# Scalable Hyperparameter Selection for Latent Dirichlet Allocation

Wei Xia and Hani Doss *
Department of Statistics
University of Florida

**Abstract**

Latent Dirichlet Allocation (LDA) is a heavily-used Bayesian hierarchical model used in machine learning for modelling high-dimensional sparse count data, for example, text documents. As a Bayesian model, it incorporates a prior on a set of latent variables. The prior is indexed by some hyperparameters, which have a big impact on inference regarding the model. The ideal estimate of the hyperparameters is the empirical Bayes estimate which is, by definition, the maximizer of the marginal likelihood of the data with all the latent variables integrated out. This estimate cannot be obtained analytically. In practice, the hyperparameters are chosen either in an ad-hoc manner, or through some variants of the EM algorithm for which the theoretical basis is weak. We propose an MCMC-based fully-Bayesian method for obtaining the empirical Bayes estimate of the hyperparameter. We compare our method with other existing approaches both on synthetic and real data. The comparative experiments demonstrate that the LDA model with hyperparameters specified by our method outperforms models with the hyperparameters estimated by other methods. Supplemental materials for the paper are available online.

*Key words and phrases:* Empirical Bayes inference, Hamiltonian Monte Carlo, Markov chain Monte Carlo, model selection, topic modelling.

*Research supported by NSF Grant DIIS-17-24174

# 1 Introduction

Latent Dirichlet Allocation (LDA, Blei et al. 2003) is a general probabilistic framework for modelling high-dimensional sparse count data represented by feature counts. It was introduced by Blei et al. (2003) to discover topics in text documents. Since its introduction, it and its extensions have been successfully applied to many other data types, such as image-caption data (Blei and Jordan, 2003) and author-document data (Rosen-Zvi et al., 2004), and applications to new problems continue to arise (the paper has been cited over $23{,}000$ times according to Google Scholar, and the yearly citation rate is currently increasing). In this paper, we focus on data consisting of a collection of documents. Suppose we have a corpus of documents, which span several different topics, such as sports, medicine, politics, etc. We imagine that for each word in each document, there is a latent (i.e. unobserved) variable indicating a topic from which that word is drawn. The main objectives in using the LDA model are usually to obtain an interpretable set of topics for the corpus, and to make inference on the latent topic variables for each document.

To describe the LDA model, we first set up some terminology and notation. There is a vocabulary of $V$ words; typically, this is taken to be the union of all the words in all the documents of corpus, after removing stop (i.e. uninformative) words. There are $D$ documents in the corpus, and for $d = 1, \ldots, D$, document $d$ has $n_d$ words, $w_{d1}, \ldots, w_{dn_d}$. In total, the corpus has $N = \sum_{d=1}^{D} n_d$ words. The order of the words is considered uninformative, and so is neglected. Each word is represented as a $V$-dimensional index vector with a $1$ at the $v^{\text{th}}$ element, where $v$ denotes the term selected from the vocabulary. Thus, document $d$ is represented by the matrix $\boldsymbol{w}_d = (w_{d1}, \ldots, w_{dn_d})$ and the corpus is represented by the list $\boldsymbol{w} = (\boldsymbol{w}_1, \ldots, \boldsymbol{w}_D)$. A topic is, by definition, a distribution over the vocabulary, i.e. a point in $\mathbb{S}_{V-1}$, the $(V-1)$-dimensional simplex. The number of topics, $T$, is finite and known. For each word $w_{di}$, $i = 1, \ldots, n_d$ and $d = 1, \ldots, D$, $z_{di}$ is a $T$-dimensional index vector which represents the latent variable that denotes the topic from which $w_{di}$ is drawn. Let $\boldsymbol{z}_d = (z_{d1}, \ldots, z_{dn_d})$ and $\boldsymbol{z} = (\boldsymbol{z}_1, \ldots, \boldsymbol{z}_D)$. The distribution of $\boldsymbol{z}_d$ will depend on a document-specific variable $\theta_d$ which indicates a distribution on the topics for document $d$. We let $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_D)$. We will use $\mathrm{Dir}_L(a_1, \ldots, a_L)$ to denote the finite-dimensional Dirichlet distribution on the simplex $\mathbb{S}_{L-1}$ and $\mathrm{Mult}_L(b_1, \ldots, b_L)$ to denote the multinomial distribution with number of trials equal to $1$ and probability vector $(b_1, \ldots, b_L)$. We will form a $T \times V$ matrix $\boldsymbol{\beta}$, whose $t^{\text{th}}$ row is the $t^{\text{th}}$ topic (how $\boldsymbol{\beta}$ is formed will be described shortly). Thus, $\boldsymbol{\beta}$ will consist of vectors $\beta_1, \ldots, \beta_T$, all lying in $\mathbb{S}_{V-1}$. Formally, the LDA model is described by the following Bayesian hierarchical model, in which $\eta, \alpha_1, \alpha_2, \ldots, \alpha_T > 0$ are hyperparameters:

1. $\beta_t \overset{\text{iid}}{\sim} \text{Dir}_V(\eta, \ldots, \eta)$, $t = 1, \ldots, T$.
2. $\theta_d \overset{\text{iid}}{\sim} \text{Dir}_T(\alpha_1, \ldots, \alpha_T)$, $d = 1, \ldots, D$, and the $\theta_d$'s are independent of the $\beta_t$'s.
3. Given $\theta_1, \ldots, \theta_D$, $z_{di} \overset{\text{iid}}{\sim} \text{Mult}_T(\theta_d)$, for $i = 1, \ldots, n_d$, $d = 1, \ldots, D$, and the $D$ matrices $(z_{11}, \ldots, z_{1n_1}), \ldots, (z_{D1}, \ldots, z_{Dn_D})$ are independent.
4. Given $\boldsymbol{\beta}$ and the $z_{di}$'s, $w_{di}$ is is drawn from the row of $\boldsymbol{\beta}$ indicated by $z_{di}$, independently for $i = 1, \ldots, n_d$, $d = 1, \ldots, D$.

The parameters in the model are $\boldsymbol{\beta}$, $\boldsymbol{\theta}$, and $\boldsymbol{z}$, and we let $\boldsymbol{\psi} = (\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{z})$ denote the entire set of parameters. We also let $h = (\alpha_1, \ldots, \alpha_T, \eta)$ denote the hyperparameter vector and $\mathcal{H} = (0, \infty)^{T+1}$ denote the hyperparameter space. Lines 1–3 of the LDA model description define a prior distribution on the parameters, which we will denote by $\nu^{(h)}$, and line 4 gives the likelihood, which we denote by $\ell_{\boldsymbol{w}}(\boldsymbol{\psi})$. We observe the words in the documents of the corpus, and inference regarding the parameter vector $\boldsymbol{\psi}$ is based on its posterior distribution, which we denote by $\nu^{(h)}_{\boldsymbol{\psi}\,|\,\boldsymbol{w}}$. In order to use the LDA model, one needs to specify $h$. This hyperparameter has a big impact on inference drawn from the model. For example, consider $\int \|\beta_i - \beta_j\|_2 \, d\nu^{(h)}_{\boldsymbol{\psi}\,|\,\boldsymbol{w}}(\boldsymbol{\psi})$, the posterior expectation of the $L_2$ norm between topics $i$ and $j$, and for some small $\epsilon$, $\nu^{(h)}_{\boldsymbol{\psi}\,|\,\boldsymbol{w}}(\|\theta_i - \theta_j\|_2 \leq \epsilon)$, the posterior probability that the topic vectors for documents $i$ and $j$ are nearly the same. These are standard objects of interest, and George and Doss (2018, pages 16–17) have shown that on real corpora these vary considerably with $h$. George (2015) showed that in single-membership situations, $h$ affects the model's classification performance and its ability to correctly cluster the documents in the corpus. Therefore, it is important to choose the hyperparameter carefully.

Current schemes for specifying the hyperparameter fall into three groups. The first group consists of very simple ad-hoc rules that do not depend on the data. These are trivial to implement but are not based on any statistical principle; and they perform poorly. They are reviewed briefly in Section 4.1. The second and third group consist of methods that are based on the following idea. Let $m_{\boldsymbol{w}}(h)$ be the marginal likelihood of the corpus. This is the likelihood of the corpus with all the latent variables integrated/summed out, i.e. $m_{\boldsymbol{w}}(h) = \int \ell_{\boldsymbol{w}}(\boldsymbol{\psi}) \, d\nu^{(h)}(\boldsymbol{\psi})$. A principled way of choosing $h$ is to use $\hat{h} = \arg\max_h m_{\boldsymbol{w}}(h)$ which is, by definition, the empirical Bayes estimate of $h$. Unfortunately, it is impossible to obtain $\hat{h}$ explicitly: the function $m(h)$ (we drop the subscript $\boldsymbol{w}$ for simplicity) is analytically intractable.

In approaches from the second group, for each $h$ over a fine grid in $\mathcal{H}$, we run a Monte Carlo experiment to form an estimate $\widehat{m}(h)$ of $m(h)$; we do this separately for each $h$, and we estimate $\arg\max_h m(h)$ via $\arg\max_h \widehat{m}(h)$. Papers that proceed in this way include Chib (1995) and Chib

and Jeliazkov (2001). We also mention the "harmonic mean estimator" introduced by Newton and Raftery (1994). There are several significant problems associated with this approach. One is that convergence can be slow. For example, the harmonic mean estimator typically converges at a rate which is much slower than $n^{1/2}$, where $n$ is the Monte Carlo sample size (Wolpert and Schmidler, 2012). Also, the fact that a separate Monte Carlo experiment needs to be run for every $h$ over a grid in $\mathcal{H}$ makes the method very time consuming. Although methods in this group do not work well in the LDA model (in fact Newton and Raftery (1994) expressed reservations regarding the harmonic mean estimator in general when they introduced it), we mention them because they are the ones that are the most frequently used in the machine learning literature; see Wallach et al. (2009) for a discussion.

The third group consists of methods that use the EM algorithm to find the maximizer of the marginal likelihood function. Here, the "complete data likelihood" $p_h(\boldsymbol{\psi}, \boldsymbol{w})$ is available directly from lines 1–4 of the LDA model description, so $\boldsymbol{w}$ may be viewed as "observed data," and $\boldsymbol{\psi}$ may be viewed as "missing data." Unfortunately, the E-step, which is an expectation with respect to the intractable distribution $\nu^{(h)}_{\boldsymbol{\psi}\,|\,\boldsymbol{w}}$, cannot be carried out exactly, and two variants of the EM algorithm have been used. One of these is Monte Carlo EM, in which the expectation is approximated by MCMC. This has been carried out by Wallach (2006), who faced the additional problem that in her implementation the maximization in the M-step cannot be done in closed form either and an approximation is used instead. (She used the Collapsed Gibbs Sampler (CGS) of Griffiths and Steyvers (2004) as her Markov chain and dubbed her scheme "Gibbs-EM.") George and Doss (2018) have shown that the performance of Gibbs-EM is mixed: it sometimes gives accurate approximations, but there are classes of cases where the algorithm converges to a value of $h$ which is not $\arg\max_h m(h)$. Another variant is "variational EM" (VEM), in which the E-step is approximated through variational methods (see Jordan et al. (1999) for an introduction to variational methods, and Blei et al. (2017) for a recent review). This is the approach that was originally used by Blei et al. (2003), and it is extremely fast. However, in empirical studies this approach has not performed well for a variety of corpora. For Gibbs-EM, there are no useful bounds on the approximation used in the M-step, and for VEM there are no useful bounds on the approximation used in the E-step. Because the approximations are used at every iteration of the algorithm, there are no results regarding the theoretical properties of either Gibbs-EM or VEM. A more thorough discussion of both these variants of the EM algorithm is given in Section 4.1.

A different approach for estimating $\hat{h} = \arg\max_h m(h)$ was developed by George and Doss

(2018). They devised an algorithm based on a combination of MCMC and importance sampling for forming an estimate of the entire marginal likelihood surface $m(h)$ up to a multiplicative constant. More specifically, using a single Markov chain—a so-called serial tempering chain—they form an importance sampling estimate $\widehat{M}(\cdot)$ with the property that $\widehat{M}(h) \xrightarrow{\text{a.s.}} c\, m(h)$ simultaneously for all $h$, where $c$ is an unknown constant, and the convergence is as the Markov chain length tends to infinity. (For the purpose of estimating $\arg\max_h m(h)$, the fact that $c$ is unknown is immaterial: $\arg\max_h c\, m(h) = \arg\max_h m(h)$.) Additionally, they show that $\arg\max_h \widehat{M}(h) \xrightarrow{\text{a.s.}} \arg\max_h m(h)$. Although their method works well for moderate-size corpora, for large corpora it requires considerable tuning because for such corpora the importance sampling weights are highly variable. A more detailed discussion of the method and its limitations is given in Section 4.1.

In this paper we use a "fully-Bayes approach," not for the purpose of doing a fully-Bayes analysis, but rather for the purpose of selecting a single value of the hyperparameter $h$. The approach is based on ideas in the recent paper by Doss and Linero (2018). Their method, which we now review, is very general, i.e. it is not developed for any particular model, and whether or not it is successful is determined by how it is implemented. Very briefly, in the context of the LDA model the method is as follows. Let $\mathcal{H}$ denote the hyperparameter space, and for simplicity we temporarily assume that this is a compact subset of $\mathbb{R}^{T+1}$, for example we assume that $\mathcal{H}$ is a large hypercube. We can then put a uniform prior on $\mathcal{H}$, which we denote by $U$, and we let $u$ be its density. This induces a joint distribution on $(\boldsymbol{w}, \boldsymbol{\psi}, h)$, which we will denote by $\nu$. Let $\nu_{(\boldsymbol{\psi},h)\,|\,\boldsymbol{w}}$ denote the posterior distribution of $(\boldsymbol{\psi}, h)$ given $\boldsymbol{w}$, and let $\nu_{h\,|\,\boldsymbol{w}}$ denote the marginal posterior distribution of $h$ given $\boldsymbol{w}$. Regarding $\nu_{h\,|\,\boldsymbol{w}}$, the statement "the posterior is proportional to the likelihood times the prior" reads as $\nu_{h\,|\,\boldsymbol{w}}(h) \propto m(h)u(h)$. Since $u$ is the uniform distribution this may be rewritten as $\nu_{h\,|\,\boldsymbol{w}}(h) \propto m(h)$, so the mode of $\nu_{h\,|\,\boldsymbol{w}}$ is $\arg\max_h m(h)$. Now, suppose that we can construct an ergodic Markov chain $(\boldsymbol{\psi}^{(1)}, h^{(1)}), (\boldsymbol{\psi}^{(2)}, h^{(2)}), \ldots$ whose invariant distribution is $\nu_{(\boldsymbol{\psi},h)\,|\,\boldsymbol{w}}$. The marginal sequence $h^{(1)}, h^{(2)}, \ldots$ then has invariant distribution equal to $\nu_{h\,|\,\boldsymbol{w}}$. Any method for estimating the mode of $\nu_{h\,|\,\boldsymbol{w}}$ from the sequence $h^{(1)}, h^{(2)}, \ldots$ gives rise to an estimate of $\arg\max_h m(h)$. (The Doss and Linero (2018) paper is unpublished, but we do not rely on it except for the idea stated in the present paragraph.)

Now, generally speaking, estimation of the mode of a density is a hard problem. If $f$ is a density on $\mathbb{R}^p$, the rate of convergence of typical nonparametric estimators of the mode based on an iid sample of size $n$ is $n^{1/(4+p)}$ (Tsybakov, 1990; Donoho and Liu, 1991) so even in the sim-

plest case where $p = 1$, this is the very slow rate of $n^{1/5}$. In our LDA setup, we are able to construct an augmentation random vector $A$ such that when we consider the vector $(\boldsymbol{w}, \boldsymbol{\psi}, h, A)$, the marginal conditional distribution of $(\boldsymbol{\psi}, h)$ given $\boldsymbol{w}$ is equal to $\nu_{(\boldsymbol{\psi},h)\,|\,\boldsymbol{w}}$, and we are able to construct a uniformly geometrically ergodic Markov chain $(\boldsymbol{\psi}^{(1)}, h^{(1)}, A^{(1)}), (\boldsymbol{\psi}^{(2)}, h^{(2)}, A^{(2)}), \ldots$ with invariant distribution $\nu_{(\boldsymbol{\psi},h,A)\,|\,\boldsymbol{w}}$. Moreover, Rao-Blackwellization is possible, i.e. the conditional density of $h$ given $(\boldsymbol{\psi}, A)$ and $\boldsymbol{w}$ is available in closed form, so $\nu_{h\,|\,\boldsymbol{w}}$ may be estimated by $\hat{\nu}_{h\,|\,\boldsymbol{w}}(h) = (1/n)\sum_{i=1}^{n} \nu_{h\,|\,(\boldsymbol{\psi}=\boldsymbol{\psi}^{(i)}, A=A^{(i)}, \boldsymbol{w})}(h)$. This is simply an average, so we have a central limit theorem that says that for any fixed $h$, $n^{1/2}\big(\hat{\nu}_{h\,|\,\boldsymbol{w}}(h) - \nu_{h\,|\,\boldsymbol{w}}(h)\big)$ converges in distribution to a mean-zero normal random vector. We view $\hat{\nu}_{h\,|\,\boldsymbol{w}}(\cdot)$ and $\nu_{h\,|\,\boldsymbol{w}}(\cdot)$ as functions, and using tools from empirical process theory, we establish uniformity in the convergence, i.e. $n^{1/2}\big(\hat{\nu}_{h\,|\,\boldsymbol{w}}(\cdot) - \nu_{h\,|\,\boldsymbol{w}}(\cdot)\big)$ converges in distribution to a mean-zero Gaussian process indexed by $h$, and this entails that $n^{1/2}\big(\arg\max_h \hat{\nu}_{h\,|\,\boldsymbol{w}}(h) - \arg\max_h \nu_{h\,|\,\boldsymbol{w}}(h)\big)$ converges in distribution to a mean-zero normal random vector; in particular, $\arg\max_h \hat{\nu}_{h\,|\,\boldsymbol{w}}(h)$ converges to $\arg\max_h \nu_{h\,|\,\boldsymbol{w}}(h)$ at the rate of $n^{1/2}$. This gives a successful implementation of the approach in Doss and Linero (2018). To recapitulate, the method involves two distinct steps: (1) construct a Markov chain whose invariant distribution is the posterior distribution of $(\boldsymbol{\psi}, h)$ given $\boldsymbol{w}$, and (2) develop a procedure for using the output of the chain to efficiently estimate the marginal posterior density of $h$. We develop two ways of carrying out Step 1. The first is based on a combination of Hamiltonian Monte Carlo and the CGS of Griffiths and Steyvers (2004), and the second is based on an implementation of data augmentation. We also develop two ways of carrying out Step 2. Then we compare and contrast the various combinations and make a recommendation for which overall procedure to use.

The paper is organized as follows. In Section 2, we develop two Markov chains on $(\boldsymbol{z}, h)$ with invariant distribution equal to the marginal posterior distribution of $(\boldsymbol{z}, h)$ given $\boldsymbol{w}$ (we argue that it is possible to deal with $(\boldsymbol{z}, h)$ instead of with $(\boldsymbol{\psi}, h)$ and that doing so is more convenient). In Section 3, we present two methods for estimation of the marginal posterior density of $h$ given $\boldsymbol{w}$ from the output of the Markov chains. Also in Section 3 we give results on consistency and asymptotic normality of the resulting estimates of $\arg\max_h m(h)$, and explain how to construct confidence sets for $\arg\max_h m(h)$. In Section 4 we give the results of experiments on synthetic and real data sets, compare and contrast the various methods we propose, also compare them with other methods in the current literature, and make our recommendations.

# 2 Two Markov Chains Whose Invariant Distribution Is the Posterior Distribution of $(\boldsymbol{z}, h)$

This section consists of two parts. In Section 2.1 we show how we can use Hamiltonian Monte Carlo (HMC) in conjunction with the CGS of Griffiths and Steyvers (2004) to develop a Markov chain with invariant distribution equal to $\nu_{(\boldsymbol{z},h)\,|\,\boldsymbol{w}}$. In Section 2.2 we introduce an augmentation vector $A$ and develop a chain that runs on the triple $(\boldsymbol{z}, h, A)$, and for which the marginal sequence $(\boldsymbol{z}^{(1)}, h^{(1)}), (\boldsymbol{z}^{(2)}, h^{(2)}), \ldots$ also has invariant distribution equal to $\nu_{(\boldsymbol{z},h)\,|\,\boldsymbol{w}}$; we also provide a theorem that states that this chain is uniformly ergodic. We compare these two chains in Section 4, where we shall see that which chain is preferable depends on certain features of the corpus, such as its size. Note that we are dealing with the pair $(\boldsymbol{z}, h)$, whereas the development in Doss and Linero (2018) deals with the pair $(\boldsymbol{\psi}, h)$, where $\boldsymbol{\psi} = (\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{z})$ is the full parameter. However, there is no problem in working with $(\boldsymbol{z}, h)$ as long as we can efficiently estimate the marginal posterior density of $h$ given $\boldsymbol{w}$ from the pairs $(\boldsymbol{z}^{(1)}, h^{(1)}), (\boldsymbol{z}^{(2)}, h^{(2)}), \ldots$ by Rao-Blackwellization or some other method which, as we shall see, is the case.

Before proceeding, we need to establish our notation for distributions, since there are many distributions involved in our development. When $h$ is not random, we use $\nu^{(h)}$ for the prior distribution of $\boldsymbol{\psi}$ and add subscripts as necessary to denote conditional and marginal distributions. Thus for example, $\nu^{(h)}_{\boldsymbol{\psi}\,|\,\boldsymbol{w}}$ denotes the posterior distribution of $\boldsymbol{\psi}$ in the LDA model indexed by $h$. When $h$ is random, we use $\nu$ for the joint prior distribution of $(h, \boldsymbol{\psi})$ and again use subscripts as necessary. So for example, $\nu_h$ denotes the prior distribution of $h$, and $\nu_{h\,|\,\boldsymbol{w}}$ denotes the marginal posterior distribution of $h$. Also, note that $\nu^{(h)}$, $\nu$, $\nu_{h\,|\,\boldsymbol{w}}$, etc. are probability measures; however, we will on occasion slightly abuse notation and use the same symbol to denote both the probability measure and its density, if this does not cause confusion. Thus, when we write $\nu_{h\,|\,\boldsymbol{w}}(h) \propto m(h)u(h)$, this will be understood to be a statement regarding densities.

In Section 1 we took the prior on $h$, $\nu_h$, to be the uniform distribution on $\mathcal{H}$, which we temporarily assumed was a compact set. In fact, we prefer to take $\mathcal{H} = (0, \infty)^{T+1}$, and it turns out that the posterior corresponding to the uniform prior on $(0, \infty)^{T+1}$ is improper. This is shown in Section 1 of Xia and Doss (2019). Therefore we will take $\nu_h$ to be a proper prior. When $\nu_h$ is a proper prior, we need to take the statement $\nu_{h\,|\,\boldsymbol{w}}(h) \propto m(h)\nu_h(h)$ and rewrite it as $\nu_{h\,|\,\boldsymbol{w}}(h)/\nu_h(h) \propto m(h)$, so we will need to estimate the maximizer of $\nu_{h\,|\,\boldsymbol{w}}/\nu_h$, rather than the mode of $\nu_{h\,|\,\boldsymbol{w}}$, but this does not create any difficulties either with the theory or in practice. We will use gamma priors

because of their conjugacy properties; specifically, we will take $\nu_h(h) = g_{a,b}(\eta)\prod_{t=1}^{T} g_{a,b}(\alpha_t)$, where $g_{a,b}(u) \propto u^{a-1}\exp(-bu)$. For small $b$, $g_{1,b}$ is nearly uniform (in the sense that as $b \to 0$, $g_{1,b}$ restricted to a bounded set converges to the uniform distribution on that set). So in our experiments in Section 4 we will use $g_{1,b}$ with a small $b$; however, because we are adjusting via division by $\nu_h$, any gamma distribution would work.

We digress briefly and consider the following natural question: Since we are dealing with a fully-Bayes approach, why not stop there, i.e. why do we need to go on and maximize the marginal likelihood of $h$ to implement an empirical Bayes approach? The fully-Bayes approach, which goes under the general name of "Bayesian model averaging," can be very useful. On the other hand, there are several good reasons why one may want to avoid it. First, to use a fully-Bayes approach, we must specify the prior on $h$, and as mentioned in Section 1, this choice can have a great influence on the analysis. Thus, two different analysts can reach different conclusions. In contrast, the empirical Bayes approach consists of maximizing the marginal likelihood, which does not involve any prior on $h$. Our implementation of the empirical Bayes approach does involve putting a prior on $h$, but this is just a mechanism for obtaining the maximizer. To clarify, any prior on $h$ would yield the same estimate of $h$, so we are free to use any prior we want, and our choice is based on convenience. Second, one may wish to do Bayesian model selection, as opposed to Bayesian model averaging, because the subsequent inference is then more parsimonious and interpretable. The issues surrounding the choice of empirical Bayes and fully-Bayes inference are discussed more fully in George and Foster (2000) and Robert (2001, Chapter 7).

Before developing our Markov chain algorithms, we will express the joint distribution of $(\boldsymbol{z}, h)$ up to a normalizing constant, and in order to do that we need to review some notation which is standard when using the LDA model. Let $n_{dt} = \sum_{i=1}^{n_d} z_{dit}$ denote the number of words in document $d$ assigned to topic $t$; let $m_{dtv} = \sum_{i=1}^{n_d} z_{dit} w_{div}$ denote the number of words in document $d$ for which the latent topic is $t$ and the index of the word in the vocabulary is $v$; let $m_{\cdot tv} = \sum_{d=1}^{D} m_{dtv}$ denote the number of words in the corpus for which the latent topic is $t$ and the vocabulary element is $v$; and let $m_{\cdot t\cdot} = \sum_{v=1}^{V} m_{\cdot tv}$ denote the number of words in the corpus for which the latent topic is $t$.

For the model in which $h$ is not random, the prior distribution of $\boldsymbol{\psi}$ is given by lines 1–3 of the LDA model description, and is

$$\nu^{(h)}(\boldsymbol{\psi}) = \left[\prod_{d=1}^{D}\left(\frac{\Gamma\left(\sum_{t=1}^{T}\alpha_t\right)}{\prod_{t=1}^{T}\Gamma(\alpha_t)}\prod_{t=1}^{T}\theta_{dt}^{n_{dt}+\alpha_t-1}\right)\right]\left[\prod_{t=1}^{T}\left(\frac{\Gamma(V\eta)}{\Gamma(\eta)^V}\prod_{v=1}^{V}\beta_{tv}^{\eta-1}\right)\right]. \tag{2.1}$$

When $h$ is random, the joint prior distribution of $(\boldsymbol{\psi}, h)$ is obtained by multiplying the expression for $\nu^{(h)}(\boldsymbol{\psi})$ given in (2.1) by $\big[\prod_{t=1}^{T} \alpha_t^{a-1} \exp(-b\alpha_t)\big][\eta^{a-1} \exp(-b\eta)]$. The joint posterior distribution of $(\boldsymbol{\psi}, h)$ is obtained (up to a normalizing constant) by further multiplying by the likelihood (given by line 4 of the LDA model description) and, finally, the joint posterior distribution of $(\boldsymbol{z}, h)$ is obtained by integrating out $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$. This gives

$$\begin{aligned}\nu_{(\boldsymbol{z},h)\,|\,\boldsymbol{w}}(\boldsymbol{z}, h) \propto &\left[\prod_{d=1}^{D}\left(\frac{\Gamma\big(\sum_{t=1}^{T}\alpha_t\big)}{\prod_{t=1}^{T}\Gamma(\alpha_t)}\frac{\prod_{t=1}^{T}\Gamma(n_{dt}+\alpha_t)}{\Gamma\big(n_d+\sum_{t=1}^{T}\alpha_t\big)}\right)\right]\left[\prod_{t=1}^{T}\alpha_t^{a-1}\exp(-b\alpha_t)\right]\\ &\times\left[\prod_{t=1}^{T}\left(\frac{\Gamma(V\eta)}{\Gamma(\eta)^V}\frac{\prod_{v=1}^{V}\Gamma(m_{\cdot tv}+\eta)}{\Gamma(m_{\cdot t\cdot}+V\eta)}\right)\right]\eta^{a-1}\exp(-b\eta).\end{aligned} \tag{2.2}$$

We will construct two MCMC algorithms for sampling from this distribution, and our general approach is as follows. Write $(\boldsymbol{z}, h) = (z_{11}, \ldots, z_{1n_1}, \ldots, z_{D1}, \ldots, z_{Dn_D}, h)$. The CGS of Griffiths and Steyvers (2004) runs on the $N$-dimensional vector $(z_{11}, \ldots, z_{1n_1}, \ldots, z_{D1}, \ldots, z_{Dn_D})$, updating one variable at a time, with $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ integrated out. Let $Q_h(\boldsymbol{z}, \boldsymbol{z}')$ denote the Markov transition function for the CGS for the LDA model indexed by $h$. The Markov transition function $Q_h$ leaves the posterior distribution $\nu^{(h)}_{\boldsymbol{z}\,|\,\boldsymbol{w}}$ invariant; equivalently, $Q_h$ leaves $\nu_{\boldsymbol{z}\,|\,(h,\boldsymbol{w})}$ invariant. Now, suppose that $P_{\boldsymbol{z}}(h, h')$ is a Markov transition function that leaves $\nu_{h\,|\,(\boldsymbol{z},\boldsymbol{w})}$ invariant. It then follows that the composition of $Q_h$ and $P_{\boldsymbol{z}}$ leaves $\nu_{(\boldsymbol{z},h)\,|\,\boldsymbol{w}}$ invariant. In other words, if we update $\boldsymbol{z}$ using $Q_h$ (in $N$ steps) and then update $h$ using $P_{\boldsymbol{z}}$, then the result is one cycle of a Markov chain whose invariant distribution is $\nu_{(\boldsymbol{z},h)\,|\,\boldsymbol{w}}$. We will construct two Markov transition functions which leave $\nu_{h\,|\,(\boldsymbol{z},\boldsymbol{w})}$ invariant, one based on Hamiltonian Monte Carlo (Section 2.1), and the other based on data augmentation (Section 2.2). Either of these can be used, in conjunction with the CGS, to produce a Markov chain on $(\boldsymbol{z}, h)$ with the desired invariant distribution.

Let $\boldsymbol{w}_{(-di)}$ denote the collection of all the words in the corpus except for $w_{di}$, and let $\boldsymbol{z}_{(-di)}$ denote the vector consisting of all the $z_{kl}$'s except for $z_{di}$. Let $n_{dt(-di)}$, $m_{\cdot tv(-di)}$, and $m_{\cdot t\cdot(-di)}$ be the variables $n_{dt}$, $m_{\cdot tv}$, and $m_{\cdot t\cdot}$, respectively, except that they are based on $\boldsymbol{w}_{(-di)}$ and $\boldsymbol{z}_{(-di)}$, instead of $\boldsymbol{w}$ and $\boldsymbol{z}$. The conditional distributions needed to run the CGS are given by

$$\nu_{z_{di}\,|\,(\boldsymbol{z}_{-(di)},h,\boldsymbol{w})}(e_t) \propto \left(\frac{n_{dt(-di)}+\alpha_t}{n_d-1+\sum_{t'=1}^{T}\alpha_{t'}}\right)\left(\frac{m_{\cdot tv(-di)}+\eta}{m_{\cdot t\cdot(-di)}+V\eta}\right), \tag{2.3}$$

where $e_t$ denotes the $t^{\text{th}}$ unit vector, i.e. the vector with a 1 at the $t^{\text{th}}$ position and 0's elsewhere. Formula (2.3) is given without proof in Griffiths and Steyvers (2004). Its derivation is not trivial and is given, for example, in Chen (2015). In the next two subsections we turn to our methods for sampling from $\nu_{h\,|\,(\boldsymbol{z},\boldsymbol{w})}$.

We now pause to consider the big picture. We have chosen to use the CGS to sample $\boldsymbol{z}$. Another possibility is to approximate $\nu^{(h)}_{\boldsymbol{z}\,|\,\boldsymbol{w}}$ (equivalently $\nu_{\boldsymbol{z}\,|\,(h,\boldsymbol{w})}$) via variational methods; see the Appendix to Blei et al. (2003) for a description. Variational inference has the advantage that it is very fast, and so can handle very large corpora. Unfortunately, there are no useful theoretical bounds on the approximation error. If we were to use variational inference to estimate $\nu_{\boldsymbol{z}\,|\,(h,\boldsymbol{w})}$ then, because an approximation would be used in every sweep through $(\boldsymbol{z}, h)$, we would have no guarantee that the resulting sequence $(\boldsymbol{z}^{(1)}, h^{(1)}), (\boldsymbol{z}^{(2)}, h^{(2)}), \ldots$ has $\nu_{(\boldsymbol{z},h)\,|\,\boldsymbol{w}}$ as its limiting distribution, and in fact we would have no guarantee that the sequence even has a limiting distribution. For this reason, we did not use the variational approximation. A viable alternative to using the CGS is to use the "Grouped Gibbs Sampler," which is a two-cycle Gibbs sampler that runs on the *pair* $(\boldsymbol{z}, (\boldsymbol{\beta}, \boldsymbol{\theta}))$. This sampler, discussed in Section 5, can handle very large corpora because it is amenable to distributed computing: given $\boldsymbol{z}$ and $\boldsymbol{w}$, the $\theta_d$'s and $\beta_t$'s are all independent, so can be updated simultaneously by different processors; and given $\boldsymbol{\beta}$, $\boldsymbol{\theta}$, and $\boldsymbol{w}$, the components of $\boldsymbol{z}$ are independent, so can also be updated simultaneously by different processors. Although we did not use it in the present paper, this alternative method of sampling is potentially a good choice, especially when dealing with large corpora.

## 2.1 A Markov Chain Based on Hamiltonian Monte Carlo

It is not really possible to explain how HMC can be used to sample $h$ without first explaining what it is. HMC is a highly-developed methodology, and because the tutorials on it that currently exist in the literature are quite lengthy and detailed, in Section 2 of Xia and Doss (2019) we review HMC, and our description is the simplest that enables the reader to understand how we apply it. The reader who is interested in a more detailed description of HMC is referred to Neal (2011), on which most of our review is based. Below, we explain how we use HMC to construct a Markov transition function whose invariant distribution is $\nu_{h\,|\,(\boldsymbol{z},\boldsymbol{w})}$.

### Application of HMC to Sampling the Hyperparameters in the LDA Model

We now show how HMC can be used to make a draw from $\nu_{h\,|\,(\boldsymbol{z},\boldsymbol{w})}$, and for this we need an expression for $\nu_{h\,|\,(\boldsymbol{z},\boldsymbol{w})}$, at least up to a normalizing constant. Note that $\nu_{h\,|\,(\boldsymbol{z},\boldsymbol{w})}$ has the same form as $\nu_{(\boldsymbol{z},h)\,|\,\boldsymbol{w}}$, for which an expression is given in (2.2), except that $n_d$, $n_{dt}$, $m_{\cdot tv}$, and $m_{\cdot t\cdot}$, which are functions of $\boldsymbol{z}$, are now viewed as constants (these quantities are defined in the paragraph above (2.1)). The leapfrog algorithm implicitly assumes that the support of $\nu_{h\,|\,(\boldsymbol{z},\boldsymbol{w})}$ is $\mathbb{R}^{T+1}$,

and in the values for $h$ that it returns, some components may be negative. In fact, the support of $\nu_{h\,|\,(\boldsymbol{z},\boldsymbol{w})}$ is $(0,\infty)^{T+1}$. One way of handling this problem is to simply allow the Metropolis acceptance probability to deal with it: values of $h$ with negative components are automatically rejected. Unfortunately, when $\dim(h)$ is large, this can lead to an excessively high rejection rate. An alternative solution, which we have taken instead, is to simply apply a component-wise log transformation. Let $\tilde{h} = (A_1, \ldots, A_T, B)$, where $A_t = \log(\alpha_t)$, $t = 1, \ldots, T$, and $B = \log(\eta)$. We work with the induced distribution on $\tilde{h}$, which is given by

$$\begin{aligned}\nu_{\tilde{h}\,|\,(\boldsymbol{z},\boldsymbol{w})}(\tilde{h}) \propto & \left[\prod_{d=1}^{D}\left(\frac{\Gamma\big(\sum_{t=1}^{T}\exp(A_t)\big)}{\prod_{t=1}^{T}\Gamma(\exp(A_t))}\,\frac{\prod_{t=1}^{T}\Gamma(n_{dt}+\exp(A_t))}{\Gamma\big(n_d+\sum_{t=1}^{T}\exp(A_t)\big)}\right)\right] \\ & \times\left[\prod_{t=1}^{T}\exp\{(a-1)A_t\}\exp\{-b\exp(A_t)\}\right]\exp\big(\textstyle\sum_{t=1}^{T}A_t\big) \\ & \times\left[\prod_{t=1}^{T}\left(\frac{\Gamma(V\exp(B))}{\Gamma(\exp(B))^V}\,\frac{\prod_{v=1}^{V}\Gamma(m_{\cdot tv}+\exp(B))}{\Gamma(m_{\cdot t\cdot}+V\exp(B))}\right)\right] \\ & \times\exp\{(a-1)B\}\exp\{-b\exp(B)\}\exp(B).\end{aligned} \tag{2.4}$$

Let $U(\tilde{h}) = -\log\big(\nu_{\tilde{h}\,|\,(\boldsymbol{z},\boldsymbol{w})}(\tilde{h})\big)$, which is the function whose gradient we need in order to run the leapfrog algorithm. With $C$ denoting the normalizing constant in (2.4), we have

$$\begin{aligned}U(\tilde{h}) = -\sum_{d=1}^{D}\Bigg[\log\Bigg(\Gamma\Bigg(\sum_{t=1}^{T}\exp(A_t)\Bigg)\Bigg) - \sum_{t=1}^{T}\log\big(\Gamma(\exp(A_t))\big) + \sum_{t=1}^{T}\log\big(\Gamma(n_{dt}+\exp(A_t))\big) \\ - \log\big\{\Gamma\big(n_d+\textstyle\sum_{t=1}^{T}\exp(A_t)\big)\big\}\Bigg] - a\textstyle\sum_{t=1}^{T}A_t + b\sum_{t=1}^{T}\exp(A_t) \\ - \textstyle\sum_{t=1}^{T}\Big[\log\big(\Gamma(V\exp(B))\big) - V\log\big(\Gamma(\exp(B))\big) + \sum_{v=1}^{V}\log\big(\Gamma(m_{\cdot tv}+\exp(B))\big) \\ - \log\big(\Gamma(m_{\cdot t\cdot}+V\exp(B))\big)\Big] - aB + b\exp(B) + \log(C).\end{aligned} \tag{2.5}$$

We can now get a closed-form expression for the gradient of $U$ (the constant $C$ has no effect on the gradient of $U$), which involves the digamma function $\Psi$, defined by $\Psi(x) = \partial\log(\Gamma(x))/\partial x$.

From (2.5) we get

$$\begin{aligned}
\frac{\partial U}{\partial A_t} &= -\exp(A_t)\Big\{D\Big[\Psi\big(\textstyle\sum_{t'=1}^{T}\exp(A_{t'})\big) - \Psi(\exp(A_t))\Big] \\
&\qquad + \sum_{d=1}^{D}\Big[\Psi(n_{dt}+\exp(A_t)) - \Psi\big(n_d + \textstyle\sum_{t'=1}^{T}\exp(A_{t'})\big)\Big] - b\Big\} - a, \\
\frac{\partial U}{\partial B} &= -\exp(B)\Big\{TV\big[\Psi(V\exp(B)) - \Psi(\exp(B))\big] \\
&\qquad + \sum_{t=1}^{T}\sum_{v=1}^{V}\Psi(m_{\cdot tv}+\exp(B)) - V\sum_{t=1}^{T}\Psi(m_{\cdot t\cdot} + V\exp(B)) - b\Big\} - a.
\end{aligned} \tag{2.6}$$

With formulas (2.5) and (2.6) in place, we can now combine Algorithm S-1 in Section 2 of Xia and Doss (2019) and the CGS to obtain Algorithm 1, which describes a complete scheme for generating a Markov chain with invariant distribution equal to $\nu_{(\boldsymbol{z},h)\,|\,\boldsymbol{w}}$. (Note that although HMC is used to generate from $\nu_{\tilde{h}\,|\,(\boldsymbol{z},\boldsymbol{w})}$, because we transform $\tilde{h}$ back to the original scale, the final output of the algorithm is a sequence whose invariant distribution is $\nu_{h\,|\,(\boldsymbol{z},\boldsymbol{w})}$.)

## 2.2 A Markov Chain Based on Data Augmentation

HMC requires the selection of two tuning parameters. In contrast, data augmentation does not involve any tuning parameters. The method is described as follows. Suppose that $f_X$ is an intractable density on a space $\mathsf{X}$, and suppose that $f$ is a density on the space $\mathsf{X}\times\mathsf{Y}$ with the following two properties: (i) $\int_{\mathsf{Y}} f(x,y)\,dy = f_X(x)$ for all $x\in\mathsf{X}$, and (ii) simulation from the associated conditional pdf's $f_{X\,|\,Y}$ and $f_{Y\,|\,X}$ is feasible. The data augmentation algorithm works by successively generating from $f_{X\,|\,Y}$ and $f_{Y\,|\,X}$, obtaining $(X_1,Y_1),(X_2,Y_2),\ldots$. This sequence of pairs is a Markov chain with invariant density $f$, and the marginal sequence $X_1,X_2,\ldots$ is a Markov chain with invariant density $f_X$. Transparently, the procedure described above is nothing more than a two-cycle Gibbs sampler. Given an intractable density $f_{X^{(0)}}$, it is sometimes possible to devise an augmentation scheme involving $k$ variables $X^{(1)},\ldots,X^{(k)}$, and data augmentation is then simply a $(k+1)$-cycle Gibbs sampler on the vector $(X^{(0)},X^{(1)},\ldots,X^{(k)})$. The marginal sequence $X_1^{(0)},X_2^{(0)},\ldots$ is then not necessarily a Markov chain, but if the Markov chain $(X_1^{(0)},X_1^{(1)},\ldots,X_1^{(k)}),(X_2^{(0)},X_2^{(1)},\ldots,X_2^{(k)}),\ldots$ is ergodic, then the marginal sequence $X_1^{(0)},X_2^{(0)},\ldots$ has limiting density equal to $f_{X^{(0)}}$. In the above, $f_X$ and $f$ are densities with respect to Lebesgue measure, but the description can be extended to more general settings. A nice exposition to data augmentation is Hobert (2011), which also discusses conditions under which one

**Algorithm 1:** Sampling $(\boldsymbol{z}, h)$

**Data:** Observed words $\boldsymbol{w}$

**Result:** A Markov chain $(\boldsymbol{z}^{(1)}, h^{(1)}), (\boldsymbol{z}^{(2)}, h^{(2)}), \ldots$ with invariant distribution equal to $\nu_{(\boldsymbol{z},h)\,|\,\boldsymbol{w}}$

Initialize $\boldsymbol{z}^{(0)}, h^{(0)}, \epsilon, L, n$. Let $\tilde{h}^{(0)} = \log(h^{(0)})$, $U(\tilde{h})$ be defined by (2.5), and $\nabla_{\tilde{h}} U(\tilde{h})$ be defined by (2.6);

**for** $i = 1, \ldots, n$ **do**

`// Update` $\boldsymbol{z}$ `given` $h$ `by the CGS`

  **for** *each* $z_{di}$ *in* $\boldsymbol{z}$ **do**

    update $z_{di}$ using the multinomial distribution $\nu_{z_{di}\,|\,(\boldsymbol{z}_{-(di)}, h, \boldsymbol{w})}$ given by (2.3);

`// Update` $h$ `given` $\boldsymbol{z}$ `by HMC`

  generate $y^{(0)} \sim \mathcal{N}(0, M)$;

  set $\tilde{h}^{(i)} \leftarrow \tilde{h}^{(i-1)}$, $h^* \leftarrow \tilde{h}^{(i-1)}$, $y^* \leftarrow y^{(0)}$;

  **for** $j = 1, \ldots, L$ **do**

    set $y^* \leftarrow y^* - \epsilon \nabla_h U(h^*)/2$;

    set $h^* \leftarrow h^* + \epsilon M^{-1} y^*$;

    set $y^* \leftarrow y^* - \epsilon \nabla_h U(h^*)/2$;

  set $r = \exp\{-U(h^*) - (y^*)^\top M^{-1} y^*/2 + U(\tilde{h}^{(i-1)}) + (y^{(0)})^\top M^{-1} y^{(0)}/2\}$;

  with probability $\min\{1, r\}$ set $\tilde{h}^{(i)} \leftarrow h^*$;

  set $h^{(i)} = \exp(\tilde{h}^{(i)})$;

can establish the needed ergodicity. In our situation, the variable $h$ will play the role of $X^{(0)}$, and $\nu_{h\,|\,(\boldsymbol{z},\boldsymbol{w})}$ will correspond to $f_{X^{(0)}}$. It is worth emphasizing that data augmentation plays a role only in the local generation of $h$ in the $(\boldsymbol{z}, h)$ pair.

Our data augmentation scheme is based on a strategy originally used in Escobar and West (1995) to estimate the posterior distribution of the precision parameter in a mixture of Dirichlet processes problem in which there is a prior on this parameter. A related strategy was used in Teh et al. (2006) in the context of hierarchical Dirichlet processes, and the strategy was also used by Newman et al. (2009) in a version of the hierarchical Dirichlet processes model suitable for distributed computing. Our scheme is based on the two facts below. Recall that the Beta function is defined by $B(a, b) = \Gamma(a)\Gamma(b)/\Gamma(a + b)$, for any $a, b > 0$, and gives the normalizing constant for the un-normalized density $q^{a-1}(1 - q)^{b-1}$ on $(0, 1)$.

**Fact 1** For any $u > 0$ and positive integer $n$,

$$\frac{\Gamma(u)}{\Gamma(u+n)} = \frac{B(u,n)}{\Gamma(n)} = \frac{1}{\Gamma(n)} \int_0^1 q^{u-1}(1-q)^{n-1}\, dq.$$

For $n$ a positive integer, consider the product $u(u+1)\cdots(u+n-1)$, which is a polynomial (in $u$) of degree $n$, call it $p_n(u)$. The coefficients of this polynomial are called "unsigned Stirling numbers of the first kind," and are denoted $S(n,i)$, $i = 0, \ldots, n$, i.e.

$$p_n(u) = \sum_{i=0}^{n} S(n,i) u^i. \tag{2.7}$$

**Fact 2** For any $u > 0$ and positive integer $n$,

$$\frac{\Gamma(u+n)}{\Gamma(u)} = u(u+1)\cdots(u+n-1) = \sum_{i=0}^{n} S(n,i) u^i.$$

Note that Fact 2 is a tautology.

Let $u > 0$ and consider the discrete random variable with values in $\{0, 1, \ldots, n\}$ and probability mass function $\pi_u(i) \propto S(n,i)u^i/c$, where $c$ is a normalizing constant. Fact 2 states that $\Gamma(u+n)/\Gamma(u)$ is the normalizing constant, i.e.

$$\pi_u(i) = \frac{\Gamma(u)}{\Gamma(u+n)} S(n,i) u^i. \tag{2.8}$$

The probability mass function $\pi_u$ appears in consideration of samples associated with the Dirichlet process, as follows. Consider the Dirichlet process $\mathcal{D}(G,u)$, where $G$ is the base probability measure, assumed continuous, and $u > 0$ is the concentration parameter. Suppose $F \sim \mathcal{D}(G,u)$ and $\xi_1, \ldots, \xi_n \overset{\text{iid}}{\sim} F$. As we will soon see, $\pi_u$ arises as the distribution of the number of distinct values among $\xi_1, \ldots, \xi_n$. For $l = 1, \ldots, n$, let $Y_l$ be the indicator that $\xi_l$ is not equal to any of its predecessors. As is well known, $Y_l \sim \text{Bernoulli}\big(p = u/(u+l-1)\big)$. Therefore, denoting $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ and $\boldsymbol{y} = (y_1, \ldots, y_n)$, we have

$$P(\boldsymbol{Y} = \boldsymbol{y}) = \frac{\Gamma(u)}{\Gamma(u+n)} u^{\sum_{l=1}^n y_l} \prod_{l=1}^{n} (l-1)^{1-y_l}.$$

It follows that for $I \coloneqq \sum_{l=1}^n Y_l$, for $i = 0, 1, \ldots, n$, we have

$$\begin{aligned} P(I=i) = \sum_{\sum_{l=1}^n y_l = i} P(\boldsymbol{Y} = \boldsymbol{y}) &= \frac{\Gamma(u)}{\Gamma(u+n)} \sum_{\boldsymbol{y}: \sum_{l=1}^n y_l = i} u^{\sum_{l=1}^n y_l} \prod_{l=1}^{n} (l-1)^{1-y_l} \\ &= \frac{\Gamma(u)}{\Gamma(u+n)} u^i \sum_{\boldsymbol{y}: \sum_{l=1}^n y_l = i} \prod_{l=1}^{n} (l-1)^{1-y_l} \\ &= \frac{\Gamma(u)}{\Gamma(u+n)} u^i S(n,i), \end{aligned} \tag{2.9}$$

where the last equality in (2.9) comes from the fact that $\sum_{\boldsymbol{y}:\sum_{l=1}^{n} y_l = i} \prod_{l=1}^{n} (l-1)^{1-y_l}$ is precisely the coefficient of $u^i$ for the polynomial $p_n$ defined in (2.7). From (2.9) we see that the distribution of the random variable $I$ is equal to the distribution $\pi_u$ defined in (2.8). The significance of this is that if we wish to generate a random variable from the distribution $\pi_u$, for some $u$, then we can do this by simply generating $n$ independent Bernoullis and taking their sum, instead of dealing with the Stirling numbers, which are computationally expensive to obtain.

We now return to the conditional distribution $\nu_{h\,|\,(\boldsymbol{z},\boldsymbol{w})}$ which, as mentioned in the beginning of Section 2.1, is the same as $\nu_{(\boldsymbol{z},h)\,|\,\boldsymbol{w}}$, except that $n_d$, $n_{dt}$, $m_{\cdot tv}$, and $m_{\cdot t\cdot}$ are viewed as fixed constants. For convenience, we write it explicitly here, so we can notice that for this distribution, $\boldsymbol{\alpha}$ and $\eta$ are mutually independent given $\boldsymbol{z}$ and $\boldsymbol{w}$. We have

$$\begin{aligned}
\nu_{h\,|\,(\boldsymbol{z},\boldsymbol{w})}(h) \propto & \left[\prod_{d=1}^{D}\left(\frac{\Gamma\big(\sum_{t=1}^{T}\alpha_t\big)}{\prod_{t=1}^{T}\Gamma(\alpha_t)}\frac{\prod_{t=1}^{T}\Gamma(n_{dt}+\alpha_t)}{\Gamma\big(n_d+\sum_{t=1}^{T}\alpha_t\big)}\right)\right]\left[\prod_{t=1}^{T}\alpha_t^{a-1}\exp(-b\alpha_t)\right] \\
& \times\left[\prod_{t=1}^{T}\left(\frac{\Gamma(V\eta)}{\Gamma(\eta)^V}\frac{\prod_{v=1}^{V}\Gamma(m_{\cdot tv}+\eta)}{\Gamma(m_{\cdot t\cdot}+V\eta)}\right)\right]\eta^{a-1}\exp(-b\eta) \qquad (2.10)\\
\propto\ & \nu_{\boldsymbol{\alpha}\,|\,(\boldsymbol{z},\boldsymbol{w})}(\alpha)\times\nu_{\eta\,|\,(\boldsymbol{z},\boldsymbol{w})}(\eta),
\end{aligned}$$

in self-explanatory notation. This conditional independence makes joint sampling of $\boldsymbol{\alpha}$ and $\eta$ simple: we sample $\boldsymbol{\alpha}$ and $\eta$ separately from $\nu_{\boldsymbol{\alpha}\,|\,(\boldsymbol{z},\boldsymbol{w})}(\alpha)$ and $\nu_{\eta\,|\,(\boldsymbol{z},\boldsymbol{w})}(\eta)$, respectively. We may write

$$\begin{aligned}
\nu_{\boldsymbol{\alpha}\,|\,(\boldsymbol{z},\boldsymbol{w})}(\boldsymbol{\alpha}) \propto & \left[\prod_{d=1}^{D}\left(\frac{\Gamma\big(\sum_{t=1}^{T}\alpha_t\big)}{\prod_{t=1}^{T}\Gamma(\alpha_t)}\frac{\prod_{t=1}^{T}\Gamma(n_{dt}+\alpha_t)}{\Gamma\big(n_d+\sum_{t=1}^{T}\alpha_t\big)}\right)\right]\left[\prod_{t=1}^{T}\alpha_t^{a-1}\exp(-b\alpha_t)\right] \\
= & \left[\prod_{d=1}^{D}\frac{\Gamma\big(\sum_{t=1}^{T}\alpha_t\big)}{\Gamma\big(n_d+\sum_{t=1}^{T}\alpha_t\big)}\right]\left[\prod_{d=1}^{D}\prod_{t=1}^{T}\frac{\Gamma(n_{dt}+\alpha_t)}{\Gamma(\alpha_t)}\right]\left[\prod_{t=1}^{T}\alpha_t^{a-1}\exp(-b\alpha_t)\right]. \qquad (2.11)
\end{aligned}$$

Applying Fact 1 to the first term in brackets in (2.11), we get

$$\prod_{d=1}^{D}\frac{\Gamma\big(\sum_{t=1}^{T}\alpha_t\big)}{\Gamma\big(n_d+\sum_{t=1}^{T}\alpha_t\big)} = \prod_{d=1}^{D}\left[\frac{1}{\Gamma(n_d)}\int_0^1 q_d^{\sum_{t=1}^{T}\alpha_t-1}(1-q_d)^{n_d-1}\,dq_d\right], \qquad (2.12)$$

and applying Fact 2 to the second term in brackets in (2.11), we get

$$\prod_{d=1}^{D}\prod_{t=1}^{T}\frac{\Gamma(n_{dt}+\alpha_t)}{\Gamma(\alpha_t)} = \prod_{d=1}^{D}\prod_{t=1}^{T}\sum_{i=0}^{n_{dt}} S(n_{dt},i)\alpha_t^i. \qquad (2.13)$$

In view of equations (2.12) and (2.13), we see that if we introduce the augmentation variables $\boldsymbol{I} = (I_{11},\ldots,I_{1T},\ldots,I_{D1},\ldots,I_{DT})$ and $\boldsymbol{Q} = (Q_1,\ldots,Q_D)$, then $\nu_{\boldsymbol{\alpha}\,|\,(\boldsymbol{z},\boldsymbol{w})}(\boldsymbol{\alpha})$ may be re-expressed in

an augmented form up to a normalized constant as

$$\nu_{(\boldsymbol{\alpha},\boldsymbol{I},\boldsymbol{Q})\,|\,(\boldsymbol{z},\boldsymbol{w})}(\boldsymbol{\alpha},\boldsymbol{i},\boldsymbol{q}) \propto \left[\prod_{d=1}^{D} q_d^{\sum_{t=1}^{T}\alpha_t-1}(1-q_d)^{n_d-1}\right]\left[\prod_{d=1}^{D}\prod_{t=1}^{T} S(n_{dt},i_{dt})\alpha_t^{i_{dt}}\right]\left[\prod_{t=1}^{T}\alpha_t^{a-1}\exp(-b\alpha_t)\right]. \quad (2.14)$$

We will now show that (2.14) enables us to obtain closed-form expressions for $\nu_{\boldsymbol{\alpha}\,|\,(\boldsymbol{I},\boldsymbol{Q},\boldsymbol{z},\boldsymbol{w})}$, $\nu_{\boldsymbol{I}\,|\,(\boldsymbol{\alpha},\boldsymbol{Q},\boldsymbol{z},\boldsymbol{w})}$, and $\nu_{\boldsymbol{Q}\,|\,(\boldsymbol{\alpha},\boldsymbol{I},\boldsymbol{z},\boldsymbol{w})}$, which will allow us to successively sample $\boldsymbol{\alpha}$, $\boldsymbol{I}$, and $\boldsymbol{Q}$.

Regarding $\boldsymbol{\alpha}$, from (2.14) we see that

$$\begin{aligned}\nu_{\boldsymbol{\alpha}\,|\,(\boldsymbol{I},\boldsymbol{Q},\boldsymbol{z},\boldsymbol{w})}(\boldsymbol{\alpha}) &\propto \left[\prod_{d=1}^{D} q_d\right]^{\sum_{t=1}^{T}\alpha_t}\left[\prod_{d=1}^{D}\prod_{t=1}^{T}\alpha_t^{i_{dt}}\right]\left[\prod_{t=1}^{T}\alpha_t^{a-1}\exp(-b\alpha_t)\right]\\ &= \prod_{t=1}^{T}\left[\left[\prod_{d=1}^{D} q_d\right]^{\alpha_t}\alpha_t^{\sum_{d=1}^{D} i_{dt}+a-1}\exp(-b\alpha_t)\right] \qquad (2.15)\\ &= \prod_{t=1}^{T}\Big[\alpha_t^{\sum_{d=1}^{D} i_{dt}+a-1}\exp\big\{-\big(b-\textstyle\sum_{d=1}^{D}\log(q_d)\big)\alpha_t\big\}\Big],\end{aligned}$$

which we recognize as a product of $T$ gamma densities. Thus, to sample $\boldsymbol{\alpha}$, we generate $\alpha_t$ from the gamma distribution with shape parameter $a+\sum_{d=1}^{D} i_{dt}$ and rate parameter $b-\sum_{d=1}^{D}\log(q_d)$, independently for $t=1,\ldots,T$.

Regarding $\boldsymbol{I}$, from (2.14) we see that

$$\nu_{\boldsymbol{I}\,|\,(\boldsymbol{\alpha},\boldsymbol{Q},\boldsymbol{z},\boldsymbol{w})}(\boldsymbol{i}) \propto \prod_{d=1}^{D}\prod_{t=1}^{T} S(n_{dt},i_{dt})\alpha_t^{i_{dt}}, \quad (2.16)$$

from which we note that the $I_{dt}$'s are independent given $(\boldsymbol{\alpha},\boldsymbol{Q},\boldsymbol{z},\boldsymbol{w})$. It is easy to see that if $n_{dt}=0$, then $I_{dt}$ is deterministically equal to 0. If $n_{dt}>0$ then, as we discussed earlier, $I_{dt}$ may be represented as $I_{dt}=\sum_{l=1}^{n_{dt}} I_{dt}^{(l)}$ where

$$I_{dt}^{(l)} \overset{\text{indep}}{\sim} \text{Bernoulli}\left(\frac{\alpha_t}{\alpha_t+l-1}\right), \qquad l=1,\ldots,n_{dt},$$

and this enables us to easily generate $I_{dt}^{(l)}$.

Regarding $\boldsymbol{Q}$, from (2.14) we see that

$$\nu_{\boldsymbol{Q}\,|\,(\boldsymbol{\alpha},\boldsymbol{I},\boldsymbol{z},\boldsymbol{w})}(\boldsymbol{q}) \propto \prod_{d=1}^{D}\Big[q_d^{\sum_t^T\alpha_t-1}(1-q_d)^{n_d-1}\Big], \quad (2.17)$$

which we recognize as a product of $D$ beta densities. Thus, to sample $\boldsymbol{Q}$, we generate $Q_d \sim \text{Beta}\big(\sum_{t=1}^{T}\alpha_t, n_d\big)$ independently for $d=1,\ldots,D$.

We now step back and review the big picture. Let us temporarily act as if $\eta$ is an unknown constant, i.e. in our fully-Bayes model we need to estimate the posterior distribution of $(\boldsymbol{z}, \boldsymbol{\alpha})$. Recall that $N = \sum_{d=1}^{D} n_d$. What we have described is a Gibbs sampler that runs on the $(N+T+DT+D)$-dimensional vector $(\boldsymbol{z}, \boldsymbol{\alpha}, \boldsymbol{I}, \boldsymbol{Q})$, where $\boldsymbol{z}$ is updated according to the CGS of Griffiths and Steyvers (2004), and $\boldsymbol{\alpha}$, $\boldsymbol{I}$, and $\boldsymbol{Q}$ are updated as described in the three preceding paragraphs. From the Markov chain $(\boldsymbol{z}^{(1)}, \boldsymbol{\alpha}^{(1)}, \boldsymbol{I}^{(1)}, \boldsymbol{Q}^{(1)}), (\boldsymbol{z}^{(2)}, \boldsymbol{\alpha}^{(2)}, \boldsymbol{I}^{(2)}, \boldsymbol{Q}^{(2)}), \ldots$, we may estimate the posterior distribution of $(\boldsymbol{z}, \boldsymbol{\alpha})$ given $\boldsymbol{w}$ by considering the marginal sequence $(\boldsymbol{z}^{(1)}, \boldsymbol{\alpha}^{(1)}), (\boldsymbol{z}^{(2)}, \boldsymbol{\alpha}^{(2)}), \ldots$. To estimate the posterior density of $\boldsymbol{\alpha}$, we may use Rao-Blackwellization, which uses the sequence $(\boldsymbol{z}^{(1)}, \boldsymbol{I}^{(1)}, \boldsymbol{Q}^{(1)}), (\boldsymbol{z}^{(2)}, \boldsymbol{I}^{(2)}, \boldsymbol{Q}^{(2)}), \ldots$ through (2.15). The inclusion of $\eta$, discussed next, does not make any conceptual changes to the big picture.

We now turn to sampling from $\nu_{\eta \,|\, (\boldsymbol{z}, \boldsymbol{w})}$. For this purpose we introduce the augmentation variables $\boldsymbol{J} = (J_{11}, \ldots, J_{1V}, \ldots, J_{T1}, \ldots, J_{TV})$ and $R = (R_1, \ldots, R_T)$, in order to re-express $\nu_{\eta \,|\, (\boldsymbol{z}, \boldsymbol{w})}$ in augmented form as

$$\nu_{(\eta, \boldsymbol{J}, \boldsymbol{R}) \,|\, (\boldsymbol{z}, \boldsymbol{w})}(\eta, \boldsymbol{j}, \boldsymbol{r}) \propto \left[\prod_{t=1}^{T} r_t^{V\eta - 1}(1 - r_t)^{m_{\cdot t \cdot} - 1}\right] \left[\prod_{t=1}^{T} \prod_{v=1}^{V} S(m_{\cdot tv}, j_{tv}) \eta^{j_{tv}}\right] \eta^{a-1} \exp(-b\eta), \tag{2.18}$$

which is analogous to (2.14).

For sampling $\eta$, from (2.18) we see that

$$\nu_{\eta \,|\, (\boldsymbol{R}, \boldsymbol{J}, \boldsymbol{z}, \boldsymbol{w})}(\eta) \propto \eta^{a + \sum_{t=1}^{T} \sum_{v=1}^{V} j_{tv} - 1} \exp\Big\{-\Big(b - V \sum_{t=1}^{T} \log(r_t)\Big)\eta\Big\}, \tag{2.19}$$

which we recognize as a gamma distribution with shape parameter $a + \sum_{t=1}^{T} \sum_{v=1}^{V} j_{tv}$ and rate parameter $b - V \sum_{t=1}^{T} \log(r_t)$.

For sampling $\boldsymbol{J}$, from (2.18) we see that

$$\nu_{\boldsymbol{J} \,|\, (\eta, \boldsymbol{R}, \boldsymbol{z}, \boldsymbol{w})}(\boldsymbol{j}) \propto \left[\prod_{t=1}^{T} \prod_{v=1}^{V} S(m_{\cdot tv}, j_{tv}) \eta^{j_{tv}}\right].$$

This distribution is analogous to $\nu_{\boldsymbol{I} \,|\, (\boldsymbol{\alpha}, \boldsymbol{Q}, \boldsymbol{z}, \boldsymbol{w})}$, which is given by (2.16). The $J_{tv}$'s are independent, and can be generated as follows. If $m_{\cdot tv} = 0$, then $J_{tv}$ is deterministically equal to 0. If $m_{\cdot tv} > 0$, then $J_{tv}$ may be represented as $J_{tv} = \sum_{l=1}^{m_{\cdot tv}} J_{tv}^{(l)}$, where

$$J_{tv}^{(l)} \overset{\text{indep}}{\sim} \text{Bernoulli}\left(\frac{\eta}{\eta + l - 1}\right), \qquad l = 1, \ldots, m_{\cdot tv}.$$

For sampling $\boldsymbol{R}$, from (2.18) we see that

$$\nu_{\boldsymbol{R} \,|\, (\eta, \boldsymbol{J}, \boldsymbol{z}, \boldsymbol{w})}(\boldsymbol{r}) \propto \left[\prod_{t=1}^{T} r_t^{V\eta - 1}(1 - r_t)^{m_{\cdot t \cdot} - 1}\right],$$

which we recognize as a product of $T$ beta densities. Consequently, to sample $\boldsymbol{R}$ we generate $R_t \overset{\text{indep}}{\sim} \text{Beta}(V\eta,\, m_{\cdot t\cdot}),\, t = 1, \ldots, T$.

The data augmentation algorithm runs on the parameter $\lambda = (\boldsymbol{z}, \boldsymbol{I}, \boldsymbol{Q}, \boldsymbol{J}, \boldsymbol{R}, h)$. Let $P$ denote the Markov transition function for the algorithm, i.e. $P(\lambda_0, \cdot)$ is the distribution of $\lambda_1$ given $\lambda_0$, and let $P^k(\lambda_0, \cdot)$ denote the $k$-step Markov transition function. Also, let $\Lambda$ denote the set of all possible values of $\lambda$, and $\mathcal{B}_\Lambda$ be the associated Borel sigma-field. Theorem 1 establishes *uniform ergodicity*, which is the very strong condition that there exist constants $M > 0$ and $c > 0$ such that $\|P^k(\lambda_0, \cdot) - \nu_{\lambda \,|\, \boldsymbol{w}}(\cdot)\|_{\text{TV}} \leq M(1-c)^k$ for all initial $\lambda_0 \in \Lambda$, where the total variation distance $\|\cdot\|_{\text{TV}}$ denotes the supremum over $\mathcal{B}_\Lambda$ (the geometric rate of convergence does not depend on the initial starting point $\lambda_0$).

**Theorem 1** *Let $\mathcal{H}_0$ be a bounded hyper-rectangle, and assume that the support of the prior on $h$ is contained in $\mathcal{H}_0$. Then the data augmentation chain is uniformly ergodic.*

The proof of the theorem is in Section 3 of Xia and Doss (2019). We believe that the HMC chain is geometrically ergodic, but this chain is very difficult to analyze, and we have not been able to establish the result.

# 3 Efficient Estimation of the Empirical Bayes Choice of the Hyperparameter

This section is structured as follows. In Section 3.1 we develop two methods for estimation of the marginal posterior density $\nu_{h \,|\, \boldsymbol{w}}(h)$; one is based on Rao-Blackwellization, and the other is based on an extension of Rao-Blackwellization, introduced by Chen (1994), and which is applicable when Rao-Blackwellization is not feasible. Each of these gives rise to an estimator of $\arg\max_h m(h)$. In Section 3.2 we show how to use these methods to obtain confidence sets for $\arg\max_h m(h)$, and our general approach is as follows. Recall from Section 2 that $\nu_{h \,|\, \boldsymbol{w}}(h)/\nu_h(h) \propto m(h)$. To avoid distracting minor complications, in this preamble we will assume that $\nu_h$ is the uniform prior, so that the preceding relation becomes simply $\nu_{h \,|\, \boldsymbol{w}}(h) \propto m(h)$. Suppose that for each fixed $h$, $\hat{\nu}_{h \,|\, \boldsymbol{w}}(h)$, our estimate of $\nu_{h \,|\, \boldsymbol{w}}(h)$, takes the form of an average, so that by the strong law of large numbers in the form of the ergodic theorem and by a central limit theorem for Markov chains we have

$$\hat{\nu}_{h \,|\, \boldsymbol{w}}(h) \xrightarrow{\text{a.s.}} \nu_{h \,|\, \boldsymbol{w}}(h) \qquad \text{and} \qquad n^{1/2}\big(\hat{\nu}_{h \,|\, \boldsymbol{w}}(h) - \nu_{h \,|\, \boldsymbol{w}}(h)\big) \xrightarrow{d} \mathcal{N}(0, \sigma^2(h)). \tag{3.1}$$

Now, generally speaking, consistency and asymptotic normality of $\hat{\nu}_{h\,|\,\boldsymbol{w}}(h)$ for each fixed $h$ is not sufficient to entail that $\arg\max_h \hat{\nu}_{h\,|\,\boldsymbol{w}}(h)$ converges to $\arg\max_h \nu_{h\,|\,\boldsymbol{w}}(h)$ in any sense at all. In fact, even for deterministic real-valued functions $f_n$ and $f$ defined on $\mathcal{H}$, the pointwise convergence condition $f_n(h) \to f(h)$ for each $h \in \mathcal{H}$ does not imply that $\arg\max_h f_n(h) \to \arg\max_h f(h)$, and a simple counterexample to show this is given in the Appendix of George and Doss (2018). To obtain consistency and asymptotic normality of $\arg\max_h \hat{\nu}_{h\,|\,\boldsymbol{w}}(h)$ as an estimator of $\arg\max_h \nu_{h\,|\,\boldsymbol{w}}(h)$, one needs very strong regularity conditions. These are discussed in detail in Section 3.2, but here we mention that the main ones are geometric ergodicity of the Markov chain being used, and a significant strengthening of (3.1) to the uniform versions

$$\sup_h |\hat{\nu}_{h\,|\,\boldsymbol{w}}(h) - \nu_{h\,|\,\boldsymbol{w}}(h)| \xrightarrow{\text{a.s.}} 0 \tag{3.2}$$

and

$$n^{1/2}\big(\hat{\nu}_{h\,|\,\boldsymbol{w}}(\cdot) - \nu_{h\,|\,\boldsymbol{w}}(\cdot)\big) \xrightarrow{d} G(\cdot), \tag{3.3}$$

where $G(\cdot)$ is a mean 0 Gaussian process indexed by $h \in \mathcal{H}$. In Section 4 of Xia and Doss (2019) we establish (3.2) and (3.3) for the estimate based on Rao-Blackwellization, using methods from empirical process theory. Combining this with the uniform ergodicity of the data augmentation chain asserted by Theorem 1, we obtain consistency and asymptotic normality of the estimate of $\arg\max_h \nu_{h\,|\,\boldsymbol{w}}(h)$ based on the data augmentation chain and Rao-Blackwellization. This is stated formally as Theorem 2. A corresponding result for the case of the estimate based on Chen's (1994) method is given by Theorem 3.

## 3.1 Two Methods for Estimation of the Marginal Posterior Density of the Hyperparameter

**Estimation of $\boldsymbol{\nu_{h\,|\,w}}$ via Rao-Blackwellization**

Rao-Blackwellization is immediate from the data augmentation scheme described in Section 2.2. Recall that we use $g_{a,b}$ to denote the gamma density with shape parameter $a$ and rate parameter $b$. The data augmentation scheme gives the sequence $(\boldsymbol{z}^{(k)}, \boldsymbol{\alpha}^{(k)}, \boldsymbol{I}^{(k)}, \boldsymbol{Q}^{(k)}, \eta^{(k)}, \boldsymbol{J}^{(k)}, \boldsymbol{R}^{(k)})$, $k = 1, \ldots, n$, from which we may form

$$\begin{aligned}\hat{\nu}_{h\,|\,\boldsymbol{w}}(h) &= \frac{1}{n}\sum_{k=1}^{n} \nu_{h\,|\,(\boldsymbol{w},\boldsymbol{z}^{(k)},\boldsymbol{I}^{(k)},\boldsymbol{Q}^{(k)},\boldsymbol{J}^{(k)},\boldsymbol{R}^{(k)})}(\boldsymbol{\alpha},\eta) \\ &= \frac{1}{n}\sum_{k=1}^{n}\left\{\left[\prod_{t=1}^{T} g_{a+\sum_{d=1}^{D} I_{dt}^{(k)},\, b-\sum_{d=1}^{D}\log(Q_d^{(k)})}(\alpha_t)\right] g_{a+\sum_{t=1}^{T}\sum_{v=1}^{V} J_{tv}^{(k)},\, b-V\sum_{t=1}^{T}\log(R_t^{(k)})}(\eta)\right\}\end{aligned} \tag{3.4}$$

(see (2.15) and (2.19)), which uses only the $\boldsymbol{I}$, $\boldsymbol{Q}$, $\boldsymbol{J}$, and $\boldsymbol{R}$ components of the sequence. Note that if we use the data augmentation chain, then the $\boldsymbol{I}, \boldsymbol{Q}, \boldsymbol{J}, \boldsymbol{R}$ variables are already available. If we use the HMC chain, it is still possible to do Rao-Blackwellization: from the sequence $\{(\boldsymbol{z}^{(k)}, h^{(k)}),\ k = 1, \ldots, n\}$ produced by HMC, we can generate the augmentation variables $\{(\boldsymbol{I}^{(k)}, \boldsymbol{Q}^{(k)}, \boldsymbol{J}^{(k)}, \boldsymbol{R}^{(k)}),\ k = 1, \ldots, n\}$, as explained in Section 2.2, and use these to compute the estimator in (3.4).

**Estimation of $\nu_{h \mid \boldsymbol{w}}$ Through the Importance Weighted Marginal Density Method of Chen**

Suppose that $(X_1, Y_1), (X_2, Y_2), \ldots$ is a Markov chain with invariant density $f_{X,Y}$ on a space $\mathsf{X} \times \mathsf{Y}$ where $\mathsf{Y}$ is Euclidean. For the purpose of estimating the marginal density $f_Y$, Chen (1994) introduced a generic procedure, the so-called Importance Weighted Marginal Density Estimation Method, which is described as follows for our context, in which $\boldsymbol{z}$ corresponds to $X$, $h$ corresponds to $Y$, and our Markov chain is $(\boldsymbol{z}^{(1)}, h^{(1)}), (\boldsymbol{z}^{(2)}, h^{(2)}), \ldots$. Let $\mathcal{Z}$ be the set of possible values of $\boldsymbol{z}$ and let $\{\omega_{\boldsymbol{z}}(\cdot),\ \boldsymbol{z} \in \mathcal{Z}\}$ be a family of densities on $\mathcal{H}$. To estimate $\nu_{h \mid \boldsymbol{w}}$ we use the estimator $\hat{\nu}_{h \mid \boldsymbol{w}}$ whose value at $h^*$ is given by

$$\hat{\nu}_{h \mid \boldsymbol{w}}(h^*) = \frac{1}{n} \sum_{i=1}^{n} \omega_{\boldsymbol{z}^{(i)}}(h^{(i)}) \frac{\nu_{(\boldsymbol{z},h) \mid \boldsymbol{w}}(\boldsymbol{z}^{(i)}, h^*)}{\nu_{(\boldsymbol{z},h) \mid \boldsymbol{w}}(\boldsymbol{z}^{(i)}, h^{(i)})}. \tag{3.5}$$

Note that to calculate (3.5), we need only to know $\nu_{(\boldsymbol{z},h) \mid \boldsymbol{w}}$ up to a normalizing constant, and this is given by the expression on the right side of (2.2). A proof that for every $h^*$, $\hat{\nu}_{h \mid \boldsymbol{w}}(h^*)$ converges almost surely to $\nu_{h \mid \boldsymbol{w}}(h^*)$ is given in Section 5 of Xia and Doss (2019). In principle, any family $\{\omega_{\boldsymbol{z}},\ \boldsymbol{z} \in \mathcal{Z}\}$ of densities can be used in (3.5), but Chen (1994) showed that the choice $\omega_{\boldsymbol{z}} = \nu_{h \mid (\boldsymbol{z},\boldsymbol{w})}$ is optimal in the sense of minimizing the asymptotic variance and, moreover, for this choice the estimator reduces to the Rao-Blackwellized estimate $\hat{\nu}^{\mathrm{RB}}_{h \mid \boldsymbol{w}}(h^*) = (1/n) \sum_{i=1}^{n} \nu_{h \mid (\boldsymbol{z}^{(i)},\boldsymbol{w})}(h^*)$. Thus, the general estimate (3.5) is to be used only in cases where $\nu_{h \mid (\boldsymbol{z},\boldsymbol{w})}$ is unknown, so that ordinary Rao-Blackwellization is not possible. (In our situation, $\nu_{h \mid (\boldsymbol{z},\boldsymbol{w})}$ is analytically intractable—see (2.10)—and we are able to do Rao-Blackwellization only because we have available a scheme for data augmentation.)

For the case where $\nu_{h \mid (\boldsymbol{z},\boldsymbol{w})}$ is not known or is analytically intractable, Chen (1994) suggested that we consider a parametric family $\{f^{\phi},\ \phi \in \Phi\}$ of distributions on $\mathcal{Z} \times \mathcal{H}$, run a pilot Markov chain with invariant distribution $\nu_{(h,\boldsymbol{z}) \mid \boldsymbol{w}}$, and use it to estimate the mean and covariance matrix of $\nu_{(h,\boldsymbol{z}) \mid \boldsymbol{w}}$. The parameter $\phi$ is then chosen so that the mean and covariance matrix of $f^{\phi}$ match those of the estimate of $\nu_{(h,\boldsymbol{z}) \mid \boldsymbol{w}}$. We then take $\omega_{\boldsymbol{z}}$ to be $f^{\phi}_{h \mid \boldsymbol{z}}$ for each $\boldsymbol{z}$. In our situation, the very

high dimension of $\boldsymbol{z}$ precludes estimating the covariance matrix of $(\boldsymbol{z}, h)$. So instead we consider a family of distributions on $h$ (and not on $(\boldsymbol{z}, h)$) and further restrict each $f^{\phi}$ to be a product of univariate densities: $f^{\phi}(\alpha, \eta) = \left[\prod_{t=1}^{T} f^{\phi_t}(\alpha_t)\right] f^{\phi_{T+1}}(\eta)$. From our pilot chain on $(\boldsymbol{z}, h)$, we form an estimate of the mean and variance of each component of $h$, and select $\phi_t$, $t = 1, \ldots, T+1$ to match these estimates. We took $f^{\phi_t}$ to be gamma densities. Perhaps surprisingly, this quite simple procedure seems to work very well.

Xia (2018) presents a third method for estimating the marginal posterior density of $h$, based on averaging Markov transition densities. In its current implementation, this approach is very computationally intensive and is not competitive with the other methods.

## 3.2 Consistency and Asymptotic Normality of the Estimate of the Empirical Bayes Choice of the Hyperparameter

The main result in this section is Theorem 2, which establishes consistency and asymptotic normality of the estimate of $\arg\max_h m(h)$ that is based on the data augmentation chain and Rao-Blackwellization (3.4). The theorem also states that the estimate of the covariance matrix of the estimate of $\arg\max_h m(h)$ constructed through the method of batching is consistent, and this implies that we can construct asymptotically valid $95\%$ confidence sets for $\arg\max_h m(h)$. (The method of batching for the present setup is reviewed right after the statement of the theorem.) In the theorems, $p$ is the dimension of $h$: $p = 2$ if we take the distribution of the $\theta_d$'s to be a symmetric Dirichlet, and $p = T + 1$ if we allow this distribution to be an arbitrary Dirichlet. Theorem 2 refers to the regularity conditions below.

A1 The hyperparameter space $\mathcal{H}$ is compact.

A2 The maximizer of $m(\cdot)$ is unique (thus it makes sense to talk about $\arg\max_h m(h)$).

A3 The maximizer of $m(\cdot)$ is in $\mathcal{H}$.

A4 The function $m(\cdot)$ is twice continuously differentiable in $\mathcal{H}$, and the $p \times p$ Hessian matrix $\nabla_h^2 m(\arg\max_h m(h))$ is nonsingular.

**Theorem 2** *Suppose that $\lambda_1, \lambda_2, \ldots$ are generated according to the data augmentation algorithm, let $\hat{\nu}_{h\,|\,\boldsymbol{w}}(h)$ be given by* (3.4)*, let $\widehat{m}_n(h)$ be given by $\widehat{m}_n(h) = \hat{\nu}_{h\,|\,\boldsymbol{w}}(h)/\nu_h(h)$, and suppose that for each $n$, the maximizer of $\widehat{m}_n(\cdot)$ is unique. Further, assume that Conditions A1–A4 hold. Then:*

*1.* $\arg\max_h \widehat{m}_n(h) \xrightarrow{\text{a.s.}} \arg\max_h m(h)$.

2. $n^{1/2}\big(\arg\max_h \widehat{m}_n(h) - \arg\max_h m(h)\big) \xrightarrow{d} \mathcal{N}_p(0, \Sigma)$ *for some positive definite matrix* $\Sigma$.

3. *Let* $\widehat{\Sigma}_n$ *be the estimate of* $\Sigma$ *obtained by the method of batching. Then* $\widehat{\Sigma}_n \xrightarrow{\text{a.s.}} \Sigma$, *and in particular* $\widehat{\Sigma}_n$ *is invertible for large* $n$. *Consequently, the ellipse* $\mathcal{E}$ *given by*

$$\mathcal{E} = \big\{h : (\arg\max_h \widehat{m}_n(h) - h)^\top \widehat{\Sigma}_n^{-1} (\arg\max_h \widehat{m}_n(h) - h) \le \chi^2_{p,.95}/n\big\}$$

*is an asymptotic* $95\%$ *confidence set for* $\arg\max_h m(h)$. *Here,* $\chi^2_{p,.95}$ *denotes the* $0.95$ *quantile of the chi-square distribution with* $p$ *degrees of freedom.*

The proof of the theorem is in Section 4 of Xia and Doss (2019). Theorem 2 (and Theorem 3 below) may be used to determine the minimal Markov chain length that is needed to obtain an acceptably narrow confidence region for $\arg\max_h m(h)$. The method of batching for estimation of $\Sigma$ is as follows. The data augmentation scheme gives the sequence $\lambda^{(1)}, \ldots, \lambda^{(n)}$. The sequence is broken up into $J$ consecutive pieces of equal lengths called batches. For $j = 1, \ldots, J$, let $A^{[j]}$ be the estimate of $\arg\max_h m(h)$ produced from batch $j$, and let $A^{[]}$ be the estimate of $\arg\max_h m(h)$ produced from the entire sequence. The batch-based estimate is simply $\widehat{\Sigma}_n = (n/J)\big\{[1/(J-1)]\sum_{j=1}^{J}\big(A^{[j]} - A^{[]}\big)\big(A^{[j]} - A^{[]}\big)^\top\big\}$. (The quantity inside the braces is essentially sample covariance matrix of $A^{[1]}, \ldots, A^{[J]}$, except that we use $A^{[]}$ instead of the average of $A^{[1]}, \ldots, A^{[J]}$ as the centering value; and the term $n/J$ is a correction to account for the fact that the $A^{[j]}$'s are each formed from a sample of size $n/J$, not $n$.) Estimates of the covariance matrix based on batching are consistent under very general conditions which include that $J \to \infty$ as $n \to \infty$. The literature recommends taking $J = n^{1/2}$; see Flegal et al. (2008) and also Jones et al. (2006).

**Theorem 3** *The conclusions of Theorem 2 remain true if* $\hat{\nu}_{h\,|\,\boldsymbol{w}}$ *is given by* (3.5).

The proof of the theorem is in Section 6 of Xia and Doss (2019).

# 4 Evaluation of the Fully-Bayes Empirical Bayes Method

This section consists of three parts. In Section 4.1 we review current methods for approximating $\arg\max_h m(h)$. Because of the need to understand the strengths and deficiencies of these methods, it is essential to have a clear understanding of how these methods work, so our review is necessarily fairly detailed. In Section 4.2 we compare our approach with these methods, and evaluate it on synthetic and real data sets. In Section 4.3 we compare the empirical Bayes and fully Bayes methods. This is a comparison of two statistical procedures, as opposed to a comparison of two

numerical methods. The reader who is interested in understanding and using our methodology but is not interested in a review and evaluation of other approaches can read only Section 4.2.1 without loss.

## 4.1 Existing Methods for Approximating the Maximum Marginal Likelihood Estimator of the Hyperparameter

As mentioned in Section 1, the maximizer of the marginal likelihood of the hyperparameter can be expected to have good statistical properties, and here we review the literature on approximations of this estimator. But before we do this, we mention various ad-hoc rules for choosing $h$ that have been presented in the literature; these deal with the case where the distribution of the $\theta_d$'s is a symmetric Dirichlet, indexed by a single parameter $\alpha$, so that $h = (\alpha, \eta)$, i.e. $\dim(h) = 2$. The rules are as follows: $h_{\text{DG}} = (0.1, 50/T)$, used in Griffiths and Steyvers (2004); $h_{\text{DA}} = (0.1, 0.1)$, used in Asuncion et al. (2009); and $h_{\text{DR}} = (1/T, 1/T)$, used in the `Gensim` topic modelling package (Řehůřek and Sojka, 2010), a well-known package used in the topic modelling community.

*Gibbs-EM* In Gibbs-EM, the E-step of the EM algorithm is approximated by the CGS of Griffiths and Steyvers (2004). There are several problems with Gibbs-EM (at least for the version implemented by Wallach (2006)): (1) the approximation is used in the E-step at every iteration of the algorithm; (2) as with all EM-based methods, the algorithm can converge to a local maximum; and (3) an approximation is used in the M-step. Of these, the third appears to be the most serious, and we now discuss it in more detail and explain the problem. At the $k^{\text{th}}$ iteration, we must maximize with respect to $h$ the expectation $E_{h^{(k)}}\big(\log(p_h(\boldsymbol{z}, \boldsymbol{w}))\big)$, where $p_h(\boldsymbol{z}, \boldsymbol{w})$ is the joint distribution of $(\boldsymbol{z}, \boldsymbol{w})$ under the LDA model indexed by $h$, and the subscript to the expectation indicates that the expectation is taken with respect to $\nu^{h^{(k)}}_{\boldsymbol{z}\,|\,\boldsymbol{w}}$. A Markov chain $\boldsymbol{z}_1, \ldots, \boldsymbol{z}_{m_k}$ with invariant distribution equal to the posterior distribution of $\boldsymbol{z}$ given $\boldsymbol{w}$ is generated, and we want to maximize the function $G(h) = (1/m_k) \sum_{i=1}^{m_k} \log(p_h(\boldsymbol{z}_i, \boldsymbol{w}))$, which is a proxy for the expectation above. The maximization is done by solving the equation $\nabla G(h) = 0$ using fixed-point iteration, and because $\nabla G(h)$ is computationally intractable, Minka's (2003) approximation is used (in effect, a lower bound to $G(h)$ is found, and the lower bound is what is maximized). George and Doss (2018) have shown that when both components of $\arg\max_h m(h)$ are bigger than 1, the Minka (2003) approximation is very poor, and in Section 4.2 we show that in this case, even when a huge sample size is used for the CGS, Gibbs-EM converges to a value which is far from $\arg\max_h m(h)$.

*VEM* Conceptually, the estimate of $\arg\max_h m(h)$ given by VEM is obtained as follows. If $h^{(k)}$ is the current value of $h$, the E-step of the EM algorithm is to calculate $E_{h^{(k)}}\big(\log(p_h(\boldsymbol{\psi}, \boldsymbol{w}))\big)$, where $p_h(\boldsymbol{\psi}, \boldsymbol{w})$ is the joint distribution of $(\boldsymbol{\psi}, \boldsymbol{w})$ under the LDA model indexed by $h$, and the subscript to the expectation indicates that the expectation is taken with respect to $\nu^{h^{(k)}}_{\boldsymbol{\psi}\,|\,\boldsymbol{w}}$. This step is infeasible because $\nu^{h^{(k)}}_{\boldsymbol{\psi}\,|\,\boldsymbol{w}}$ is analytically intractable. We consider $\{q_\phi,\, \phi \in \Phi\}$, a (finite-dimensional) parametric family of analytically tractable distributions on $\boldsymbol{\psi}$, and within this family, we find the distribution, say $q_{\phi_*}$, which is "closest" to $\nu^{h^{(k)}}_{\boldsymbol{\psi}\,|\,\boldsymbol{w}}$. Let $Q(h)$ be the expected value of $\log(p_h(\boldsymbol{\psi}, \boldsymbol{w}))$ with respect to $q_{\phi_*}$. We view $Q(h)$ as a proxy for $E_{h^{(k)}}\big(\log(p_h(\boldsymbol{\psi}, \boldsymbol{w}))\big)$, and the M-step is then to maximize $Q(h)$ with respect to $h$, to produce $h^{(k+1)}$. The maximization is done analytically. While VEM can handle very large corpora with many topics, there is no theoretical reason to expect the sequence $h^{(k)}$ to converge to $\arg\max_h m(h)$. And if the likelihood surface is multimodal, then it can fail to find the global maximum (as is the case for all EM-type algorithms and also gradient-based approaches).

*Importance Sampling and Serial Tempering* This approach was developed by George and Doss (2018), who showed that it greatly outperforms both Gibbs-EM and VEM. It is based on the observation that if $c$ is a constant, then the information regarding $h$ given by the two functions $m(h)$ and $c\,m(h)$ is the same: the same value of $h$ maximizes both functions, and the second derivative matrices of the logarithm of these two functions are identical. In particular, the Hessians of the logarithm of these two functions at the maximum (i.e. the observed Fisher information) are the same and, therefore, the standard point estimates and confidence regions based on $m(h)$ and $c\,m(h)$ are identical. The relevance of this observation is as follows. Let $h_1 \in \mathcal{H}$ be fixed but arbitrary, and suppose that $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_n$ are the initial segment an ergodic Markov chain with invariant distribution $\nu^{(h_1)}_{\boldsymbol{\psi}\,|\,\boldsymbol{w}}$. Recall that $\ell_{\boldsymbol{w}}(\boldsymbol{\psi})$ is the likelihood function. Note that $m(h)$, which is given by $m(h) = \int \ell_{\boldsymbol{w}}(\boldsymbol{\psi})\, d\nu^{(h)}(\boldsymbol{\psi})$, is the normalizing constant in the statement "the posterior is proportional to likelihood times the prior," i.e. $\nu^{(h)}_{\boldsymbol{\psi}\,|\,\boldsymbol{w}}(\boldsymbol{\psi}) = \ell_{\boldsymbol{w}}(\boldsymbol{\psi})\, \nu^{(h)}(\boldsymbol{\psi})/m(h)$. For any $h \in \mathcal{H}$, consider the quantity $(1/n)\sum_{i=1}^n \nu^{(h)}(\boldsymbol{\psi}_i)/\nu^{(h_1)}(\boldsymbol{\psi}_i)$. As $n \to \infty$, we have

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\nu^{(h)}(\boldsymbol{\psi}_i)}{\nu^{(h_1)}(\boldsymbol{\psi}_i)} \xrightarrow{\text{a.s.}} \int \frac{\nu^{(h)}(\boldsymbol{\psi})}{\nu^{(h_1)}(\boldsymbol{\psi})} \nu^{(h_1)}_{\boldsymbol{\psi}\,|\,\boldsymbol{w}}(\boldsymbol{\psi})\, d\boldsymbol{\psi} \tag{4.1a}$$

$$= \frac{m(h)}{m(h_1)} \int \frac{\ell_{\boldsymbol{w}}(\boldsymbol{\psi})\nu^{(h)}(\boldsymbol{\psi})/m(h)}{\ell_{\boldsymbol{w}}(\boldsymbol{\psi})\nu^{(h_1)}(\boldsymbol{\psi})/m(h_1)} \nu^{(h_1)}_{\boldsymbol{\psi}\,|\,\boldsymbol{w}}(\boldsymbol{\psi})\, d\boldsymbol{\psi} \tag{4.1b}$$

$$= \frac{m(h)}{m(h_1)} \int \frac{\nu^{(h)}_{\boldsymbol{\psi}\,|\,\boldsymbol{w}}(\boldsymbol{\psi})}{\nu^{(h_1)}_{\boldsymbol{\psi}\,|\,\boldsymbol{w}}(\boldsymbol{\psi})} \nu^{(h_1)}_{\boldsymbol{\psi}\,|\,\boldsymbol{w}}(\boldsymbol{\psi})\, d\boldsymbol{\psi} = \frac{m(h)}{m(h_1)}. \tag{4.1c}$$

The significance of (4.1) is that it shows that we can estimate the entire family $\{m(h)/m(h_1),\, h \in \mathcal{H}\}$ with a single Markov chain run. Since $m(h_1)$ is a fixed constant, as noted above, the two functions $m(h)$ and $m(h)/m(h_1)$ give exactly the same information about $h$. The usefulness of (4.1) stems from the fact that the average on the left side involves *only the priors*, so we effectively bypass having to deal with the posterior distributions.

Actually, the statement that (4.1) shows that we can estimate all of $\{m(h)/m(h_1),\, h \in \mathcal{H}\}$ with a single Markov chain run is too good to be true, and in reality the estimate on the left side of (4.1a) has a serious defect: unless $h$ is close to $h_1$, $\nu^{(h)}$ can be nearly singular with respect to $\nu^{(h_1)}$ over the region where the $\boldsymbol{\psi}_i$'s are likely to be, resulting in a very unstable estimate. From a practical point of view, this means that there is effectively a "radius" around $h_1$ within which one can safely move, and there may not exist a single value of $h_1$ that gives rise to estimates that are stable for all $h \in \mathcal{H}$. One way of dealing with this problem is to select $J$ fixed points $h_1, \ldots, h_J \in \mathcal{H}$ that "cover" $\mathcal{H}$ in the sense that for every $h \in \mathcal{H}$, $\nu^{(h)}$ is "close to" at least one of $\nu^{(h_1)}, \ldots, \nu^{(h_J)}$. We then replace $\nu^{(h_1)}$ in the denominator by $\sum_{j=1}^{J} b_j \nu^{(h_j)}$, for some suitable choice of positive constants $b_1, \ldots, b_J$. Operating intuitively, we say that for any $h \in \mathcal{H}$, because there exists at least one $j$ for which $\nu^{(h)}$ is close to $\nu^{(h_j)}$, the variance of $\nu^{(h)}(\boldsymbol{\psi}) \big/ \big[\sum_{j=1}^{J} b_j \nu^{(h_j)}(\boldsymbol{\psi})\big]$ is small; hence the variance of $\nu^{(h)}(\boldsymbol{\psi}) \big/ \big[\sum_{j=1}^{J} b_j \nu^{(h_j)}(\boldsymbol{\psi})\big]$ is small simultaneously for all $h \in \mathcal{H}$. Whereas for the estimates (4.1a) we need a Markov chain with invariant distribution is $\nu^{(h_1)}_{\boldsymbol{\psi}\,|\,\boldsymbol{w}}$, in the present situation we need a Markov chain whose invariant distribution is a mixture of $\nu^{(h_1)}_{\boldsymbol{\psi}\,|\,\boldsymbol{w}}, \ldots, \nu^{(h_J)}_{\boldsymbol{\psi}\,|\,\boldsymbol{w}}$. This approach may be implemented by a methodology called serial tempering (Marinari and Parisi (1992); Geyer and Thompson (1995)). Serial tempering produces an estimator $\widehat{M}_n(h)$ with the property that for each $h$, as $n \to \infty$, $\widehat{M}_n(h) \xrightarrow{\text{a.s.}} c\, m(h)$ for some constant $c$. So to estimate $\arg\max_h m(h)$, we use $\arg\max_h \widehat{M}_n(h)$. A key issue with the methodology involves the choice of the "skeleton points" $h_1, \ldots, h_J$: in order that $\arg\max_h \widehat{M}_n(h)$ accurately estimate $\arg\max_h m(h)$, it is necessary that $h_1, \ldots, h_J$ be close to $\arg\max_h m(h)$, but $\arg\max_h m(h)$ is unknown, leading to a circular problem. George and Doss (2018) propose an iterative scheme for selecting the skeleton points. While the scheme works well for moderate-size corpora, whether it works for large corpora is not clear, and how well it works for corpora with a large number of topics is also not clear.

## 4.2 Evaluation and Comparison with Existing Methods

This section consists of four parts. In Section 4.2.1 we compare Markov chains on $(\boldsymbol{z}, h)$ in terms of mixing rates and execution time. The chains we discuss are the two developed in this paper

(based on HMC and data augmentation), and also a chain based on slice sampling, which we include for the sake of completeness. Very briefly, the HMC chain turns out to be the overall winner. Therefore, in our subsequent sections it is the one on which we focus when we compare the fully-Bayes empirical Bayes method (for which we use the acronym FBEB) with other methods. In Sections 4.2.2 and 4.2.3 we compare our FBEB method with existing methods, and before doing so, we first discuss the criteria we use for the comparison. Our first criterion is simply estimation accuracy: our FBEB approach gives one way to estimate $\arg\max_h m(h)$, and we compare it with the methods described in Section 4.1, namely Gibbs-EM, VEM, and the method based on importance sampling via serial tempering. For unity of notation, we will denote these estimators by $\hat{\hat{h}}_{\text{FBEB}}$, $\hat{\hat{h}}_{\text{GEM}}$, $\hat{\hat{h}}_{\text{VEM}}$, and $\hat{\hat{h}}_{\text{ISST}}$, respectively, and we will use the acronyms in the subscripts to denote the corresponding methods. (As a remark on notation, we note that $\arg\max_h m(h)$ is an estimate of the true value of $h$, i.e. the $h$ used to generate the corpus, so this empirical Bayes estimate should be called $\hat{h}$; and $\hat{\hat{h}}_{\text{FBEB}}$, $\hat{\hat{h}}_{\text{GEM}}$, $\hat{\hat{h}}_{\text{VEM}}$, and $\hat{\hat{h}}_{\text{ISST}}$ are all estimates of $\hat{h}$, hence the need for the "double hat": it reminds us that we are estimating an estimator.) Our first goal is to compare these as estimators of $\hat{h} = \arg\max_h m(h)$. This requires us to know the true value of $\arg\max_h m(h)$, or at least have a tight confidence region for it with theoretically guaranteed coverage probability. Our second criterion is model fit (or predictive accuracy): we wish to select the value of $h$, say $h_{\text{opt}}$, for which the LDA model indexed by $h_{\text{opt}}$ outperforms LDA models indexed by any other value of $h$. These two criteria are not the same. (This is analogous to a variable selection situation in linear regression. One goal is to identify those regression coefficients which are exactly zero, and a distinct goal is to select a set of variables for which the corresponding model has the best predictive ability. See Yang (2005) for a discussion of these points.) In Sections 4.2.2 and 4.2.3 we compare our FBEB method with existing methods under our first and second criteria, respectively. Finally, in Section 4.2.4 we discuss scalability and the advantages of FBEB over ISST for large corpora. With the exception of the last part of Section 4.2.2, in our evaluations we will always take the prior distribution of the $\theta_d$'s to be a symmetric Dirichlet. In Section 5 we return to the issue of feasibility of FBEB for large corpora.

Before we start on these four subsections, we illustrate the role of the hyperparameter in inference, by considering two corpora with different characters. These corpora are taken from the 20Newsgroups dataset, which is often used for benchmarking in topic modelling. Corpus C-A consists of articles from six of the 20 topics, while Corpus C-B consists of articles from five subcategories of the single topic *Computers* (and when fitting the LDA model, we took the number

of topics to be $T = 6$ for corpus C-A, and $T = 5$ for Corpus C-B). Thus, corpus C-A consists of documents which are easy to distinguish from each other, while C-B consists of documents which are difficult to distinguish from each other. Table 1 gives relevant information on these two corpora. We applied the methodology developed in this paper to produce plots of estimates of the marginal likelihood surface (up to a constant) for each corpus, and Figure 1 gives the results. Recall that if $U \sim \mathrm{Dir}_T(\epsilon, \ldots, \epsilon)$, then in the limiting case where $\epsilon \to 0$, $U$ tends to be a vector with one of its components being equal to 1, and the rest 0, and the location of the 1 is uniform over $\{1, \ldots, T\}$. The surface for corpus C-A suggests a value of $\alpha$ equal to .06, correctly reflecting the fact that documents in C-A will have a single topic. On the other hand, for corpus C-B the topics are close to each other, which makes the topic for a given word more uncertain; as a consequence, a given document may have several topics. The larger value for $\alpha$ suggested by the plot for corpus C-B ($\alpha = .093$) reflects that fact. To produce Figure 1, we used HMC chains of lengths 10,000, number of leapfrog steps $L = 2$, and stepsizes $\epsilon = 0.0035$ and $0.005$ for corpora C-A and C-B, respectively. Here and in the rest of the paper, we arrived at the values for $L$ and $\epsilon$ by following the recommendations in Neal (2011, Section 5.4.2). Specifically, in preliminary runs involving short chains, we set $L = 2$ and adjusted $\epsilon$ so that the acceptance rate was about $0.65$; increasing $L$ did not help much, and only slowed down the algorithm.

| Corpus | Categories | $T$ | $D$ | $V$ | $N$ |
|---|---|---|---|---|---|
| C-A | comp.sys.ibm.pc.hardware (189), misc.forsale (167), soc.religion.christian (192), talk.politics.guns (196), rec.sport.baseball (184), sci.space (186) | 6 | 1114 | 6,394 | 74,607 |
| C-B | comp.graphics (167), comp.os.ms-windows.misc (188), comp.sys.ibm.pc.hardware (192), comp.windows.x (194), comp.sys.mac.hardware (187) | 5 | 928 | 3,567 | 45,571 |

Table 1: Two corpora created from the 20Newsgroups dataset. The columns labeled $T$, $D$, $V$, and $N$ give the number of topics, number of documents, vocabulary size, and total number of words, respectively, for each corpus, and the numbers shown in parentheses next to the category names are the number of documents associated with the corresponding categories.

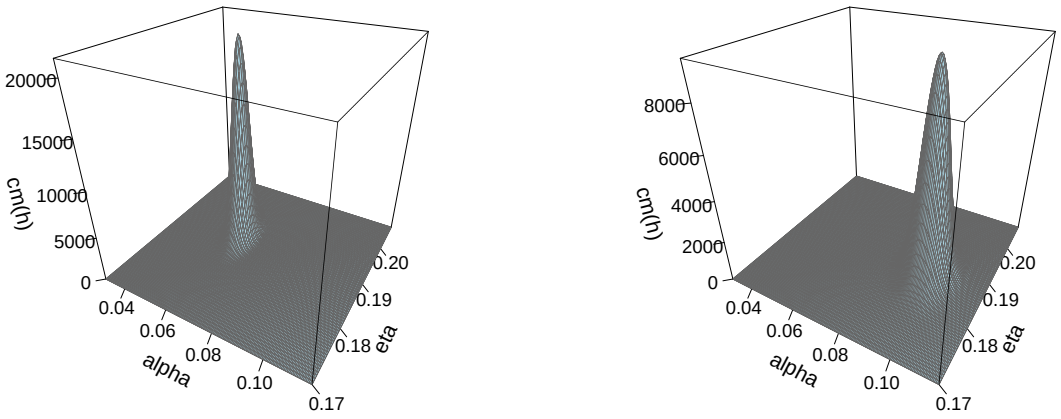



Figure 1: Estimate of marginal likelihood surface (up to a constant) for two corpora. The surface on left is for a corpus for which topics are very distinct, and the surface on the right is for a corpus for which topics are closer to each other. The $\alpha$-values suggested by the plots correctly reflect the fact that for the first corpus one expects each document to have a single topic, whereas for the second corpus one expects the documents to have several topics.

### 4.2.1 Statistical and Computational Efficiency of Three Markov Chains on $(\boldsymbol{z}, h)$

Recall that the Markov chain developed in Section 2.1 updates $(\boldsymbol{z}, h)$ by updating $\boldsymbol{z}$ via the CGS and updating $h$ via HMC. An alternative for the $h$-update is slice sampling (Neal, 2003), and because slice sampling has been used for the LDA model before (Wallach, 2008)—although for a purpose different than ours—it is natural to ask how it would perform in the present setting, so we include it in our comparison. In this section we compare the HMC-based chain, the augmented collapsed Gibbs sampler developed in Section 2.2, and the chain based on slice sampling (we will use the acronyms HMC, ACGS, and SS, respectively) based on two criteria: mixing rate and execution time. To this end, we generated artificial corpora from LDA models with configuration parameters set as follows. All of them had $V = 200$ and $n_d = 80$, and the corpora were in three groups:

- $T = 5$; $D = 100, 300, 500$; nine hyperparameters given by Table 2 below;
- $T = 10$; $D = 100$; nine hyperparameters given by Table 2;
- $T = 15$; $D = 100$; nine hyperparameters given by Table 2.

So there were $45$ corpora. Each chain was run for $10{,}000$ cycles for each corpus.

To compare mixing rates of several Markov chains, quantities such as asymptotic variances and auto-correlation functions (ACF's) are often used. Unfortunately, the very high dimension of $(\boldsymbol{z}, h)$

| HP name | $(\alpha, \eta)$ | HP name | $(\alpha, \eta)$ | HP name | $(\alpha, \eta)$ |
|---|---|---|---|---|---|
| $h_1$ | $(0.5, 0.5)$ | $h_4$ | $(1.0, 0.5)$ | $h_7$ | $(2.0, 0.5)$ |
| $h_2$ | $(0.5, 1.0)$ | $h_5$ | $(1.0, 1.0)$ | $h_8$ | $(2.0, 1.0)$ |
| $h_3$ | $(0.5, 2.0)$ | $h_6$ | $(1.0, 2.0)$ | $h_9$ | $(2.0, 2.0)$ |

Table 2: Names and values for nine hyperparameters.

precludes computing these for each component of this parameter. An attractive alternative is to consider, for a chain of length $n$, the posterior densities $\nu_{(\boldsymbol{z},h)\,|\,\boldsymbol{w}}(\boldsymbol{z}^{(1)}, h^{(1)}), \ldots, \nu_{(\boldsymbol{z},h)\,|\,\boldsymbol{w}}(\boldsymbol{z}^{(n)}, h^{(n)})$, and compute these quantities for this sequence (on the log scale). Thus, letting $\pi_1, \ldots, \pi_n$ be the logarithms of these posterior densities, we can compare the ACF's of $\pi_1, \ldots, \pi_n$ for the chains. Also, letting $\bar{\pi}_{(n)} = (1/n) \sum_{i=1}^{n} \pi_i$, we can compare the asymptotic variance of $\bar{\pi}_{(n)}$ across the chains. Actually, we will consider the so-called "efficiency factor," defined as follows. Suppose that $n^{1/2}(\bar{\pi}_{(n)} - E(\pi_1)) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$, and let $\tau^2$ be the variance that we would get if the chain was an iid sequence. The ratio $\tau^2/\sigma^2$ is called the efficiency factor. It may be estimated by standard spectral methods, and this is implemented, for example, by the R package `mcmcse` (Flegal et al., 2016). The posterior density is a single univariate quantity, and is known except for a normalizing constant. The fact that we don't know this constant is immaterial, since on the log scale the constant becomes an additive constant, which affects neither the ACF's nor the variances.

Figure 2 gives plots of the efficiency factor for each of the three Markov chains, and for the $45$ corpora. The plots show that the HMC chain is the clear winner: it has the largest efficiency factor in all cases, often by a large margin. A general pattern is that the superiority of the HMC chain greatly increases as the number of documents in the corpus increases (see plots a, b, and c), while its superiority decreases slightly as the number of topics increases (see plots c, d and e). The superiority of the HMC chain is understated by the plots, as these are on the log scale. The SS chain is nearly uniformly the worst, sometimes by a large margin. Figure 3 gives plots of the ACF for the three chains and two of the corpora. The message here mirrors the message given by the plots of the efficiency factors for these two corpora and for the other corpora also (plots not shown). The HMC chain wins overall, sometimes by a very large margin; and the SS chain is the worst, its ACF sometimes dying down very slowly.

Figure 4 gives results on execution time for the three algorithms and the five $(T, D)$ combinations (it turned out that the hyperparameter has essentially no effect on execution time), on a 3.70GHz quad core Intel Xeon Processor E5-1630V3. A summary of the results is as follows.

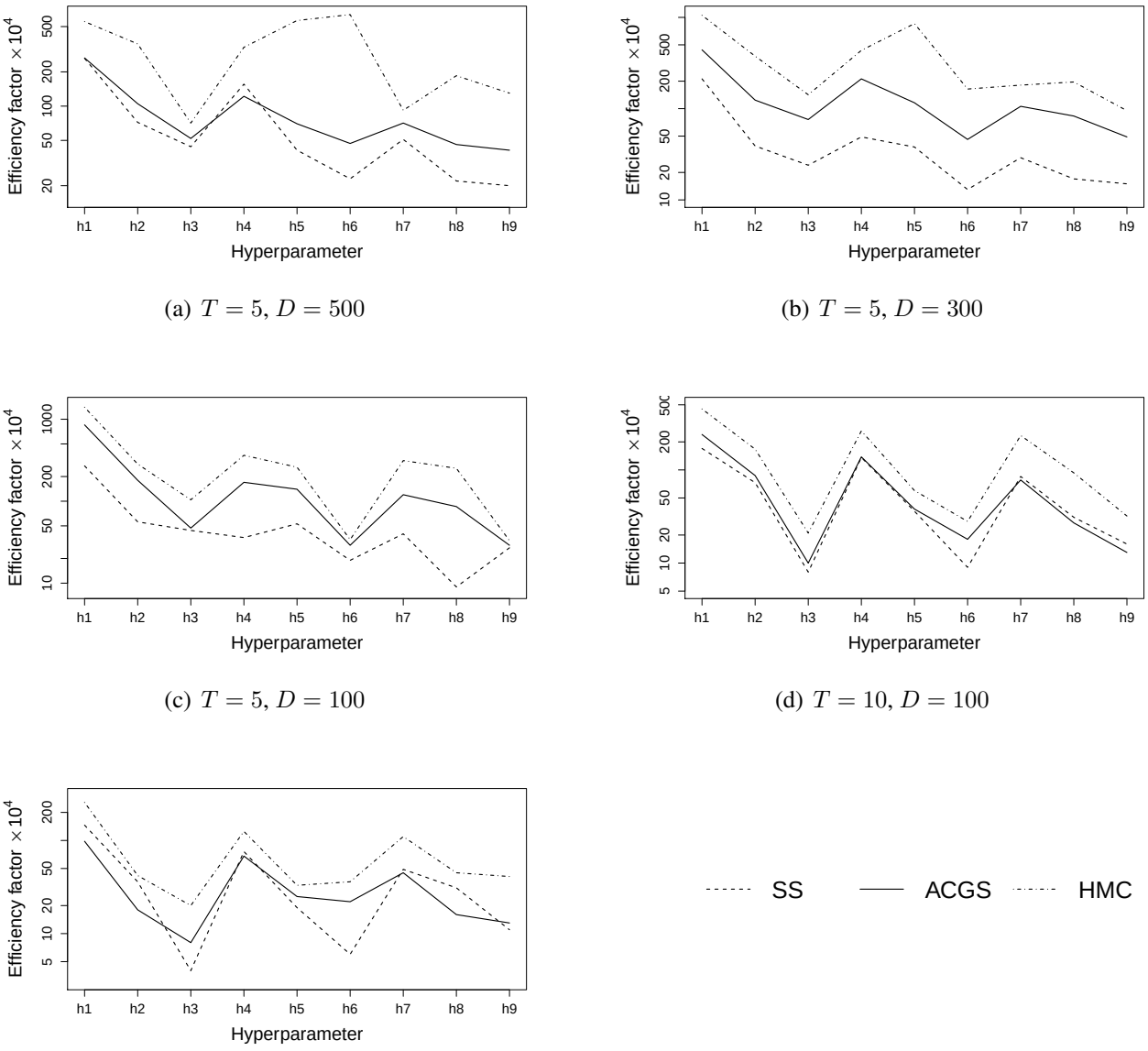
Execution times are smallest for SS uniformly. However, the times for the other two are not much bigger, and the ratio of largest to smallest is always less than 2; therefore, the effect of execution time is rather small when compared to the effect of mixing rate. When we take both mixing rate and execution time into account, the HMC chain turns out to be the overall winner, and the SS chain is worst overall. In some experiments involving a large number of documents and a large number of topics, the ACGS chain matched or slightly outperformed the HMC chain in terms of both efficiency factor and ACF. The HMC chain requires tuning, whereas the DA chain does not. The amount of time required for tuning is document specific, so our recommendation for large corpora is to try the HMC chain first, and in the event that tuning is time consuming, use the DA chain.

#### 4.2.2 Comparison of All Methods for Estimating $\arg\max_h m(h)$ Based on Estimation Accuracy

To compare $\hat{\hat{h}}_{\text{FBEB}}$, $\hat{\hat{h}}_{\text{ISST}}$, $\hat{\hat{h}}_{\text{GEM}}$, and $\hat{\hat{h}}_{\text{VEM}}$, we created a synthetic corpus generated according to the LDA model with $D = 100$, $T = 4$, $V = 20$, $n_d = 80$ for all $d$, and $h_{\text{true}} = (\alpha_{\text{true}}, \eta_{\text{true}}) = (.2, .8)$. This is a very small corpus, and we chose a corpus of this size in order to be able to include the ISST method, whose implementation is very time consuming. (The relative merits of the four estimators do not change much as we change the size of the corpus.) The estimates $\hat{\hat{h}}_{\text{FBEB}}$, $\hat{\hat{h}}_{\text{ISST}}$, $\hat{\hat{h}}_{\text{GEM}}$, and $\hat{\hat{h}}_{\text{VEM}}$ were computed using the following specifications.

$\hat{\hat{h}}_{\text{FBEB}}$ We used an ACGS chain of length $10^5$, 10 times, using 10 different seeds, obtaining 10 estimates, which we call $\hat{\hat{h}}^{[1]}_{\text{FBEB}}, \ldots, \hat{\hat{h}}^{[10]}_{\text{FBEB}}$. These were obtained from the Rao-Blackwellized estimates (3.4), as described in the statement of Theorem 2. According to Theorem 2, the independent variables $\hat{\hat{h}}^{[1]}_{\text{FBEB}}, \ldots, \hat{\hat{h}}^{[10]}_{\text{FBEB}}$ are approximately bivariate normally distributed with mean vector $\arg\max_h m(h)$. Therefore, they can be used to form a $95\%$ confidence ellipse for $\arg\max_h m(h)$, based on Hotelling's $T^2$ distribution (this ellipse is simply the two-dimensional analogue of the standard $t$-interval, which is based on the $t$-distribution).

We repeated this, but using the HMC chain, with number of leapfrog steps $L = 2$ and step size $\epsilon = 0.025$, and (3.5) instead of (3.4), obtaining a second confidence ellipse. Strictly speaking, the theoretical validity of this confidence ellipse requires that we have a version of Theorem 2 that applies to the case where we use the HMC chain, and (3.5) instead of (3.4), and we have not established such a theorem. Nevertheless, it is useful to consider this confidence ellipse.




(a) $T = 5$, $D = 500$

(b) $T = 5$, $D = 300$

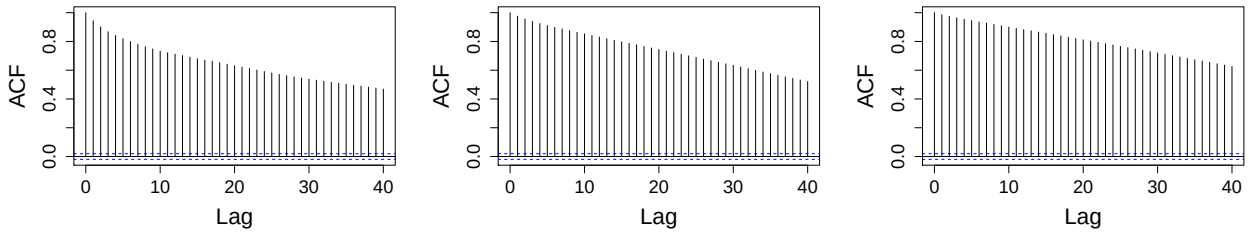
(c) $T = 5$, $D = 100$

(d) $T = 10$, $D = 100$

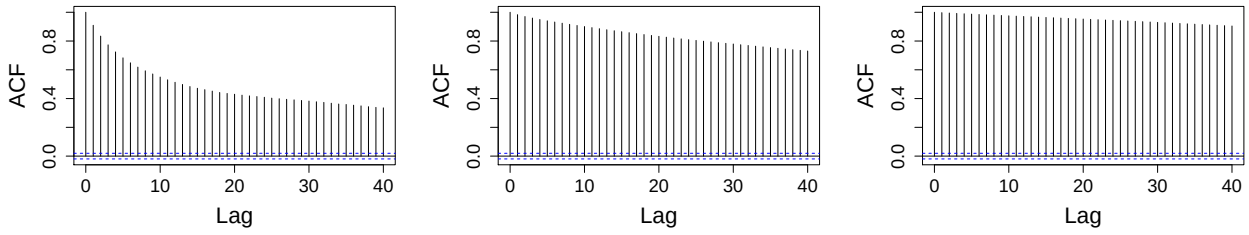
(e) $T = 15$, $D = 100$

Figure 2: Efficiency factors for the HMC, ACGS, and SS chains on the log scale, for $45$ corpora. The HMC chain has the largest efficiency factor almost uniformly, sometimes by large margins.

$\hat{\hat{h}}_{\text{ISST}}$ We used the serial tempering chain, as described in George and Doss (2018). We used $20$ iterations of their scheme for choosing the set of skeleton points, and this set turned out to be a $7 \times 9$ grid of $63$ values over the region $(\eta, \alpha) \in [.6, .9] \times [.1, .3]$. We used a Markov chain length of $100{,}000$, and as for FBEB, we repeated this a total of $10$ times, obtaining $10$ estimates, which we call $\hat{\hat{h}}^{[1]}_{\text{ISST}}, \ldots, \hat{\hat{h}}^{[10]}_{\text{ISST}}$, and we used these to construct a $95\%$ confidence ellipse. (The theoretical validity of this confidence ellipse is supported by Theorem 1 and Remark 3 of that




(a) Corpus config. 1, HMC (b) Corpus config. 1, ACGS (c) Corpus config. 1, SS




(d) Corpus config. 2, HMC (e) Corpus config. 2, ACGS (f) Corpus config. 2, SS

Figure 3: ACF's for the three Markov chains and two corpora configurations. Configuration 1: $T = 5$, $D = 500$, $\alpha = 2$, $\eta = .5$; configuration 2: $T = 5$, $D = 300$, $\alpha = 1$, $\eta = 2$. The ACF's die down fastest for the HMC chain and slowest for the SS chain.

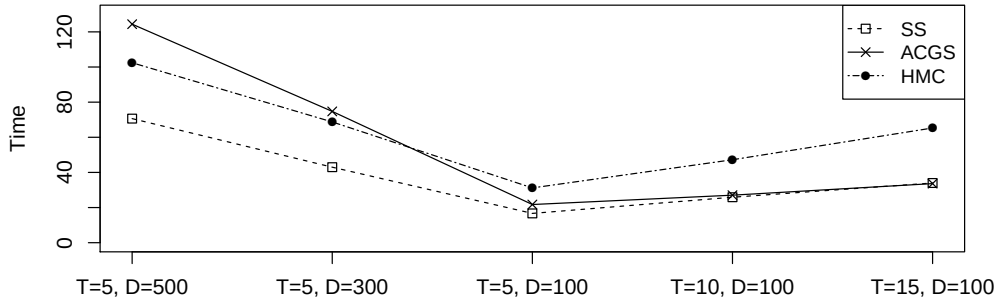



Figure 4: Execution time in seconds for $10{,}000$ iterations for HMC, ACGS, and SS chains, for five corpus configurations.

paper.)

$\hat{\hat{h}}_{\text{GEM}}$ We used $50$ iterations of the EM algorithm, and within each iteration, we ran a CGS of length $120{,}000$ (discarding the first $2000$ as burn-in) to approximate the E-step. We repeated this a total of $10$ times, using $10$ different starting values and $10$ different seeds.

$\hat{\hat{h}}_{\text{VEM}}$ We used $100$ iterations of the EM algorithm, and within each iteration, the E-step was approximated via $100$ variational inference iterations. We repeated this a total of $10$ times, using $10$

different starting values.

Figure 5 gives the results, but before looking at them we clarify what is the target of estimation. The following facts are obvious, but it is perhaps worthwhile to state them explicitly. Here, $h_{\text{true}} = (\alpha_{\text{true}}, \eta_{\text{true}}) = (.2, .8)$ is the value used to generate the synthetic corpus. The maximum marginal likelihood estimate $\hat{h} \coloneqq \arg\max_h m(h)$ depends on the corpus, and in general is not equal to $h_{\text{true}}$ (although it is likely to be close to it). The quality of the estimates $\hat{\hat{h}}_{\text{FBEB}}$, $\hat{\hat{h}}_{\text{ISST}}$, $\hat{\hat{h}}_{\text{GEM}}$, and $\hat{\hat{h}}_{\text{VEM}}$ is determined by how close these are to $\hat{h}$, not by how close they are to $h_{\text{true}}$. The left panel of Figure 5 shows the estimates produced by the four methods: FBEB (through both HMC and ACGS), ISST, GEM, and VEM, $10$ points for each, for a total of $50$ points, and we see that the $30$ points produced by FBEB and ISST are so close to each other that they are visually indistinguishable. The right panel gives a zoomed version of the plot, magnifying the region which contains the FBEB and ISST points, and gives the three confidence ellipses. The two panels together show the following. GEM greatly outperforms VEM. Nevertheless, the GEM estimates are not close to being within any of the $95\%$ confidence ellipses. It appears that the problem is with the maximization step which, as explained in Section 4.1, is done through an approximation. VEM does poorly, and the estimates strongly depend on the starting value. The estimates are so poor that they do not even appear in the zoomed version of the plot. The HMC ellipse is more narrow than the ACGS ellipse (which is consistent with the results of Section 4.2.1 that show that the HMC chain has better mixing), and the ISST ellipse is more narrow than both of these. However, the ellipses are comparable in size, so that the decision of which method is preferable depends on computational efficiency. This issue is discussed in detail in Section 4.2.4.

**Estimation of $\arg\max_h m(h)$ When the Prior on the $\theta_d$'s Is an Asymmetric Dirichlet** Our initial motivation for this work was our desire to handle the case where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_T)$ and the $\alpha_t$'s are not assumed to be equal, so that $\dim(h) = T + 1$. When $\dim(h)$ is large and the corpus is large, the EM algorithm is expected to have poor performance: the rate of convergence is determined by the amount of missing information (Meng and Van Dyk, 1997), and in our situation this is $\boldsymbol{\psi}$, which has high dimension. And for serial tempering to work, we need that every $h$ value in the relevant part of the hyperparameter space be close to at least one point in the skeleton set $\{h_1, \ldots, h_J\}$, which forces the size of this set to be astronomical (this is the curse of dimensionality). On the other hand, for the FBEB method to work well, the main requirement is that we can devise a Markov chain on $(\boldsymbol{z}, h)$ (or on $(\boldsymbol{\psi}, h)$) for which the $h$ component mixes well. There

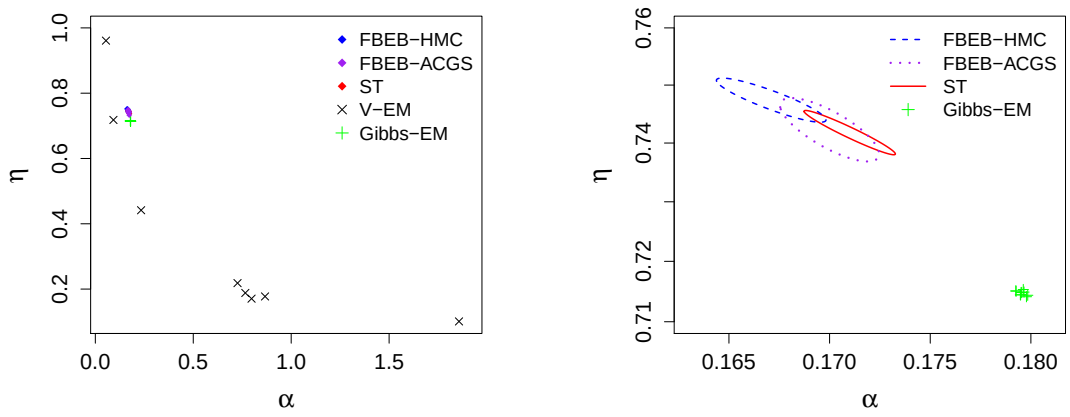



Figure 5: Estimates and confidence ellipses for $\arg\max_h m(h)$ produced by the FBEB, ISST, GEM, and VEM methods. Left panel shows the point estimates. The FBEB and ISST points are so close that they seem to merge into a single point. Right panel is a zoomed version of a small region of the left panel that contains these 30 points, and gives the confidence ellipses. The GEM points are not close to being in any of the ellipses, and the VEM points are not even in the plot.

is now a highly-developed literature, spanning over more than two decades, on the construction of Markov chains on high-dimensional state spaces with good mixing properties, and we can take advantage of this literature.

Care must be exercised, however, when dealing with the case where the $\alpha_t$'s are not assumed to be equal. Suppose that the prior on $(\alpha_1, \ldots, \alpha_T)$ is a symmetric function of $(\alpha_1, \ldots, \alpha_T)$, i.e. it is invariant under permutations of $\alpha_1, \ldots, \alpha_T$ (this is the situation for the prior we use in this paper, which is a product of gammas, all of them equal). In this case, since the likelihood function is symmetric in $(\alpha_1, \ldots, \alpha_T)$, so is the marginal likelihood, and therefore so is the posterior. Thus any reasonable estimator of $(\alpha_1, \ldots, \alpha_T)$ obtained from the posterior will be symmetric in $(\alpha_1, \ldots, \alpha_T)$; so if the corpus is generated according to an LDA model for which the components of $\boldsymbol{\alpha}_{\text{true}}$ are not all equal, then $\boldsymbol{\alpha}_{\text{true}}$ cannot be "retrieved" from the posterior, even if the number of documents in the corpus is arbitrarily large. Therefore, without a mechanism for ensuring identifiability, it is not appropriate to use an asymmetric Dirichlet as the prior on the $\theta_d$'s. In practical terms, suppose that there are only two topics, call them $\beta_S$ and $\beta_M$, where $\beta_S$ gives most of its mass to sports-like words, and $\beta_M$ gives most of its mass to medical words. It is not the case that we can claim that $\beta_S = \beta_1$ and $\beta_M = \beta_2$ any more than we can claim $\beta_S = \beta_2$ and $\beta_M = \beta_1$. One way of handling this situation when dealing with the $\boldsymbol{\beta}$-part of a Markov chain with the posterior as its invariant distribution is to "align" the $\beta$'s. (This general situation, referred to as the "label-

switching problem," is a well-known issue in Bayesian mixture modelling, and there is a literature on this topic; see Celeux et al. (2000), and also Jasra et al. (2005) for a review.) The alignment can be done as follows. Let $\boldsymbol{\beta}^{(b)}$ denote the first $\boldsymbol{\beta}$-value after the burn-in period. We take the ordering of the $\beta$'s in $\boldsymbol{\beta}^{(b)}$ as our "baseline." If $\boldsymbol{\beta}^{(i)}$ is any subsequent $\boldsymbol{\beta}$-value, we form the discrepancy matrix $D$, whose $(r, s)$ entry is $\|\beta_r^{(i)} - \beta_s^{(b)}\|_1$, where $\|\cdot\|_1$ is the $\ell_1$ norm in $\mathbb{R}^V$. Let $D_{r_1 s_1}$ be the smallest element of this matrix. We set $s_1$ to be the topic label for $\beta_{r_1}^{(i)}$. We eliminate the $r_1$ row and $s_1$ column of this matrix, and repeat the procedure above on the reduced matrix. This process continues, and we sequentially determine the labels for all the topics in $\boldsymbol{\beta}^{(i)}$.

We now evaluate the performance of the FBEB method for the case of a multidimensional $\boldsymbol{\alpha}$. To this end, we formed three synthetic corpora, generated according to LDA models for which the components of $\boldsymbol{\alpha}$ are not all equal. The dimensions of $\boldsymbol{\alpha}$ were $4$, $8$, and $12$. Table 3 gives the configuration parameters for the corpora, including the true values of the hyperparameter used to generate the corpora. We implemented the FBEB method, using ACGS chains of length $11{,}000$, from which we discarded $1{,}000$ as burn-in, and thinned the remaining $10{,}000$ by a factor of $5$, so effectively having a sample of size $2{,}000$. The argmax of the estimate of the marginal likelihood function was obtained using the R package `optimx` (Nash and Varadhan, 2011), and Table 3 also gives these estimates. In the table, $h_{\text{true}}$ denotes the value of $h$ used to generate the synthetic corpus, and $\hat{\hat{h}}_{\text{FBEB}}$ is the estimator of $\hat{h} = \arg\max_h m(h)$. From the table, we see that $\hat{\hat{h}}_{\text{FBEB}}$ is plausibly performing remarkably well. (We wrote "plausibly" because, as remarked earlier, $\hat{\hat{h}}_{\text{FBEB}}$ is really an estimate of $\hat{h}$, which is unknown, and not of $h_{\text{true}}$. Of course, because the three corpora are large, we expect $\hat{h}$ to be very close to $h_{\text{true}}$.) The experiments needed to produce the data for corpora 1–3 took $15.1$, $24.0$, and $55.6$ minutes, respectively.

We did not do experiments with large $T$, as it is inappropriate to use asymmetric Dirichlets with a large $T$. When $T$ is large, to estimate $\arg\max_h m(h)$ accurately would require very large Markov chain lengths, but this is not the main issue, which is that even if we were to be able to calculate $\arg\max_h m(h)$ exactly, $\arg\max_h m(h)$ itself would not be an accurate estimate of $h_{\text{true}}$ unless the number of documents was huge. (To be clear, we are not saying that it is inappropriate to use an LDA model with a large $T$; we are saying that it is inappropriate to use asymmetric Dirichlets with a very large $T$.) Xia (2018) discusses this issue in more detail.

| *Corpus 1* $D = 2000, V = 1000, n_d = 80$ for all $d, T = 4$ | | | | | |
|---|---|---|---|---|---|
| $h$ | $\alpha_1$ | $\alpha_2$ | $\alpha_3$ | $\alpha_4$ | $\eta$ |
| $h_{\text{true}}$ | 0.2 | 0.4 | 0.6 | 0.8 | 0.5 |
| $\hat{\hat{h}}_{\text{FBEB}}$ | 0.185 | 0.386 | 0.590 | 0.787 | 0.513 |

| *Corpus 2* $D = 4000, V = 1000, n_d = 80$ for all $d, T = 8$ | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| $h$ | $\alpha_1$ | $\alpha_2$ | $\alpha_3$ | $\alpha_4$ | $\alpha_5$ | $\alpha_6$ | $\alpha_7$ | $\alpha_8$ | $\eta$ |
| $h_{\text{true}}$ | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.5 |
| $\hat{\hat{h}}_{\text{FBEB}}$ | 0.102 | 0.202 | 0.299 | 0.424 | 0.491 | 0.605 | 0.690 | 0.833 | 0.499 |

| *Corpus 3* $D = 8000, V = 1000, n_d = 80$ for all $d, T = 12$ | | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $h$ | $\alpha_1$ | $\alpha_2$ | $\alpha_3$ | $\alpha_4$ | $\alpha_5$ | $\alpha_6$ | $\alpha_7$ | $\alpha_8$ | $\alpha_9$ | $\alpha_{10}$ | $\alpha_{11}$ | $\alpha_{12}$ | $\eta$ |
| $h_{\text{true}}$ | 0.1 | 0.2 | 0.3 | 0.4 | 0.5 | 0.6 | 0.7 | 0.8 | 0.9 | 1.0 | 1.1 | 1.2 | 0.5 |
| $\hat{\hat{h}}_{\text{FBEB}}$ | 0.117 | 0.230 | 0.273 | 0.386 | 0.484 | 0.620 | 0.686 | 0.817 | 0.939 | 0.965 | 1.135 | 1.223 | 0.491 |

Table 3: Estimates of $\hat{h} = \arg\max_h m(h)$ for three synthetic corpora generated according to LDA models with components of $\boldsymbol{\alpha}$ not all equal.

#### 4.2.3 Comparison of all Methods for Selecting the Hyperparameter Based on Posterior Predictive Likelihood

The criterion for model fit (or predictive accuracy) that we will use is the "Posterior Predictive Likelihood" (PPL) score. It is inversely related to the so-called perplexity score which is sometimes used in the machine learning literature. When applied in the LDA context, the PPL score is obtained as follows. For $d = 1, \ldots, D$, let $\boldsymbol{w}_{(-d)}$ denote the corpus consisting of all the documents except for document $d$. To evaluate a given model (in our case the LDA model indexed by a given $h$), in essence we see how well the model based on $\boldsymbol{w}_{(-d)}$ predicts document $d$, the held-out document. We do this for $d = 1, \ldots, D$, and take the geometric mean (Wallach et al., 2009). We formalize this as follows. The predictive likelihood of $h$ for the held-out document is

$$L_d(h) = \int \ell_{\boldsymbol{w}_d}(\boldsymbol{\psi})\, d\nu^{(h)}_{\boldsymbol{\psi}\,|\,\boldsymbol{w}_{(-d)}}(\boldsymbol{\psi}), \tag{4.2}$$

where $\ell_{\boldsymbol{w}_d}(\boldsymbol{\psi})$ is the likelihood of $\boldsymbol{\psi}$ for the held-out document $d$, and $\nu^{(h)}_{\boldsymbol{\psi}\,|\,\boldsymbol{w}_{(-d)}}$ is the posterior distribution of $\boldsymbol{\psi}$ given $\boldsymbol{w}_{(-d)}$. We form the score $S(h) = \left[\prod_{d=1}^{D} L_d(h)\right]^{1/D}$ (the reason for tak-

ing the geometric mean is that this keeps the score stable as document length changes). Two different values of hyperparameter $h$ are compared via their scores. Of course, $L_d(h)$ is analytically intractable, but it may be estimated by MCMC if we can generate an ergodic Markov chain $\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \ldots$ with invariant distribution $\nu^{(h)}_{\boldsymbol{\psi}\,|\,\boldsymbol{w}_{(-d)}}$. The CGS gives only a Markov chain $\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots$ with invariant distribution $\nu^{(h)}_{\boldsymbol{z}\,|\,\boldsymbol{w}_{(-d)}}$. However, the sequence $\boldsymbol{z}_1, \boldsymbol{z}_2, \ldots$ may be easily augmented to a sequence $(\boldsymbol{z}_1, \boldsymbol{\beta}_1, \boldsymbol{\theta}_1), (\boldsymbol{z}_2, \boldsymbol{\beta}_2, \boldsymbol{\theta}_2), \ldots$ with invariant distribution $\nu^{(h)}_{\boldsymbol{\psi}\,|\,\boldsymbol{w}_{(-d)}}$, because the conditional distribution of $(\boldsymbol{\beta}, \boldsymbol{\theta})$ given $\boldsymbol{z}$ and $\boldsymbol{w}_{(-d)}$ is available in closed form as a product of Dirichlets. Explicit expressions are given in George and Doss (2018). We then approximate the integral by $(1/n)\sum_{i=1}^{n} \ell_{\boldsymbol{w}_d}(\boldsymbol{\psi}_i)$, where $\boldsymbol{\psi}_i = (\boldsymbol{z}_i, \boldsymbol{\beta}_i, \boldsymbol{\theta}_i)$. Care needs to be exercised, because in (4.2) the variable $\boldsymbol{\psi}$ in the term $\ell_{\boldsymbol{w}_d}(\boldsymbol{\psi})$ has a dimension that is different than that of the variable $\boldsymbol{\psi}$ in the rest of the integral. Chen (2015) gives a careful description of an MCMC scheme for estimating the integral in (4.2), and his scheme is the one that we use.

For our comparisons, we consider three real corpora. Two of these are from Wikipedia, and one is a subset of the 20Newsgroups dataset, which is often used for benchmarking in topic modelling.

Corpus C-1 This is a subset of the articles under the Wikipedia category *Birds of Prey*, and consists of all articles under the seven subcategories *Eagles*, *Falco (genus)*, *Falconry*, *Harriers*, *Hawks*, *Kites*, and *Owls*. When fitting the LDA model, we took the number of topics $T$ to be seven.

Corpus C-2 This is a subset of the articles under the Wikipedia category *Whales*, and consists of all articles under the six subcategories *Baleen Whale*, *Dolphins*, *Killer Whale*, *Oceanic Dolphins*, *Whaling*, and *Whale Products*. When fitting the LDA model, we set $T = 6$.

Corpus C-3 This is a subset of the articles under the 20Newsgroups super-category *Computers* and consists of all articles from the five categories *comp.graphics*, *comp.os.ms-windows.misc*, *comp.windows.x*, *comp.sys.ibm.pc.hardware*, and *comp.sys.mac.hardware*. When fitting the LDA model, we set $T = 5$.

Table 4 gives some information on these three corpora.

We computed an estimate $\hat{S}(h)$ of $S(h)$, where $h$ ranges over the seven cases

$$\hat{\hat{h}}_{\text{FBEB}}, \hat{\hat{h}}_{\text{ISST}}, \hat{\hat{h}}_{\text{GEM}}, \hat{\hat{h}}_{\text{VEM}}, h_{\text{GS}}, h_{\text{A-etal}}, \text{ and } h_{\text{RS}}, \tag{4.3}$$

where the last three refer to the values used in Griffiths and Steyvers (2004), Asuncion et al. (2009), and Řehůřek and Sojka (2010), respectively, and $\hat{\hat{h}}_{\text{FBEB}}$ refers to the value computed when we use HMC in our FBEB scheme. The specifications used to compute the first four estimates in (4.3) are similar to those described in Section 4.2.2, except for the following: for $\hat{\hat{h}}_{\text{FBEB}}$ and $\hat{\hat{h}}_{\text{ISST}}$ we

| Corpus | Categories | $T$ | $D$ | $V$ | $N$ |
|---|---|---|---|---|---|
| C-1 | Eagles (62), Falco (genus) (45), Falconry (52), Harriers (21), Hawks (16), Kites (22), Owls (76) | 7 | 294 | 1,369 | 114,056 |
| C-2 | Baleen Whale (40), Dolphins (10), Killer Whale (11), Oceanic Dolphins (50), Whaling (32), Whale Products (10) | 6 | 153 | 712 | 52,107 |
| C-3 | comp.graphics (50), comp.os.ms-windows.misc (49), comp.sys.ibm.pc.hardware (49), comp.windows.x (47), comp.sys.mac.hardware (49) | 5 | 244 | 1,114 | 8,829 |

Table 4: Corpora created from Wikipedia pages and the 20Newsgroups dataset. The columns labeled $T$, $D$, $V$, and $N$ give the number of topics, number of documents, vocabulary size, and total number of words, respectively, for each corpus, and the numbers shown in parentheses next to the category names are the number of documents associated with the corresponding categories.

used $10{,}000$ iterations after discarding $1000$ as burn-in; for $\hat{\hat{h}}_{\text{GEM}}$ we used $100$ EM iterations, and within each we used $2000$ iterations of the CGS to approximate the E-step; and for $\hat{\hat{h}}_{\text{VEM}}$ we used $100$ EM iterations, and within each we used $20$ variational inference iterations to approximate the E-step. (The specifications used in Section 4.2.2 were a bit extravagant, and because we now need to make comparisons of execution time, we are using more realistic numbers.) Table 5 gives $S(h)$ as $h$ ranges over the last six values in (4.3), for the three corpora. The ratios are standardized by $S\big(\hat{\hat{h}}_{\text{FBEB}}\big)$, i.e. the table actually gives the ratios $S(h)/S\big(\hat{\hat{h}}_{\text{FBEB}}\big)$. The main message from the table is as follows. Generally, $\hat{\hat{h}}_{\text{FBEB}}$ and $\hat{\hat{h}}_{\text{ISST}}$ do best and are comparable, so that the choice of which to use should be based on computational considerations; $\hat{\hat{h}}_{\text{GEM}}$ and $\hat{\hat{h}}_{\text{VEM}}$ do worse; and all ad-hoc choices have very poor performances.

### 4.2.4 Comparison in Terms of Scalability

As we saw in Sections 4.2.2 and 4.2.3, by both our criteria, the ad-hoc choices perform very poorly, so they should not be used; the EM-based methods do not perform well (and this is especially true of VEM); and the two MCMC-based methods perform well and are comparable, so which one to use boils down to computational considerations, and this is what we discuss next.

There are two problems that cause ISST to be slow. First, for the method to work well, $\arg\max_h m(h)$ should be close to at least one point in the skeleton set; second, at the same time,

| Corpus | $\hat{\hat{h}}_{\text{ISST}}$ | $\hat{\hat{h}}_{\text{GEM}}$ | $\hat{\hat{h}}_{\text{VEM}}$ | $h_{\text{GS}}$ | $h_{\text{A-etal}}$ | $h_{\text{RS}}$ |
|---|---|---|---|---|---|---|
| C-1 | $2.35\,e{-01}$ | $2.78\,e{-01}$ | $1.70\,e{-01}$ | $1.23\,e{-13}$ | $5.23\,e{-06}$ | $6.81\,e{-03}$ |
| C-2 | $4.43\,e{-01}$ | $1.18\,e{-01}$ | $8.37\,e{-02}$ | $2.34\,e{-13}$ | $2.26\,e{-05}$ | $3.72\,e{-03}$ |
| C-3 | $1.44\,e{+01}$ | $1.18\,e{-01}$ | $5.04\,e{-01}$ | $6.64\,e{-03}$ | $2.41\,e{-02}$ | $8.09\,e{-01}$ |

Table 5: Ratios $S(h)/S(\hat{\hat{h}}_{\text{FBEB}})$ for six choices of $h$, for three corpora. A small number indicates a lack of fit, thus a poor choice of $h$. On the whole, the MCMC methods do best, the EM-based methods are worse, and all ad-hoc choices are abysmal. Notation: $6.59\,e{-01} = 6.59 \times 10^{-01}$.

the points in the skeleton set cannot be far from each other, or else the chain does not mix well (George and Doss, 2018). A look at Figure 1 enables us to appreciate the problem. If for corpus C-A (for which the marginal likelihood surface is on the left) we position the skeleton set in, say, the region $(\alpha, \eta) \in (.1, .17) \times (.17, .18)$ (the Southeast part of the displayed $(\alpha, \eta)$ region), the near singularity of $\nu^{(h)}$ for $h$ in the vicinity of $\arg\max_h m(h)$ and all the $\nu^{(h_j)}$'s will cause the estimator of the marginal likelihood surface for $h$ in the vicinity of $\arg\max_h m(h)$ to have an extremely large variance; and this problem would be far worse if the skeleton grid was in a small region surrounding the "distant" point $(\alpha, \eta) = (1, 1)$. The iterative scheme of George and Doss (2018) initializes a wide skeleton set, and using this skeleton set computes an estimate of the marginal likelihood surface $m(h)$ over the convex hull of the skeleton set. For the next iteration, we form a new skeleton set, centered at the current value of the estimate of $\arg\max_h m(h)$, and the skeleton set is made more narrow. This process continues until the estimate of $\arg\max_h m(h)$ stabilizes. In contrast, for the FBEB method, assuming that our Markov chain mixes well, in essence we get a sample $h^{(1)}, \ldots, h^{(n)}$ approximately distributed according to $\nu^{(h)}_{h \,|\, \boldsymbol{w}}(\cdot) \propto m(\cdot)$, so the general location of $\arg\max_h m(h)$ can be obtained by inspection (visual inspection, in fact, if $\dim(h) = 2$), and there is no need for an iterative scheme. The second reason why ISST can be slow is that after the Markov chain $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_n$ has been obtained, the method requires the calculation of the ratios $\nu^{(h)}(\boldsymbol{\psi}_i) \big/ \big[\sum_{j=1}^{J} b_j \nu^{(h_j)}(\boldsymbol{\psi}_i)\big]$, $i = 1, \ldots, n$ (see Section 4.1), and here the constants $b_1, \ldots, b_J$ are tuning parameters, which must be obtained via a time-consuming iterative procedure.

Table 6 gives the times needed for the FBEB and ISST methods, and also for the GEM and VEM methods, for corpora C-1–C-3. The left part of the table gives the actual times, and the right part gives the times standardized by the FBEB time, for ease of comparison. The variable $n_{\text{iter}}$ is the number of iterations needed for ISST; this is typically about $20$. As can be seen from the table,

the time needed to implement FBEB is very substantially less than the time needed for ISST, and is only about a single order of magnitude larger than the time needed for VEM. As is discussed in Section 5, when suitably adjusted, FBEB can handle large corpora.

| Corpus | $\hat{\hat{h}}_{\text{FBEB}}$ | $\hat{\hat{h}}_{\text{ISST}}$ | $\hat{\hat{h}}_{\text{GEM}}$ | $\hat{\hat{h}}_{\text{VEM}}$ |
|---|---|---|---|---|
| C-1 | 14.33 | $1016.11 \times n_{\text{iter}}$ | 56.95 | 1.21 |
| C-2 | 5.65 | $185.83 \times n_{\text{iter}}$ | 24.38 | 0.89 |
| C-3 | 3.81 | $70.33 \times n_{\text{iter}}$ | 6.78 | 0.34 |

| Corpus | $\hat{\hat{h}}_{\text{FBEB}}$ | $\hat{\hat{h}}_{\text{ISST}}$ | $\hat{\hat{h}}_{\text{GEM}}$ | $\hat{\hat{h}}_{\text{VEM}}$ |
|---|---|---|---|---|
| C-1 | 1 | $71 \times n_{\text{iter}}$ | 4.0 | 0.084 |
| C-2 | 1 | $33 \times n_{\text{iter}}$ | 4.3 | 0.158 |
| C-3 | 1 | $18 \times n_{\text{iter}}$ | 1.8 | 0.089 |

Table 6: Execution times, in minutes, for four estimators of $\arg\max_h m(h)$. The table on left gives the actual times, and the table on right gives the times standardized by the time for $\hat{\hat{h}}_{\text{FBEB}}$. The time for $\hat{\hat{h}}_{\text{ISST}}$ depends on $n_{\text{iter}}$, the number of iterations needed to set the skeleton grid.

## 4.3 Comparison of the Empirical Bayes and Fully Bayes Methods

It is natural to ask how our FBEB approach compares with a fully Bayesian approach in which we put a prior on $h$. Such a comparison is quite different from those in Sections 4.2.2–4.2.4: in those sections, all the contenders were numerical implementations of the empirical Bayes method. On the other hand, a comparison of the FBEB and fully-Bayes methods is really a comparison of two different statistical procedures. Furthermore, "fully-Bayes" does not refer to a single procedure, but rather to a family of procedures, one for each prior on $h$. In this section, we consider the case where $h$ is two-dimensional, $h = (\alpha, \eta)$, and we consider the family of gamma priors: $\alpha, \eta \overset{\text{iid}}{\sim} g_{a,b}$. Our criterion for comparison is the PPL score discussed in Section 4.2.3.

Recall that for a given value of $h$, the PPL score for the model indexed by $h$ is $S(h) = \left[\prod_{d=1}^{D} L_d(h)\right]^{1/D}$, where $L_d(h)$ is given by (4.2). In order to carry out our comparison, we need to define the score for the fully-Bayesian case. For any $a$ and $b$, let $\nu^{(a,b)}_{(\boldsymbol{\psi},h)\,|\,\boldsymbol{w}_{(-d)}}$ denote the posterior distribution of $(\boldsymbol{\psi}, h)$ given $\boldsymbol{w}_{(-d)}$ when the prior on $h$ is $g_{a,b}$. The analogue of (4.2) for the fully-Bayes case is

$$\mathsf{L}_d(a,b) = \int \ell_{\boldsymbol{w}_d}(\boldsymbol{\psi})\, d\nu^{(a,b)}_{(\boldsymbol{\psi},h)\,|\,\boldsymbol{w}_{(-d)}}(\boldsymbol{\psi}),$$

and the analogue of $S(h)$ is $\mathsf{S}_d(a,b) = \left[\prod_{d=1}^{D} \mathsf{L}_d(a,b)\right]^{1/D}$. As for the fixed-$h$ case, $\mathsf{L}_d(a,b)$ is analytically intractable. We estimate it by MCMC in a way that is very similar to the way we estimated $L_d(h)$. In a little more detail, let $(\boldsymbol{\psi}^{(1)}, h^{(1)}), (\boldsymbol{\psi}^{(2)}, h^{(2)}), \ldots$ be a Markov chain whose invariant distribution is $\nu^{(a,b)}_{(\boldsymbol{\psi},h)\,|\,\boldsymbol{w}_{(-d)}}$. We estimate $\mathsf{L}_d(a,b)$ by $\widehat{\mathsf{L}}_d(a,b) = (1/n)\sum_{i=1}^{n} \ell_{\boldsymbol{w}_d}(\boldsymbol{\psi}^{(i)})$,

and estimate the score $\mathsf{S}_d(a,b)$ by $\left[\prod_{d=1}^{D} \widehat{\mathsf{L}}_d(a,b)\right]^{1/D}$. We compare the empirical Bayes and fully Bayes methods through the ratio $S(\hat{h})/\mathsf{S}_d(a,b)$, where $\hat{h}$ is the empirical Bayes estimate of $h$ or, to be more precise, the estimate of this ratio. The plots in Figure 6 give graphs, from two different angles, of this ratio, as $a$ ranges from $.1$ to $2.0$ by increments of $.1$ and $b$ ranges from $.01$ to $.5$ by increments of $.01$, for a total of $1000$ values. (Significantly, this range for $(a,b)$ "nearly includes" the value $(a,b) = (0,0)$, which corresponds to a standard non-informative prior $p(u) \propto (1/u)$, an improper prior that is a common choice.) The plot shows that the empirical Bayes method outperforms the fully Bayes method for most of the cases and when it does not, it is not by much.

We now briefly remark on how the plots were constructed. We ran $1000$ experiments, one for each $(a,b)$ pair and, because the experiments are independent, the resulting figure was very ragged: even when two $(a,b)$ pairs are close, the estimates of the ratios of the scores for these pairs don't need to be close. We can deal with this difficulty using standard regression methods, in which we use the following model: Calculated Ratio of Scores $= f(a,b) + \epsilon$, where $f$ is unknown. The function $f$ can be estimated nonparametrically by bivariate splines, or by using generalized additive models. The plots in Figure 6 actually result from applying the R function `gam` to the $1000$ experimental points.

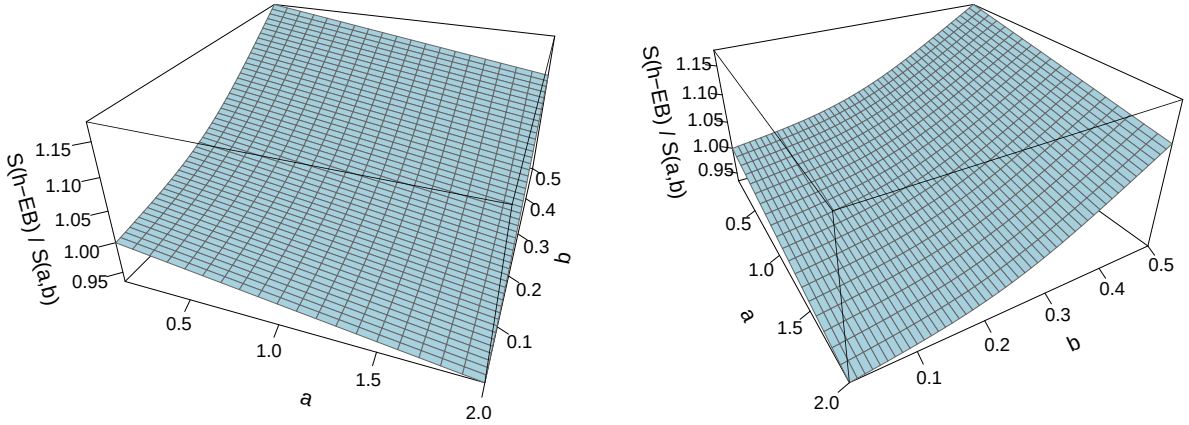



Figure 6: Smoothed estimates of the ratio of the PPL for the model based on the empirical Bayes choice of $h$ to that of the fully Bayes model using $g_{a,b}$ priors for $\alpha$ and $\eta$, from two different views. The plots show that the empirical Bayes model does better or not much worse over the entire $(a,b)$ range.

# 5 Discussion

Our methodology handles corpora for which either $D$ or $T$ or both are large, whereas as we have shown, competing methods do not. In Xia and Doss (2019, Section 7) we give the results of an experiment that shows that it gives accurate estimates of the posterior distribution for corpora of a hundred thousand documents. Whether we can handle a given corpus depends on whether MCMC methods can handle that corpus, and we now elaborate on what we mean by this. We have used the CGS, which runs through the vector $\boldsymbol{z} = (z_{11}, \ldots, z_{1n_1}, \ldots, z_{D1}, \ldots, z_{Dn_D})$, updating each variable sequentially, with $\boldsymbol{\beta}$ and $\boldsymbol{\theta}$ integrated out. So there is a node for each word in the corpus, and this makes the CGS very slow. George (2015) (see also Doss and George (2018)) considered another Markov chain for estimating the posterior distribution of $\boldsymbol{\psi} = (\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{z})$: we look at the *pair* $(\boldsymbol{z}, (\boldsymbol{\beta}, \boldsymbol{\theta}))$, and the chain is a two-cycle Gibbs sampler that alternates between updating $\boldsymbol{z}$ and updating $(\boldsymbol{\beta}, \boldsymbol{\theta})$. The conditional distributions $\nu^{(h)}_{\boldsymbol{z}\,|\,(\boldsymbol{\beta},\boldsymbol{\theta},\boldsymbol{w})}$ and $\nu^{(h)}_{(\boldsymbol{\beta},\boldsymbol{\theta})\,|\,(\boldsymbol{z},\boldsymbol{w})}$ are available in closed form; they may be found in Doss and George (2018), for example. This "Grouped Gibbs Sampler" has the very attractive feature that it can be parallelized: given $(\boldsymbol{\beta}, \boldsymbol{\theta})$ (and $\boldsymbol{w}$), the components of $\boldsymbol{z}$ are all independent, so can be updated simultaneously by different processors; and given $\boldsymbol{z}$ (and $\boldsymbol{w}$), the $\theta_d$'s and $\beta_t$'s are all independent, so also can be updated simultaneously by different processors. Moreover, contrary to a widely-held view, the mixing rate for this sampler is comparable to that of the CGS (Doss and George, 2018). When we use this Gibbs sampler, FBEB can handle corpora with up to hundreds of thousands of documents, depending on how many processors are available.

We believe that the FBEB method can be developed for other topic models, where the dimension of the hyperparameter is high enough to preclude the use of competing approaches. Hierarchical Dirichlet processes (Teh et al., 2006) and the Correlated Topics Model (Blei and Lafferty, 2007) are prominent examples of such topic models.

# Supplementary Materials

**R Code and Data** The supplemental files for this article include files containing R code and data for reproducing all the empirical studies in the paper. The Readme file contained in the zip file gives a description of all the other files in the archive. (shs-lda-code.zip, zip archive)

**Appendix** The supplemental files include a document which gives the following: (i) a review of Hamiltonian Monte Carlo, (ii) a section showing feasibility of our method on large corpora,

(iii) proofs of Theorems 1–3, and (iv) some minor theoretical details. (shs-lda-supp.pdf)

## Acknowledgments

We are grateful to the two referees for their helpful constructive criticism.

# References

Asuncion, A., Welling, M., Smyth, P. and Teh, Y. W. (2009). On smoothing and inference for topic models. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*. UAI '09, AUAI Press, Arlington, Virginia, United States.

Blei, D. M. and Jordan, M. I. (2003). Modeling annotated data. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '03, ACM, New York, NY.

Blei, D. M., Kucukelbir, A. and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association* **112** 859–877.

Blei, D. M. and Lafferty, J. D. (2007). A correlated topic model of *Science*. *The Annals of Applied Statistics* **1** 17–35.

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

Celeux, G., Hurn, M. and Robert, C. P. (2000). Computational and inferential difficulties with mixture posterior distributions. *Journal of the American Statistical Association* **95** 957–970.

Chen, M.-H. (1994). Importance-weighted marginal Bayesian posterior density estimation. *Journal of the American Statistical Association* **89** 818–824.

Chen, Z. (2015). *Inference for the Number of Topics in the Latent Dirichlet Allocation Model via Bayesian Mixture Modelling*. Ph.D. thesis, University of Florida.

Chib, S. (1995). Marginal likelihood from the Gibbs output. *Journal of the American Statistical Association* **90** 1313–1321.

Chib, S. and Jeliazkov, I. (2001). Marginal likelihood from the Metropolis-Hastings output. *Journal of the American Statistical Association* **96** 270–281.

Donoho, D. L. and Liu, R. C. (1991). Geometrizing rates of convergence, II. *The Annals of Statistics* **19** 633–667.

Doss, H. and George, C. P. (2018). Theoretical and empirical evaluation of a grouped Gibbs sampler for parallel computation in the LDA model. Tech. rep., Department of Statistics, University of Florida.

Doss, H. and Linero, A. (2018). A fully-Bayes approach to empirical Bayes inference and Bayesian sensitivity analysis. Tech. rep., Department of Statistics, University of Florida (in preparation).

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* **90** 577–588.

Flegal, J. M., Haran, M. and Jones, G. L. (2008). Markov chain Monte Carlo: Can we trust the third significant figure? *Statistical Science* **23** 250–260.

Flegal, J. M., Hughes, J. and Vats, D. (2016). *mcmcse: Monte Carlo Standard Errors for MCMC*. Riverside, CA and Minneapolis, MN. R package version 1.2-1.

George, C. P. (2015). *Latent Dirichlet Allocation: Hyperparameter Selection and Applications to Electronic Discovery*. Ph.D. thesis, University of Florida.

George, C. P. and Doss, H. (2018). Principled selection of hyperparameters in the latent Dirichlet allocation model. *Journal of Machine Learning Research* **18** 1–38.

George, E. I. and Foster, D. P. (2000). Calibration and empirical Bayes variable selection. *Biometrika* **87** 731–747.

Geyer, C. J. and Thompson, E. A. (1995). Annealing Markov chain Monte Carlo with applications to ancestral inference. *Journal of the American Statistical Association* **90** 909–920.

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

Hobert, J. P. (2011). The data augmentation algorithm: Theory and methodology. In *Handbook of Markov Chain Monte Carlo* (S. P. Brooks, A. Gelman, G. L. Jones and X.-L. Meng, eds.). CRC Press, Boca Raton, 253–293.

Jasra, A., Holmes, C. C. and Stephens, D. A. (2005). Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Statistical Science* **20** 50–67.

Jones, G. L., Haran, M., Caffo, B. S. and Neath, R. (2006). Fixed-width output analysis for Markov chain Monte Carlo. *Journal of the American Statistical Association* **101** 1537–1547.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S. and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine Learning* **37** 183–233.

Marinari, E. and Parisi, G. (1992). Simulated tempering: A new Monte Carlo scheme. *Europhysics Letters* **19** 451–458.

Meng, X.-L. and Van Dyk, D. (1997). The EM algorithm—an old folk-song sung to a fast new tune. *Journal of the Royal Statistical Society,* Series B **59** 511–567.

Minka, T. P. (2003). Estimating a Dirichlet distribution.
URL `http://research.microsoft.com/~minka/papers/dirichlet/`

Nash, J. C. and Varadhan, R. (2011). Unifying optimization algorithms to aid software system users: optimx for R. *Journal of Statistical Software* **43** 1–14.

Neal, R. M. (2003). Slice sampling. *The Annals of Statistics* **31** 705–741.

Neal, R. M. (2011). MCMC using Hamiltonian dynamics. In *Handbook of Markov Chain Monte Carlo* (S. P. Brooks, A. Gelman, G. L. Jones and X.-L. Meng, eds.). CRC Press, Boca Raton, 113–162.

Newman, D., Asuncion, A., Smyth, P. and Welling, M. (2009). Distributed algorithms for topic models. *Journal of Machine Learning Research* **10** 1801–1828.

Newton, M. and Raftery, A. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap (with discussion). *Journal of the Royal Statistical Society,* Series B **56** 3–48.

Řehůřek, R. and Sojka, P. (2010). Software framework for topic modelling with large corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. ELRA, Valletta, Malta.

Robert, C. P. (2001). *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. Springer-Verlag, New York.

Rosen-Zvi, M., Griffiths, T., Steyvers, M. and Smyth, P. (2004). The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*. UAI '04, AUAI Press, Arlington, Virginia, United States.

Teh, Y. W., Jordan, M. I., Beal, M. J. and Blei, D. M. (2006). Hierarchical Dirichlet processes. *Journal of the American Statistical Association* **101** 1566–1581.

Tsybakov, A. B. (1990). Recursive estimation of the mode of a multivariate distribution. *Problemy Peredachi Informatsii* **26** 38–45.

Wallach, H. M. (2006). Topic modeling: Beyond bag-of-words. In *Proceedings of the 23rd International Conference on Machine Learning*. ICML '06, ACM, New York, NY, USA.

Wallach, H. M. (2008). *Structured Topic Models for Language*. Ph.D. thesis, University of Cambridge.

Wallach, H. M., Murray, I., Salakhutdinov, R. and Mimno, D. (2009). Evaluation methods for topic models. In *Proceedings of the 26th Annual International Conference on Machine Learning*. ACM.

Wolpert, R. L. and Schmidler, S. C. (2012). $\alpha$-stable limit laws for harmonic mean estimators of marginal likelihoods. *Statistica Sinica* **22** 1233–1251.

Xia, W. (2018). *Scalable Hyperparameter Selection for Latent Dirichlet Allocation*. Ph.D. thesis, University of Florida.

Xia, W. and Doss, H. (2019). Supplement to "Scalable hyperparameter selection for latent Dirichlet allocation".

Yang, Y. (2005). Can the strengths of AIC and BIC be shared? A conflict between model indentification and regression estimation. *Biometrika* **92** 937–950.