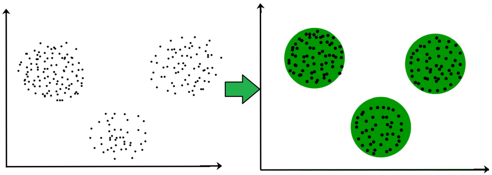
Clustering consistency with Dirichlet process mixtures

Giovanni Rebaudo

(Joint work with Ascolani, Lijoi and Zanella)

https://giovannirebaudo.github.io/Publications/Slides_Consistency.pdf



DPM

- Dirichlet process mixtures (**DPM**) (Lo, 1984) are the most popular Bayesian nonparametric method for density estimation and probabilistic clustering

$$X_i|\theta_i \overset{\text{ind}}{\sim} k(\cdot|\theta_i), \quad \theta_i \mid \tilde{P} \overset{\text{iid}}{\sim} \tilde{P}, \quad \tilde{P} = \sum_{j=1}^{\infty} \tilde{p}_j \delta_{\tilde{\theta}_j} \sim \text{DP}(\alpha, Q_0).$$

- **Validation:** one of the most popular ways to validate inferential procedure is via frequentist properties. **Consistency** is a natural minimal requirement.
- **Density estimation:** ideal **data generating truth:**

$$X_i \overset{\text{iid}}{\sim} f^*$$

in several relevant cases and metrics, the posterior distribution concentrates at the true data-generating density (at the minimax-optimal rate, up to a logarithmic factor) (Ghosal et al., 1999; Ghosal & Van der Vaart, 2007).

DPM for Probabilistic Clustering

- Obs are clustered together if they arise from the same k (e.g., Gaussian). # clusters in a sample = # of occupied mixture components $K_n \leq n$.
- Let $\tau_s(n)$ = set of unordered **partitions** of $\{1, \dots, n\}$ in s non empty subsets.
- The DPM model can be rewritten with respect to the random partitions:

$$\mathrm{p}(A \mid \alpha) = \frac{\alpha^s}{\alpha^{(n)}} \prod_{j=1}^{s} (a_j - 1)!, \quad A \in \tau_s(n) \quad \rightarrow \text{Partition}$$

$$\mathrm{p}(\hat{\theta}_{1:s} \mid A, s, \alpha) = \prod_{j=1}^{s} Q_0(\hat{\theta}_j) \quad \rightarrow \text{Unique parameters}$$

$$\mathrm{p}(X_{1:n} \mid \hat{\theta}_{1:s}, A) = \prod_{j=1}^{s} \prod_{i \in A_j} k(X_i \mid \hat{\theta}_j) \quad \rightarrow \text{Observations}$$

Validation: Probabilistic Clustering

- **Validation:** ideal data-generating truth is a finite mixture model

$$X_i \overset{\text{iid}}{\sim} \sum_{j=1}^{t} p_j^* k(\cdot \mid \theta_j^*)$$

- $t \in \mathbb{N}$ is the **true** number of mixture components. Some mis-specification: DPM has ∞ components! However, DPM is often used in practice when we believe that $t \in \mathbb{N}$ for any n to avoid fixing an upper bound for t (Miller and Harrison, 2013).
- Def: clusters = occupied mixture components. t is also the true # of clusters in the ideal population.

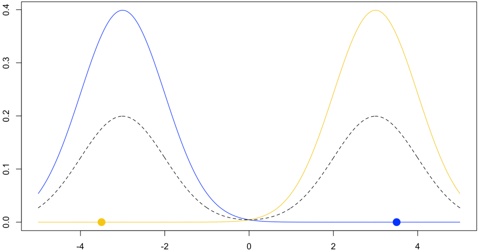
  More precisely, we can sample from the truth as first sampling the true clustering memberships $Z_i \in \{1, \dots, t\}$. $K_n^\star = \#$ clusters $:= \#\{Z_1, \dots, Z_n\}$

  Under the truth, $K_n^\star = t$ eventually almost surely.

Validation: Clustering Consistency under DPM

- Question: can we hope to learn the true partition?
  No! Without other info (e.g., repeated measurements with the same clustering).

$$f^* = 0.5\, \mathrm{N}(\cdot \mid \texttt{mean} = -3, \texttt{sd} = 1) + 0.5\, \mathrm{N}(\cdot \mid \texttt{mean} = 3, \texttt{sd} = 1)$$



- Question: can we learn the true t?

Related Results

- **Related interesting consistency results:**
  - **Wasserstein** distance posterior consistency of the mixing distribution under general conditions (Nguyen, 2013);
  - Under over-fitted finite Dirichlet mixtures (dimension $K > t$) and regularity assumptions, the additional weights can vanish or not depending on the hyperparameters of the finite Dirichlet (Rousseau and Mengersen, 2011).
- This kind of consistency does not imply consistency for t.

Further Motivations

Understanding the posterior behavior of K_n is useful for

- **Consistency/robustenss**
    - As a frequentist validation of the clustering and inference for the number of components.
    - What if? Understanding the learning and not just the prior.
- **Parsimony.** Having posterior behavior of K_n such that we don't overshoot (open too many clusters) if they are not needed to fit the data is useful for
    - Computation (e.g., better efficiency and mixing of MCMC, fewer identifiability issues).
    - better estimates (more borrowing i.e., bigger clusters thus better learning and less prior when we have enough good info).

Miller and Harrison, 2014

Consider a DPM model with **fixed** α and essentially any continuous kernel $k(\cdot)$. Assuming $(X_1, X_2, \ldots) \sim P^{*(\infty)}$ such that

$$X_i \overset{\text{iid}}{\sim} \sum_{j=1}^{t} p_j^* k(\cdot \mid \theta_j^*),$$

then

$$\limsup \mathrm{p}(K_n = t \mid X_{1:n}) < 1,$$

in $P^{*(\infty)}$-probability.

$\Rightarrow$ **inconsistency!**

Recall the notation. Two probabilities:

- p is the model.
- P^* is the data generating truth.

Gaussian Case

Assume $k(\cdot \mid \theta) = N(\cdot \mid \theta, 1)$.

Miller and Harrison, 2013

If P^* is any distribution with finite first moment, then $\mathrm{p}(K_n = 1 \mid X_{1:n})$ does not converge to 1. Even if the data are all constant.

Miller and Harrison, 2013

If $X_i \overset{\text{iid}}{\sim} \mathrm{N}(0, 1)$, then:

$$\mathrm{p}(K_n = 1 \mid X_{1:n}) \rightarrow 0,$$

as $n \to \infty$ in $P^{*(\infty)}$-probability.

Comments

- with fixed α, we always have **inconsistency**.
- finer lower bounds $\mathrm{p}(K_n = t \mid X_{1:n})$ in the DPM of Normals can be found in Yang et al. (2023+).
- inconsistency holds also for the **Pitman-Yor** process (Miller and Harrison, 2014)...
- ..and the other **Gibbs-type** priors (De Blasi et al.) with $\sigma > 0$ (Alamichel et al., 2023+).

An important Comment

The **concentration parameter** plays a **crucial role**

$$p(\theta_i \neq \theta_j) = \frac{\alpha}{1+\alpha},$$

so smaller $\alpha \Rightarrow$ less clusters.

- Fixing α is difficult.
- Usually a prior is placed, i.e. $\alpha \sim \pi(\cdot)$.

  To have a more flexible distribution on the clustering of the data, in most implementations of the DPM (e.g., Escobar & West 1995)

  $$\alpha \sim \pi \quad \to \text{Prior for concentration parameter}$$

  the mixing measure is itself a mixture in the sense of Antoniak (1974).
- Does it **change** the asymptotic behavior of K_n?

Intuition: Why Inconsistency is not Obvious from Literature

For any **fixed** $\alpha \in \mathbb{R}$

$$\limsup \mathrm{p}(K_n = t \mid X_{1:n}, \alpha) < 1 \text{ (=0 Gaussian case) }.$$

When a **prior** is placed

$$\limsup \mathrm{p}(K_n = t \mid X_{1:n}) = \limsup \int \mathrm{p}(K_n = t \mid X_{1:n}, \alpha)\pi(\alpha \mid X_{1:n}) \,\mathrm{d}\alpha$$

$$\stackrel{?}{=} \int \overbrace{\limsup \mathrm{p}(K_n = t \mid X_{1:n}, \alpha)}^{=0} \pi(\alpha \mid X_{1:n}) \,\mathrm{d}\alpha.$$

- In general the limit and the integral **cannot be exchanged**!
- If $\pi(\alpha \mid X_{1:n})$ concentrates around 0 we may achieve consistency.

Our result

- If $\pi(\alpha \mid X_{1:n})$ concentrates around 0 we may achieve consistency.

Posterior of α and K_n consistency

Under mild assumptions on π, if the model is **consistent for the number of clusters** we have

$$\pi(\alpha \mid X_{1:n}) \to \delta_0,$$

weakly as $n \to \infty$, in $P^{*(\infty)}$-probability.

- **A New Hope**: a priori $K_n \sim \alpha \log(n)$, therefore if the data are very close in terms of the kernel we expect empirical-based estimator $\hat{\alpha}(n) \to 0$ as $n \to \infty$.

Proof Technique

We have consistency if and only if

$$\sum_{s \neq t} \frac{\mathrm{p}(K_n = s \mid X_{1:n})}{\mathrm{p}(K_n = t \mid X_{1:n})} \to 0 \quad \text{as } n \to \infty\,.$$

- It suffices to work with **ratios**.
- Why is it useful?

Proof Idea

It holds

$$\frac{\mathrm{p}(K_n = s \mid X_{1:n})}{\mathrm{p}(K_n = t \mid X_{1:n})} = \underbrace{\frac{\int \frac{\alpha^s}{\alpha^{(n)}} \pi(\alpha)\,\mathrm{d}\alpha}{\int \frac{\alpha^t}{\alpha^{(n)}} \pi(\alpha)\,\mathrm{d}\alpha}}_{C(n,t,s)} \underbrace{\frac{\sum_{A \in \tau_s(n)} \prod_{j=1}^{s} (a_j - 1)! \prod_{j=1}^{s} m(X_{A_j})}{\sum_{B \in \tau_t(n)} \prod_{j=1}^{t} (b_j - 1)! \prod_{j=1}^{t} m(X_{B_j})}}_{R(n,t,s)}.$$

- The prior π impact only $C(n, t, s)$.
- If $C(n, t, s) \to 0$, for $s > t$, this may help!

The Choice of the Prior

We make the following assumptions

*A*1. **Absolute continuity**: the prior admits a density with respect to the Lebesgue measure.

*A*2. **Polynomial behaviour around the origin**: $\exists\, \epsilon,\ \delta,\ \beta$ such that $\forall \alpha \in (0, \epsilon)$ it holds $\frac{1}{\delta}\alpha^{\beta} \leq \pi(\alpha) \leq \delta\alpha^{\beta}$.

*A*3. **Subfactorial moments**: $\exists\, D, \nu, \rho > 0$ such that $\int \alpha^s \pi(\alpha)\, d\alpha < D\rho^{-s}\Gamma(\nu + s + 1)$ for every $s \geq 1$.

The following choices of $\pi(\cdot)$ satisfy assumptions $A1, A2$ and $A3$:

- Any distribution with **bounded support** that satisfies assumptions $A1$ and $A2$.
- The **Generalized Gamma** distribution with density proportional to $\alpha^{d-1} e^{-\left(\frac{\alpha}{a}\right)^p}$, provided that $p > 1$.
- The **Gamma** distribution with shape ν and rate ρ.

Main Result

Coefficients $C(n,t,s)$ can be interpreted as **posterior moments**

$$C(n,t,t+s) = \int_0^\infty \alpha^s \pi(\alpha \mid K_n = s)\, \mathrm{d}\alpha = E[\alpha^s \mid K_n = t].$$

Let π satisfy A1 and A2. Then for fixed s, that does not depend on n, we have

$$C(n,t,t+s) = E[\alpha^s \mid K_n = t] \sim \frac{1}{\log^s(n)}.$$

$\Rightarrow$ it helps consistency!

General Consequences

Informal

Under suitable assumptions on π, we may have

$$\limsup \mathrm{p}(K_n = t \mid X_{1:n}, \alpha) < 1$$

for every $\alpha > 0$ and

$$\lim \mathrm{p}(K_n = t \mid X_{1:n}) \to 1, \quad \text{in } P^{*(\infty)}\text{-probability.}$$

- **Idea:** Lower bounds $R(n, s, t)$ in the literature are enough to prove inconsistency with fixed α, but it is an open question when $\alpha \sim \pi$ (composed with our rate for $C(n, t, s)$ they go to zero).
- we have to derive new upper bounds (or tighter lower bounds) for $R(n, s, t)$ to prove consistency (or inconsistency).

A Simple Application

Let

$$P^* = \delta_{\theta^*}, \quad k(\cdot \mid \theta) = N(\cdot \mid \theta^*, 1), \quad Q_0 = N(0, 1)$$

Let π satisfies A1-A3 (with $\rho > 16$). Then

$$p(K_n = 1 \mid X_{1:n}) \to 1,$$

as $n \to \infty$ in $P^{*(\infty)}$-probability.

If α is **fixed**, this is **not true**.

A More General Class

Let

B1 θ be a location parameter, i.e. $k(x \mid \theta) = g(x - \theta)$.

B2 The support of g be bounded.

B3 The true values $(\theta_1^*, \ldots, \theta_t^*)$ be sufficiently separated.

Let π satisfies A1-A3 (with ρ high enough). Then

$$p(K_n = t \mid X_{1:n}) \to 1,$$

as $n \to \infty$ in $P^{*(\infty)}$-probability. If $\pi(\cdot) = \delta_{\alpha^*}$, then

$$\limsup p(K_n = t \mid X_{1:n}) < 1.$$

Summary

- A prior on α significantly changes the scenario.
- It makes the model more **robust**...
- ...and **adaptive**.

What's next?

- Other mixture kernel and truth.
- Impact of random α in **infinite** mixtures...
- Convergence rates.
- What about other BNP priors.
  E.g., Gibbs-type (Gnedin & Pitman 2006, De Blasi et. al., 2015).

Other Interesting Solutions

- If t is a **crucial parameter** and we think it is finite for any sample size n, better **explicitly** model it: mixture of finite mixtures (MFM) (Nobile, 1994; Richardson & Green, 1997; De Blasi et al. 2015; Miller & Harrison, 2018; Greve et al., 2022; Argiento & De Iorio, 2022).
  $\Rightarrow$ How to compare with MFM? Finite (unbounded) vs infinite # components.
- Consistent **post-processing**, even with α fixed (Guha et al., 2021; Alamichel et al., 2023+).
- Let the hyperparameter changes deterministically with n (Ohn & Lin, 2023; Zeng, Miller & Duan, 2023)

Final Comments

Problems and practical comments:

- Mis-specification of the kernel leads to inconsistency for the number of components (Cai et al., 2021).
- High-dimensional data are particularly challenging for clustering methods, which often incorrectly estimate the number of clusters (Chandra et al., 2023).
- Understanding the posterior behavior of the number of clusters in a finite sample obtained from the Bayesian estimate for the clustering under different losses (Chaumeny et al., 2023+; Franzolini & Rebaudo, 2023+).

My co-authors

Filippo Ascolani

Antonio Lijoi

Giacomo Zanella

References 1/2

- Alamichel, Bystrova, Arbel & King (2023+). Bayesian mixture models (in)consistency for the number of clusters. *ArXiv: 2210.14201*.
- Antoniak (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Stat.*, **2**.
- Argiento & De Iorio (2023). Is infinity that far? A Bayesian nonparametric perspective of finite mixture models. *Ann. Stat.*, **50**.
- Ascolani, Lijoi, Rebaudo & Zanella (2023). Clustering consistency with Dirichlet process mixtures. *Biometrika*, **110**.
- Cai, Campbell & Broderick (2021). Finite mixture models do not reliably learn the number of components. *ICML*, **139**.
- Chandra, Canale & Dunson (2023) Escaping the curse of dimensionality in Bayesian model based clustering. *J. Mach. Learn. Res.*, **24**.
- Chaumeny, Van der Molen, Anthony & Kirk (2023+) Bayesian nonparametric mixture inconsistency for the number of components: How worried should we be in practice? *ArXiv: 2207.14717*.
- De Blasi, Favaro, Lijoi, Mena, Prünster & Ruggiero (2015). Are Gibbs-type priors the most natural generalization of the Dirichlet process? *IEEE Trans. Pattern Anal. Mach. Intell.*, **37**.
- Franzolini & Rebaudo (2023) Entropy-regularized probabilistic clustering. *Submitted*.
- Gnedin & Pitman (2006). Exchangeable Gibbs partitions and Stirling triangles. *J. Math. Sci.*, **138**.
- Ghosal, Ghosh & Ramamoorthi (1999). Posterior consistency of Dirichlet mixtures in density estimation. *Ann. Stat.*, **27**.
- Ghosal & Van der Vaart (2007). Posterior convergence rates of Dirichlet mixtures at smooth densities. *Ann. Stat.*, **35**.

References 2/2

- Guha, Ho and Nguyen. (2021). On posterior contraction of parameters and interpretability in Bayesian mixture modeling. *Bernoulli*, **27**.
- Greve, Grün, Malsiner-Walli & Frühwirth-Schnatter (2022). Spying on the prior of the number of data clusters and the partition distribution in Bayesian cluster analysis. *Aust. N. Z. J. Stat.*, **64**.
- Lo (1984). On a class of Bayesian nonparametric estimates: I. Density estimates. *Ann. Stat.* **12**.
- Miller & Harrison (2013). A simple example of Dirichlet process mixture inconsistency for the number of components. *NeurIPS*.
- Miller & Harrison (2014). Inconsistency of Pitman-Yor process mixtures for the number of components. *J. Mach. Learn. Res.*, **15**.
- Miller & Harrison (2018). Mixture models with a prior on the number of components. *J. Am. Stat. Assoc.*, **113**.
- Nobile (1994). Bayesian Analysis of Finite Mixture Distributions. *Ph.D. thesis, Carnegie Mellon Univ.*
- Nguyen (2013). Convergence of latent mixing measures in finite and infinite mixture models. *Ann. Stat.*, **41**.
- Ohn & Lin (2023). Optimal Bayesian estimation of Gaussian mixtures with growing number of components. *Bernoulli*, **29**.
- Richardson & Green (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *JRSSB*, **59**.
- Rousseau & Mengersen (2011) Asymptotic behaviour of the posterior distribution in overfitted mixture models. *JRSS B*, **5**.
- Yang, Xia, Ho & Jordan (2023+). Posterior distribution for the number of clusters in Dirichlet process mixture models. *ArXiv: 1905.09959*.
- Zeng, Miller & Duan (2023). Quasi-Bernoulli stick-breaking: infinite mixture with cluster consistency. *J. Mach. Learn. Res.*, **24**.

`https://giovannirebaudo.github.io/Publications/Slides_Consistency.pdf`