Bayesian Nonparametric Mixtures Why and How?
Julyan Arbel

To cite this version:
Julyan Arbel. Bayesian Nonparametric Mixtures Why and How?. IFSS 2018 - 2nd Italian-French Statistics Seminar, Sep 2018, Grenoble, France. hal-01950664

HAL Id: hal-01950664
https://hal.archives-ouvertes.fr/hal-01950664
Submitted on 11 Dec 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Bayesian Nonparametric Mixtures
Why and How?

Introduction
Bayesian nonparametric framework
- Massively many parameters
- Inference on curves: pdf, cdf, hazard, link…
- Mixtures, exchangeable data \( X^* = (X_1, \ldots, X_n) \)

\[
X_1, \ldots, X_n | P \sim \left\{ P \sum_{k=1}^\infty k \cdot \delta_{P_k}(dt) \right\} \rightarrow 1 \rightarrow 2 \rightarrow 3
\]
- Natural uncertainty quantification
- Flexibility, avoids over-fitting by regularization (prior)
- Adapt to data complexity
- Underlying clustering
- Justify prior, expert
- Efficient posterior sampling
- Quantify truncation error

What prior for \( P \)?
- Learn about data through posterior dist.
- Discrete random probability measure prior
- Random weights \( (p_j) \) and locations \( (\theta_j) \)

\[
P = \sum_{j=1}^\infty p_j \delta_{\theta_j}
\]

\text{Dirichlet process } DP(\alpha, G_0) \text{ (Ferguson, 1973)}

Predictive: Chinese Restaurant Process

\[
P(X_{n+1} \in \cdot | X^n) = \frac{\alpha}{\alpha+n} G_0 + \frac{1}{\alpha+n} \sum_{j=1}^n (\delta_{\theta_j})
\]

\text{Or for varying } P(X_{n+1} \text{ new } \cdot ) \rightarrow

Survival Analysis
Bayesian hazard mixture (Arbel et al., 2016c)
- Data are (remission) times possibly censored
- Prior on hazard rate \( h(t) \) for every time \( t \)
- Induces prior on survival function \( S(t) \)
- Availability of posterior mean, median, mode
- Smooth estimator VS Kaplan–Meier
- Proper uncertainty quantification

Open Questions
- How to best use underlying clustering? (Wade and Ghahramani, 2015)
- Find consistent estimator of number of clusters: posterior inconsistent (Miller and Harrison, 2014), what about posterior mode?
- Devise efficient posterior sampling, truncation error (Arbel and Prünster, 2016)

Species Modeling
Data can be species, microbes, words, genes…

Discovery probabilities (Arbel et al., 2016a)
- Estimation of \( \ell \)-discovery \( D_\ell = P(X_{\ell+1} \text{ is a species seen } \ell \text{ times}) \)
- Comparison with Good-Turing estimator
- Closed form posterior and estimators
- Uncertainty quantifiable, unavailable for GT
- 2nd order (fast) approximations

Diversity in ecology (Arbel et al., 2015, 2016d)
- Assess impact of pollution on microbial community via study of diversity

\[
\text{Div} = -\sum_i p_i \log p_i
\]

- Model detects an hormetic effect
- Uncertainty quantification
- Prediction across full range of covariates

Density Estimation
Ecological risk assessment (Arbel et al., 2016b)
- Data are species critical effect concentrations (CEC), possibly censored
- Estimation of species sensitivity distribution (SSD), the density of CEC
- Safe concentration which protects most of the species: 95th percentile of the SSD (HC5)
- Very moderate sample sizes, \( \sim 10 – 50 \)
- BNP describes well variability of the data, without being prone to over-fitting
- Species clustering as an outcome

Funding
This work is supported by the European Research Council (ERC) through StG ”N-BNP” 306406.

References & Collaborators