Population MCMC for Dirichlet Diffusion Trees

Pu Wang
George Mason Univ.
pwang7@gmu.edu

Kathryn B. Laskey
George Mason Univ.
klaskey@gmu.edu

Carlotta Domeniconi
George Mason Univ.
carlotta@cs.gmu.edu

Dirichlet Diffusion Trees (DDT) [3, 4] are an interesting nonparametric Bayesian model, which defines a nonparametric Bayesian prior over binary trees, with an *a priori* unbounded depth. MCMC is a natural choise to perform inference with DDT. However, MCMC suffers from getting trapped in local optima. This presents a serious problem when performing inference with tree structures, since the resulting solution space is very complex and highly multi-modal. In this scenario, the local nature of the moves of MCMC makes convergence prohibitive. Population MCMC (PopMCMC) [2, 1] has been proposed to try to avoid local optima when using MCMC for inference. PopMCMC runs multiple chains at the same time, and exchanges information between the chains, in order to propose non-local moves for each chain. In this work, we apply PopMCMC to DDT inference to try to overcome the local optima problem.

DDT is a generative model, and data generated from DDT are exchangeable [3]. DDT generates data points sequentially, and assumes all data points diffuse from the origin for unit time, say $[0, 1]$, according to brownian motion $\mathcal{N}(x_1; 0, \vec{I}t)$. For each data point, at time t, it diverges from a branch shared by m previous points with probability $a(t)dt/m$, where $a(t)$ is the predefined *diverging function*. The way DDT generates data can be represented as a tree. Figure 1, illustrates how DDT generates four points, where the detailed diffusion paths are suppressed and replaced by straight lines between diverging points and data points.

One nice property of DDT is that the joint probability of observing a tree can be factorized into two terms, a data term and a structure term. The data term involves the concrete locations of each diverging point and the point itself; the structure term involves only the information about the tree structure, which includes the diverging time of each diverging point. One can marginalize the data term since the locations of each diverging point and each data point follow a Brownian motion. Here we focus on structure term only. According to the definition of DDT, the probability of a specific tree structure is the product of the probabilities of observing each branch and the probabilities of choosing each branch. The probability of observing a branch shared by m data points from time s to time t is given by the probability of no divergence at this branch [3]: $p(\text{no divergence}) = \exp\left(-\int_s^t \frac{a(u)}{m} du\right) = \exp\left(\frac{A(s)-A(t)}{m}\right)$, where $A(t) = \int_0^t a(u)du$ is the cumulative diverging function. The probability of observing the tree structure in Figure 1 is:

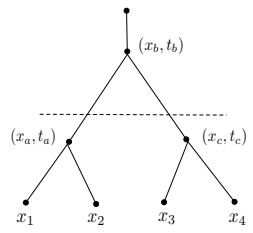

Figure 1: Illustration of DDT

$$\exp\left(-A(t_a)\right) a(t_a) \exp\left(-\frac{A(t_b)}{2}\right) \frac{a(t_b)}{2} \exp\left(-\frac{A(t_b)}{3}\right) \frac{1}{3} \exp\left(A(t_b) - A(t_c)\right) a(t_c) \tag{1}$$

To define clusters using DDT, we extend the standard DDT by adding a stopping timestamp: each subtree below the stopping timestamp constitutes a cluster of leaf nodes. For example, the dotted line in Figure 1 represents a stopping timestamp, and it defines two clusters, $\{x_1, x_2\}$ and $\{x_3, x_4\}$. We use DDT as a prior to the latent cluster structures of the observed data. As for the DDT in Figure 1, we don't use the data term, the concrete locations of internal nodes and leaf nodes, $\{x_a, x_b, x_c, x_1, x_2, x_3, x_4\}$. We only use the structure term to define cluster structures. Further, assume the observed features of the four data points are y_1, y_2, y_3, and y_4.

Gibbs sampling a DDT structure is straightforward. At every iteration, Gibbs sampling can only sample a path to one leaf node. This generates changes that are local to the DDT structure. We propose to use

PopMCMC to perform non-local changes, and therefore avoid getting trapped in local optima. When using PopMCMC for DDT, we run multiple chains; each chain has the same stationary distribution, which is the posterior DDT distribution given the same observed data. We assume that the stopping timestamps for all DDTs in different chains are the same, and randomly choose two chains to exchange DDT structure information. We exchange the structure term of the two chosen DDTs for the same subset of leaf nodes. Again, lets assume the DDT shown in Figure 1 is the current state of a chosen chain of PopMCMC, denoted as $\mathcal{T}_1$. The exchanging procedure of PopMCMC is as follows. We first assume a generation order on the leaf nodes, say $\langle 1, 3, 2, 4\rangle$ (we can make this assumption because data generated from a DDT are exchangeable). We randomly choose a subset of the leaf nodes, say $\{x_2, x_4\}$, and find the diverging time for each chosen leaf node, $t_2 = t_a$ and $t_4 = t_c$. We also remove the chosen leaf nodes from $\mathcal{T}_1$, and denote the rest of the DDT as $\mathcal{T}_1^{\neg\{x_2,x_4\}}$. The reason we assume an order before choosing a subset of the leaf nodes is that the diverging time of each leaf node depends on the order of data generation. Likewise, we denote the second chosen DDT as $\mathcal{T}_2$; we assume the same generation order on the leaf nodes, and find the diverging time for $\{x_2, x_4\}$ in $\mathcal{T}_2$, denoted as t'_2 and t'_4. We remove $\{x_2, x_4\}$ from $\mathcal{T}_2$, resulting in $\mathcal{T}_2^{\neg\{x_2,x_4\}}$. The structure terms to be exchanged between $\mathcal{T}_1$ and $\mathcal{T}_2$ are $\{t_2, t_4\}$ and $\{t'_2, t'_4\}$. We propose new paths to x_2 and x_4 in $\mathcal{T}_1^{\neg\{x_2,x_4\}}$ using the diverging time $\{t'_2, t'_4\}$. This leads to a new DDT $\mathcal{T}'_1$. We also propose new paths to x_2 and x_4 in $\mathcal{T}_2^{\neg\{x_2,x_4\}}$ using diverging time $\{t_2, t_4\}$. This leads to a new DDT $\mathcal{T}'_2$. The acceptance ratio of the exchange is:

$$\frac{\pi(\mathcal{T}'_1)\pi(\mathcal{T}'_2)Q(\mathcal{T}_1, \mathcal{T}_2|\mathcal{T}'_1, \mathcal{T}'_2)}{\pi(\mathcal{T}_1)\pi(\mathcal{T}_2)Q(\mathcal{T}'_1, \mathcal{T}'_2|\mathcal{T}_1, \mathcal{T}_2)} \tag{2}$$

where $\pi(\cdot)$ denotes the posterior distribution of DDT, and $Q(\cdot, \cdot|\cdot, \cdot)$ denotes the proposal distribution.

When proposing new DDTs, the proposal distribution needs to choose branches for each chosen leaf node. For example, lets assume we have already removed $\{x_2, x_4\}$ from $\mathcal{T}_1$, and we are now adding $\{x_2, x_4\}$ to $\mathcal{T}_1^{\neg\{x_2,x_4\}}$, conditioned on the new diverging time $\{t'_2, t'_4\}$. According to the leaf node generation order $\langle 1, 3, 2, 4\rangle$, we need to first generate a path for x_2 conditioned on $\mathcal{T}_1^{\neg\{x_2,x_4\}}$ and a new diverging time of x_2, say t'_2. Without loss of generality, assume t'_2 is greater than the stopping timestamp; then, x_2 will be either with x_1 or with x_4. A branch, between (x_b, x_1) and (x_b, x_3), needs to be chosen for x_2 to follow. Here we propose to choose a branch according to likelihood only, that is, the probability of choosing the branch (x_b, x_1) is proportional to $\mathcal{L}(\{y_1, y_2\}, \{y_3\})$, and the probability of choosing the branch (x_b, x_3) is proportional to $\mathcal{L}(\{y_1\}, \{y_2, y_3\})$. We proceed similarly when proposing a new path for x_4.

In addition, in order to increase the acceptance ratio, we use a temperature τ to adjust the proposal distribution, and the acceptance ratio becomes:

$$\frac{\pi(\mathcal{T}'_1)\pi(\mathcal{T}'_2)}{\pi(\mathcal{T}_1)\pi(\mathcal{T}_2)} \left(\frac{Q(\mathcal{T}_1, \mathcal{T}_2|\mathcal{T}'_1, \mathcal{T}'_2)}{Q(\mathcal{T}'_1, \mathcal{T}'_2|\mathcal{T}_1, \mathcal{T}_2)}\right)^{\frac{1}{\tau}} \tag{3}$$

References

[1] O. Cappé, A. Guillin, J.-M. Marin, C. P. Robert, and C. P. Robertyz. Population monte carlo. *Journal of Computational and Graphical Statistics*, 13:907–929, 2004.

[2] K. B. Laskey and J. W. Myers. Population markov chain monte carlo. *Machine Learning*, 50(1-2):175–196, 2003.

[3] R. M. Neal. Defining priors for distributions using dirichlet diffusion trees. Technical Report 0104, Dept. of Statistics, University of Toronto, 2001.

[4] R. M. Neal. Density modeling and clustering using dirichlet diffusion trees. *Bayesian Statistics*, 7:619–629, 2003.

Topic: learning topic

Preference: oral/poster