Школа анализа данных

Семинар 10. Процессы Дирихле

Курс: Байесовские методы в машинном обучении, осень 2015

1. Рассматриваются две независимые случайные величины $X \sim \mathcal{G}(a, 1)$ и $Y \sim \mathcal{G}(b, 1)$. Доказать, что случайная величина $Z = X + Y$ имеет распределение $\mathcal{G}(a + b, 1)$.

2. Рассматривается модель смеси распределений с априорным распределением из процесса Дирихле, представленным с помощью усечённой (truncated) модели stick-breaking:
$$p(X, Z, V, \Theta|\alpha, H) = \left[\prod_{n=1}^{N}\prod_{k=1}^{T} p(\boldsymbol{x}_n|\boldsymbol{\theta}_k)^{[z_n=k]}\left(v_k\prod_{j=1}^{k-1}(1-v_j)\right)^{[z_n=k]}\right]\prod_{k=1}^{T-1}\mathrm{Beta}(v_k|1,\alpha)\prod_{k=1}^{T}p_H(\boldsymbol{\theta}_k).$$
Здесь предполагается, что $p(v_T = 1) = 1$, а индикатор $[z_n = k]$ равен 1, если объект $\boldsymbol{x}_n$ принадлежит к k-ой компоненте смеси (k-ому кластеру). Требуется для данной модели выписать формулы для поиска вариационного приближения вида
$$p(Z, V, \Theta|X) \approx q(Z)q(V, \Theta).$$

3. Рассматривается модель смеси распределений с априорным распределением Дирихле:
$$p(X, Z, \boldsymbol{w}, \Theta|\boldsymbol{\alpha}, H) = \left[\prod_{n=1}^{N}\prod_{k=1}^{K}\Big(w_k p(\boldsymbol{x}_n|\boldsymbol{\theta}_k)\Big)^{[z_n=k]}\right]\mathrm{Dir}(\boldsymbol{w}|\boldsymbol{\alpha})\prod_{k=1}^{K}p_H(\boldsymbol{\theta}_k).$$
Требуется для данной модели найти коллапсированное распределение $p(Z)$ и формулу для одномерного распределения $p(z_N = k|z_{N-1}, \dots, z_1)$. Для модели $p(X, Z, \Theta|\boldsymbol{\alpha}, H)$ требуется выписать формулу генерации выборки по схеме Гиббса из апостериорного распределения $p(Z, \Theta|X, \boldsymbol{\alpha}, H)$.