Supplementary Materials for "An Alternative Prior Process for Nonparametric Bayesian Clustering"

Anonymous Author(s)
Affiliation
Address
email

1 Proof of Square Root Law for $\mathbb{E}\left(K_N \mid \mathrm{UN}\right)$

We define $T_k = \inf\left\{m > T_{k-1}; X_m \notin \{X_1, \ldots, X_{m-1}\}\right\}$. T_k is the "waiting time" (number of observations needed) until the k-th new cluster generated by the uniform process. Under the uniform process, $T_k = \sum_{i=1}^{k} \tau_i$ where $\tau_i \sim \mathrm{Geometric}\left(\theta \,/\, (\theta + i - 1\right)$ and the τ_i's are independent, so

$$\mathbb{E}\left(T_k\right) = \sum_{i=1}^{k} \frac{\theta + i - 1}{\theta} = \frac{k^2}{2\theta} + k\left(1 - \frac{1}{2\theta}\right) \tag{1}$$

$$\mathrm{Var}\left(T_k\right) = \sum_{i=1}^{k} \frac{\left(\theta + i - 1\right)\left(i - 1\right)}{\theta^2} = \frac{k^3}{3\theta^2} + k^2 \frac{1}{2\theta}\left(1 - \frac{1}{\theta}\right) + k\frac{1}{2\theta}\left(\frac{1}{3\theta} - 1\right) \tag{2}$$

In terms of T_j, $K_N = \max\{j; T_j \leq N\} = \sum_{j=1}^{N} \mathrm{I}\left(T_j \leq N\right)$. We first prove a strong law for the convergence of T_k. Let $\epsilon > 0$. From Chebychev's inequality and (2), we have

$$\mathrm{P}\,\left(\left|T_k - \mathbb{E}\left(T_k\right)\right| > \epsilon k^2\right) \leq \frac{\mathrm{Var}\left(T_k\right)}{\epsilon^2 k^4} \leq \frac{C(\theta, \epsilon)}{k} \tag{3}$$

From (3), we have that,

$$\mathrm{P}\,\left(\left|T_{k^2} - \mathbb{E}\left(T_{k^2}\right)\right| > \epsilon k^4\right) \leq \frac{C(\theta, \epsilon)}{k^2}$$

and so by the Borel-Cantelli lemma, we have $\mathrm{P}\,\left(\left|T_{k^2} - \mathbb{E}\left(T_{k^2}\right)\right| > \epsilon k^4\right) = 0$. Since $\epsilon > 0$ was chosen arbitrarily, it follows that $\frac{T_{k^2} - \mathbb{E}\left(T_{k^2}\right)}{k^4} \to 0$ almost surely and hence $\frac{T_{k^2}}{k^4} \to \frac{1}{2\theta}$ almost surely. Now, let $m = \lfloor\sqrt{k}\rfloor$. Since T_k is increasing, we have the following inequality:

$$\frac{T_{m^2}}{(m+1)^4} \leq \frac{T_k}{k^2} \leq \frac{T_{(m+1)^2}}{m^4}. \tag{4}$$

Since $\frac{m+1}{m} \to 1$, both sides of the inequality (4) converge to $(2\theta)^{-1}$ almost surely, and so

$$\frac{T_k}{k^2} \to \frac{1}{2\theta} \text{ almost surely.} \tag{5}$$

The strong law (5) implies a strong law for K_N as follows. $T_{K_N} \leq N < T_{K_N+1}$ and, consequently,

$$\frac{T_{K_N}}{K_N^2} \leq \frac{n}{K_N^2} < \frac{T_{K_N+1}}{K_N^2}.$$

Since $K_N \to \infty$ almost surely and $T_k \,/\, k^2 \to 1 \,/\, (2\theta)$ almost surely, it follows that the left and right hand side above both converge to $1/(2\theta)$ almost surely. Thus, $K_N^2 \,/\, N \to 2\theta$ almost surely and so

$$\frac{K_N}{\sqrt{N}} \to \sqrt{2\theta} \text{ almost surely.} \tag{6}$$

From the strong law (6) and the dominated convergence theorem, we have

$$\frac{\mathbb{E}\,(K_N)}{N} \to 0. \tag{7}$$

We combine (7) together with following result (derived in section 2 below),

$$\mathbb{E}\,(K_N^2) = \mathbb{E}\,(K_N) + 2\theta\,(N - \mathbb{E}\,(K_N)). \tag{8}$$

to give us

$$\frac{\mathbb{E}\,(K_N^2)}{N} \to 2\theta. \tag{9}$$

Finally, using (9) together with Fatou's lemma and Jensen's inequality, gives us

$$\sqrt{2\theta} \leq \liminf_{N\to\infty} \frac{\mathbb{E}\,(K_N)}{\sqrt{N}} \leq \limsup_{N\to\infty} \frac{\mathbb{E}\,(K_N)}{\sqrt{N}} \leq \limsup_{N\to\infty} \sqrt{\frac{\mathbb{E}\,(K_N^2)}{N}} = \sqrt{2\theta}.$$

which proves the result

$$\frac{\mathbb{E}\,(K_N)}{\sqrt{N}} \to \sqrt{2\theta}.$$

under the uniform process.

2 Result relating $\mathbb{E}\,(K_N)$ to $\mathbb{E}\,(K_N^2)$

Recall the definition of T_j from above and now define $M_N = K_N + 1$. Consider the "waiting time" T_{M_N} until the observation that creates the $(K_N+1)^{\text{th}}$ unique cluster. We relate $\mathbb{E}\,(K_N)$ to $\mathbb{E}\,(K_N^2)$ by calculating $\mathbb{E}\,(T_{M_N})$ in two different ways. First, observe that we have

$$\begin{aligned}
\mathbb{E}\,(T_{M_N}) &= \mathbb{E}\left(\sum_{k=1}^{\infty} \tau_k \cdot \mathrm{I}\,(k \leq M_N)\right) = \sum_{k=1}^{\infty} \mathbb{E}\,(\tau_k) \cdot \mathrm{P}\,(k \leq M_N) \\
&= \frac{\theta-1}{\theta} \sum_{k=1}^{\infty} \mathrm{P}\,(k \leq M_N) + \frac{1}{\theta} \sum_{k=1}^{\infty} k \cdot \mathrm{P}\,(k \leq M_N) \\
&= \frac{\theta-1}{\theta}\,\mathbb{E}\,(M_N) + \frac{1}{2\theta}\,\mathbb{E}\,(M_N(M_N+1))
\end{aligned}$$

which, since $M_N = K_N + 1$, simplifies to

$$\mathbb{E}\,(T_{M_N}) = 1 + \mathbb{E}\,(K_N)\left(1 + \frac{1}{2\theta}\right) + \mathbb{E}\,(K_N^2)\,\frac{1}{2\theta}. \tag{10}$$

$T_{M_N} = N + \sum_j \mathrm{I}\,(M_{N+j} = M_N)$ and so $\mathbb{E}\,(T_{M_N}) = N + \sum_j \mathrm{P}\,(M_{N+j} = M_N)$ where

$$\begin{aligned}
\mathrm{P}\,(M_{N+j} = M_N) &= \sum_k \mathrm{P}\,(T_k \leq N,\ N+j < T_{k+1}) \\
&= \sum_k \mathrm{P}\,(T_k \leq N < T_{k+1})\,\mathrm{P}\,(j < \tau_{k+1}) \\
&= \sum_k \mathrm{P}\,(M_N = k+1)\,\mathrm{P}\,(j < \tau_{k+1}).
\end{aligned}$$

It follows that

$$\begin{aligned}
\mathbb{E}\,(T_{M_N}) &= n + \sum_j \sum_k \mathrm{P}\,(M_N = k+1)\,\mathrm{P}\,(j < \tau_{k+1}) \\
&= N + \sum_k \mathrm{P}\,(M_N = k+1) \sum_j \mathrm{P}\,(j < \tau_{k+1}) \\
&= N + \sum_k \mathrm{P}\,(M_N = k+1)\,\mathbb{E}\,(\tau_{k+1}) \\
&= N + \sum_k \mathrm{P}\,(K_N = k)\,\frac{k+\theta}{\theta}
\end{aligned}$$

which can be simplified to

$$\mathbb{E}\left(T_{M_N}\right) = N + 1 + \mathbb{E}\left(K_N\right)\frac{1}{\theta} \tag{11}$$

Combining (10) and (11) gives (8):

$$\mathbb{E}\left(K_N^2\right) = \mathbb{E}\left(K_N\right) + 2\theta\left(N - \mathbb{E}\left(K_N\right)\right).$$

3 Graphical Model for Document Clustering

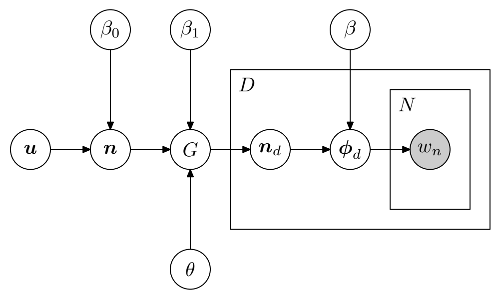

Figure 1: Word-based mixture model. G is either drawn from a Dirichlet process or a uniform process. Variables $\boldsymbol{n}$, $\boldsymbol{u}$, β_1 and β_0 comprise the hierarchical Dirichlet base measure G_0.

4 Evaluation Algorithm

The evaluation algorithm for computing $\log P(\mathcal{W}^{\text{test}} \,|\, \mathcal{W}^{\text{train}}, \boldsymbol{c}^{\text{train}}\theta, \boldsymbol{\beta})$ is based on the "left-to-right" evaluation algorithm introduced by [1], adapted to marginalize out test cluster assignments:

initialize $l := 0$
for each document d in $\mathcal{W}^{\text{test}}$ **do**
  initialize $p_d := 0$
  for each particle $r = 1$ to R **do**
    for $d' < d$ **do**
      $c_{d'}^{(r)} \sim P(c_{d'}^{(r)} \,|\, \mathcal{W}_{<d}^{\text{test}}, \{\boldsymbol{c}_{<d}^{(r)}\}_{\setminus d'}, \mathcal{W}^{\text{train}}, \boldsymbol{c}^{\text{train}}, \theta, \boldsymbol{\beta})$
    end for
    $p_d := p_d + \sum_c P(\boldsymbol{w}_d^{\text{test}}, c_d^{(r)} \!=\! c \,|\, \mathcal{W}_{<d}^{\text{test}}, \boldsymbol{c}_{<d}^{(r)}, \mathcal{W}^{\text{train}}, \boldsymbol{c}^{\text{train}}, \theta, \boldsymbol{\beta})$
    $c_d^{(r)} \sim P(c_d^{(r)} \,|\, \boldsymbol{w}_d^{\text{test}}, \mathcal{W}_{<d}^{\text{test}}, \boldsymbol{c}_{<d}^{(r)}, \mathcal{W}^{\text{train}}, \boldsymbol{c}^{\text{train}}, \theta, \boldsymbol{\beta})$
  end for
  $p_n := p_n \,/\, R$
  $l := l + \log p_n$
end for
$\log P(\mathcal{W}^{\text{test}} \,|\, \mathcal{W}^{\text{train}}, \boldsymbol{c}^{\text{train}}, \theta, \boldsymbol{\beta}) \simeq l$

References

[1] Hanna Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno. Evaluation methods for topic models. In *Proceedings of the 26th Interational Conference on Machine Learning*, 2009.