SIMULATIONS FOR ENRICHED DIRICHLET PROCESS WITH NON-CONJUGATE KERNELS

HENGRUI LUO

Enriched Dirichlet process [1, 2] is an improvement of the classical Dirichlet process used in Bayesian nonparametric inference [3, 4]. As criticized by [2], Dirichlet process models or its mixture generalizations suffer from high dimensional covariates. The estimation of the joint posterior density $(X, Y) \in \mathcal{X} \times \mathcal{Y}$ will be dominated by the space $\mathcal{X}$ of explanatory covariate X if $dim\mathcal{X} \gg dim\mathcal{Y}$. If the marginal density of X is complicated, then such a behavior of posterior density will lead to concentration to many small clusters.

We explained the algorithm implemented in [2] first in this section and pointed out what is to be modified in order to use non-conjugate kernels. And then we proposed an algorithm that implements Efromovich-Pinsker estimator when an explicit ordering of components $X_1, \cdots X_p$ of multivariate explanatory covariate X is available.

The algorithm is basically an Markov Chain-Monte Carlo (MCMC) algorithm in nonparametric setting. The major difference is the proposal function of a new partition. The proposal function of new partition is a "partition-valued" probability measure. To be precise, we need to propose a new partition from urn-scheme [2, 3] according to Dirichlet process; In Enriched Dirichlet Process, such a proposal does not only follow urn-scheme but also preserves the nested structure from the partition of data. Suppose that the response covariates $y_1, \cdots y_n$ are partitioned into k clusters on the level of space $\mathcal{Y}$. Correspondingly we can collect those explanatory covariates $x_1, \cdots x_n$ on space $\mathcal{X}$ to each cluster on space $\mathcal{Y}$, each such cluster can be partitioned further by another sub-urn-scheme. Schematically we can represent the hierarchical structure as following diagram.

with help and guidance of Prof.S.MacEachern.

$$
\begin{matrix} Dirichlet(\theta_1, \cdots \theta_k) \\ OR\,(\alpha\,(\theta_1)\,, \cdots \alpha\,(\theta_k)) \end{matrix} \sim \begin{cases} y_1 \sim f_1(Y \mid X) & , x_i \sim Dirichlet(\psi_{1|1}, \cdots \psi_{k_1|1}) \begin{cases} g_1 & x_1, x_2, x_5 \\ \vdots & \\ g_{k_1} & x_9, x_{12} \end{cases} \\ y_2, y_3 \sim f_2 f_1(Y \mid X) & , x_i \sim Dirichlet(\psi_{1|2}, \cdots \psi_{k_2|2}) \begin{cases} h_1 & x_2, x_4 \\ \vdots & \\ h_{k_2} & x_{19}, x_{21} \end{cases} \\ \vdots & \\ y_{n-1}, y_n \sim f_k f_1(Y \mid X) & , x_i \sim Dirichlet(\psi_{1|k}, \cdots \psi_{k_k|k}) \begin{cases} m_1 & x_3, x_6, x_9, x_{10} \\ \vdots & \\ m_{k_k} & x_9 \end{cases} \end{cases}
$$

The parameter θ is associated with the conditional density functions on coarser partition of y clusters on space $\mathcal{Y}$, as well as the intensity parameter α_θ of Dirichlet process priors on $M(\mathcal{Y})$ is also determined by parameters θ.

The parameter ψ is associated with the marginal density functions on finer partition of x clusters on space $\mathcal{X}$, as well as the intensity parameter $\alpha_{\psi|\theta}$ of Dirichlet process priors on $M(\mathcal{X})$ is also determined by parameters ψ when given θ on a coarser partition.

The partition parameter $\rho, S = ((s_{y,1}, s_{x,1}), \cdots (s_{y,n}, s_{x,n}))$ is also regarded as a parameter to be monitored during the MCMC.

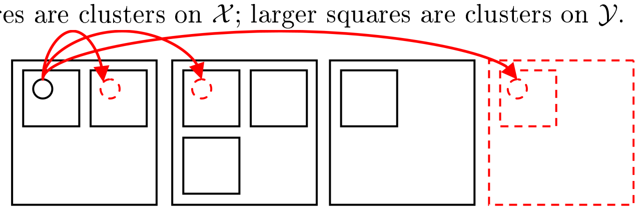
(1) Step 1: Propose a new partition s_i^* from its marginal.

For each observation $(x_i, y_i), i = 1 \cdots n$, we sample a proposal partition s_i for (x_i, y_i) based on marginal distribution of the removed- $(x, y)_{-i}$ sample $(x_1, y_1), \cdots (x_{i-1}, y_{i-1}), (x_{i+1}, y_{i+1}), \cdots (x_n, y_n)$ and $\psi_{-i} = (\psi_1, \cdots \psi_{i-1}, \psi_{i+1}, \cdots \psi_n)$ and θ_{-i}. This is the most tricky part in implementation because *unless* (x_i, y_i) is the only observation in that x, y-cell, $\theta_{-i} = \theta$. For each i we can calculate the proposal (unnormalized) probability $\omega_{j,l}$ for the i-th observation (x_i, y_i) falling in the j-th cluster on space $\mathcal{Y}$; l-th nested cluster on space $\mathcal{X}$ determined by an urn-scheme and marginal densities $f(Y \mid X), f(X)$. Due to the conjugacy of Dirichlet process we can obtained unnormalized probabilities of an observation (x_i, y_i) falling into the j-th cluster on space $\mathcal{Y}$ *and simultaneously* l-th nested cluster on space $\mathcal{X}$ $\omega_{j,l}(y_i, x_i), j = 1, 2 \cdots k^{-i}, k^{-i}+1$($k^{-i}$ is the number of clusters on space $\mathcal{Y}$ after removal of i-th observation) and $l = 1, 2, \cdots k_j^{-i}, k_j^{-i}+1$ (k_j^{-i} is the number of clusters on space $\mathcal{X}$ nested in the j-th clusters on space $\mathcal{Y}$ after removal of i-th observation). Its closed form id derived in (16) of [2] and we repeat them below for convenience of discussion.

$$\omega_{j,l}(y_i, x_i) = \begin{cases} \begin{cases} \frac{n_j^{-i} n_{l|j}^{-i}}{\alpha_\psi\left(\theta_j^{*-i}\right)+n_j^{-i}} \cdot K_{\theta_j^{*-i}}(y_i \mid x_i)\, K_{\psi_{l|j}^{*-i}}(x_i) & j = 1, 2 \cdots k^{-i} \\ \frac{n_j^{-i} \alpha_\psi\left(\theta_j^{*-i}\right)}{\alpha_\psi\left(\theta_j^{*-i}\right)+n_j^{-i}} \cdot K_{\theta_j^{*-i}}(y_i \mid x_i)\, h_x(x_i) & j = k^{-i}+1 \end{cases} & l = 1, 2, \cdots k_j^{-i} \\ \delta_{1j} \cdot \alpha_\theta h_y(y_i \mid x_i) h_x(x_i) & l = k_j^{-i}+1 \end{cases}$$

where $n_j^{-i}, n_{l|j}^{-i}$ are the number of observations in the j-th cluster on space $\mathcal{Y}$ *and simultaneously* l-th nested cluster on space $\mathcal{X}$ after removal of observation (x_i, y_i). The parameters $\theta_j^{*-i}, \psi_{l|j}^{*-i}$are the unique parameters associated with the j-th cluster on space $\mathcal{Y}$ *and simultaneously* l-th nested cluster on space $\mathcal{X}$ after removal of observation (x_i, y_i). $K_{\theta_j^{*-i}}(y_i \mid x_i)$ is the conditional density of $Y \mid X$; $K_{\psi_{l|j}^{*-i}}(x_i)$ is the marginal density of X; $h_x(x_i) = \int_\Psi \prod_{l \in S_j} K_\psi(x_l)\, dP_{0\psi|\theta}(\psi)$ where S_j is the number of observations in the j-th cluster on space $\mathcal{Y}$, $n_j = |S_j|$; $h_y(y_i \mid x_i) = \int_\Theta \prod_{l \in S_j} K_\theta(y_i \mid x_l)\, dP_{0\theta}(\theta)$. There is a minor typo on p.1046 of [2].

FIGURE 1. Possible proposals for an observation in Step 1. Small squares are clusters on $\mathcal{X}$; larger squares are clusters on $\mathcal{Y}$.



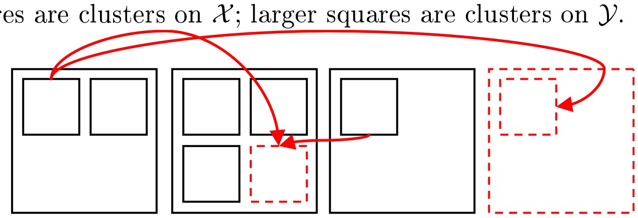
(2) Step 2: Improve mixing Metropolis-Hasting step.

This step only improves mixing, its idea is to permute all the observations in a clusters on space $\mathcal{X}$ to another new cluster (potentially a different cluster or create a new cluster depending on the random proposal) rather than permuting the observations themselves as in Step 1. This step is proposing a new partition on a space $\mathcal{X}$ structure while preserving a coarser structure on space $\mathcal{Y}$. This step reveals the essence of the hierarchical structure of Enriched Dirichlet process.

This step improves mixing, intuitively it moves clusters instead of a single observation while preserving the coarser structure. As commented by [2], "To improve mixing, we include an additional Metropolis-Hastings step; at each iteration, after performing the n Gibbs updates for each s_i, we propose a shuffle of the nested partition structure obtained by moving a ψ-cluster to be nested within a different or new θ-cluster. This move greatly improves mixing. "

This shuffle algorithm is by proposing a new cluster structure on space $\mathcal{X}$, Are there other $\psi - clusters$ within this $\theta - cluster$?

FIGURE 2. Possible proposals for an observation in Step 2. Small squares are clusters on $\mathcal{X}$; larger squares are clusters on $\mathcal{Y}$.



$$
\begin{cases} Yes & \begin{cases} \frac{1}{2}(Move1 + Move2) & \text{Move this } \psi - cluster \text{ to a new } \theta - cluster \\ \frac{1}{2}(Move1) & \text{Move this } \psi - cluster \text{ to other existing } \theta - clusters \end{cases} \\ No & \begin{cases} \frac{1}{2} & \text{Nothing happen.} \\ \frac{1}{2}(Move3) & \text{Move this } \psi - cluster \text{ to other existing } \theta - clusters \end{cases} \end{cases}
$$

(3) Step 3: Sampling hyper-parameters $\theta = (\theta_1, \cdots \theta_k), \psi = \left(\psi_{1|1}, \cdots \psi_{k_1|1}, \cdots \psi_{1|k}, \cdots \psi_{k_k|k}\right)$
We can sample from known posterior likelihoods if the prior base measures on Θ, Ψ $P_{0\theta}, P_{0\psi|\theta}$ (The priors on $M(\Theta), M(\Psi)$ are always chosen to be Dirichlet process prior as we assumed above) and $K_\theta\left(y_i \mid x_i\right), K_\psi\left(x_i\right)$ on $\mathcal{Y}, \mathcal{X}$ are conjugate family because they have closed form posterior probability densities.

As implemented in [2], they proposed Inverse Gamma-Gaussian conjugate family as a reasonable model for simulation as well as Alzheimer disease data; however, if the base measure $P_{0\theta}, P_{0\psi|\theta}$ of Dirichlet process prior does not form a conjugate family with $K_\theta\left(y_i \mid x_i\right), K_\psi\left(x_i\right)$ then we need additional MCMC-Metropolis-Hasting steps in Step 3 to sample from posterior likelihood of parameters θ, ψ. This does not cause too serious difficulty in computational aspect, therefore we can actually generalize Enriched Dirichlet process to non-conjugate family. A flexible suggestion is to use mixture of Dirichlet processes prior as [4], however it also suffers from high dimensional phenomena we mentioned above.

References

[1] Wade, Sara, Silvia Mongelluzzo, and Sonia Petrone. "An enriched conjugate prior for Bayesian nonparametric inference." Bayesian Analysis 6.3 (2011): 359-385.

[2] Wade, Sara, et al. "Improving prediction from dirichlet process mixtures via enrichment." Journal of Machine Learning Research 15.1 (2014): 1041-1071.

[3] Ghosh, Jayanta K., R. V. Ramamoorthi. Bayesian nonparametrics. Springer, 2003.

[4] Antoniak, Charles E. "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems." The annals of statistics (1974): 1152-1174.

[5] Wu, Yuefeng, and Subhashis Ghosal. "Kullback Leibler property of kernel mixture priors in Bayesian density estimation." Electronic Journal of Statistics 2 (2008): 298-331.

[6] Muller, Peter, Alaattin Erkanli, and Mike West. "Bayesian curve fitting using multivariate normal mixtures." Biometrika (1996): 67-79.

[7] Efromovich, Sam. "Conditional density estimation in a regression setting." The Annals of Statistics (2007): 2504-2535.