Supplement to "Inference for the Number of Topics in the Latent Dirichlet Allocation Model via Bayesian Mixture Modelling"

Zhe Chen
Department of Statistics
University of Florida

Hani Doss
Department of Statistics
University of Florida

Abstract

This document consists of three parts. The first derives an expression for the conditional distributions needed to run the Collapsed Gibbs Sampler of Griffiths and Steyvers (2004). The second part provides a version of Theorem 3 of "Inference for the Number of Topics in the Latent Dirichlet Allocation Model via Bayesian Mixture Modelling" without the condition that $\mathcal{T}$ is finite. The third part provides additional simulations to compare the PMMH algorithm with other methods for choosing the number of topics.

Throughout, equations, conditions, etc. refer to the main paper; also, notation is as in the main paper.

Derivation of the Conditional Distribution of $\boldsymbol{z}_{di}$ Given $\boldsymbol{z}_{(-di)}$ and $\boldsymbol{w}$

In order to obtain $p(z_{dit} = 1 \mid \boldsymbol{z}_{(-di)}, \boldsymbol{w})$, we will obtain, up to a constant of proportionality, an expression for $p\bigl(z_{dit} = 1, \boldsymbol{\beta}, \boldsymbol{\theta} \bigm| \boldsymbol{z}_{(-di)}, \boldsymbol{w}\bigr)$, from which we will integrate out $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$. We have

$$\begin{aligned} p\bigl(z_{dit} = 1, \boldsymbol{\beta}, \boldsymbol{\theta} \bigm| \boldsymbol{z}_{(-di)}, \boldsymbol{w}\bigr) &= p\bigl(z_{dit} = 1, \boldsymbol{\beta}, \boldsymbol{\theta} \bigm| w_{di}, \boldsymbol{z}_{(-di)}, \boldsymbol{w}_{(-di)}\bigr) \\ &\propto p\bigl(w_{di}, z_{dit} = 1, \boldsymbol{\beta}, \boldsymbol{\theta} \bigm| \boldsymbol{z}_{(-di)}, \boldsymbol{w}_{(-di)}\bigr) \\ &= p\bigl(w_{di} \bigm| z_{dit} = 1, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{z}_{(-di)}, \boldsymbol{w}_{(-di)}\bigr) \\ &\qquad \times p\bigl(z_{dit} = 1 \bigm| \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{z}_{(-di)}, \boldsymbol{w}_{(-di)}\bigr) p\bigl(\boldsymbol{\beta}, \boldsymbol{\theta} \bigm| \boldsymbol{z}_{(-di)}, \boldsymbol{w}_{(-di)}\bigr). \end{aligned} \tag{1}$$

Now we consider the three quantities in the right side of (1), in reverse order. First, using the results from (2.5) but for the reduced corpus $\boldsymbol{w}_{(-di)}$, we have that given $\boldsymbol{z}_{(-di)}$ and $\boldsymbol{w}_{(-di)}$,

$$\begin{aligned}
&\theta_1, \dots, \theta_D \text{ and } \beta_1, \dots, \beta_T \text{ are all independent,} \\
&\theta_{d'} \overset{\text{indep}}{\sim} \mathrm{Dir}_T\big(n_{d'1(-di)} + \alpha, \dots, n_{d'T(-di)} + \alpha\big), \text{ for } d' = 1, \dots, D, \\
&\beta_t \overset{\text{indep}}{\sim} \mathrm{Dir}_V\big(m_{\cdot t1(-di)} + \eta, \dots, m_{\cdot tV(-di)} + \eta\big), \text{ for } t = 1, \dots, T.
\end{aligned}$$

Hence we can express $p\big(\boldsymbol{\beta}, \boldsymbol{\theta} \,\big|\, \boldsymbol{z}_{(-di)}, \boldsymbol{w}_{(-di)}\big)$ as

$$p\big(\boldsymbol{\beta}, \boldsymbol{\theta} \,\big|\, \boldsymbol{z}_{(-di)}, \boldsymbol{w}_{(-di)}\big) = \left(\prod_{t=1}^{T} p\big(\beta_t \,\big|\, \boldsymbol{z}_{(-di)}, \boldsymbol{w}_{(-di)}\big)\right)\left(\prod_{d'=1}^{D} p\big(\theta_{d'} \,\big|\, \boldsymbol{z}_{(-di)}, \boldsymbol{w}_{(-di)}\big)\right), \tag{2}$$

in self-explanatory notation, where $p\big(\beta_t \,\big|\, \boldsymbol{z}_{(-di)}, \boldsymbol{w}_{(-di)}\big)$ and $p\big(\theta_{d'} \,\big|\, \boldsymbol{z}_{(-di)}, \boldsymbol{w}_{(-di)}\big)$ denote the densities of $\mathrm{Dir}_V\big(m_{\cdot t1(-di)} + \eta, \dots, m_{\cdot tV(-di)} + \eta\big)$ and $\mathrm{Dir}_T\big(n_{d'1(-di)} + \alpha, \dots, n_{d'T(-di)} + \alpha\big)$, respectively. Second, from the nature of the LDA model we know that given θ_d, $z_{di} \sim \mathrm{Mult}_T(\theta_d)$, where d indexes the document which contains w_{di}. That is, for any $t = 1, \dots, T$,

$$p\big(z_{dit} = 1 \,\big|\, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{z}_{(-di)}, \boldsymbol{w}_{(-di)}\big) = p\big(z_{dit} = 1 \,\big|\, \theta_d\big) = \theta_{dt}. \tag{3}$$

Lastly, from Line 4 of the LDA model statement, we know that given $\boldsymbol{\beta}$ and t such that $z_{dit} = 1$, $w_{di} \sim \mathrm{Mult}_V(\beta_t)$. Hence,

$$p\big(w_{di} \,\big|\, z_{dit} = 1, \boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{z}_{(-di)}, \boldsymbol{w}_{(-di)}\big) = p(w_{di} \,|\, z_{dit} = 1, \boldsymbol{\beta}) = \prod_{v=1}^{V} \beta_{tv}^{w_{div}}. \tag{4}$$

Plugging (2), (3) and (4) into (1), we get

$$\begin{aligned}
&p\big(z_{dit} = 1, \boldsymbol{\beta}, \boldsymbol{\theta} \,\big|\, \boldsymbol{z}_{(-di)}, \boldsymbol{w}\big) \\
&\quad \propto \left(\prod_{v=1}^{V} \beta_{tv}^{w_{div}}\right)\theta_{dt}\left(\prod_{d'=1}^{D} p\big(\theta_{d'} \,\big|\, \boldsymbol{z}_{(-di)}, \boldsymbol{w}_{(-di)}\big)\right)\left(\prod_{t'=1}^{T} p\big(\beta_{t'} \,\big|\, \boldsymbol{z}_{(-di)}, \boldsymbol{w}_{(-di)}\big)\right) \\
&\quad = \left(\prod_{v=1}^{V} \beta_{tv}^{w_{div}}\right) p\big(\beta_t \,\big|\, \boldsymbol{z}_{(-di)}, \boldsymbol{w}_{(-di)}\big)\left(\prod_{t' \neq t} p\big(\beta_{t'} \,\big|\, \boldsymbol{z}_{(-di)}, \boldsymbol{w}_{(-di)}\big)\right) \\
&\qquad \times \theta_{dt} p\big(\theta_d \,\big|\, \boldsymbol{z}_{(-di)}, \boldsymbol{w}_{(-di)}\big)\left(\prod_{d' \neq d} p\big(\theta_{d'} \,\big|\, \boldsymbol{z}_{(-di)}, \boldsymbol{w}_{(-di)}\big)\right).
\end{aligned} \tag{5}$$

Recall that for any integer $K \geq 1$, if $X = (X_1, \dots, X_K) \sim \mathrm{Dir}_K(a_1, \dots, a_K)$, then for any non-negative constants $r_1, \dots, r_K$,

$$E\left(\prod_{k=1}^{K} X_k^{r_k}\right) = \frac{\Gamma\big(\sum_{k=1}^{K} a_k\big)}{\Gamma\big(\sum_{k=1}^{K} a_k + \sum_{k=1}^{K} r_k\big)}\left(\prod_{k=1}^{K} \frac{\Gamma(a_k + r_k)}{\Gamma(a_k)}\right).$$

Integrating $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ out from (5) using the fact above, we see that for any $t = 1, \dots, T$, the conditional posterior probability that $z_{dit} = 1$ given $\boldsymbol{z}_{(-di)}$ and $\boldsymbol{w}$ satisfies

$$\begin{aligned}
p(z_{dit} = 1 \,|\, \boldsymbol{z}_{(-di)}, \boldsymbol{w}) &\propto \left(\frac{\Gamma\big(m_{\cdot t \cdot (-di)} + V\eta\big)}{\Gamma\big(m_{\cdot t \cdot (-di)} + V\eta + 1\big)}\right)\left(\prod_{v=1}^{V}\frac{\Gamma\big(m_{\cdot tv(-di)} + \eta + w_{div}\big)}{\Gamma\big(m_{\cdot tv(-di)} + \eta\big)}\right) \\
&\quad \times \left(\frac{\Gamma\big(n_d - 1 + T\alpha\big)}{\Gamma\big(n_d - 1 + T\alpha + 1\big)}\right)\left(\frac{\Gamma\big(n_{dt(-di)} + \alpha + 1\big)}{\Gamma\big(n_{dt(-di)} + \alpha\big)}\right) \\
&= \left(\frac{1}{m_{\cdot t \cdot (-di)} + V\eta}\right)\left(\prod_{v=1}^{V}(m_{\cdot tv(-di)} + \eta)^{w_{div}}\right) \\
&\quad \times \left(\frac{1}{n_d - 1 + T\alpha}\right)(n_{dt(-di)} + \alpha) \\
&= \prod_{v=1}^{V}\left(\frac{m_{\cdot tv(-di)} + \eta}{m_{\cdot t \cdot (-di)} + V\eta}\right)^{w_{div}}\left(\frac{n_{dt(-di)} + \alpha}{n_d - 1 + T\alpha}\right) \\
&= \left(\frac{m_{\cdot tv(-di)} + \eta}{m_{\cdot t \cdot (-di)} + V\eta}\right)\left(\frac{n_{dt(-di)} + \alpha}{n_d - 1 + T\alpha}\right),
\end{aligned}$$

where v denotes the term which w_{di} is observed to take, i.e. v is such that $w_{div} = 1$.

Another derivation of (2.8) is given in Carpenter (2010).

Ergodicity of the PMMH Algorithm When $\mathcal{T}$ Is Possibly Infinite

We first set up some notation. For any $t \in \mathcal{T}$ and any $\epsilon \in (0, 1]$, let $N(\epsilon, t) = \inf\{n : \|P^n_{\text{MH}}(t, \cdot) - \nu_{T\,|\,\boldsymbol{w}}(\cdot)\| \leq \epsilon\}$, where $P_{\text{MH}}(\cdot, \cdot)$ denotes the one-step Markov transition function for the ideal Metropolis-Hastings algorithm running on $\mathcal{T}$. Also, for any $t \in \mathcal{T}$, let $\rho(t)$ denote the probability of staying at t under the ideal Metropolis-Hastings algorithm, that is,

$$\rho(t) = 1 - \sum_{t' \in \mathcal{T}} \min\{r(t, t'), 1\} q_T(t, t'),$$

where recall that q_T is the Markov transition function running on $\mathcal{T}$.

Theorem 3′ *Assume Conditions A2 and A3 in the context of the LDA model, and that the mechanism for generating $\boldsymbol{\zeta}$ is the CGS (see* (4.8)*). Then:*

1. *For any $\epsilon, \ell > 0$ and $t_0 \in \mathcal{T}$, there exists a positive integer $M(\epsilon, \ell, t_0)$ such that for any $m > M(\epsilon, \ell, t_0)$ and $\boldsymbol{\zeta}_0 \in \mathcal{Z}^m_{t_0}$ such that $|\log(\tilde{\nu}^{(m)}_{T\,|\,\boldsymbol{w}}(t_0)) - \log(\nu_{T\,|\,\boldsymbol{w}}(t_0))| < \ell\epsilon/(24N(\epsilon, t_0))$, for any*

$n > N(\epsilon, t_0)$ *we have*

$$\left\| P^{m,n}(t_0, \boldsymbol{\zeta}_0; \cdot, \cdot) - \nu^{(m)}_{T,\boldsymbol{\zeta} \mid \boldsymbol{w}}(\cdot, \cdot) \right\| \le (1 + \ell)\epsilon + \rho^n(t_0).$$

2.
$$\left\| \mu^{m,n}_T(\cdot) - \nu_{T \mid \boldsymbol{w}}(\cdot) \right\| \to 0 \qquad \text{as } m, n \to \infty.$$

3.
$$\left\| \mu^{m,n}_{T,\boldsymbol{z}}(\cdot, \cdot) - \nu_{T,\boldsymbol{z} \mid \boldsymbol{w}}(\cdot, \cdot) \right\| \to 0 \qquad \text{as } m, n \to \infty.$$

4. For any $t \in \mathcal{T}$, $\boldsymbol{z} \in \mathcal{Z}_t$,

$$\left\| \mu^{m,n}_{T,\boldsymbol{z},\boldsymbol{\beta},\boldsymbol{\theta}}(t, \boldsymbol{z}, \cdot, \cdot) - \nu_{T,\boldsymbol{z},\boldsymbol{\beta},\boldsymbol{\theta} \mid \boldsymbol{w}}(t, \boldsymbol{z}, \cdot, \cdot) \right\| \to 0 \qquad \text{as } m, n \to \infty.$$

Proof To prove Part 1, note that under Conditions A2 and A3, the ideal Metropolis-Hastings algorithm is ergodic, i.e. $\lim_{n \to 0} \| P^n_{\text{MH}}(t_0, \cdot) - \nu_{T \mid \boldsymbol{w}}(\cdot) \| = 0$ for each $t_0 \in \mathcal{T}$. Also, in Theorem 2 of Chen and Doss (2017) we have shown that $\tilde{\nu}^{(m)}_{T \mid \boldsymbol{w}}(t) \xrightarrow{\text{a.s.}} \nu_{T \mid \boldsymbol{w}}(t)$ as $m \to \infty$ for each $t \in \mathcal{T}$. Hence, $| \log(\tilde{\nu}^{(m)}_{T \mid \boldsymbol{w}}(t)) - \log(\nu_{T \mid \boldsymbol{w}}(t)) |$ converges to zero in probability as $m \to \infty$ for each $t \in \mathcal{T}$. Therefore, we can apply Theorem 6 of Andrieu and Roberts (2009) to get the result.

The proofs of Parts 2, 3, 4 are the same as the proofs of Parts 2, 3, 4 of Theorem 3 in Chen and Doss (2017). □

We now remark on the differences between Theorem 3 and Theorem 3′. The proof of Theorem 3′ is deceptively simple. It seems short and trivial, but it is not, because it relies on Theorem 6 of Andrieu and Roberts (2009), the proof of which is extremely complex. In contrast, the proof of Theorem 3 is developed essentially from first principles, and the external result on which it is based is only Part 2 of Theorem 8 of Andrieu and Roberts (2009), which is nearly trivial.

Additional Illustrations on Synthetic Data

Here we evaluate the performance of our PMMH algorithm on three corpora generated synthetically. The analysis here goes beyond that given in Section 5.2 in two directions. First, we study the effect of "corpus difficulty" on the effectiveness of the algorithm. Intuitively, a corpus is difficult to analyze—more specifically, the number of topics is difficult to estimate—if the topics are close to each other. Second, we include the PPC criterion in our empirical evaluation. To be able to do this, we found it necessary to reduce the number of topics and the number of documents, because calculation of the PPC criterion is very computationally demanding.

For the first corpus, which we call S-1, we used the following specifications of the LDA model: the number of topics is $T = 6$, the hyperparameter is $h = (\alpha, \eta) = (0.1, 0.1)$, the vocabulary size is $V = 100$, the number of documents is $D = 300$, and the document lengths are $n_d = 300$, $d = 1, \ldots, D$. We generated the latent variables and the documents according to the LDA model with these specifications. We also generated corpora S-2 and S-3, in the same way, except that changed $\eta = .1$ to $\eta = 1$ and $\eta = 5$, respectively. The reason for this is that when η is increased, for each t, the components of β_t tend to be closer to their mean, and hence the topic vectors $\beta_1, \ldots, \beta_T$ tend to be closer to each other; hence when η is increased, estimation of T is more difficult. Thus, estimation of T is easiest for S-1 and hardest for S-3.

For each corpus we took the prior on T to be the uniform distribution on $\{2, \ldots, 100\}$, as in Chen and Doss (2017), and for our PMMH algorithm, we took the Markov transition function q_T to be as in Chen and Doss (2017). For each corpus we ran the algorithm with $m = 30$, 50, and 75, respectively, each time for $n = 10{,}000$ iterations, and taking the starting value for the number of topics to be $T_0 = 30$. The reason for considering multiple values of m is to evaluate the effect of m on the performance of the algorithm, and in particular to determine whether a value of m as small as 30 gives good results.

Figure 1 gives plots of running means for T produced by the PMMH algorithm using three different values of m, for corpora S-1, S-2 and S-3. The two upper panels give running means for T for S-1, for $m = 30$, 50, and 75. The figures differ only in the scaling of the x-axis. As mentioned earlier, because of the size of corpus, we presume that the posterior distribution is essentially a point mass at the true value of T (which is 6) and also that the marginal likelihood $m_{\boldsymbol{w}}(\cdot)$ is maximized at the true value. The plots show that even when $m = 30$, it takes less than 2000 iterations for the mean to reach 6; and once it reaches 6 the mean essentially remains there. The plots also show that convergence is faster when m is larger, but that the gain in efficiency is relatively minor, and that a value of m as small as 30 produces good results.

The bottom two panels give running means produced by the PMMH algorithm using $m = 30$

for corpora S-1, S-2 and S-3. The two panels differ only in the scaling of the x-axis. The reason we created these plots is to evaluate the change in the rate of convergence when the topics in the corpus are harder to distinguish from each other. In this case the posterior is more diffuse, and the sample mean takes longer to converge to the mean of the posterior. The bottom two panels show that, indeed, convergence is slower for S-2 and S-3; but this effect is not strong, and we conclude that the algorithm is successful even for challenging corpora.

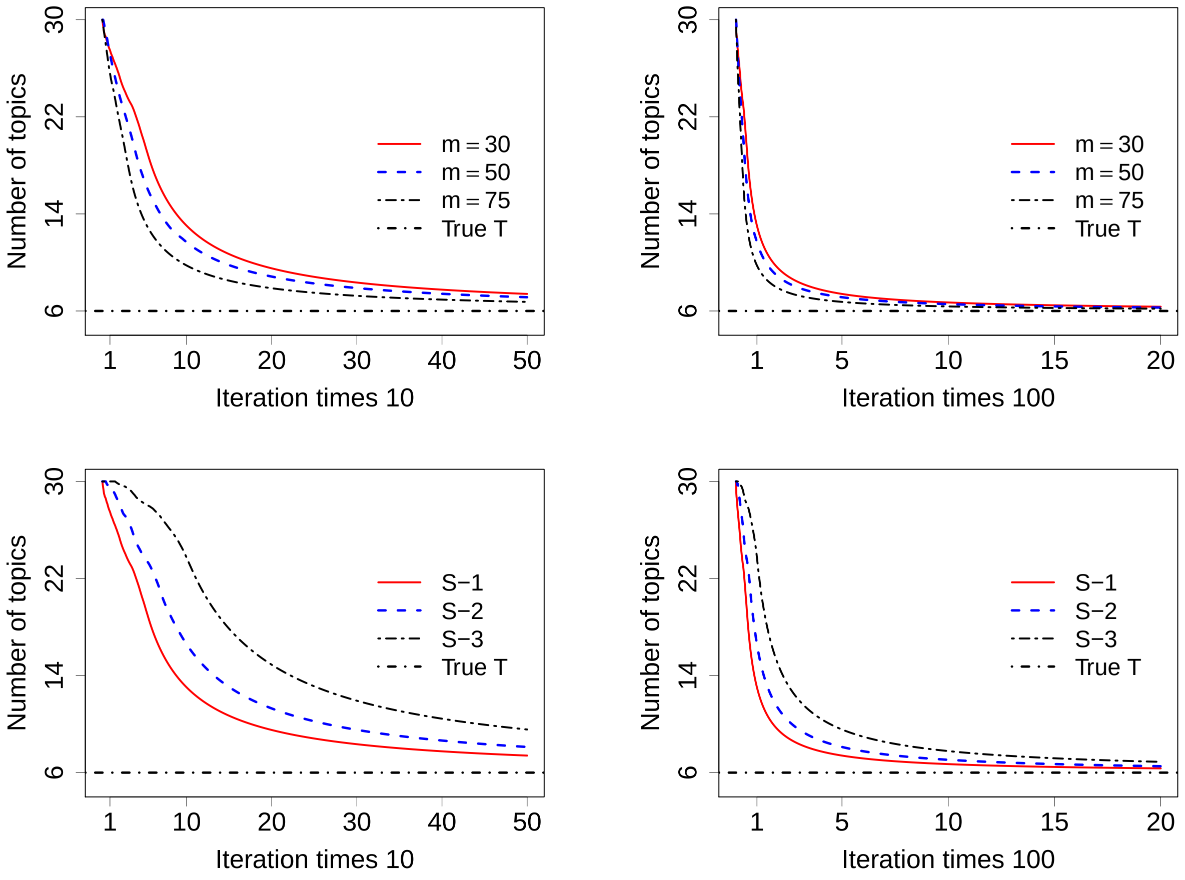
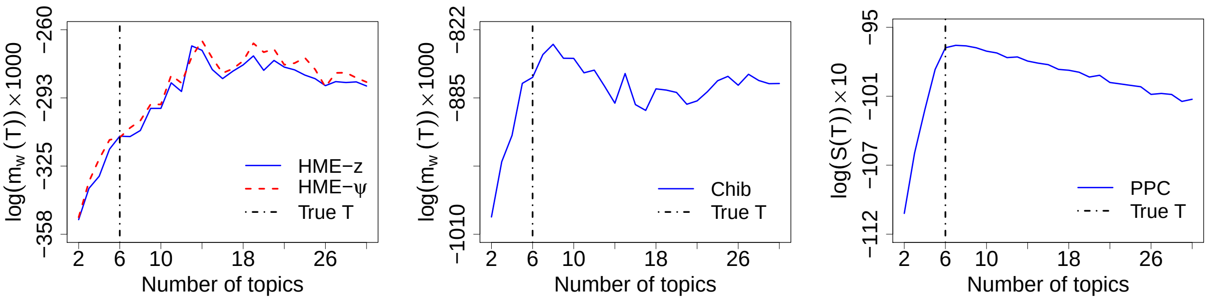


Figure 1: Running means for T produced by the PMMH algorithm on artificially-generated corpora. The upper plots, which differ only in the scale of the x-axis, show the running means using $m = 30$, 50, and 75 for corpus S-1. The lower plots compare the running means for corpora S-1, S-2, and S-3 using $m = 30$.

We now compare the performances of the PMMH algorithm, the two harmonic mean estimators, and the Chib estimator. To this end for corpus S-1 we calculated the HME-$\boldsymbol{z}$, HME-$\boldsymbol{\psi}$, and Chib estimates for $T = 2, 3, \ldots, 30$, using Markov chains of lengths $10{,}000$, $10{,}000$, and $3{,}600$, respectively. For the Chib method, we used a Markov chain of length $1{,}200$ for the pilot study needed to select the high-density point that is used in the main simulation. We also calculated the estimate of the PPC score $S(T)$ for $T = 2, 3, \ldots, 30$ using Markov chains of length 100. With these Markov chain lengths, the running times of all these methods are approximately equal, and this common time is about five times the running time of the PMMH algorithm using $m = 75$.

Figure 2 shows the results. From Figure 2(a) we see that the HME-$\boldsymbol{z}$ and HME-$\boldsymbol{\psi}$ estimates of $\arg\max_T m_{\boldsymbol{w}}(T)$ are 13 and 14, respectively. While the HME-$\boldsymbol{z}$ does better than HME-$\boldsymbol{\psi}$, both badly miss the true value of T. Figure 2(b) gives a plot of Chib's estimate of the marginal likelihood on the log scale. The maximum is reached at $T = 8$. While this is not very far off from the true value of 6, it should be noted that the ratio of estimate of the marginal likelihood at 8 to the estimate at 6 is $\exp(30427)$ (the ratio is understated by the appearance of the plot, which is on the log scale); thus, Chib's method effectively rules out the true value of T. Figure 2(c) gives a plot of the estimate of the PPC score, again on the log scale. If we use the PPC criterion, we would estimate T to be 7, although $T = 6$ and $T = 8$ would also be deemed plausible ($\widehat{S}(7) = \exp(-963), \widehat{S}(6) = \exp(-965.1)$ and $\widehat{S}(8) = \exp(-963.5)$). So the PPC criterion gives a reasonable performance for this data set.



(a) HME's of $\log(m_{\boldsymbol{w}}(T))$ (b) Chib estimate of $\log(m_{\boldsymbol{w}}(T))$ (c) Estimate of PPC score

Figure 2: HME-$\boldsymbol{\psi}$, HME-$\boldsymbol{z}$, and Chib estimates of $\log(m_{\boldsymbol{w}}(T))$, and estimates of the PPC score $S(T)$ (on the log scale) for the models indexed by different T's. The vertical line at 6 in each plot represents the true value of T.

References

Andrieu, C. and Roberts, G. O. (2009). The pseudo-marginal approach for efficient Monte Carlo computations. *The Annals of Statistics* **37** 697–725.

Carpenter, B. (2010). Integrating out multinomial parameters in latent Dirichlet allocation and naive Bayes for collapsed Gibbs sampling. Tech. rep., LingPipe, Inc.
URL `http://lingpipe.files.wordpress.com/2010/07/lda3.pdf`

Chen, Z. and Doss, H. (2017). Inference for the number of topics in the latent Dirichlet allocation model via Bayesian mixture modelling. Tech. rep., Department of Statistics, University of Florida.

Griffiths, T. L. and Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences* **101** 5228–5235.