Estimating Likelihoods for Topic Models*

Wray Buntine**

NICTA and Australian National University
Locked Bag 8001, Canberra, 2601, ACT Australia
wray.buntine@nicta.com.au

Abstract. Topic models are a discrete analogue to principle component analysis and independent component analysis that model *topic* at the word level within a document. They have many variants such as NMF, PLSI and LDA, and are used in many fields such as genetics, text and the web, image analysis and recommender systems. However, only recently have reasonable methods for estimating the likelihood of unseen documents, for instance to perform testing or model comparison, become available. This paper explores a number of recent methods, and improves their theory, performance, and testing.

1 Introduction

Topic models are a discrete analogue to principle component analysis (PCA) and independent component analysis (ICA) that model *topic* at the word level within a document. They have many variants such as NMF [LS] PLSI [Hof] and LDA [BNJ], and are used in many fields such as genetics [PSD], text and the web, image analysis and recommender systems. A unifying treatment of these models and their relationship to PCA and ICA is given by Buntine and Jakulin [BJ2]. The first Bayesian treatment was due to Pritchard, Stephens and Donnelly [PSD] and the broadest model is the Gamma-Poisson model of Canny [Can].

A variety of extensions exist to the basic models incorporating various forms of bierarchies [BGJT,MLM], combining topical and syntactic information [GSBT], jointly modelling text and citations/links [NAXC], models of information retrieval [AGvR]. The literature is extensive, especially in genetics following [PSD] and using NMF, and we cannot hope to cover the breadth here.

A continuing problem with these methods is how to do unbiased evaluation of different models, for instance using different topic dimensions or different variants and hierarchical models. The basic problem is that the likelihood of a single document (or datum/image) incorporates a large number of latent variables and thus its exact calculation is intractable. In this paper we present the problem according to the theory in its Dirichlet formulation such as LDA, but following the theory of [BJ2] our results apply more broadly to the wider class of topic models.

* Prepared for *The 1st Asian Conference on Machine Learning*, 2009, Nanjing China. The original publication is available at www.springerlink.com.

** Also a fellow at Helsinki Institute of IT.

Innovative methods for estimating the likelihood given a model and a test set of documents have been tried. The first approach suggested was the harmonic mean [GS,BJ1], but has proven to be considerably biased. Li and McCallum noted the unstable performance of this and proposed Empirical Likelihood [LM], a non-probabilistic method that is not able to give individual document estimates but rather broad scores. The approach is not widely used. Another approach suggested is to hold out a set of words [RZGSS], rather than just a set of documents. Since training is not optimising for the scores of the held out words, unbiased perplexity scores can be computed for these words. While this approach is reasonable, it still does not address the whole problem, the quality of individual documents.

Recently, an advance has been made with a number of algorithms presented and tested in the groundbreaking breaking paper of [WMSM]. Testing of Wallach's left-to-right algorithm [Wal, Algorithm 3.3] under small controlled tests (reported in Section 4.1) indicates that it is still biased. In this paper we present one new method for likelihood evaluation, based on the left-to-right algorithm, and revisit an importance sampling one. For both, improved theory is presented. The algorithms are tested rigorously in a number of controlled situations to demonstrate that significant improvements are made over previous methods.

2 Notation and Problem

This section introduces the notation used and then the problem being considered.

2.1 Documents and Topics

In our data reduction approach, one normally has a collection of documents and are estimating the model parameters for the topic model from the documents. We will, however, consider only a single document, one whose likelihood we wish to estimate. The document has L terms (or words or tokens), and these are indexed $l = 0, \ldots, L-1$. The l-th term in the document has dictionary value j_l. Assume the dictionary has J entries numbered $0, \ldots, J-1$. The values of terms can be stored as a sequence in vector $\boldsymbol{j}$, or "bagged" into a vector of sparse counts.

Now we will associate the terms with K topics, aspects or components, and give each document a vector of propensities for seeing the topics, represented as a K-dimensional probability vector $\boldsymbol{q}$. This propensity vector is sampled per document.

We will also assign to each term indexed by (l) a hidden topic (also called aspect or component) denoted $k_l \in \{0, \ldots, K-1\}$. This is modelled as a latent variable.

2.2 Model

The full probability for a document is given by a product of generative probabilities for each term. For the vectors of latent variables $\boldsymbol{q}, \boldsymbol{k}$ and the data vector

$\boldsymbol{j}$.

$$\begin{aligned}\boldsymbol{q} &\sim \text{Dirichlet}_K(\boldsymbol{\alpha}) \ , \\ k_l &\sim \text{Discrete}_K(\boldsymbol{q}) \qquad \text{for } l=0,\ldots,L-1 \ , \\ j_l &\sim \text{Discrete}_J(\boldsymbol{\theta}_{k_l}) \qquad \text{for } l=0,\ldots,L-1 \ .\end{aligned}$$

Here the subscripts K, J indicate the dimensions of the distributions. Note the component indicators $\boldsymbol{k}$ can be aggregated in total counts per class, to become a K-dimensional counts vector $\boldsymbol{C}$ with entries

$$C_k \ = \sum_{l=0,\ldots,L-1} 1_{k_l=k} \ .$$

This aggregate $\boldsymbol{C}$ corresponds to entries in the *score matrix* in conventional PCA. The parameter matrix $\boldsymbol{\Theta}$ corresponds to the *loading matrix* in conventional PCA. Here the model parameters are

- $\boldsymbol{\alpha}$ **:** A K dimensional vector of Dirichlet parameters generating probabilities for the topic.
- $\boldsymbol{\Theta}$ **:** A $K \times J$ dimensional matrix defines term probabilities, with column vectors $\boldsymbol{\theta}_k$ giving them for topic k. This is the loading matrix in conventional PCA.

One can extend topic models in all sorts of ways, for instance placing a hierarchical prior on $\boldsymbol{\alpha}$ or $\boldsymbol{\Theta}$ or splitting the terms up into separate semantic parts, such as citations and words, and modelling them with separate processes. The literature here is extensive.

2.3 The Problem

The full likelihood for a document, after marginalising $\boldsymbol{q}$ takes the form

$$p(\boldsymbol{k}, \boldsymbol{j} \,|\, \boldsymbol{\alpha}, \boldsymbol{\Theta}) \ = \ \frac{Z_K(\boldsymbol{C}+\boldsymbol{\alpha})}{Z_K(\boldsymbol{\alpha})} \prod_{l=0,\ldots,L} \theta_{k_l, j_l} \ , \tag{1}$$

where $Z_K()$ is the normalising constant for the Dirichlet distribution. Assume the model parameters are given, then the task we are considering is how to evaluate the marginal of this, $p(\boldsymbol{j} \,|\, \boldsymbol{\alpha}, \boldsymbol{\Theta})$, which means summing out over all K^L values for $\boldsymbol{k}$ in Equation (1). For L taking values in the 100's and K taking values in the 10's, the exact calculation is clearly impractical, and no alternative exact algorithms to brute force are known. Also note, that following [BJ2], the methods should be readily adapted to related models such as NMF.

3 Sampling Methods

We now consider several algorithms for the task of estimating document likelihoods for these basic topic models. Note that these should extend to more sophisticated topic models based on the same framework.

3.1 Importance Sampling

The method of importance sampling works as follows. Suppose one wishes to estimate $\mathcal{E}_{p(v)}\left[f(v)\right]$, and use a sampler $q(v)$ instead of $p(v)$, then importance sampling uses the N samples $\{v_n : n = 1, ..., N\}$ to make the unbiased estimate $\sum_n f(v_n)\frac{p(v_n)}{q(v_n)}$. Now if $q(v)$ is approximated via Gibbs, so its normalising constant is not known, then one uses

$$\frac{\sum_n f(v_n)\frac{p(v_n)}{q(v_n)}}{\sum_n \frac{p(v_n)}{q(v_n)}} ,$$

where the denominator estimates the inverse of the normaliser of $q(v)$.

3.2 The Harmonic Mean

The first method suggested for estimating the likelihood uses the second form of importance sampling with $v \to \boldsymbol{k}$. $f(v) \to p(\boldsymbol{j} \,|\, \boldsymbol{k}, \boldsymbol{\alpha}, \boldsymbol{\Theta})$, $p(v) \to p(\boldsymbol{k} \,|\, \boldsymbol{\alpha}, \boldsymbol{\Theta})$, and $q(v) \to p(\boldsymbol{j}, \boldsymbol{k} \,|\, \boldsymbol{\alpha}, \boldsymbol{\Theta})$. The renumerator simplifies dramatically to the sample count. Then one takes N samples of $\boldsymbol{k}$ using Gibbs sampling from $p(\boldsymbol{k}, \boldsymbol{j} \,|\, \boldsymbol{\alpha}, \boldsymbol{\Theta})$, and forms the estimate $\hat{p}(\boldsymbol{j} \,|\, \boldsymbol{\alpha}, \boldsymbol{\Theta})$

$$\frac{N}{\sum_n 1/p(\boldsymbol{j} \,|\, \boldsymbol{k}_n, \boldsymbol{\alpha}, \boldsymbol{\Theta})}$$

This formula is a harmonic mean, the inverse of the mean of the inverses, suggested in [GS,BJ1].

Unfortunately, the convergence is not stable in general [CC], the variance of the approximation can be large (since $\boldsymbol{k}$ is finite discrete, and no probabilities are zero, it must be finite). In practice, one can see this because the importance weights $w_n = \frac{p(v_n)}{q(v_n)} / \sum_n \frac{p(v_n)}{q(v_n)}$ are mostly near zero and usually only one or two significantly nonzero. Thus the estimate is usually dominated by the least $p(\boldsymbol{j} \,|\, \boldsymbol{k}_n, \boldsymbol{\alpha}, \boldsymbol{\Theta})$ seen so far. This makes it highly unstable.

3.3 Mean Field Approximation

The weakness in the last approach occurs because of the need to estimate the normalising constant, in the second form of importance sampling. Instead, use a proposal distribution $q(v)$ for which a normaliser is known. Then the estimate becomes

$$\frac{1}{N}\sum_n p(\boldsymbol{j}, \boldsymbol{k}_n \,|\, \boldsymbol{\alpha}, \boldsymbol{\Theta})\frac{1}{q(\boldsymbol{k}_n)} \tag{2}$$

Since the optimum $q(v)$ is proportional to $f(v)p(v)$ (when $f(v) \geq 0$), one could develop $q(v)$ by a Kullback-Leibler minimiser. So set $q(\boldsymbol{k}) = \prod_{l<L} q_l(k_l)$ and find $q()$ to minimise

$$\mathcal{E}_{\boldsymbol{k}\sim q(\boldsymbol{k})}\left[\log\frac{q(\boldsymbol{k})}{p(\boldsymbol{j}, \boldsymbol{k}_n \,|\, \boldsymbol{\alpha}, \boldsymbol{\Theta})}\right]$$

This yields the system of rewrite rules [GB]

$$q_l(k_l) \;\propto\; \exp\left(\mathcal{E}_{\boldsymbol{k}_{-l}\sim\prod_{m\neq l} q_m(k_m)}\left[\log p(\boldsymbol{j}, \boldsymbol{k}_n \,|\, \boldsymbol{\alpha}, \boldsymbol{\Theta})\right]\right)$$

where $\boldsymbol{k}_{-l}$ is $\boldsymbol{k}$ with the entry k_l removed. The rewriting system will converge due to the general properties of Kullback-Leibler minimisation. This simpifies to (in the first proportion, note $C_{k'}$ is a function of $\boldsymbol{k}$ and thus includes k_l),

$$\begin{aligned} q_l(k_l) &\propto \theta_{k_l,j_l} \exp\left(\sum_{k'} \mathcal{E}_{\boldsymbol{k}_{-l}\sim\prod_{m\neq l} q_m(k_m)}\left[\log \Gamma(C_{k'} + \alpha_{k'})\right]\right) \;, \\ q_l(k) &\propto \theta_{k,j_l} \exp\left(\mathcal{E}_{\boldsymbol{k}_{-l}\sim\prod_{m\neq l} q_m(k_m)}\left[\log(C'_k + \alpha_k)\right]\right) \;, \end{aligned} \tag{3}$$

where $C'_k = C_k - 1_{k_l=k}$ (which is independent of k_l since its effect is removed). Now $\mathcal{E}_{u\sim p(u)}\left[g(u)\right]$ can be approximated by $g(\overline{u})$ or $g(\overline{u}) - \sigma_u^2 \frac{1}{2\overline{u}^2}$, where $\overline{u}$ and σ_u^2 are the mean and variance by $p(u)$. Thus we have two options for the rewrite rules then, the simpler first order version is

$$\begin{aligned} q_l(k) &\propto \theta_{k,j_l}\left(\mathcal{E}_{\boldsymbol{k}_{-l}\sim\prod_{m\neq l} q_m(k_m)}\left[C'_k\right] + \alpha_k\right) \\ &\propto \theta_{k,j_l}\left(\sum_{m\neq l} q_m(k) + \alpha_k\right) \end{aligned} \tag{4}$$

Note, after all this theory, we have derived an approach virtually identical to the iterated pseudo-counts importance sampler (IS-IP) of [WMSM]. That this method performed far better than its related importance sampling algorithms [WMSM, Figure 1] comes as no surprise, given its derivation here as a mean-field approximation to the optimal importance sampler. For the second order version we subtract the variance term, which is computed similarly. Since this is part of a larger approximation, either version could work to construct a proposal distribution.

3.4 Left-to-Right Samplers

Wallach [Wal] suggests a particular sampler that breaks the problem into a series of parts:

$$p(\boldsymbol{j} \,|\, \boldsymbol{\alpha}, \boldsymbol{\Theta}) \;=\; \prod_{l<L} p(j_l \,|\, j_1, \ldots, j_{l-1}, \boldsymbol{\alpha}, \boldsymbol{\Theta}) \;. \tag{5}$$

Each term is estimated seperately using vector samples $(k_1, \ldots, k_{l-1}) \sim p(k_1, \ldots, k_{l-1} \,|\, j_1, \ldots, j_{l-1}, \boldsymbol{\alpha}, \boldsymbol{\Theta})$:

$$\begin{aligned} & p(j_l \,|\, j_1, \ldots, j_{l-1}, \boldsymbol{\alpha}, \boldsymbol{\Theta}) \\ & \approx \frac{1}{|\text{Sample}|} \sum_{(k_1,\ldots,k_{l-1})\in\text{Sample}} p(j_l \,|\, j_1, \ldots, j_{l-1}, k_1, \ldots, k_{l-1}, \boldsymbol{\alpha}, \boldsymbol{\Theta}) \end{aligned} \tag{6}$$

where the probability in the mean in Equation (6) is calculated using

$$p(j_l \,|\, j_1,\ldots,j_{l-1},k_1,\ldots,k_{l-1},\boldsymbol{\alpha},\boldsymbol{\Theta}) \;=\; \sum_k \theta_{k_l,j_l} p(k_l \,|\, k_1,\ldots,k_{l-1},\boldsymbol{\alpha},\boldsymbol{\Theta}) \;. \quad (7)$$

3.5 Left-to-Right Particle Sampler

Wallach's approach generates the vector samples $(k_1,\ldots,k_{l-1})$ for different l independently as follows:

1. For $l = 0,...,L-1$,
   (a) For $l' = 0,...,l-1$, resample $k_{l'}$ using

$$k_{l'} \;\sim\; p(k_{l'} \,|\, j_1,\ldots,j_{l-1},k_1,\ldots,k_{l'-1},k_{l'+1},\ldots,k_{l-1},\boldsymbol{\alpha},\boldsymbol{\Theta}) \;.$$

   (b) Sample k_l using $p(k_l \,|\, j_1,\ldots,j_l,k_1,\ldots,k_{l-1},\boldsymbol{\alpha},\boldsymbol{\Theta})$.
   (c) Record sample details using the current values $(k_0,\ldots,k_l)$ and Formula (7).

This is done R times as a so-called particle sampler. One particle's run generates L vectors from size 1 to L:

$$\{k_0\},\{k_0,k_1\},\{k_0,k_1,k_3\},\{k_0,k_1,k_3,k_4\},\ldots,\{k_0,\ldots,,k_{L-1}\} \;.$$

All R particles takes $RL^2/2$ multinomial samples and generates RL vectors used to generate estimates for the L terms in Equation (5).

3.6 Left-to-Right Sequential Sampler

Alternatively, generate the samples sequentially, so instead of R independent operations,

1. For $l = 0,...,L-1$,
   (a) Repeat R times:
      i. For $l' = 0,...,l$, resample $k_{l'}$ using

$$k_{l'} \;\sim\; p(k_{l'} \,|\, j_1,\ldots,j_l,k_1,\ldots,k_{l'-1},k_{l'+1},\ldots,k_l,\boldsymbol{\alpha},\boldsymbol{\Theta}) \;.$$

      ii. Record detail using the sample $(k_0,\ldots,k_l)$ and Formula (7).
   (b) Form the mean from the R samples to estimate Formula (6).
2. Form the estimate of Equation (5) from the L means.

This has the same complexity as the particle version but in contrast is easily seen to produce an unbiased estimate of Equation (5) as R approaches infinity for fixed L.

Lemma 1. *The left to right sampler gives unbiased estimates of* $p(\boldsymbol{j} \,|\, \boldsymbol{\alpha},\boldsymbol{\Theta})$ *for sufficiently large sample size* R*.*

Proof. For sufficiently large R, the L estimates for $p(j_l \,|\, j_1,\ldots,j_{l-1},\boldsymbol{\alpha},\boldsymbol{\Theta})$ become independent, and thus the mean of their product is the product of their means. Since each of these are unbiased, their product is unbiased. □

The particle sampler does not produce unbiased estimates regardless of sample size R because each particle still only runs a finite time.

4 Experiments

There are four different algorithms to compare:

Harmonic mean (HM): the early method know to be biased.

Mean-field importance sampler (MFI): the importance sampler of Equation (2) using the mean-field approximation built using rewrites rules of Equation (4).

Left-to-right particle sampler (LR): Wallach's method.

Left-to-right sequential sampler (LRS): Wallach's method modified to run sequentially.

Note the first two are linear in the document size, and the second two are quadratic.

Two different experiments are performed. The first does exact computation of the document likelihood formula for small K, L in single artificial cases. The second generates samples of artificial data with realistic K, L from a known, larger scale model taken from a real problem. This allows testing of the algorithms under conditions where the truth is known, and thus the results can be properly evaluated.

4.1 Comparison with Exact Calculation

For K^L in the trillions, the exact calculation is feasible. We therefore evaluate the four different sampling algorithms in the context of a specific topic model and a specific document. For this, we generate a model according to a Dirichlet posterior, and then generate a document according to the model, and then do the evaluation. C code for this is available from the DCA distribution[1].

The topic model has K topics for different values (usually 3,4,5), a vocabulary size of $J = 1000$ and a document length of $L = 12, 14, 16, 18$. The model, the $\boldsymbol{\Theta}$ matrix is generated with each topics $\boldsymbol{\theta}_k$ generated by a symmetric Dirichlet with parameters uniformly γ (varied below). For the document, we take a fixed length L and then generate word indexes $j_1, ..., j_{L-1}$ according to the model with Dirichlet prior on the components having $\alpha = 0.1$.

First, a calibration test is done. For fixed $L = 14$, $K = 4$ and a sample size of 200, we generate 100 different model-document pairs then run the different likelihood estimation algorithms and compare them with the exact likelihood. In the MFI algorithm, 10 full cycles of Equation (4) for the initial mean field approximator are run. Times for the computation (averaged over 1000 runs) are about 0.7 milliseconds for HM and the two MFIs, and 4.3 milliseconds for LR and LRS. The exact computation takes 5 minutes and 46 seconds, for a 2.16GHz Intel Core Duo machine.

The sample mean and standard-deviation are given in Table 1, along with the resultant Student-t value (for 99 df) for whether the estimate is unbiased (so the mean is zero). We see that LRS is the clear winner, consistently more precise,

[1] DCA is available from NICTA, and this small evaluator is in `doc/Approx`.

$\boldsymbol{\Theta}$ prior parameter $\gamma = 0.2$			
Method	Mean	Std.Dev.	St's t
HM	-0.3357	0.2345	-14.3
MFI (1st ord)	0.0018	0.0114	1.58
MFI (2nd ord)	0.0016	0.0204	0.815
LR	0.0032	0.0256	1.26
LRS	0.00072	0.0156	0.46

$\boldsymbol{\Theta}$ prior parameter $\gamma = 0.5$			
Method	Mean	Std.Dev.	St's t
HM	-0.210	0.120	-17.5
MFI (1st ord)	0.00131	0.0347	0.377
MFI (2nd ord)	0.00624	0.0244	2.55
LR	0.0128	0.0429	2.98
LRS	0.00002	0.0233	-0.0079

$\boldsymbol{\Theta}$ prior parameter $\gamma = 1.0$			
Method	Mean	Std.Dev.	St's t
HM	-0.109	0.0878	-12.4
MFI (1st ord)	0.0181	0.0668	2.70
MFI (2nd ord)	0.0261	0.0380	6.87
LR	0.000796	0.0502	1.58
LRS	0.00457	0.0317	1.44

$\boldsymbol{\Theta}$ prior parameter $\gamma = 3.0$			
Method	Mean	Std.Dev.	St's t
HM	-0.0440	0.0819	-5.37
MFI (1st ord)	0.0694	0.0797	8.71
MFI (2nd ord)	0.0553	0.0646	8.55
LR	0.00926	0.0417	2.22
LRS	0.00510	0.0259	1.97

Table 1. Estimator Precision for L=14, K=4

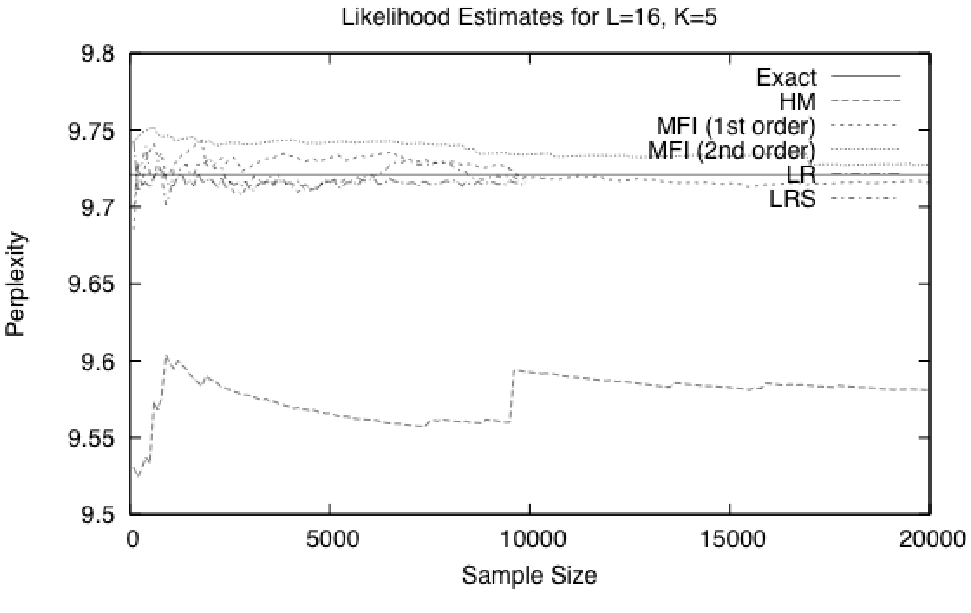

Fig. 1. Estimates for L=14, K=4 for increasing samples

where as all other methods are significantly different from the exact value at least once with a p-value of greater than 0.995^2. LRS is mostly not significantly different at a p-value of 0.9. The two MFI approximations are comparable to Wallach's LR method. Timewise, LR and LRS are an order of magnitude slower for these comparable sample sizes.

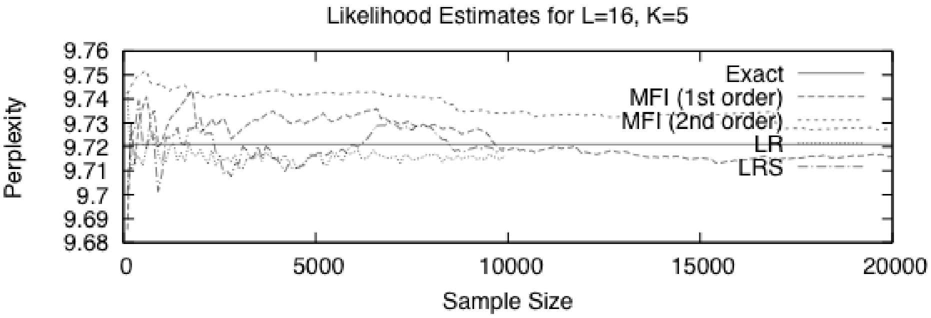

Fig. 2. Estimates for L=14, K=4 for increasing samples, closeup

Letting the different methods run out past 200 samples for a single model (sampled with $\gamma = 0.5$) plus document instance yields the plot given in Figures 1 and 2. This illustrates convergence to the exact value.

4.2 Comparison on Realistic Models

We built a topic model on a problem taken from a search engine news domain. In this domain, major stop words are removed, and less frequent words are removed leaving a total of $L = 50,182$ documents, $J = 28,251$ words and about 151 words on average per document. Then $K = 10$ and $K = 40$ topic models are built using standard LDA. Inspection shows these to be good looking models with clear separation between the topics and a clear semantics. From these known topic models, three data sets are generated according to the standard probabilistic LDA model: **s10** uses $K = 10$ and the existing number of words, $J = 28,251$; **s10s** uses $K = 10$ and a reduced $J = 5,000$ word count; and **s40** uses $K = 40$ and the existing number of words, $J = 28,251$. Thus three different data sets are created with known models and distinct topics and dictionary sizes.

Topic models were then built using vanilla LDA from varying subsets of the newly generated data sets, and their likelihood estimated on a hold out set of $10,000$ generated documents using the four different algorithms. Training set sizes used were $I = 1000, 2500, 5000, 10000, 20000, 30000, 40000$, and topics used were $K = 5, 8, 9, 10, 11, 12, 20, 40$ when the true $K = 10$ and $K = 20, 32, 35, 38, 40, 45, 50, 60$ when the true $K = 40$.

[2] The cutoff t-value is 2.58 for 0.995 and 1.28 for 0.90.

Variables are held fixed as much as possible, so when comparing MFI, LR and LRS at a given data set size, the same estimated model is used and the same hold out set is used to estimate likelihood. Note the HM method was not included in these comparisons because it is known to be poor. The LR and LRS methods were run with $R = 100$ particles/samples, and the MFI method was run with $R = 200$ samples. Note that LRS was also tested with $R = 2000$ samples, and the results where indistinguishable from that for $R = 100$ samples. For these estimates to be done on the 10000 test cases in the **s10** data set, MFI took 55 seconds and LR and LRS took 25 minutes and 30 seconds, approximately 20 times slower. The likelihood estimation for this is integrated into the DCA distribution[3].

Two views of the results are useful to look at. The first view compares the methods for different K as the data set sizes increases. Each method appears on a different plot. Two sets of plots are given, the first, Figure 3, shows the results for the **s10** data set with the large dictionary size, $J = 28,251$ and the **s10s** data set with the small dictionary size, $J = 5,000$. Like methods are presented side by side in the figure. The second set of plots, Figure 4, shows the results for the **s40** data set. Notice that the plots for the LR algorithm here indicates the perplexity starts to increase as data set sizes increase, in contrast to both MFI and LRS. This indicates a bias exists in the LR method.

The second view looks at how each method performs in selecting the "right" number of topics $K = 10$ for the **s10s** data set. The plots given in Figure 5 differ in the training sizes used, $I = 1000, 5000, 10000, 20000$. One can clearly see both MFI and LRS converging to the truth here, which has the "true" number of components at 10. The fast MFI method tracks LRS remarkably well. The LR method also does indicate the truth, but it is not as clear, and the distinction between different K is much finer, and thus harder to distinguish.

5 Discussion and Conclusion

A new method for estimating the likelihood of topic models has been developed that yielded significant improvements over those presented in [WMSM]. A second importance sampling method is revisited and shown to perform well in the kinds of model selection scenarios required in practice. Both methods, while similar to LR and IS-IP of Wallace *et al.*, come with an improved theory, and for one a proof that it can be used as a "gold standard". Our experiments also compare the approximations with an exact calculation, and demonstrate the use of the methods in a realistic model selection scenario using test set sizes typically required in practice (10,000 versus 50 in [WMSM]), as well as a variety of training set sizes. This more rigorous testing showed that Wallach's left-to-right algorithm is slightly biased.

[3] DCA is available from NICTA, and this functionality is the "-X" option to command `mphier`. The methods HM, MFI (1st), LR and LRS with S samples correspond to using the flags "-XS,G", "-XS,I", "-XS,L", "-XS,M" respectively.

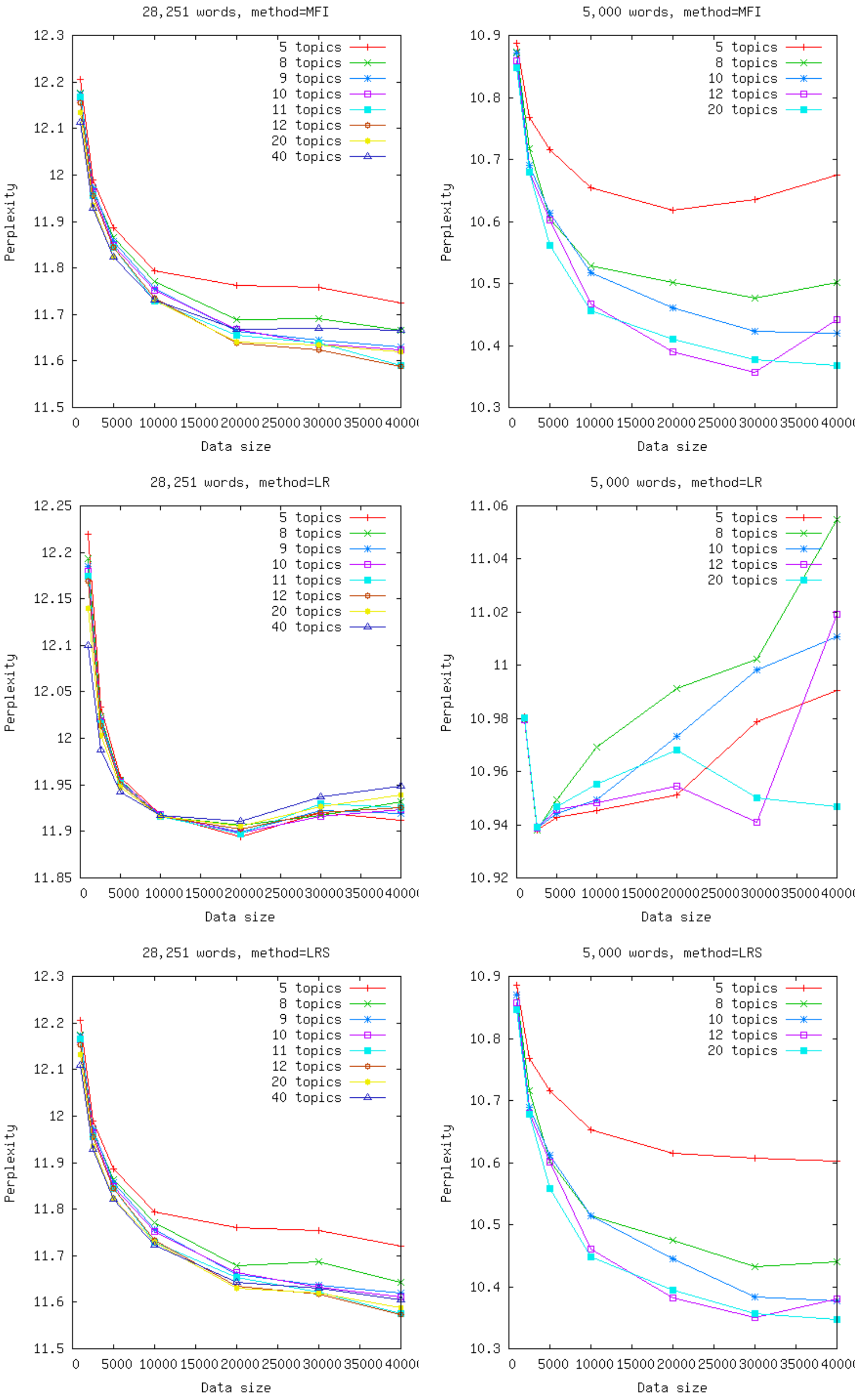

Fig. 3. Estimated test likelihood for different methods for data **s10** and **s10s**.

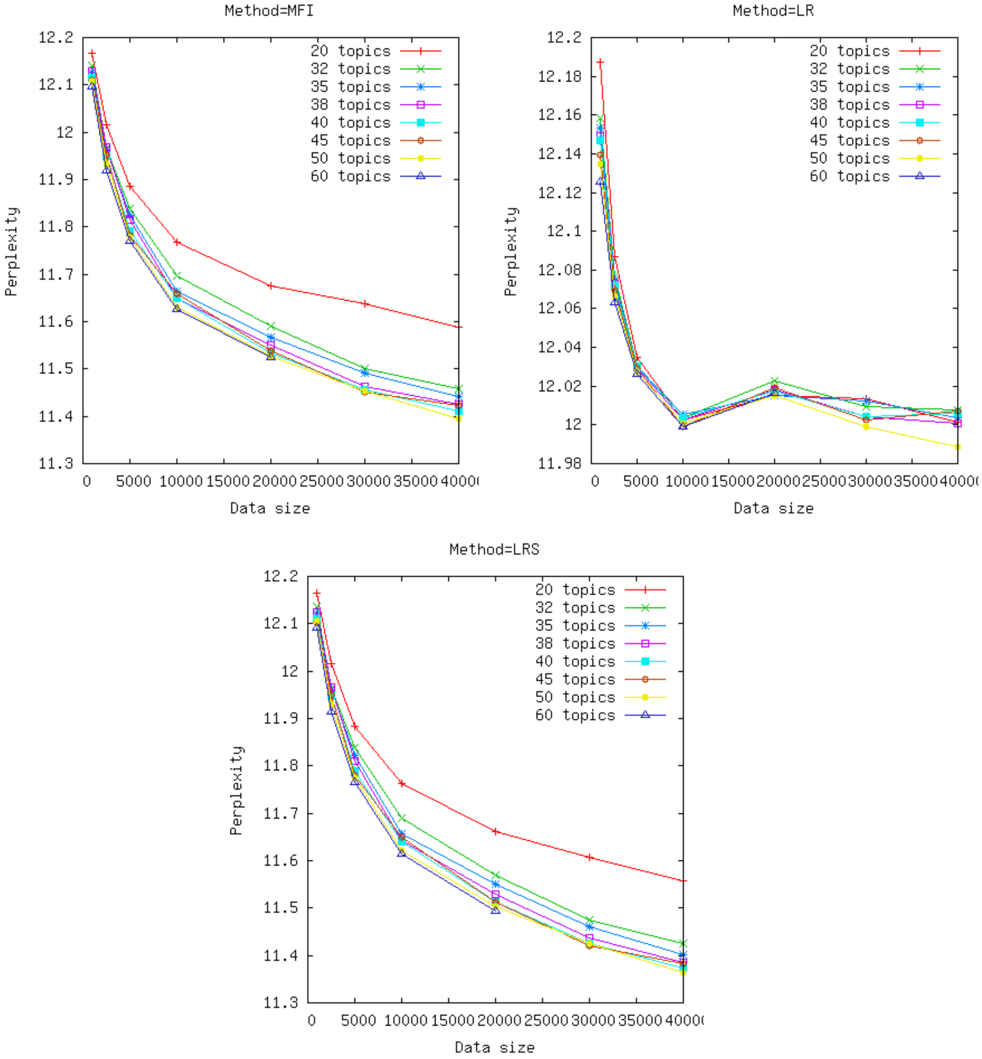

Fig. 4. Estimated test likelihood for different methods for data **s40**

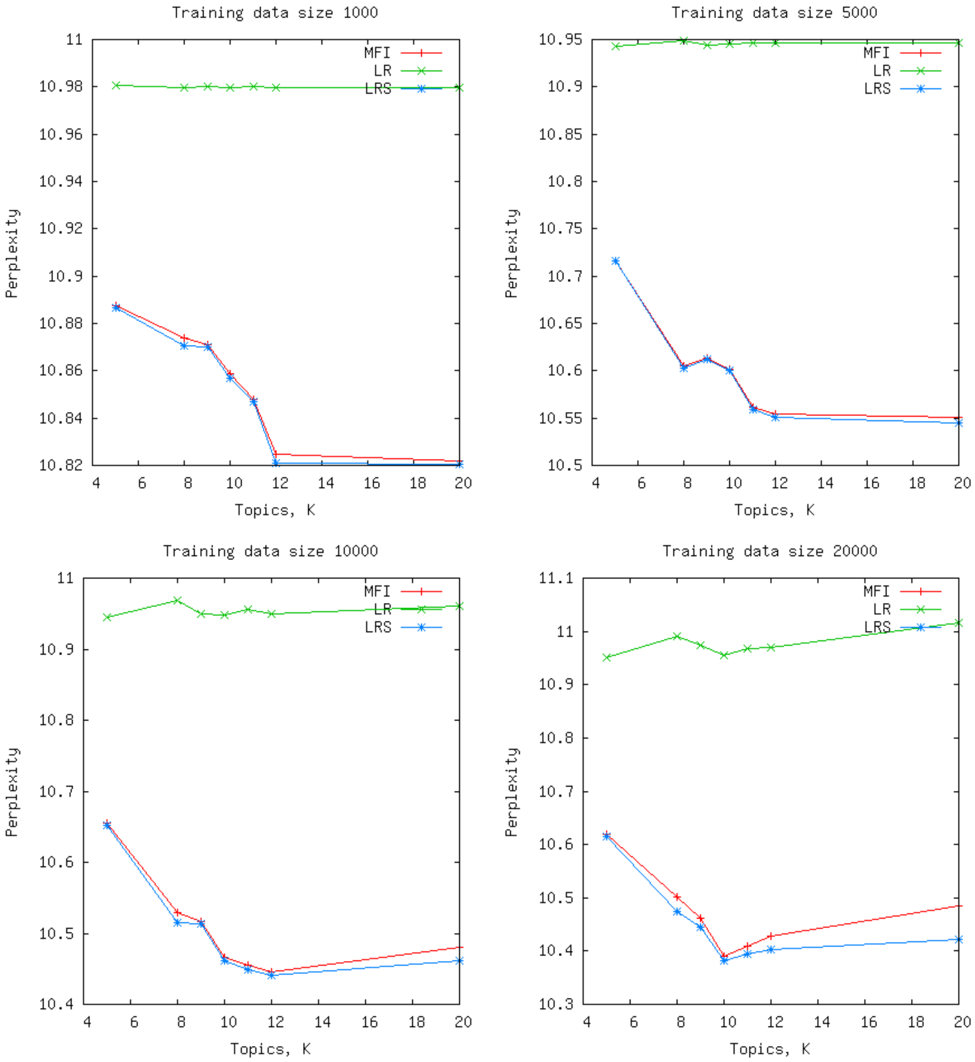

Fig. 5. Estimated test likelihood for different training set sizes on data **s10s**

By converting Wallach's left-to-right algorithm for estimating likelihoods from a particle sampler to a sequential sampler (labelled LRS), the method becomes provably unbiased, and can thus be used as a gold standard for estimation with large enough number of samples. Moreover, the sequential sampler clearly picks the "right" number of topics using the tests sets, whereas Wallach's original particle samper does not do so as distinctly, and also provides a biased estimate of the likelihood. However, this LRS method is quadratic in the document size, and thus may not be realistic in practice where one wants to test thousands of documents.

A second method (labelled MFI) uses importance sampling with a proposal distribution based on a mean field approximation to the optimal importance sampler. This method is linear in the document size, and thus an order of magnitude faster than the other methods for comparable sample sizes, and while performing acceptably on the exact tests, seems to perform almost as well as LRS when applied to a collection of documents (where many document likelihood results are averaged over the collection). This suggests the second method can be used in place of LRS for efficiency.

Acknowledgements. NICTA is funded by the Australian Government as represented by the Department of Broadband, Communications and the Digital Economy and the Australian Research Council through the ICT Centre of Excellence program. Wray Buntine also acknowledges support from the EU project CLASS (IST project 027978), and thanks Helsinki Institute of Information Technology for providing some of the software and computer facilities on which the experimental work is based.

References

[AGvR] L. Azzopardi, M. Girolami, and K. van Risjbergen. Investigating the relationship between language model perplexity and IR precision-recall measures. In *SIGIR '03*, pages 369–370, 2003.

[BGJT] D. Blei, T.L. Griffiths, M.I. Jordan, and J.B. Tenenbaum. Hierarchical topic models and the nested Chinese restaurant process. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*. MIT Press, Cambridge, MA, 2004.

[BJ1] W. Buntine and A. Jakulin. Applying discrete PCA in data analysis. In *UAI-2004*, Banff, Canada, 2004.

[BJ2] W.L. Buntine and A. Jakulin. Discrete components analysis. In C. Saunders, M. Grobelnik, S. Gunn, and J. Shawe-Taylor, editors, *Subspace, Latent Structure and Feature Selection Techniques*. Springer-Verlag, 2006.

[BNJ] D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[Can] J. Canny. GaP: a factor model for discrete data. In *SIGIR 2004*, pages 122–129, 2004.

[CC] B.P. Carlin and S. Chib. Bayesian model choice via MCMC. *Journal of the Royal Statistical Society B*, 57:473–484, 1995.

[GB] Z. Ghahramani and M.J. Beal. Propagation algorithms for variational Bayesian learning. In *NIPS*, pages 507–513, 2000.

[GS] T.L. Griffiths and M. Steyvers. Finding scientific topics. *PNAS Colloquium*, 2004.

[GSBT] T.L. Griffiths, M. Steyvers, D.M. Blei, and J.B. Tenenbaum. Integrating topics and syntax. In L.K. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems 17*, pages 537–544. MIT Press, Cambridge, MA, 2005.

[Hof] T. Hofmann. Probabilistic latent semantic indexing. In *Research and Development in Information Retrieval*, pages 50–57, 1999.

[LM] W. Li and A. McCallum. Pachinko allocation: DAG-structured mixture models of topic correlations. In *ICML '06: Proc. of the 23rd Int. Conf. on Machine learning*, pages 577–584, New York, NY, USA, 2006. ACM.

[LS] D. Lee and H. Seung. Learning the parts of objects by non-negative matrix factorization. *Nature*, 401:788–791, 1999.

[MLM] D. Mimno, W. Li, and A. McCallum. Mixtures of hierarchical topics with Pachinko allocation. In *ICML '07: Proceedings of the 24th international conference on Machine learning*, pages 633–640. ACM, 2007.

[NAXC] R. Nallapati, A. Ahmed, E.P. Xing, and W.W. Cohen. Joint latent topic models for text and citations. In *Proc. of the 14th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining*, pages 542–550, Las Vegas, 2008. ACM.

[PSD] J.K. Pritchard, M. Stephens, and P.J. Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155:945–959, 2000.

[RZGSS] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In *Proc. of the 20th Annual Conf. on Uncertainty in Artificial Intelligence (UAI-04)*, pages 487–49, Arlington, Virginia, 2004. AUAI Press.

[Wal] H. Wallach. *Structured Topic Models for Language.* PhD thesis, University of Cambridge, 2008.

[WMSM] H.M. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno. Evaluation methods for topic models. In L. Bottou and M. Littman, editors, *Proceedings of the 26th International Conference on Machine Learning (ICML 2009)*, 2009.