Dirichlet Process Report

Ziyi Song

May 2020

This report is comprised of what I think is most fundamental in Bayesian Nonparametric for novices. The report goes in a sequence of topics: Dirichlet Processes, Dirichlet Process Mixtures, Markov Chain Sampling for Dirichlet Process Mixture models, Hierarchical Dirichlet Processes, applications on autonomous multi-vehicle interaction scenarios modeling, deficiencies of this application, and potential way to improve this application modeling so as we can work in the coming months.

The report content follows books and papers I read as shown in reference, and includes my personal derivations and understandings.

Contents

1 Dirichlet Processes 3

1.1 Some basic knowledge on Dirichlet Process 3

1.2 Stick-Breaking Construction 6

1.3 Mixtures of Dirichlet Processes (MDP) 9

2 Dirichlet Process Mixtures (DPM) 13

3 Markov Chain Sampling for Dirichlet Process Mixture 17

3.1 limit of finite mixture models 17

3.2 Gibbs Samplings for Dirichlet Process Mixture 19

4 Hierarchical Dirichlet Processes (HDP) 23

4.1 HDP 23

4.2 Stick-Breaking Construction 24

4.3 The Chinese Restaurant Franchise 25

5 autonomous multi-vehicle interaction scenarios 26

6 Reference 29

7 Appendix 29

1 Dirichlet Processes

1.1 Some basic knowledge on Dirichlet Process

Definition 1.1. Let (Θ, B) be a measurable space, with P a random probability measure over this space. A *Dirichlet process* is defined as the distribution of the random probability measure P such that, for any finite measurable partition $(A_1, A_2, \ldots, A_k)$ of Θ, with *base measure* α,

$$(P(A_1), \ldots, P(A_k)) \sim Dir(k; \alpha(A_1), \ldots, \alpha(A_k)) \tag{1}$$

denoted by $P \sim DP(\alpha)$.

constant $M = |\alpha| = \alpha(\Theta) =$ total mass, called *prior precision*

probability measure $\bar{\alpha} = \alpha/|\alpha|$, called center measure

Remember probability measure P is random

Followings are two basic propositions about Dirichlet prior P without proof. Just take them for granted.

Proposition 1.1. *If $P \sim DP(\alpha)$, for any measurable sets A and B,*

$$\mathrm{E}(P(A)) = \bar{\alpha}(A) \tag{2}$$

$$\operatorname{var}(P(A)) = \frac{\bar{\alpha}(A)\bar{\alpha}\left(A^c\right)}{1+|\alpha|} \tag{3}$$

Proposition 1.2. *If $P \sim DP(\alpha)$, then for any measurable functions ψ and ϕ,*

$$\mathrm{E}(P\psi) = \int \psi d\bar{\alpha} \tag{4}$$

$$\operatorname{var}(P\psi) = \frac{\int \left(\psi - \int \psi d\bar{\alpha}\right)^2 d\bar{\alpha}}{1+|\alpha|} \tag{5}$$

The models in this report are under the setting of Bayes. So prior, posterior, and marginal distributions are very important to us. They are involved in almost every model through this report.

Here comes one about marginal.

For a random probability measure P possessing a Dirichlet process distribution, and an observation X from the probability measure P, using equation (2), we have

$$\mathrm{P}\left(X \in A\right) = \mathrm{E}\mathrm{P}\left(X \in A|P\right) = \mathrm{E}P(A) = \bar{\alpha}(A) \tag{6}$$

i.e., marginal $X \sim \bar{\alpha}$

So when we say an observation X from P, it means

$$\text{if } P \sim \mathrm{DP}(\alpha) \text{ and } X|P \sim P, \text{ then } X \sim \bar{\alpha} \tag{7}$$

It naturally extends to the question: what is the joint marginal distribution of observations $X_1, X_2, \ldots$ from a Dirichlet process, i.e., $P \sim DP(\alpha)$ and $X_1, X_2, \ldots |P \overset{\text{iid}}{\sim} P$, then $(X_1, X_2, \ldots) \sim$?

Before moving on, we need to know the posterior distribution of the random probability measure P:

Theorem 1.3. *If prior $P \sim DP(\alpha)$, $X_1, X_2, \ldots, X_n|P \overset{iid}{\sim} P$, then posterior $P|X_1, \ldots, X_n \sim \mathrm{DP}\left(\alpha + \sum_{i=1}^{n} \delta_{X_i}\right)$*

so updating rules

$$\text{base measure } \alpha \mapsto \alpha + \sum_{i=1}^{n} \delta_{X_i}$$

$$\text{total mass } M \mapsto M + n$$

$$\text{center measure } \bar{\alpha} \mapsto \frac{M}{M+n}\bar{\alpha} + \frac{1}{M+n}\sum_{i=1}^{n}\delta_{X_i}$$

so by equation (2), we directly see that

$$\mathrm{E}\left(P(A)|X_1,\ldots,X_n\right) = \frac{|\alpha|}{|\alpha|+n}\bar{\alpha}(A) + \frac{1}{|\alpha|+n}\sum_{i=1}^{n}\delta_{X_i}(A)$$

Now we go back to the question: if $P \sim DP(\alpha)$ and $X_1, X_2, \ldots |P \overset{\text{iid}}{\sim} P$, then $(X_1, X_2, \ldots) \sim$?

$(X_1, X_2, \ldots) \overset{d}{=} X_1 \times X_2|X_1 \times X_3|X_2, X_1 \times \ldots\ldots$

$X_1 \sim \bar{\alpha}$

$X_2|\left(P, X_1\right) \sim P$ and $P|X_1 \sim \mathrm{DP}\left(\alpha + \delta_{X_1}\right)$ by **Theorem 1.3**

$$\begin{aligned} P(X_2 \in A|X_1) &= E[\mathbb{1}_A(X_2)|X_1] = E[E(\mathbb{1}_A(X_2)|X_1, P)] \\ &= E[P(X_2 \in A)|X_1] = E[P(A)|X_1] \\ &= \frac{|\alpha|}{|\alpha|+1}\bar{\alpha}(A) + \frac{1}{|\alpha|+1}\delta_{X_1}(A) \end{aligned}$$

$$\text{so } X_2|X_1 \sim \begin{cases} \delta_{X_1}, & \text{with probability } 1/(|\alpha|+1) \\ \bar{\alpha}, & \text{with probability } |\alpha|/(|\alpha|+1) \end{cases}$$

Similarly,

$$X_n|X_1,\ldots,X_{n-1} \sim \begin{cases} \delta_{X_1}, & \text{with probability } 1/(|\alpha|+n-1) \\ \vdots & \vdots \\ \delta_{X_{n-1}}, & \text{with probability } 1/(|\alpha|+n-1) \\ \bar{\alpha}, & \text{with probability } |\alpha|/(|\alpha|+n-1) \end{cases} \tag{8}$$

The above model, **equation (8)**, is called generalized Polya urn scheme. This generalized Polya urn scheme is very similar to Chinese

Restaurant Process. These famous schemes are used a lot, and their detailed story not being illustrated here.

Now the joint marginal

$$\begin{aligned}(X_1,\ldots,X_n) &= X_1 \times (X_2|X_1) \times (X_3|X_2,X_1) \times \ldots\cdots \times (X_n|X_1,\ldots,X_{n-1}) \\ &= \prod_{j=1}^{n}(X_j|X_1,\ldots,X_{j-1}) \\ &\sim \prod_{j=1}^{n}(\bar{\alpha}\frac{|\alpha|}{|\alpha|+j-1}+\frac{\delta_{X_1}}{|\alpha|+j-1}+\cdots+\frac{\delta_{X_{j-1}}}{|\alpha|+j-1}) \\ &\sim \prod_{j=1}^{n}\frac{(\alpha+\delta_{X_1}+\cdots+\delta_{X_{j-1}})}{|\alpha|+j-1}\end{aligned} \tag{9}$$

So we have the joint marginal distribution of $(X_1,\ldots,X_n)$ if $P \sim DP(\alpha)$ and $X_1,X_2,\ldots|P \overset{\text{iid}}{\sim} P$. This equation (9) is important and will be used later in this report.

1.2 Stick-Breaking Construction

Sticking-Breaking is one of the most famous representation of Dirichlet process. Its proof in reference book is too sketchy and over-concise. I complete it and will provide a detailed proof of it in the Appendix. Before doing that, let me first introduce **5** fundamental properties of Dirichlet distribution, which will be used in the proof of Stick-Breaking. But readers can just skip them and go straight to **Theorem 1.9** directly.

Proposition 1.4 (Representations). *For random variables $Y_1,\ldots,Y_k$ and $Y=\sum_{i=1}^{k}Y_i$ (i) If $Y_i \overset{ind}{\sim} \text{Gamma}(\alpha_i,1)$, then $(Y_1,\ldots,Y_k)/Y \sim \text{Dir}(k;\alpha_1,\ldots,\alpha_k)$, and is independent of Y*

(ii) If $Y_i \overset{ind}{\sim} \text{Beta}(\alpha_i,1)$, then $((Y_1,\ldots,Y_k)|Y=1) \sim \text{Dir}(k;\alpha_1,\ldots,\alpha_k)$

(iii) If $Y_i \overset{ind}{\sim} \text{Exp}(\alpha_i)$, then $\Big((e^{-Y_1},\ldots,e^{-Y_k})\,|\sum_{i=1}^{k}e^{-Y_i}=1\Big) \sim \text{Dir}(k;\alpha_1,\ldots,\alpha_k)$

Proposition 1.5 (Aggregation). *If $X \sim \text{Dir}\,(k; \alpha_1, \ldots, \alpha_k)$ and $Z_j = \sum_{i \in I_j} X_i$ for a given partition $I_1, \ldots, I_m$ of $\{1, \ldots, k\}$, then*
(i) $(Z_1, \ldots, Z_m) \sim \text{Dir}\,(m; \beta_1, \ldots, \beta_m)$, where $\beta_j = \sum_{i \in I_j} \alpha_i$, for $j = 1, \ldots, m$
(ii) $(X_i/Z_j : i \in I_j) \overset{ind}{\sim} \text{Dir}\,(\#I_j; \alpha_i, i \in I_j)$, for $j = 1, \ldots, m$
(iii) $(Z_1, \ldots, Z_m)$ and $(X_i/Z_j : i \in I_j)$, $j = 1, \ldots, m$, are independent

Proposition 1.6 (Conjugacy). *If $p \sim \text{Dir}(k; \alpha)$ and $N|p \sim \text{Multinomial}_k(n; p)$, then $p|N \sim \text{Dir}(k; \alpha + N)$, where N is a vector of $(N_1, \ldots, N_k)$ and $N_1 + \ldots N_k = n$*

Proof.

$$\underbrace{p_1^{\alpha_1 - 1} \cdots p_k^{\alpha_k - 1}}_{\text{Dirichlet density}} \times \underbrace{p_1^{N_1} \cdots p_k^{N_k}}_{\text{multinomial likelihood}} = p_1^{\alpha_1 + N_1 - 1} \cdots p_k^{\alpha_k + N_k - 1}$$

so $p|N \sim \text{Dir}(k; \alpha_1 + N_1, \ldots, \alpha_k + N_k)$ □

Proposition 1.7. *For $k \in \mathbb{N} = 1, 2, 3, \ldots$ and any α with $|\alpha| > 0$,*

$$\sum_{i=1}^{k} \frac{\alpha_i}{|\alpha|} \text{Dir}\,(k; \alpha_1, \ldots \alpha_{i-1}, \alpha_i + 1, \alpha_{i+1}, \ldots, \alpha_k) = \text{Dir}\,(k; \alpha_1, \ldots, \alpha_k)$$

Proof. If $p \sim \text{Dir}(k; \alpha)$ and $N|p \sim \text{Multinomial}_k(1; p)$, then $\text{P}(N = i) = \alpha_i/|\alpha|$. Then by **Proposition 1.6**, $p|\{N = i\} \sim \text{Dir}\,(k; \alpha + e_i)$, where e_i is the i th unit vector.

And $P \sim \sum_{i=1}^{k} (P, N = i) \sim \sum_{i=1}^{k} p|\{N = i\} \times \{N = i\}$
$= \sum_{i=1}^{k} \frac{\alpha_i}{|\alpha|} \text{Dir}\,(k; \alpha + e_i)$
$= \sum_{i=1}^{k} \frac{\alpha_i}{|\alpha|} \text{Dir}\,(k; \alpha_1, \ldots \alpha_{i-1}, \alpha_i + 1, \alpha_{i+1}, \ldots, \alpha_k)$ □

Proposition 1.8. *If* $p \sim \mathrm{Dir}(k;\alpha), N \sim \mathrm{Multinomial}_k(1;\alpha)$ *and* $V \sim \mathrm{Beta}(1,|\alpha|)$ *are independent, then* $VN + (1-V)p \sim \mathrm{Dir}(k;\alpha)$

Proof. If $Y_0, Y_1, \ldots, Y_k \overset{\text{ind}}{\sim} \text{Gamma}\left(\alpha_i, 1\right), i = 0, 1, \ldots, k$, where $\alpha_0 = 1$, then by **Proposition 1.4, 1.5**, the vector (Y_0, Y) for $Y = \sum_{i=1}^k Y_i$ is independent of $p := (Y_1/Y, \ldots, Y_k/Y) \sim \mathrm{Dir}\left(k, \alpha_1, \ldots, \alpha_k\right)$ and $V = Y_0/\left(Y_0 + Y\right) \sim \mathrm{Beta}(1, |\alpha|)$. Furthermore

$$(V, (1-V)p) = \left(Y_0, Y_1, \ldots, Y_k\right) / \left(Y_0 + Y\right) \sim \mathrm{Dir}(k+1; 1, \alpha)$$

thus $(Ve_i, (1-V)P) \sim Dir(e_i, \alpha)$, and by **Proposition 1.5**,

$$(Ve_i + (1-V)p) \sim \mathrm{Dir}\left(k; \alpha + e_i\right), \quad i = 1, \ldots, k$$

so

$$\begin{aligned} VN + (1-N)P &\overset{d}{=} \sum_{i=1}^k \frac{\alpha_i}{|\alpha|}(Ve_i + (1-V)P) \sim \sum_{i=1}^k \frac{\alpha_i}{|\alpha|} Dir(k; \alpha + e_i) \\ &= Dir(k; \alpha_1, \ldots, \alpha_k) \end{aligned}$$

□

Now let us look at the Stick-Breaking construction.

Theorem 1.9 (Stick-Breaking)**.** *If* $\theta_1, \theta_2, \ldots \overset{iid}{\sim} \bar{\alpha}$ *and* $V_1, V_2, \ldots \overset{iid}{\sim}$ $\text{Beta}(1, M)$ *are independent random variables and* $W_j = V_j \prod_{l=1}^{j-1}(1 - V_l) = V_j(1 - V_1)(1 - V_2)\ldots(1 - V_{j-1})$, *then* $P := \sum_{j=1}^{\infty} W_j \delta_{\theta_j} \sim \text{DP}(M\bar{\alpha}) \sim DP(\alpha)$

For $W = (W_j)_{j=1}^{\infty}$, we also write $W \sim GEM(M)$

Proof. see Appendix □

my personal insights:
By stick-breaking, the random probability measure

$$P := W_1\delta_{\theta_1} + W_2\delta_{\theta_2} + \ldots \sim DP(\alpha)$$

then $(P(A_1), \ldots, P(A_k), \ldots\ldots) \sim Dir(\infty; \alpha(A_1), \ldots, \alpha(A_k), \ldots\ldots)$
By definition of Dirichlet distribution,
all parameters $\alpha(A_1), \ldots, \alpha(A_k), \ldots > 0$. What is more, since $P(A) > 0$ almost surely for every measurable set A with $\alpha(A) > 0$, so
$P(A_1), \ldots, P(A_k), \ldots, P(A_\infty)$ have to be strictly larger than 0

By stick-breaking, probability measure P has infinite terms, and has infinitely many locations $\theta_1, \theta_2, \ldots\ldots$ So for any partition A_∞, measure $P(A_\infty) > 0$. If the random probability measure P consists of only finite terms of weighted Dirac measure instead of the infinite terms in stick-breaking, then $P(A_\infty)$ may not larger than 0 for a random partition A_∞, therefore such random probability measure P doesn't possess Dirichlet processes. So the stick-breaking, especially its infinite terms of weighted Dirac measures, guarantees P follow Dirichlet process.

1.3 Mixtures of Dirichlet Processes (MDP)

Notice that Mixtures of Dirichlet Processes (MDP) is very different from the famous Dirichlet Process Mixtures (DPM), which we will mainly discussed later in this report.

MDP means we need to have prior for the base measure α. We usually assign respective priors to center measure $\bar{\alpha}$ and precision parameter $|\alpha|$ separately. Here, for generality, we say base measure α_ξ depends on a parameter ξ, and ξ follows a prior. So just like **equation 7**, the MDP model is :

$$\xi \sim \pi$$

$$P|\xi \sim \mathrm{DP}(\alpha_\xi) \tag{10}$$

$$X|P, \xi \overset{\text{iid}}{\sim} P$$

and $X|\xi \sim \bar{\alpha_\xi}$, analog to **equation 7**

For multiple observations, the MDP model is:

$$\xi \sim \pi, \quad P|\xi \sim \mathrm{DP}\left(\alpha_\xi\right), \quad X_1, \ldots, X_n|P, \xi \overset{\text{iid}}{\sim} P, \tag{11}$$

Now we are curious about the prior and posterior distribution, and their mean and variance, i.e., prior: $E(P(A)) =?$ $Var(P(A)) =?$
posterior: $P|\xi, X_1, \ldots, X_n \sim ?$ $P|X_1, \ldots, X_n \sim ?$
$E(P|X_1, \ldots, X_n) =?$ $Var(P|X_1, \ldots, X_n) =?$
Let me show you one by one, using law of total variance $Var(X) = E[Var(X|Y)] + Var(E[X|Y])$

prior mean

$$\begin{aligned} E(P(A)) &= E[E(P(A)|\xi)] = E[\bar{\alpha_\xi}(A)] \\ &= \int \bar{\alpha_\xi}(A) d\pi(\xi) =: \bar{\alpha}_\pi(A) \end{aligned}$$

prior variance

$$\begin{aligned} Var(P(A)) &= E[Var(P(A)|\xi)] + Var(E[P(A)|\xi]) \\ &= E[\frac{\bar{\alpha}_\xi(A)\bar{\alpha}_\xi\,(A^c)}{1+|\alpha_\xi|}] + Var(\bar{\alpha}_\xi(A)) \\ &= \int \frac{\bar{\alpha}_\xi(A)\bar{\alpha}_\xi\,(A^c)}{1+|\alpha_\xi|} d\pi(\xi) + \int \left(\bar{\alpha}_\xi(A) - \bar{\alpha}_\pi(A)\right)^2 d\pi(\xi) \end{aligned}$$

posterior distribution $P|\xi, X_1, \ldots, X_n$:

by **Theorem 1.3**

$$\begin{aligned} P|\xi, X_1, \ldots, X_n &\sim \mathrm{DP}\left(\alpha_\xi + n\mathbb{P}_n\right) \\ &=: DP(\alpha_{\xi,n}) \end{aligned}$$

where empirical distribution

$$\mathbb{P}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$$

posterior distribution $P|X_1, \ldots, X_n$:

$$\begin{aligned} P|X_1, \ldots, X_n &\sim \int (P, \xi)|X_1, \ldots, X_n d\pi(\xi) \\ &\sim \int P|\xi, X_1, \ldots, X_n \times \xi|X_1, \ldots, X_n d\pi(\xi) \end{aligned}$$

where $P|\xi, X_1, \ldots, X_n \sim DP(\alpha_{\xi,n})$ and $\xi|X_1, \ldots, X_n \propto P(X_1, \ldots, X_n|\xi)\pi(\xi)$, where $(X_1, \ldots, X_n|\xi) \sim \prod_{j=1}^{n} \frac{(\alpha_\xi + \delta_{X_1} + \cdots + \delta_{X_{j-1}})}{|\alpha_\xi| + j - 1}$ by **equation 9** posterior mean:

$$\begin{aligned}\mathrm{E}\left(P(A)|X_1,\ldots,X_n\right) &= E[\mathrm{E}\left(P(A)|\xi,X_1,\ldots,X_n\right)]\\ &= E[\bar{\alpha}_{\xi,n}(A)]\\ &= \int \bar{\alpha}_{\xi,n}(A)d\pi\left(\xi|X_1,\ldots,X_n\right) =: \tilde{P}_n(A)\end{aligned}$$

posterior variance:

$$\begin{aligned}Var(P(A)|X_1,\ldots,X_n) &= E[Var(P(A)|\xi,X_1,\ldots,X_n)] + Var(E[P(A)|\xi,X_1,\ldots,X_n]\\ &= \int \frac{\bar{\alpha}_{\xi,n}(A)\bar{\alpha}_{\xi,n}\left(A^c\right)}{1+|\alpha_\xi|+n}d\pi\left(\xi|X_1,\ldots,X_n\right) + Var(\bar{\alpha}_{\xi,n}(A)))\\ &= \int \frac{\bar{\alpha}_{\xi,n}(A)\bar{\alpha}_{\xi,n}\left(A^c\right)}{1+|\alpha_\xi|+n}d\pi\left(\xi|X_1,\ldots,X_n\right) +\\ &\int \left(\bar{\alpha}_{\xi,n}(A)-\tilde{P}_n(A)\right)^2 d\pi\left(\xi|X_1,\ldots,X_n\right)\end{aligned}$$

Now the above 6 basic quantities are finished.

2 Dirichlet Process Mixtures (DPM)

Dirichlet Process Mixtures (DPM) model can be used to estimate a density. Loosely speaking, it assumes that the density is a mixture of infinite many densities, and each observation is from one of the density independently. The DPM model is :

$$\begin{aligned} G &\sim DP(\alpha) \\ \theta_i|G &\sim G \\ X_i|\theta_i &\sim F(\theta_i) \end{aligned} \tag{12}$$

and the estimated density of X is $\int F(x,\theta)dG(\theta)$, equivalent to the above **model 12**. As we can see, infinitely many parameters θ_i can be simulated from the Dirichlet process prior G, thus these θ_i constructs infinite densities, and we want the mixture of them. I think this is one the reason that DPM model doesn't need to assume the number of different classes embedded in data, and can adapt to increasingly large data.

Naturally, we want the posterior conditional expectation of the density,

$$E[\int FdG|X_1,\ldots,X_n] = E[E(\int FdG|\theta_1,\ldots,\theta_n,X_1,\ldots,X_n)] \tag{13}$$

where

$$\begin{aligned} E(\int FdG|\theta_1,\ldots,\theta_n,X_1,\ldots,X_n) &= E(\int FdG|\theta_1,\ldots,\theta_n) \\ &= E[\int FdDP(\alpha+\sum_{j=1}^{n}\delta_{\theta_j})] \\ &= \frac{1}{|\alpha|+n}\left[\int Fd\alpha+\sum_{j=1}^{n}F\left(\theta_j\right)\right] \end{aligned} \tag{14}$$

the second line is by **Theorem 1.3**, and the last line is by **Proposition 1.2**

Now we want to know

$$\begin{aligned} E[\int FdG|X_1,\ldots,X_n] &= E_{\theta_{1:n}|X_{1:n}}[\frac{1}{|\alpha|+n}\left[\int Fd\alpha+\sum_{j=1}^{n}F\left(\theta_j\right)\right]] \\ &= \frac{1}{|\alpha|+n}\int Fd\alpha+\frac{1}{|\alpha|+n}E_{\theta_{1:n}|X_{1:n}}[\sum_{j=1}^{n}F\left(\theta_j\right)] \end{aligned}$$

where the expectation, in last line, is with respect to the posterior distribution of $\theta_1,\ldots,\theta_n|X_1,\ldots,X_n$

The analytic formula of above will be shown in Appendix. The analytic formula is way too complicated and of limited practical importance. So in practice, to get $E[\int FdG|X_1,\ldots,X_n]$, we use a simulated technique. The general scheme is as followings:

1. we repeatedly draw realizations from the posterior distribution of $\theta_1,\ldots,\theta_n|X_1,\ldots,X_n$

2. evaluate **equation 14** for each realization of $(\theta_1,\ldots,\theta_n)$ from step 1, and average out over many evaluated results, then we get what we want: the estimation for the posterior conditional expectation of the density, i.e., the estimation of $E[\int FdG|X_1,\ldots,X_n]$

Now, we have a very important problem to solve: how to simulate $(\theta_1,\ldots,\theta_n)$ from posterior distribution $\theta_1,\ldots,\theta_n|X_1,\ldots,X_n$???

Gibbs Sampling

Then, what is $\theta_i|\theta_-i,X_1,\ldots,X_n\sim$???

Recall the DPM model:

$$\begin{aligned} G &\sim DP(\alpha) \\ \theta_i|G &\sim G \\ X_i|\theta_i &\sim F(\theta_i) \end{aligned}$$

For measurable sets A and B,

$$\begin{aligned}
\mathrm{E}\left(\mathbb{1}_A\left(X_i\right)\mathbb{1}_B\left(\theta_i\right)|\theta_{-i},X_{-i}\right) &= \mathrm{E}\left(\mathrm{E}\left(\mathbb{1}_A\left(X_i\right)\mathbb{1}_B\left(\theta_i\right)|G,\theta_{-i},X_{-i}\right)|\theta_{-i},X_{-i}\right)\\
&= \mathrm{E}\left(\mathrm{E}\left(\mathbb{1}_A\left(X_i\right)\mathbb{1}_B\left(\theta_i\right)|G\right)|\theta_{-i},X_{-i}\right)\\
&= E(\int\int \mathbb{1}_A\left(X_i\right)\mathbb{1}_B(\theta)F(X_i;\theta)d\mu(x)dG(\theta)|\theta_-i,X_-i)\\
&= \frac{\iint \mathbb{1}_A(x_i)\mathbb{1}_B(\theta)F(x_i;\theta)d\mu(x)d\left(\alpha+\sum_{j\neq i}\delta_{\theta_j}\right)(\theta)}{|\alpha|+n-1}
\end{aligned}$$

so the conditional distribution function of $X_i,\theta_i|X_{-i},\theta_{-i}$ is:

$$X_i,\theta_i|X_{-i},\theta_{-i}\sim\frac{1}{|\alpha|+n-1}F(x_i;\theta)\mu\times(\alpha+\sum_{j\neq i}\delta_{\theta_j})$$

then probability measure

$$\begin{aligned}
\mathrm{P}\left(\theta_i\in B|X_i,\theta_{-i},X_{-i}\right) &= \frac{P(\theta_i\in B,x_i|\theta_{-i},x_{-i})}{P(x_i|\theta_{-i},x_{-i})}\\
&= \frac{P(\theta_i\in B,x_i|\theta_{-i},x_{-i})}{\int P(\theta_i,x_i|\theta_{-i},x_{-i})}\\
&= \frac{\int_B\frac{1}{|\alpha|+n-1}F\left(X_i;\theta\right)\mu d\left(\alpha+\sum_{j\neq i}\delta_{\theta_j}\right)(\theta)}{\int\frac{1}{|\alpha|+n-1}F\left(X_i;\theta\right)\mu d\left(\alpha+\sum_{j\neq i}\delta_{\theta_j}\right)(\theta)}\\
&= \frac{\int_B F\left(X_i;\theta\right)d\left(\alpha+\sum_{j\neq i}\delta_{\theta_j}\right)(\theta)}{\int F\left(X_i;\theta\right)d\left(\alpha+\sum_{j\neq i}\delta_{\theta_j}\right)(\theta)}
\end{aligned}$$

so

$$\theta_i|X_i,\theta_{-i},X_{-i}\sim\frac{F\left(X_i;\theta\right)\left(\alpha+\sum_{j\neq i}\delta_{\theta_j}\right)}{\int F\left(X_i;\theta\right)d\left(\alpha+\sum_{j\neq i}\delta_{\theta_j}\right)(\theta)}$$

then

$$\begin{aligned}\theta_i|X_i,\theta_{-i},X_{-i} &\sim \sum_{j\neq i} F(X_i;\theta_j)\delta_{\theta_j} + F(X_i;\theta)\alpha \\ &\sim \sum_{j\neq i} F(X_i;\theta_j)\delta_{\theta_j} + \int F(X_i;\theta)d\alpha(\theta)\frac{F(X_i;\theta)\alpha}{\int F(X_i;\theta)d\alpha(\theta)}\end{aligned}$$

Since distribution of $\theta|X_i \overset{d}{=}$ distribution of$\frac{X_i|\theta\times\theta}{\int X_i|\theta\times\theta d\theta}$,
so density:

$$\begin{aligned}dH_i &:= dH(\theta|X_i) \\ &= \frac{F(X_i;\theta)d\alpha(\theta)}{\int F(X_i;\theta)d\alpha(\theta)} \\ &\propto F(X_i;\theta)d\alpha(\theta)\end{aligned}$$

so conditional distribution for use in Gibbs Sampling:

$$\theta_i|\theta_{-i},X_i \sim \sum_{j\neq i} q_{i,j}\delta_{\theta_j} + r_iH_i \qquad (15)$$

where

$$\begin{aligned}Hi &= F(X_i;\theta)\alpha \\ q_{i,j} &= bF(X_i;\theta_j) \\ r_i &= b\alpha\int F(X_i;\theta)d\alpha(\theta)\end{aligned}$$

and b is a constant such that $\sum_{j\neq i} q_{i,j} + r_i = 1$
This equation (15) is a very important part in Gibbs Sampling when we simulate posterior distribution in DPM.

3 Markov Chain Sampling for Dirichlet Process Mixture

3.1 limit of finite mixture models

Recall the Dirichlet Process Mixtures (DPM) model:

$$\begin{aligned} G &\sim DP(G_0, \alpha) \\ \theta_i | G &\sim G \\ X_i | \theta_i &\sim F(\theta_i) \end{aligned} \tag{16}$$

where G_0 is center measure, α is precision parameter.

We have shown the conditional prior distribution of θ_i:

$$\theta_i | \theta_1, \ldots, \theta_{i-1} \ \sim \frac{1}{\alpha + i - 1} \sum_{j=1}^{i-1} \delta_{\theta_j} + \frac{\alpha}{\alpha + i - 1} G_0 \tag{17}$$

Before talking about Gibbs sampling for DPM, I'd like to introduce limit of finite mixture models, and show its equivalence to DPM, because this representation can help us construct Gibbs sampling procedures.

Here is the finite mixture models with components $K \to \infty$

$$\begin{aligned} \boldsymbol{p} &\sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K) \\ \phi_c &\sim G_0 \\ c_i | \boldsymbol{p} &\sim \text{Discrete}\left(p_1, \ldots, p_K\right) \\ x_i | c_i, \boldsymbol{\phi} &\sim F\left(\phi_{c_i}\right) \end{aligned} \tag{18}$$

where $\boldsymbol{\phi}$ is the collection of all ϕ_c, and the number of elements in this collection is infinite.

$$\begin{aligned}
&P\left(c_i = c|c_1,\ldots,c_{i-1}\right)\\
&\quad= \frac{P\left(c_1,\ldots,c_{i-1},c_i=c\right)}{P\left(c_1,\ldots,c_{i-1}\right)}\\
&\quad= \frac{\int P\left(c_1,\ldots,c_{i-1},c_i=c,\boldsymbol{p}\right)d\boldsymbol{p}}{\int P\left(c_1,\ldots,c_{i-1},\boldsymbol{p}\right)d\boldsymbol{p}}\\
&\quad= \frac{\int p_{c_1}\ldots p_{c_{i-1}}p_c\Gamma(\alpha)\Gamma(\alpha/K)^{-K}p_1^{(\alpha/K)-1}\ldots p_K^{(\alpha/K)-1}d\boldsymbol{p}}{\int p_{c_1}\ldots p_{c_{i-1}}\Gamma(\alpha)\Gamma(\alpha/K)^{-K}p_1^{(\alpha/K)-1}\ldots p_K^{(\alpha/K)-1}d\boldsymbol{p}}\\
&\quad= \frac{\Gamma(\frac{\alpha}{K}+n_{i,c}+1)}{\Gamma(\frac{\alpha}{K}K+i)}/\frac{\Gamma(\frac{\alpha}{K}+n_{i,c})}{\Gamma(\frac{\alpha}{K}K+i-1)}\\
&\quad= \frac{n_{i,c}+\alpha/K}{i-1+\alpha}
\end{aligned}$$

where n_i,c is the number of c_j for $j < i$ that are equal to c

As $K \rightarrow \infty$,

$$P\left(c_i = c|c_1,\ldots,c_{i-1}\right) \rightarrow \frac{n_{i,c}}{i-1+\alpha} \tag{19}$$

$$\begin{aligned}
P\left(c_i \neq c_j \text{ for all } j < i|c_1,\ldots,c_{i-1}\right) &= 1 - \sum_c P\left(c_i = c|c_1,\ldots,c_{i-1}\right)\\
&\rightarrow 1 - \frac{\sum_c n_{i,c}}{\alpha+i-1}\\
&= 1 - \frac{i-1}{\alpha+i-1}\\
&= \frac{\alpha}{i-1+\alpha}
\end{aligned} \tag{20}$$

So **equation 19** is the probability that a new observation comes from a existing category; **equation 20** is the probability that a new observation comes from a new category that has never been showed up before.

Compare DPM model (16) and finite mixture models (18), as $K \to \infty$, and let $\theta_i = \phi_{c_i}$, with equation (17), we can see that

1. $P\left(c_i = c | c_1, \ldots, c_{i-1}\right) \to \frac{n_{i,c}}{i-1+\alpha}$ corresponds to $\theta_i | \theta_1, \ldots, \theta_{i-1} \sim \delta_{\theta_j}$, with probability $\frac{1}{\alpha+i-1}$, $j = 1, \ldots, i-1$
2. $P\left(c_i \neq c_j \text{ for all } j < i | c_1, \ldots, c_{i-1}\right) \to \frac{\alpha}{i-1+\alpha}$ corresponds to $\theta_i | \theta_1, \ldots, \theta_{i-1} \sim G_0$, with probability $\frac{\alpha}{\alpha+i-1}$. Since distribution (center measure) G_0 is atomless, $\theta_i | \theta_1, \ldots, \theta_{i-1} \sim G_0$ means θ_i is different from $\theta_1, \ldots, \theta_{i-1}$

Therefore, DPM model (16) and finite mixture model (18) are equivalent, if $K \to \infty$.

3.2 Gibbs Samplings for Dirichlet Process Mixture

Recall the Dirichlet Process Mixtures (DPM) model:

$$\begin{aligned} G &\sim DP(G_0, \alpha) \\ \theta_i | G &\sim G \\ X_i | \theta_i &\sim F(\theta_i) \end{aligned}$$

and the finite mixture model, $\theta_i = \phi_{c_i}$:

$$\begin{aligned} \boldsymbol{p} &\sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K) \\ \phi_c &\sim G_0 \\ c_i | \boldsymbol{p} &\sim \text{Discrete}\left(p_1, \ldots, p_K\right) \\ x_i | c_i, \boldsymbol{\phi} &\sim F\left(\phi_{c_i}\right) \end{aligned}$$

and also the important conditional distribution for use in Gibbs sampling, from equation (15) in Section 2:

$$\theta_i | \theta_{-i}, X_i \sim \sum_{j \neq i} q_{i,j} \delta_{\theta_j} + r_i H_i$$

where $Hi = F(X_i;\theta)\alpha; q_{i,j} = bF(X_i;\theta_j); r_i = b\alpha\int F(X_i;\theta)d\alpha(\theta)$, and b is a constant such that $\sum_{j\neq i} q_{i,j} + r_i = 1$

For Gibbs sampling to be feasible, computing integral $\int F(X_i;\theta)d\alpha(\theta)$ and sampling from distribution $F(X_i;\theta)\alpha$ must be feasible. If G_0 is the conjugate prior for the likelihood given by F, it is feasible.

So first method to simulate simply follows equation (15):

Algorithm 3.1. Repeatedly sample as follows:

- For $i = 1, \ldots, n$: draw a new value from $\theta_i|\theta_{-i}, X_i \sim \sum_{j\neq i} q_{i,j}\delta_{\theta_j} + r_i H_i$

By this algorithm, we have many realizations of $(\theta_1, \ldots, \theta_n)$ from posterior distribution $\theta_1, \ldots, \theta_n|X_1, \ldots, X_n$, and substitute them into the general algorithm in Section 2 in this report, then we can get an estimated density, i.e., $E[\int FdG|X_1, \ldots, X_n]$

This **Algorithm 3.1** is straightforward, but rather slow and inefficient. Because there are often groups of observation X_i that with high probability are associated with the same θ. But in every loop of this Algorithm 3.1, we still simulate from θ_1 to θ_n one by one. It we take θ_i shared by a group of observation as a block, then it would be more efficient if we simulate θ block by block in every loop of the algorithm.

To achieve it, we need to consider the indicator variables c

$$\begin{aligned} P\left(c_i = c|c_{-i}, x_i, \boldsymbol{\phi}\right) &\propto P(x_i|c_{-i}, c_{-i} = c, \boldsymbol{\phi}) \times P(c_i = c|c_{-i}, \boldsymbol{\phi}) \\ &\propto F(x_i, \phi_c)\frac{n_{-i,c} + \alpha/K}{n - 1 + \alpha} \\ &= bF\left(x_i, \phi_c\right)\frac{n_{-i,c} + \alpha/K}{n - 1 + \alpha} \end{aligned}$$

where $n_{-i,c}$ is the number of c_j for $j \neq i$ that are equal to c, and b is a normalizing constant.

When $K \to \infty$,
if $c = c_j$ for some $j \neq i$, $P\left(c_i = c | c_{-i}, x_i, \boldsymbol{\phi}\right) = b\frac{n_{-i,c}}{n-1+\alpha} F\left(x_i, \phi_c\right)$;
and
$\mathrm{P}(c_i \neq c_j$ for all $j \neq i | c_{-i}, x_i, \boldsymbol{\phi})$
$= P\left(c_i \neq c_j \text{ for all } j \neq i | c_{-i}, x_i\right)$
$\propto P(x_i | c_i \neq c_j \text{ for all } j \neq i, c_{-i}) \times P(c_i \neq c_j \text{ for all } j \neq i | c_{-i})$
$\propto \frac{\alpha}{n-1+\alpha} \int F\left(x_i, \phi\right) dG_0(\phi)$
$= b\frac{\alpha}{n-1+\alpha} \int F\left(x_i, \phi\right) dG_0(\phi)$

Note that $\theta_i = \phi_{c_i}$, so the following algorithm simulates ϕ_{c_i} instead of simulating θ_i like in Algorithm 3.1

Algorithm 3.2. $\boldsymbol{c} = (c_1, \ldots, c_n)$ and $\boldsymbol{\phi} = (\phi_c : c \in \{c_1, \ldots, c_n\})$
Repeatedly sample as follows:

1. For $i = 1, \ldots, n$:
    - If $n_{-i,c_i} = 0$, remove ϕ_{c_i} from $\boldsymbol{\phi}$
    - Draw a new value for c_i from $c_i | c_{-i}, y_i, \boldsymbol{\phi}$ as
      $P\left(c_i \neq c_j \text{ for all } j \neq i | c_{-i}, x_i, \boldsymbol{\phi}\right) = b\frac{\alpha}{n-1+\alpha} \int F\left(x_i, \phi\right) dG_0(\phi)$
    - If the new c_i isn't associated with any other observation, draw a value for $\phi_{c_i} \sim H_i = F(x_i, \theta) G_0$, and add it to $\boldsymbol{c}$ and $\boldsymbol{\phi}$
2. For all $c \in \{c_1, \ldots, c_n\}$: draw a new value from
   $\phi_c | \{\text{all } x_i \text{ s.t } c_i = c\} \sim \prod_{i \text{ s.t } c_i = c} F(\phi_{c_i}) G_0$

As was the case for Gibbs sampling **Algorithm 3.1**, **Algorithm 3.2** is feasible if G_0 is the conjugate prior for the likelihood given by F.

Sometimes we don't care about ϕ_c, but only care about $c_1, \ldots, c_n$. For example, in clustering, $c_1, \ldots, c_n$ can tell us partitions of observations. So we can integrate over ϕ_c and eliminate them.

If $c = c_j$ for some $j \neq i$:

$$\begin{aligned}
P\left(c_i = c | c_{-i}, x_i\right) &\propto \int P(c_i = c, \phi_c | c_{-i}, x_i) d\mu \\
&\propto \int P(c_i = c | \phi_c, c_{-i}, x_i) P(\phi_c | c_{-i}, x_i) d\mu \\
&\propto \int \frac{n_{-i,c}}{n-1+\alpha} F(x_i, \phi_c) P(\phi_c | c_{-i}, x_i) d\mu \\
&\propto \frac{n_{-i,c}}{n-1+\alpha} \int F\left(y_i, \phi\right) dH_{-i,c}(\phi)
\end{aligned}$$

where $H_{-i,c}$ is the posterior distribution of ϕ based on the prior G_0 and all observations x_j for which $j \neq i$ and $c_j = c$; in other words

$$H_{-i,c} \sim \phi_c | \{c_{-i}, y_i\} \sim \prod_{i \text{ s.t } c_i = c} F(\phi_{c_i}) G_0$$

And as we just showed,

$$P\left(c_i \neq c_j \text{ for all } j \neq i | c_{-i}, x_i\right) = b \frac{\alpha}{n-1+\alpha} \int F\left(x_i, \phi\right) dG_0(\phi)$$

Algorithm 3.3. Repeatedly sample as follows:

- For $i = 1, \ldots, n$: draw a new value from $c_i | c_{-i}, x_i$ as
  - If $c = c_j$ for some $j \neq i : P\left(c_i = c | c_{-i}, x_i\right) = b \frac{n_{-i,c}}{n-1+\alpha} \int F\left(x_i, \phi\right) dH_{-i,c}(\phi)$
  - $P\left(c_i \neq c_j \text{ for all } j \neq i | c_{-i}, x_i\right) = b \frac{\alpha}{n-1+\alpha} \int F\left(x_i, \phi\right) dG_0(\phi)$

4 Hierarchical Dirichlet Processes (HDP)

4.1 HDP

Before showing the model of Hierarchical Dirichlet Process (HDP), I'd like to describe it with some simple words to have some sense.

Dirichlet Processes mixture (DPM) works like: given a group, we figure out clusters within this group and don't know the number of clusters in advance.

For HDP, imagine we have a large number of global elements, or say, factors. There are several groups in these elements and each group has multiple clusters. So groups may share some same factors with groups, but with different proportions of clusters within different groups. In the hierarchical DP, the value of the factors are shared between the groups as well as within the groups. This is a key property of hierarchical DP.

Hierarchical DP:

$$\begin{aligned} G_0|\gamma, H &\sim DP(\gamma, H) \\ G_j|\alpha_0, G_0 &\sim DP(\alpha_0, G_0) \\ \theta_{ji}|G_j &\sim G_j \\ X_{ji}|\theta_{ji} &\sim F(\theta_{ji}) \end{aligned} \tag{21}$$

G_0 is a global random probability measure
H is base probability measure
γ and α_0 are concentration parameters
local random probability random measure G_j is one for each group
we often put vague gamma priors on γ and α_0

The baseline probability measure H provides the prior distribution for all factors θ_{ji}

The global distribution G_0 varies around the prior H. The actual distribution G_j in the j-th group deviates from G_0, with the amount of variability governed by α_0. If we expect the variability in different groups to be different, then we can use a separate precision parameter α_j for each group j

4.2 Stick-Breaking Construction

Hierarchical DP:

$$\begin{aligned} G_0|\gamma, H &\sim DP(\gamma, H) \\ G_j|\alpha_0, G_0 &\sim DP(\alpha_0, G_0) \\ \theta_{ji}|G_j &\sim G_j \\ X_{ji}|\theta_{ji} &\sim F(\theta_{ji}) \end{aligned}$$

By **Theorem 1.9**(Stick-Breaking), we have

$$\begin{aligned} G_0 &= \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k} \\ G_j &= \sum_{k=1}^{\infty} \pi_{jk} \delta_{\phi_k} \end{aligned}$$

where $\phi_k \overset{iid}{\sim} H$ and $\boldsymbol{\beta} = (\beta_k)_{k=1}^{\infty} \overset{iid}{\sim} GEM(\gamma)$

The above two representation equations indicate that value of factors are shared between groups, because each G_j shares the same Dirac measures δ_{ϕ_k}

Sticking-Breaking construction for HDP:

$$\begin{aligned} \boldsymbol{\beta}|\gamma &\sim \mathrm{GEM}(\gamma) \\ \boldsymbol{\pi}_j|\alpha_0, \boldsymbol{\beta} &\sim \mathrm{DP}\left(\alpha_0, \beta\right) \\ z_{ji}|\boldsymbol{\pi}_j &\sim \boldsymbol{\pi}_j \\ \phi_k|H &\sim H \\ x_{ji}|z_{ji}, (\phi_k)_{k=1}^{\infty} &\sim F\left(\phi_{zj}\right) \end{aligned} \tag{22}$$

and $\theta_{ji} = \phi_{z_{ji}}$

4.3 The Chinese Restaurant Franchise

Hierarchical DP:

$$\begin{aligned} G_0|\gamma, H &\sim DP(\gamma, H) \\ G_j|\alpha_0, G_0 &\sim DP(\alpha_0, G_0) \\ \theta_{ji}|G_j &\sim G_j \\ X_{ji}|\theta_{ji} &\sim F(\theta_{ji}) \end{aligned}$$

$\theta_{ji} = \phi_{z_{ji}}$
$G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k}$
$G_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\phi_k}$ and $\phi_k \overset{iid}{\sim} H$

Imagine there is a very long street with infinite many restaurants. A new customer may enter a restaurant already having customers, or may enter an empty restaurant. These restaurants share a global menu of infinite many dishes on it. Each restaurant can have infinite many tables. The dish on a table is ordered by the first customer sitting at this table, and all people sitting at a table share the same dish. For a new customer entering into a restaurant, he can choose to sit at a table already with people, or to sit at an empty new table and order a dish from the global menu. This is basically the Chinese Restaurant Franchise.

$\phi_1, \ldots, \phi_k \overset{iid}{\sim} H$ and $\phi_1, \ldots, \phi_k, \ldots$ are global menu of dishes

ψ_{jt} is the dish served at table t in restaurant j

t_{ji} is the index of ψ_{jt} associated with θ_{ji}, i.e., $\psi_{jt_{ji}}$
k_{jt} is the index of ϕ_k associated with ψ_{jt}, i.e., $\phi_{k_{jt}}$

number of customers: n_{jtk}, $n_{jt\cdot}$, $n_{j\cdot k}$
number of tables: m_{jk}, $m_{j\cdot}$, $m_{\cdot k}$ $m_{\cdot\cdot}$

$$\theta_{ji}|\theta_{j1}, \ldots, \theta_{j,i-1}, \alpha_0, G_0 \sim \sum_{t=1}^{m_j} \frac{n_{jt}}{i-1+\alpha_0} \delta_{\psi_{jt}} + \frac{\alpha_0}{i-1+\alpha_0} G_0 \tag{23}$$

$$\psi_{jt}|\psi_{11}, \psi_{12}, \ldots, \psi_{21}, \ldots, \psi_{jt-1}, \gamma, H \sim \sum_{k=1}^{K} \frac{m \cdot k}{m_{\cdot\cdot} + \gamma} \delta_{\phi_k} + \frac{\gamma}{m_{\cdot\cdot} + \gamma} H \tag{24}$$

To obtain samples of θ_{ji}: for each j, i, first sample θ_{ji} using equation (23); if a new sample from G_0 is needed, then use equation (24) to obtain a new sample ψ_{jt} and set $\theta_{ji} = \psi_{jt}$

For the inference and posterior sampling method in Chinese Restaurant Franchise, Section 5.1 of paper Hierarchical Dirichlet Processes (Teh, 2006) has a well-written and detailed explanation.

5 autonomous multi-vehicle interaction scenarios

Utilizing the Dirichlet Processes Mixture (DPM) knowledge covered in this report, a paper from Carnegie Mellon University talks about how to cluster motion patterns of multi-vehicle and predict trajectories of multi-vehicles, without any restriction on the number of motion patterns in the dataset.

Let me introduce their work briefly:

The dataset consists of trajectories of many vehicles on a road during a time slot. Segment this time slot into many pieces with tiny time-intervals. For example, a dataset has 1000 frames of time-sequence data with discretization of 0.5 second. Denote a dataset of N frames as $\mathcal{S} = \{s_i | i = 1, 2, \ldots, N\}$

Each frame is an observation. It contains location and velocity of every vehicle in that frame. The number of vehicles in a frame is unknown and uncertain. Thus a frame is actually a motion pattern, modeled by Gaussian Process, which isn't our focus in this report.

In the paper, the proposed multi-vehicle motion model is defined as a mixture G of infinite motion patterns

$$G = \sum_{k=1}^{\infty} \pi_k g_k \tag{25}$$

where each mixture component g_k is a motion pattern and is defined by a Gaussian Process (GP). So for any frame, s_i is generated by G. Each of these N frames (observations) is assigned to one of the motion patterns. An indicator variable z_i is introduced where $z_i = k$ means the frame s_i is associated with the latent motion pattern g_k.

To cluster, simulate, and predict motion patterns, the paper constructs a Dirichlet Processes Mixture (DPM) model, but doesn't exhibit the model explicitly in the paper. So let me write it out here. I use the limit of finite mixture model representation.

$$\begin{aligned}\boldsymbol{p} &\sim \text{Dirichlet}(\alpha/K, \ldots, \alpha/K)\\ g_z &\sim G\\ z_i|\boldsymbol{p} &\sim \text{Discrete}\left(p_1, \ldots, p_K\right)\\ s_i|z_i, \boldsymbol{g} &\sim g_{z_i}\end{aligned} \tag{26}$$

where $\boldsymbol{g}$ is the collection of g_z, and $K \rightarrow \infty$

With materials of Section 3 in this report, the predictive distribution of g_{z_i} conditioned on the other motion patterns $g_{z_{-i}} = \{g_{z_k}|z_k \in z_{-i}\}$, where $z_{-i} = \{z_k|k = 1, 2, \cdots, n_k, k \neq i\}$, is

$$g_{z_i}|g_{z_{-i}}, z_{-i} \sim \frac{1}{\alpha + N - 1}\left(\alpha G + \sum_{z_k \in z_{-i}} \Delta\left(g_{z_k}\right)\right)$$

where α is the concentration parameter and $\Delta\left(g_{z_k}\right)$ is the point mass at g_{z_k}. The following work again accords with what we talk about in Section 3 of this report.

The performance on two realistic dataset is good in the paper. However, I think there is a problem embedded in the Dirichlet Processes Mixture model constructed in the paper, i.e. **equation 26**

In **equation 26**, motion patterns $g_z \overset{iid}{\sim} G$ and
frames $s_i|z_i, \boldsymbol{g} \overset{iid}{\sim} g_{z_i}$. But in reality, I don't think multi-vehicle frames at consecutive time points are independent. So this is where we can work on. A potential direction is to consider a Sticky HDP-HMM model (Fox, 2011), involving hierarchical Dirichlet Process and hidden Markov model.

6 Reference

Fundamentals of Nonparametric Bayesian Inference, by Subhashis Ghosal & Aad van der Vaart

Markov Chain Sampling Methods for Dirichlet Process Mixture Models, by Radford M. Neal

Hierarchical Dirichlet Processes, by Yee Whye Teh, et al.

Infinite Mixtures of Gaussian Process Experts, by Carl Edward Rasmussen & Zoubin Ghahramani

Modeling Multi-Vehicle Interaction Scenarios Using Gaussian Random Field, by Yaohui Guo, Ding Zhao, et al.

A Sticky HDP-HMM with Application to Speaker Diarization, by Emily B. Fox, et al.

7 Appendix

Still working on it. Latex takes me too much time.