CS 698S

Homework - 4

Pralik Mishra (14493)

Problem 1

Let z_d denote the topic assignment of all the words of the document d.

$$\therefore p(z_d|\alpha) = p(z_d|\theta_d) \cdot p(\theta_d|\alpha)$$

$$\therefore p(z_d|\theta_d) \cdot p(\theta_d|\alpha) = \frac{\Gamma(\alpha)}{\Gamma(\alpha/k)^k} \cdot \prod_{k=1}^{k} \Gamma\left(\alpha/k + \sum_{n=1}^{N_d} z_{nk}^d\right)$$

$$= \frac{\Gamma(\alpha)}{\Gamma(\alpha/k)^k} \cdot \frac{\prod_{k=1}^{k} \Gamma(\alpha/k + N_d^{(k)})}{\Gamma(\alpha + N_d)}$$

Let $z_{d\setminus n}$ denote the topic assignment to every word in d document except the n^{th} word. So, since the documents are independent,

$$p(z_{d,n} = k | Z_{-(d,n)}) = p(z_{d,n} = k | z_{d\setminus n})$$

$$\therefore p(z_{d,n} = k | z_{d\setminus n}, \alpha) = \frac{p(z_d|\alpha)}{p(z_{d\setminus n}|\alpha)}$$

$$\therefore p(z_{d,n}=k \mid z_{d,-n}, \alpha) = \frac{N_d^{(k)} - 1 + \alpha/k}{N_d - 1 + \alpha}$$

This makes intuitive sense as probability that the n^{th} word of a document belongs to the k^{th} topic depends on number of words in the document that belong to the same topic. Thus more words in a document of same topic implies higher probability of an unknown word to belong to that topic.

- Let W_k denote the vocabulary assignment of all the words in all the documents that belong to the k^{th} topic.

$$\therefore p(W_k \mid \eta) = p(W_k \mid \beta_k) \cdot p(\beta_k \mid \eta)$$

$$= \frac{\Gamma(\eta)}{\Gamma(\eta/V)^V} \cdot \frac{\prod_{v=1}^{V} \Gamma(\eta/V + M_k^{(v)})}{\Gamma(\eta + M_k)}$$

Here, $M_k^{(v)}$ denotes the total number of words in all the documents same as the v^{th} word in the vocabulary and M_k denotes the total number

of words in all the documents assigned the k^{th} topic.

$$\therefore p(w_{d,n} = v \mid z_{d,n} = k, w_{-(d,n)}, z_{-(d,n)})$$

$$= p(w_{d,n} = v \mid z_{d,n} = k, w_{k-(d,n)})$$

$$\therefore p(w_{d,n} = v \mid z_{d,n} = k, w_{k-(d,n)}) = \frac{p(w_k \mid n)}{p(w_{k-(d,n)} \mid n)}$$

$$= \frac{M_k^{(v)} - 1 + n/V}{M_k - 1 + n}$$

This makes intuitive sense because if more words of k^{th} topic are same as the v^{th} word in vocabulary V, then probability of the unknown word to be the same would be higher.

PROBLEM 2

We know,

$$p(z_n = k | z_n) = \frac{n_k + \alpha | k-1}{\alpha + N - 1}$$

$$\therefore p(x_n | z_n = k, W_{-n}, Z_{-n})$$

$$= \int p(x_n | z_n = k, \mu_k, \Sigma_k) \cdot p(\mu_k, \Sigma_k | W_{-n}, Z_{-n}) d\mu_k d\Sigma_k$$

$$= T\left(x | m_N, \frac{k_N + 1}{k_N (\nu_N - D + 1)}, S_N, \nu_N - D + 1\right)$$

where N is the number of observations in cluster k and $\bar{x}$ is the mean of those observations and,

$$m_N = \frac{k_0 m_0 + N\bar{x}}{k_N}$$

$$k_N = k_0 + N$$

$$\nu_N = \nu_0 + N$$

$$S_N = S_0 + \sum_{i=1}^{N} x_i x_i^T + k_0 m_0 m_0^T - k_N m_N m_N^T$$

so, when

$K \to \infty$,

$$P(z_n = k \mid z_n) = \frac{n_k - 1}{\alpha + N - 1}$$

where, k is the cluster already seen so far.

PROBLEM 3

$$P(\vec{z} \mid \alpha) = \int_0^1 \int_0^1 \cdots \int_0^1 \left(\prod_{n=1}^{N} P(z_{nk} \mid \pi_k) \right) P(\pi_k \mid \alpha) \, d\pi_k \, d\pi_1 \, d\pi_2 \cdots$$

$$= \int_0^1 \int_0^1 \cdots \int_0^1 \left(\prod_{n=1}^{N} \pi_k^{z_{nk}} (1 - \pi_k)^{1 - z_{nk}} \right) \cdot \frac{\pi_k^{\alpha/k - 1} \Gamma(\alpha/k + 1)}{\Gamma(\alpha/k)} \cdot d\pi_k \, d\pi_1 \, d\pi_2 \cdots$$

$$= \int_0^1 \int_0^1 \cdots \int_0^1 \frac{\alpha}{k} \, \pi_k^{\sum_{n=1}^{N} z_{nk} + \alpha/k - 1} (1 - \pi_k)^{N - \sum_{n=1}^{N} z_{nk}} \, d\pi_k \, d\pi_1 \, d\pi_2 \cdots$$

$$= \frac{\alpha^k}{k^k} \prod_{k=1}^{k} \left(\frac{\Gamma\left(\sum_n z_{nk} + \alpha/k\right) \Gamma\left(N + 1 - \sum_n z_{nk}\right)}{\Gamma(N + 1 + \alpha/k)} \right)$$

$$P(\vec{z} \mid \alpha) = \frac{\alpha^k}{\left(k \times \Gamma(N + 1 + \alpha/k)\right)^k} \cdot \prod_{k=1}^{k} \left(\Gamma\left(\sum_n z_{nk} + \alpha/k\right) \times \Gamma\left(N + 1 - \sum_n z_{nk}\right) \right.$$

- Let z_k be the entries of the k^{th} column

$$p(z_{nk} \mid z_{-nk}) = \frac{p(z_k \mid \alpha)}{p(z_{-nk} \mid \alpha)}$$

$$= \frac{1}{\left(N + \frac{\alpha}{K}\right)} \times \frac{\Gamma\left(\sum_n z_{nk} + \alpha/K\right)\Gamma\left(N + 1 - \sum_n z_{nk}\right)}{\Gamma\left(\sum_l z_{lk} - z_{nk} + \alpha/K\right)\Gamma\left(N - \sum_l z_{lk} + z_{nk}\right)}$$

$$= \frac{1}{\left(N + \frac{\alpha}{K}\right)} \times \left(\sum_n z_{nk} + \alpha/K - 1\right)^{z_{nk}} \left(\frac{N - \sum_n z_{nk}}{N + \alpha/K}\right)^{1 - z_{nk}}$$

- $$E\left[\sum_{n=1}^{N} z_{nik}\right] = \sum_{n=1}^{N} E[z_{nk}]$$

$$= \sum_{n=1}^{N} 1 \cdot p(z_{nk} = 1)$$

$$= \sum_{n=1}^{N} \int p(z_{nk} = 1 \mid \pi_k) \cdot p(\pi_k \mid \alpha) \, d\pi_k$$

$$= \sum_{n=1}^{N} \int \pi_k \cdot \frac{\alpha}{K} \cdot \pi_k^{\alpha/K - 1} \, d\pi_k$$

$$= \sum_{n=1}^{N} \frac{\alpha}{K} \cdot \frac{1}{\alpha/K + 1} = \frac{N\alpha}{\alpha + K}$$

Total

$$= \sum_{k=1}^{K}$$

Total number of 1's

$$= \sum_{k=1}^{K} \left(E\left[\sum_{n=1}^{N} z_{nk} \right] \right) = \sum_{k=1}^{K} \frac{N\alpha}{\alpha + k}$$

$$= \frac{NK\alpha}{\alpha + K}$$