mixture models. Statistical Science **28**(3), 313 – 334

[6] Catalano, M., De Blasi, P., Lijoi, A., Prünster, I.: Posterior asymptotics for boosted hierarchical Dirichlet process mixtures. The Journal of Machine Learning Research **23**(1), 3471–3493 (2022)

[7] Catalano, M., Lijoi, A., Prünster, I., Rigon, T.: Bayesian modeling via discrete nonparametric priors. Japanese Journal of Statistics and Data Science pp. 1–18 (2023)

[8] Chandra, N.K., Canale, A., Dunson, D.B.: Escaping the curse of dimensionality in Bayesian model-based clustering. Journal of Machine Learning Research **24**(144), 1–42 (2023)

[9] De Blasi, P., Favaro, S., Lijoi, A., Mena, R.H., Prünster, I., Ruggiero, M.: Are Gibbs-type priors the most natural generalization of the Dirichlet process? IEEE transactions on pattern analysis and machine intelligence **37**(2), 212–229 (2013)

[10] Ferguson, T.S.: Bayesian density estimation by mixtures of normal distributions. In: Recent advances in statistics, pp. 287–302. Elsevier (1983)

[11] Franzolini, B.: Non-naïve Bayes classifiers via Bayesian mixture models. Working paper (2024)

[12] Franzolini, B., De Iorio, M., Eriksson, J.: Conditional partial exchangeability: a probabilistic framework for multi-view clustering. arXiv preprint arXiv:2307.01152 (2023)

[13] Frühwirth-Schnatter, S.: Finite mixture and Markov switching models. Springer (2006)

[14] Ghosal, S., Ghosh, J.K., Ramamoorthi, R.: Posterior consistency of Dirichlet mixtures in density estimation. The Annals of Statistics **27**(1), 143–158 (1999)

[15] Ghosal, S., Van Der Vaart, A.: Posterior convergence rates of Dirichlet mixtures at smooth densities. The Annals of Statistics **35**(2), 1556–1593 (2007)

[16] Ghosal, S., Van Der Vaart, A.W.: Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. The Annals of Statistics **29**(5), 1233–1263 (2001)
[17] Hastie, T., Tibshirani, R.: Discriminant analysis by Gaussian mixtures. Journal of the Royal Statistical Society Series B: Statistical Methodology **58**(1), 155–176 (1996)
[18] Ishwaran, H., James, L.F.: Gibbs sampling methods for stick-breaking priors. Journal of the American statistical Association **96**(453), 161–173 (2001)
[19] Lijoi, A., Prünster, I., Walker, S.G.: On consistency of nonparametric normal mixtures for Bayesian density estimation. Journal of the American Statistical Association **100**(472), 1292–1296 (2005)
[20] Lo, A.Y.: On a class of Bayesian nonparametric estimates: I. Density estimates. The Annals of Statistics **12**(1), 351–357 (1984)
[21] McLachlan, G.J., Lee, S.X., Rathnayake, S.I.: Finite mixture models. Annual Review of Statistics and Its Application **6**, 355–378 (2019)
[22] Miller, J.W., Harrison, M.T.: A simple example of Dirichlet process mixture inconsistency for the number of components. Advances in neural information processing systems **26** (2013)
[23] Miller, J.W., Harrison, M.T.: Inconsistency of Pitman-Yor process mixtures for the number of components. The Journal of Machine Learning Research **15**(1), 3333–3370 (2014)
[24] Miller, J.W., Harrison, M.T.: Mixture models with a prior on the number of components. Journal of the American Statistical Association **113**(521), 340–356 (2018)
[25] Nobile, A.: Bayesian analysis of finite mixture distributions. Carnegie Mellon University (1994)
[26] Pati, D., Dunson, D.B., Tokdar, S.T.: Posterior consistency in conditional distribution estimation. Journal of multivariate analysis **116**, 456–472 (2013)
[27] Quintana, F.A., Müller, P., Jara, A., MacEachern, S.N.: The dependent Dirichlet process and related models. Statistical Science **37**(1), 24–41 (2022)
[28] Tokdar, S.T.: Posterior consistency of Dirichlet location-scale mixture of normals in density estimation and regression. Sankhyā: The Indian Journal of Statistics pp. 90–110 (2006)
[29] Wade, S.: Bayesian cluster analysis. Philosophical Transactions of the Royal Society A **381**(2247), 20220,149 (2023)
[30] Wade, S., Inacio, V., Petrone, S.: Bayesian dependent mixture models: A predictive comparison and survey. arXiv preprint arXiv:2307.16298 (2023)
[31] Walker, S.G., Lijoi, A., Prünster, I.: On rates of convergence for posterior distributions in infinite-dimensional models. The Annals of Statistics **35**(2), 738–746 (2007)
[32] Wu, Y., Ghosal, S.: Kullback Leibler property of kernel mixture priors in Bayesian density estimation. Electronic Journal of Statistics **2**(1), 298–331 (2008)
[33] Wu, Y., Ghosal, S.: The L1-consistency of Dirichlet mixtures in multivariate Bayesian density estimation. Journal of Multivariate Analysis **101**(10), 2411–2419 (2010)