Principled Selection of Hyperparameters in the Latent Dirichlet Allocation Model

Clint P. George*
Informatics Institute
University of Florida
clintpg@ufl.edu

Hani Doss†
Department of Statistics
University of Florida
doss@stat.ufl.edu

Abstract

Latent Dirichlet Allocation (LDA) is a well known topic model that is often used to make inference regarding the properties of collections of text documents. LDA is a hierarchical Bayesian model, and involves a prior distribution on a set of latent topic variables. The prior is indexed by certain hyperparameters, and even though these have a large impact on inference, they are usually chosen either in an ad-hoc manner, or by applying an algorithm whose theoretical basis has not been firmly established. We present a method, based on a combination of Markov chain Monte Carlo and importance sampling, for estimating the maximum likelihood estimate of the hyperparameters. The method may be viewed as a computational scheme for implementation of an empirical Bayes analysis. It comes with theoretical guarantees, and a key feature of our approach is that we provide theoretically-valid error margins for our estimates. Experiments on both synthetic and real data show good performance of our methodology.

Running Title: Hyperparameter Selection in LDA Model

Key words and phrases: Empirical Bayes inference, latent Dirichlet allocation, Markov chain Monte Carlo, model selection, topic modelling.

*Research supported by the International Center for Automated Research at the UF Levin College of Law

†Research supported by NSF Grant DMS-11-06395 and NIH grant P30 AG028740

1 Introduction

Latent Dirichlet Allocation (LDA, Blei et al. 2003) is a model that is used to describe high-dimensional sparse count data represented by feature counts. Although the model can be applied to many different kinds of data, for example collections of annotated images and social networks, for the sake of concreteness, here we focus on data consisting of a collection of documents. Suppose we have a corpus of documents, say a collection of news articles, and these 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. There are several goals, but two principal ones are to recover an interpretable set of topics, 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 $\mathcal{V}$ of V words; typically, this is taken to be the union of all the words in all the documents of the corpus, after removing stop (i.e. uninformative) words. (Throughout, we use "word" to refer to either an actual word, or to a phrase, such as "heart attack"; LDA has implementations that deal with each of these.) There are D documents in the corpus, and for $d = 1, \dots, D$, document d has n_d words, $w_{d1}, \dots, w_{dn_d}$. The order of the words is considered uninformative, and so is neglected. Each word is represented as an index vector of dimension V with a 1 at the s^{th} element, where s denotes the term selected from the vocabulary. Thus, document d is represented by the matrix $\boldsymbol{w}_d = (w_{d1}, \dots, w_{dn_d})$ and the corpus is represented by the list $\boldsymbol{w} = (\boldsymbol{w}_1, \dots, \boldsymbol{w}_D)$. The number of topics, K, is finite and known. By definition, a topic is a distribution over $\mathcal{V}$, i.e. a point in the simplex $\mathbb{S}_V = \{a \in \mathbb{R}^V : a_1, \dots, a_V \geq 0, \sum_{j=1}^V a_j = 1\}$. For $d = 1, \dots, D$, for each word w_{di}, z_{di} is an index vector of dimension K which represents the latent variable that denotes the topic from which w_{di} is drawn. The distribution of $z_{d1}, \dots, z_{dn_d}$ will depend on a document-specific variable θ_d which indicates a distribution on the topics for document d.

We will use $\text{Dir}_L(a_1, \dots, a_L)$ to denote the finite-dimensional Dirichlet distribution on the simplex $\mathbb{S}_L$. Also, we will use $\text{Mult}_L(b_1, \dots, b_L)$ to denote the multinomial distribution with number of trials equal to 1 and probability vector $(b_1, \dots, b_L)$. We will form a $K \times V$ matrix $\boldsymbol{\beta}$, whose t^{th} row is the t^{th} topic (how $\boldsymbol{\beta}$ is formed will be described shortly). Thus, $\boldsymbol{\beta}$ will consist of vectors $\beta_1, \dots, \beta_K$, all lying in $\mathbb{S}_V$. The LDA model is indexed by hyperparameters $\eta \in (0, \infty)$ and $\boldsymbol{\alpha} \in (0, \infty)^K$. It is represented graphically in Figure 1, and described formally by the following hierarchical model:

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

From the description of the model, we see that there is a latent topic variable for every word that appears in the corpus. Thus it is possible that a document spans several topics. Also, because $\boldsymbol{\beta}$ is chosen once, at the top of the hierarchy, it is shared among the D documents. Thus the model encourages different documents to share the same topics, and moreover, all the documents in the corpus share a single set of topics defined by $\boldsymbol{\beta}$.

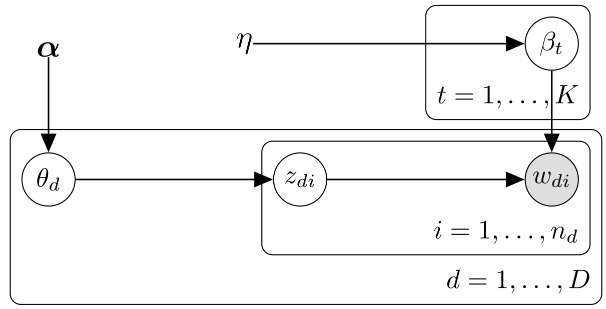


Figure 1: Graphical model representation for LDA. Nodes denote random variables, shaded nodes denote observed variables, edges denote conditional dependencies, and plates denote replicated processes.

Let $\boldsymbol{\theta} = (\theta_1, \dots, \theta_D)$, $\boldsymbol{z}_d = (z_{d1}, \dots, z_{dn_d})$ for $d = 1, \dots, D$, $\boldsymbol{z} = (\boldsymbol{z}_1, \dots, \boldsymbol{z}_D)$, and let $\boldsymbol{\psi} = (\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{z})$. The model is indexed by the hyperparameter vector $h = (\eta, \boldsymbol{\alpha}) \in (0, \infty)^{K+1}$. For any given h, lines 1–3 induce a prior distribution on $\boldsymbol{\psi}$, which we will denote by ν_h. Line 4 gives the likelihood. The words $\boldsymbol{w}$ are observed, and we are interested in $\nu_{h,\boldsymbol{w}}$, the posterior distribution of $\boldsymbol{\psi}$ given $\boldsymbol{w}$ corresponding to ν_h. In step 2 it is common to take the distribution of the θ_d's to be a symmetric Dirichlet, although arbitrary Dirichlets are sometimes used. Our model allows for arbitrary Dirichlets, for the sake of generality, but in all our examples we use symmetric Dirichlets because a high-dimensional hyperparameter can cause serious problems. We return to this point at the end of Section 2.1.

The hyperparameter vector h is not random, and must be selected in advance. It has a strong effect on the distribution of the parameters of the model. For example, when η is large, the topics tend to be probability vectors which spread their mass evenly among many words in the vocabulary, whereas when η is small, the topics tend to put most of their mass on only a few words. Also, in the special case where $\boldsymbol{\alpha} = (\alpha, \dots, \alpha)$, so that $\mathrm{Dir}_K(\boldsymbol{\alpha})$ is a symmetric Dirichlet indexed by the single parameter α, when α is large, each document tends to involve many different topics; on the other hand, in the limiting case where $\alpha \to 0$, each document involves a single topic, and this topic is randomly chosen from the set of all topics.

The preceding paragraph is about the effect of h on the prior distribution of the parameters. We may think about the role of h on statistical inference by considering posterior distributions. Let g be a function of the parameter $\boldsymbol{\psi}$. For example, $g(\boldsymbol{\psi})$ might be the indicator of the event $\|\theta_i - \theta_j\| \le \epsilon$, where i and j are the indices of two particular documents, ϵ is some user-specified small number, and $\| \cdot \|$ denotes ordinary Euclidean distance in $\mathbb{R}^K$. In this case, the value of $g(\boldsymbol{\psi})$ gives a way of determining whether the topics for documents i and j are nearly the same ($g(\boldsymbol{\psi}) = 1$), or not ($g(\boldsymbol{\psi}) = 0$). Of interest then is the posterior probability $\nu_{h,\boldsymbol{w}}(\|\theta_i - \theta_j\| \le \epsilon)$, which is given by the integral $\int g(\boldsymbol{\psi}) \, d\nu_{h,\boldsymbol{w}}(\boldsymbol{\psi})$. In another example, the function g might be taken to measure the distance between two topics of interest. In Section 2.4 we demonstrate empirically that the posterior expectation given by the integral $\int g(\boldsymbol{\psi}) \, d\nu_{h,\boldsymbol{w}}(\boldsymbol{\psi})$ can vary considerably with h.

To summarize: the hyperparameter h can have a strong effect not only on the prior distribution of the parameters in the model, but also on their posterior distribution; therefore it is important to choose it carefully. Yet in spite of the very widespread use of LDA, there is no method for choosing the hyperparameter that has a firm theoretical basis. In the literature, h is sometimes selected in some ad-hoc or arbitrary manner. A principled way of selecting it is via maximum likelihood: we let $m_{\boldsymbol{w}}(h)$ denote the marginal likelihood of the data as a function of h, and use $\hat{h} = \arg\max_h m_{\boldsymbol{w}}(h)$ which is, by definition, the empirical Bayes choice of h. We will write $m(h)$ instead of $m_{\boldsymbol{w}}(h)$ unless we need to emphasize the dependence on $\boldsymbol{w}$. Unfortunately, the function $m(h)$ is analytically intractable: $m(h)$ is the likelihood of the data with all latent variables integrated or summed out, and from the hierarchical nature of the model, we see that $m(h)$ is a very large sum, because we are summing over all possible values of $\boldsymbol{z}$. Blei et al. (2003) propose estimating $\arg\max_h m(h)$ via a combination of the EM algorithm and "variational inference" (VI-EM). Very briefly, $\boldsymbol{w}$ is viewed as "observed data," and $\boldsymbol{\psi}$ is viewed as "missing data." Because the "complete data likelihood" $p_h(\boldsymbol{\psi}, \boldsymbol{w})$ is available, the EM algorithm is a natural candidate for estimating $\arg\max_h m(h)$, since $m(h)$ is the "incomplete data likelihood." But the E-step in the algorithm is infeasible because it requires calculating an expectation with respect to the intractable distribution $\nu_{h,\boldsymbol{w}}$. Blei et al. (2003) substitute an approximation to this expectation. Unfortunately, because there are no useful bounds on the approximation, and because the approximation is used at every iteration of the algorithm, there are no results regarding the theoretical properties of this method. Wallach (2006) (see also Wallach (2008)) proposed a "Gibbs-EM" algorithm, in which the E-step is approximated by a Markov chain Monte Carlo estimate. This method can perform well empirically but, as for the VI-EM algorithm, its theoretical validity has not been established. The advantages and disadvantages of these two approximations to the EM algorithm are discussed in Section 5.1.

Wallach et al. (2009) give an overview of a class of methods for estimating a combination of some parameter components and some hyperparameter components and, in principle, these procedures could be adapted to the problem of estimating h. The methods they present differ from EM-based approaches in two fundamental respects: (1) they work with an objective function which is not the marginal likelihood function $m(h)$, but rather a measure of the "predictive performance of the LDA model indexed by h," and (2) evaluation of their objective function at hyperparameter value h requires running a Markov chain, and this has to be done "for each value of h" before doing the maximization of the objective function, which can impose a heavy computational burden. This paper is discussed further in Section 5.

Another approach for dealing with the problem of having to make a choice of the hyperparameter vector is the fully Bayes approach, in which we simply put a prior on the hyperparameter vector, that is, add one layer to the hierarchical model. For example, we can either put a flat prior on each component of the hyperparameter, or put a gamma prior instead. While this approach can be useful, there are reasons why one may want to avoid it. On the one hand, if we put a flat prior then one problem is that we are effectively skewing the results towards large values of the hyperparameter components. A more serious problem is that the posterior may be improper. In this case, insidiously, if we use Gibbs sampling to estimate the posterior, it is possible that all conditionals needed to implement the sampler are proper; but Hobert and Casella (1996) have shown that the Gibbs sampler output may not give a clue that there is a problem. On the other hand, if we use a gamma prior, then at least in the case of a symmetric Dirichlet on the θ_d's, we have not made things any easier: we have to specify four gamma hyperparameters. Another reason to avoid the fully Bayes approach is that, in broad terms, the general interest in empirical

Bayes methods arises in part from a desire to select a specific value of the hyperparameter vector because this gives a model that is more parsimonious and interpretable. This point is discussed more fully (in a general context) in George and Foster (2000) and Robert (2001, Chapter 7).

In the present paper we show that while it is not possible to compute $m(h)$ itself, it is nevertheless possible, with a single MCMC run, to estimate the entire function $m(h)$ up to a multiplicative constant. Before proceeding, we note that if c is a constant, then the information regarding h given by the two functions $m(h)$ and $cm(h)$ 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 $cm(h)$ are identical. Let g be a function of $\boldsymbol{\psi}$ and let $I(h) = \int g(\boldsymbol{\psi})\, d\nu_{h,\boldsymbol{w}}(\boldsymbol{\psi})$ denote the posterior expectation of $g(\boldsymbol{\psi})$. We also show that it is possible to estimate the entire function $I(h)$ with a single MCMC run.

As we will see in Section 2, our approach for estimating $m(h)$ up to a single multiplicative constant and $I(h)$ has two requirements: (i) we need a formula for the ratio $\nu_{h_1}(\boldsymbol{\psi})/\nu_{h_2}(\boldsymbol{\psi})$ for any two hyperparameter values h_1 and h_2, and (ii) for any hyperparameter value h, we need an ergodic Markov chain whose invariant distribution is the posterior $\nu_{h,\boldsymbol{w}}$. This paper is organized as follows. In Section 2 we explain our method for estimating the function $m(h)$ up to a single multiplicative constant (and hence its argmax) and for estimating the family of posterior expectations $\{I(h),\, h \in \mathcal{H}\}$; and we also explain how to form error margins for our estimates, paying particular attention to theoretical underpinnings. Additionally, we provide the formula for the ratio $\nu_{h_1}(\boldsymbol{\psi})/\nu_{h_2}(\boldsymbol{\psi})$. In Section 3 we consider synthetic data sets generated from a simple model in which h is low dimensional and known, and we show that our method correctly estimates the true value of h. In Section 4 we describe two Markov chains which satisfy requirement (ii) above. In Section 5 we compare, both theoretically and empirically, the various methods of estimating the maximizer of the marginal likelihood function, in terms of accuracy. Then we compare the various choices of the hyperparameter that are used in the literature—those that are ad-hoc and those that estimate the maximizer of the marginal likelihood function—through a standard criterion that is used to evaluate topic models, and we show that our method performs favorably. In Section 6 we make some concluding remarks, and the Appendix contains some of the technical material that is needed in the paper.

2 Estimation of the Marginal Likelihood up to a Multiplicative Constant and Estimation of Posterior Expectations

This section consists of four parts. In Section 2.1 we show how the marginal likelihood function can be estimated (up to a constant) with a single MCMC run. Section 2.2 concerns estimation of the posterior expectation of a function g of the parameter $\boldsymbol{\psi}$, given by the integral $\int g(\boldsymbol{\psi})\, d\nu_{h,\boldsymbol{w}}(\boldsymbol{\psi})$, and which depends on h. We show how the entire family of posterior expectations $\{I(h),\, h \in \mathcal{H}\}$ can be estimated with a single MCMC run. In Section 2.3 we explain that the simple estimates given in Sections 2.1 and 2.2 can have large variances, and we present estimates which are far more reliable. In Section 2.4 we illustrate our methodology on a corpus created from Wikipedia.

Let $\mathcal{H} = (0, \infty)^{K+1}$ be the hyperparameter space. For any $h \in \mathcal{H}$, ν_h and $\nu_{h,\boldsymbol{w}}$ are prior and posterior distributions, respectively, of the vector $\boldsymbol{\psi} = (\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{z})$, for which some components are continuous and some are discrete. We will use $\ell_{\boldsymbol{w}}(\boldsymbol{\psi})$ to denote the likelihood function (which is given by line 4 of the LDA model).

2.1 Estimation of the Marginal Likelihood up to a Multiplicative Constant

Note that $m(h)$ is the normalizing constant in the statement "the posterior is proportional to the likelihood times the prior," i.e.

$$\nu_{h,\boldsymbol{w}}(\boldsymbol{\psi}) = \frac{\ell_{\boldsymbol{w}}(\boldsymbol{\psi})\nu_h(\boldsymbol{\psi})}{m(h)}.$$

Now suppose that we have a method for constructing a Markov chain on $\boldsymbol{\psi}$ whose invariant distribution is $\nu_{h,\boldsymbol{w}}$ and which is ergodic. Two Markov chains which satisfy these criteria are discussed in later in this section. Let $h_* \in \mathcal{H}$ be fixed but arbitrary, and let $\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \dots$ be an ergodic Markov chain with invariant distribution $\nu_{h*,\boldsymbol{w}}$. For any $h \in \mathcal{H}$, as $n \to \infty$ we have

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n} \frac{\nu_h(\boldsymbol{\psi}_i)}{\nu_{h_*}(\boldsymbol{\psi}_i)} &\xrightarrow{\text{a.s.}} \int \frac{\nu_h(\boldsymbol{\psi})}{\nu_{h_*}(\boldsymbol{\psi})}\, d\nu_{h_*,\boldsymbol{w}}(\boldsymbol{\psi}) \\
&= \frac{m(h)}{m(h_*)} \int \frac{\ell_{\boldsymbol{w}}(\boldsymbol{\psi})\nu_h(\boldsymbol{\psi})/m(h)}{\ell_{\boldsymbol{w}}(\boldsymbol{\psi})\nu_{h_*}(\boldsymbol{\psi})/m(h_*)}\, d\nu_{h_*,\boldsymbol{w}}(\boldsymbol{\psi}) \\
&= \frac{m(h)}{m(h_*)} \int \frac{\nu_{h,\boldsymbol{w}}(\boldsymbol{\psi})}{\nu_{h_*,\boldsymbol{w}}(\boldsymbol{\psi})}\, d\nu_{h_*,\boldsymbol{w}}(\boldsymbol{\psi}) \\
&= \frac{m(h)}{m(h_*)}.
\end{aligned} \tag{2.1}$$

The almost sure convergence statement in (2.1) follows from ergodicity of the chain. (There is a slight abuse of notation in (2.1) in that we have used $\nu_{h_*,\boldsymbol{w}}$ to denote a probability measure when we write $d\nu_{h_*,\boldsymbol{w}}$, whereas in the integrand, ν_h, ν_{h_*}, and $\nu_{h_*,\boldsymbol{w}}$ refer to probability densities.) The significance of (2.1) is that this result shows that we can estimate the entire family $\{m(h)/m(h_*),\ h \in \mathcal{H}\}$ with a single Markov chain run. Since $m(h_*)$ is a constant, the remarks made in Section 1 apply, and we can estimate $\arg\max_h m(h)$. The usefulness of (2.1) 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.

The development in (2.1) is not new (although we do not know who first noticed it), and the estimate on the left side of (2.1) is not the one we will ultimately use (cf. Section 2.3); we present (2.1) primarily for motivation. Note that (2.1) is generic, i.e. it is not specific to the LDA model: it is potentially valid for any Bayesian model for which we have a data vector $\boldsymbol{w}$, a corresponding likelihood function $\ell_{\boldsymbol{w}}(\boldsymbol{\psi})$, and parameter vector $\boldsymbol{\psi}$ having prior ν_h, with hyperparameter $h \in \mathcal{H}$. We now discuss carefully the scope of its applicability. In order to be able to use (2.1) to obtain valid estimators of the family $m(h)/m(h_*)$, $h \in \mathcal{H}$, we need the following.

C1 A closed-form expression for the ratio of densities $\nu_h(\boldsymbol{\psi})/\nu_{h_*}(\boldsymbol{\psi})$ for some fixed $h_* \in \mathcal{H}$.

C2 A method for generating an ergodic Markov chain with invariant distribution $\nu_{h_*,\boldsymbol{w}}$.

We need C1 in order to write down the estimators, and we need C2 for the estimators to be valid.

For notational convenience, let $B_n(h) = (1/n)\sum_{i=1}^n [\nu_h(\boldsymbol{\psi}_i)/\nu_{h_*}(\boldsymbol{\psi}_i)]$, and define $B(h) = m(h)/m(h_*)$. In order to use (2.1) to obtain a valid estimator, together with a confidence interval (confidence set, if $\dim(h) > 1$) for $\arg\max_h m(h)$, we need in addition the following.

C3 A result that says that the convergence in the first line of (2.1) is uniform in h.

C4 A result that says that if $G_n(h)$ is the centered and scaled version of the estimate on the left side of (2.1) given by $G_n(h) = n^{1/2}(B_n(h) - B(h))$, then $G_n(\cdot)$ converges in distribution to a Gaussian process indexed by h.

We now explain these last two conditions. Generally speaking, for 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 $\arg\max_h f_n(h) \to \arg\max_h f(h)$. Indeed, counterexamples are easy to construct, and in Section A.2 of the Appendix we provide a simple one. In order to be able to conclude that $\arg\max_h f_n(h) \to \arg\max_h f(h)$, which is a global condition, we need the convergence of f_n to f to be uniform (this is discussed rigorously in Section A.2 of the Appendix). Hence we need C3. Regarding confidence intervals (or sets) for $\arg\max_h f(h)$, we note that $B_n(h)$ is simply an average, and so under suitable regularity conditions, it satisfies a central limit theorem (CLT). However, we are not interested in a central limit theorem for $B_n(h)$, but rather in a CLT for $\arg\max_h B_n(h)$. In this regard, C4 is a "uniform in h CLT" that is necessary to obtain a CLT of the form $n^{1/2}(\arg\max_h B_n(h) - \arg\max_h B(h)) \xrightarrow{d} \mathcal{N}(0, \Sigma)$ for some positive definite matrix Σ, which is what is needed to form confidence sets for $\arg\max_h B(h)$. Again, this is discussed rigorously in Section A.2 of the Appendix.

We now discuss Conditions C1–C4. Condition C1 is satisfied by the LDA model, and in Section A.1 of the Appendix we show that the ratio of densities ν_h/ν_{h_*} is given by

$$\frac{\nu_h(\boldsymbol{\psi})}{\nu_{h_*}(\boldsymbol{\psi})} = \left[\prod_{d=1}^{D}\left(\frac{\Gamma\big(\sum_{j=1}^K \alpha_j\big)}{\prod_{j=1}^K \Gamma(\alpha_j)} \frac{\prod_{j=1}^K \Gamma(\alpha_j^*)}{\Gamma\big(\sum_{j=1}^K \alpha_j^*\big)} \prod_{j=1}^{K} \theta_{dj}^{\alpha_j - \alpha_j^*}\right)\right]\left[\prod_{j=1}^{K}\left(\frac{\Gamma(V\eta)}{\Gamma(\eta)^V} \frac{\Gamma(\eta^*)^V}{\Gamma(V\eta^*)} \prod_{t=1}^{V} \beta_{jt}^{\eta-\eta^*}\right)\right], \tag{2.2}$$

where $h_* = (\eta^*, \boldsymbol{\alpha}^*)$.

There are many extensions and variants of the LDA model described in Section 1—too many to even list them all here—and versions of (2.2) can be obtained for these models. This has to be done separately for each case. The features of the LDA model that make it possible to obtain a ratio of densities formula are that it is a hierarchical model, and at every stage the distributions are explicitly finite dimensional. For other models, a ratio of densities formula is obtainable routinely as long as these features exist: when they do, we have a closed-form expression for the prior distribution $\nu_h(\boldsymbol{\psi})$, and hence a closed-form expression for the ratio $\nu_h(\boldsymbol{\psi})/\nu_{h_*}(\boldsymbol{\psi})$.

Unfortunately, a ratio of densities formula is not always available. A prominent example is the "Hierarchical Dirichlet Processes" model introduced in Teh et al. (2006), which effectively allows infinitely many topics but with finitely many realized in any given document. Very briefly, in this model, for word i in document d, there is an unobserved topic ψ_{di}. The latent topic vector

$\boldsymbol{\psi}_d = (\psi_{d1}, \ldots, \psi_{dn_d})$ has a complicated joint distribution with strength of dependence governed by a hyperparameter h_1 (the precision parameter of the Dirichlet process in the middle of the hierarchy), and the D vectors $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_D$ also have a complicated dependence structure, with strength of dependence governed by a hyperparameter h_2 (the precision parameter of the Dirichlet process at the top of the hierarchy). The parameter vector for the model is $\boldsymbol{\psi} = (\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_D)$ and the hyperparameter is $h = (h_1, h_2)$. Unfortunately, the joint (prior) distribution of $\boldsymbol{\psi}$ is not available in closed form, and our efforts to obtain a formula for $\nu_h(\boldsymbol{\psi})/\nu_{h_*}(\boldsymbol{\psi})$ have been fruitless.

Regarding Condition C2, we note that Griffiths and Steyvers have developed a "collapsed Gibbs sampler" (CGS) which runs over the vector $\boldsymbol{z}$. The invariant distribution of the CGS is the conditional distribution of $\boldsymbol{z}$ given $\boldsymbol{w}$. The CGS cannot be used directly, because to apply (2.1) we need a Markov chain on the triple $(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{z})$, whose invariant distribution is $\nu_{h_*, \boldsymbol{w}}$. In Section 4 we obtain the conditional distribution of $(\boldsymbol{\beta}, \boldsymbol{\theta})$ given $\boldsymbol{z}$ and $\boldsymbol{w}$, and we show how to sample from this distribution. Therefore, given a Markov chain $\boldsymbol{z}^{(1)}, \ldots, \boldsymbol{z}^{(n)}$ generated via the CGS, we can form triples $(\boldsymbol{z}^{(1)}, \boldsymbol{\beta}^{(1)}, \boldsymbol{\theta}^{(1)}), \ldots, (\boldsymbol{z}^{(n)}, \boldsymbol{\beta}^{(n)}, \boldsymbol{\theta}^{(n)})$, and it is easy to see that this sequence forms a Markov chain with invariant distribution $\nu_{h_*, \boldsymbol{w}}$. We will refer to this Markov chain as the Augmented Collapsed Gibbs Sampler, and use the acronym ACGS. In Section 4 we show that the ACGS is not only geometrically ergodic, but actually is uniformly ergodic. We also show how to sample from the conditional distribution of $\boldsymbol{z}$ given $(\boldsymbol{\beta}, \boldsymbol{\theta})$ and $\boldsymbol{w}$. This enables us to construct a two-cycle Gibbs sampler which runs on the pair $(\boldsymbol{z}, (\boldsymbol{\beta}, \boldsymbol{\theta}))$. We will refer to this chain as the Grouped Gibbs Sampler, and use the acronym GGS. Either the ACGS or the GGS may be used, and we see that Condition C2 is satisfied for the LDA model.

The theorem below pertains to the LDA model and states that for this model, $\arg\max_h B_n(h)$ converges to $\arg\max_h m(h)$ almost surely, and that $\arg\max_h B_n(h)$ satisfies a CLT. The theorem also gives a procedure for constructing confidence sets for $\arg\max_h m(h)$. The result is explicit. Therefore, given a desired level of precision, we can determine the Markov chain length needed to estimate $\arg\max_h m(h)$ with that level of precision. The proof of the theorem is in Section A.2 of the Appendix. The theorem is valid under some natural and mild regularity conditions which are given in the Appendix. (We have relegated the regularity conditions and a discussion of their significance to the Appendix in order to avoid making the present section too technical.) Let p be the dimension of h. So $p = 2$ if we take the distribution of the θ_d's to be a symmetric Dirichlet, and $p = K + 1$ if we allow this distribution to be an arbitrary Dirichlet.

Theorem 1 *Let $\psi_1, \psi_2, \ldots$ be generated according to the Augmented Collapsed Gibbs Sampling algorithm described above, let $B_n(h)$ be the estimate on the left side of* (2.1)*, and assume that Conditions A1–A6 in Section A.2 of the Appendix hold. Then:*

1. $\arg\max_h B_n(h) \xrightarrow{\text{a.s.}} \arg\max_h m(h)$.
2. $n^{1/2}\big(\arg\max_h B_n(h) - \arg\max_h m(h)\big) \xrightarrow{d} \mathcal{N}_p(0, \Sigma)$ *for some positive definite matrix* Σ.
3. *Let $\widehat{\Sigma}_n$ be the estimate of Σ obtained by the method of batching described in Section A.2 of the Appendix. 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*

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

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

Remark 1 *The mathematical development in the proof requires a stipulation (Condition A5) which says that the distinguished point h_* is not quite arbitrary: if we specify $\mathcal{H} = [\eta^{(L)}, \eta^{(U)}] \times [\alpha_1^{(L)}, \alpha_1^{(U)}] \times \cdots \times [\alpha_K^{(L)}, \alpha_K^{(U)}]$, then h_* must satisfy*

$$\eta^* < 2\eta^{(L)} \quad \text{and} \quad \alpha_j^* < 2\alpha_j^{(L)},\ j = 1, \ldots, K. \tag{2.3}$$

(Condition (2.3) is replaced by the obvious simpler analogue in the case of symmetric Dirichlets.) Thus, Condition A5 provides guidelines regarding the user-selected value of h_*.

Remark 2 *Part 3 suggests that one should not use arbitrary Dirichlets when K is large. The ellipse is centered at $\arg\max_u B_n(u)$ and the lengths of its principal axes are governed by the term $\chi^2_{p,.95}$. When $p = K+1$ and K is large, $\chi^2_{K+1,.95}$ is on the order of $K+1$, and the confidence set for the empirical Bayes estimate of h is then huge. In other words, when we use an arbitrary Dirichlet, our estimates are very inaccurate.*

Actually, the problem that arises when $\dim(h) = K$ is not limited to our Monte Carlo scheme for estimating $\arg\max_h m(h)$. There is a fundamental problem, which has to do with the fact that there is not enough information in the corpus to estimate a high-dimensional h. Suppose we view the D documents as being drawn from some idealized population generated according to the LDA model indexed by h_0. Leaving aside computational issues, suppose we are able to calculate $\hat{h} = \arg\max_h m(h)$, the maximum likelihood estimate of h_0, to infinite accuracy. Standard asymptotics give $D^{1/2}(\hat{h} - h_0) \xrightarrow{d} \mathcal{N}_p(0, \Omega^{-1})$ as $D \to \infty$, where Ω is the Fisher information matrix. Therefore, a 95% confidence set for h_0 is given by the ellipse $\bigl\{h : (\hat{h} - h)^\top \Omega (\hat{h} - h) \le \chi^2_{K+1,.95}/D\bigr\}$, and we see that for high-dimensional h, D must be very large for us to be able to accurately estimate h_0.

2.2 Estimation of the Family of Posterior Expectations

Let g be a function of $\boldsymbol{\psi}$, and let $I(h) = \int g(\boldsymbol{\psi})\, d\nu_{h,\boldsymbol{w}}(\boldsymbol{\psi})$ be the posterior expectation of $g(\boldsymbol{\psi})$ when the prior is ν_h. Suppose that we are interested in estimating $I(h)$ for all $h \in \mathcal{H}$. Proceeding as we did for estimation of the family of ratios $\{m(h)/m(h_*),\ h \in \mathcal{H}\}$, let $h_* \in \mathcal{H}$ be fixed but arbitrary, and let $\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \ldots$ be an ergodic Markov chain with invariant distribution $\nu_{h*,\boldsymbol{w}}$. To estimate $\int g(\boldsymbol{\psi})\, d\nu_{h,\boldsymbol{w}}(\boldsymbol{\psi})$, the obvious approach is to write

$$\int g(\boldsymbol{\psi})\, d\nu_{h,\boldsymbol{w}}(\boldsymbol{\psi}) = \int g(\boldsymbol{\psi}) \frac{\nu_{h,\boldsymbol{w}}(\boldsymbol{\psi})}{\nu_{h_*,\boldsymbol{w}}(\boldsymbol{\psi})}\, d\nu_{h_*,\boldsymbol{w}}(\boldsymbol{\psi}) \tag{2.4}$$

and then use the importance sampling estimate $(1/n)\sum_{i=1}^n g(\boldsymbol{\psi}_i)[\nu_{h,\boldsymbol{w}}(\boldsymbol{\psi}_i)/\nu_{h_*,\boldsymbol{w}}(\boldsymbol{\psi}_i)]$. This doesn't work because we do not know the normalizing constants for $\nu_{h,\boldsymbol{w}}$ and $\nu_{h_*,\boldsymbol{w}}$. This difficulty is handled by rewriting $\int g(\boldsymbol{\psi})\, d\nu_{h,\boldsymbol{w}}(\boldsymbol{\psi})$, via (2.4), as

$$\int g(\boldsymbol{\psi}) \frac{\ell_{\boldsymbol{w}}(\boldsymbol{\psi})\nu_h(\boldsymbol{\psi})/m(h)}{\ell_{\boldsymbol{w}}(\boldsymbol{\psi})\nu_{h_*}(\boldsymbol{\psi})/m(h_*)}\, d\nu_{h_*,\boldsymbol{w}}(\boldsymbol{\psi}) = \frac{m(h_*)}{m(h)} \int g(\boldsymbol{\psi}) \frac{\nu_h(\boldsymbol{\psi})}{\nu_{h_*}(\boldsymbol{\psi})}\, d\nu_{h_*,\boldsymbol{w}}(\boldsymbol{\psi})$$

$$= \frac{\frac{m(h_*)}{m(h)} \int g(\boldsymbol{\psi}) \frac{\nu_h(\boldsymbol{\psi})}{\nu_{h_*}(\boldsymbol{\psi})}\, d\nu_{h_*,\boldsymbol{w}}(\boldsymbol{\psi})}{\frac{m(h_*)}{m(h)} \int \frac{\nu_h(\boldsymbol{\psi})}{\nu_{h_*}(\boldsymbol{\psi})}\, d\nu_{h_*,\boldsymbol{w}}(\boldsymbol{\psi})} \tag{2.5a}$$

$$= \frac{\int g(\boldsymbol{\psi}) \frac{\nu_h(\boldsymbol{\psi})}{\nu_{h_*}(\boldsymbol{\psi})}\, d\nu_{h_*,\boldsymbol{w}}(\boldsymbol{\psi})}{\int \frac{\nu_h(\boldsymbol{\psi})}{\nu_{h_*}(\boldsymbol{\psi})}\, d\nu_{h_*,\boldsymbol{w}}(\boldsymbol{\psi})}, \tag{2.5b}$$

where in (2.5a) we have used the fact that the integral in the denominator is just 1, in order to cancel the unknown constant $m(h_*)/m(h)$ in (2.5b). The idea to express $\int g(\boldsymbol{\psi})\, d\nu_{h,\boldsymbol{w}}(\boldsymbol{\psi})$ in this way was proposed in a different context by Hastings (1970). Expression (2.5b) is the ratio of two integrals with respect to $\nu_{h_*,\boldsymbol{w}}$, each of which may be estimated from the sequence $\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \ldots, \boldsymbol{\psi}_n$. We may estimate the numerator and the denominator by

$$\frac{1}{n}\sum_{i=1}^{n} g(\boldsymbol{\psi}_i)[\nu_h(\boldsymbol{\psi}_i)/\nu_{h_*}(\boldsymbol{\psi}_i)] \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n}[\nu_h(\boldsymbol{\psi}_i)/\nu_{h_*}(\boldsymbol{\psi}_i)]$$

respectively. Thus, if we let

$$w_i^{(h)} = \frac{\nu_h(\boldsymbol{\psi}_i)/\nu_{h_*}(\boldsymbol{\psi}_i)}{\sum_{e=1}^{n}[\nu_h(\boldsymbol{\psi}_e)/\nu_{h_*}(\boldsymbol{\psi}_e)]},$$

then these are weights, and we see that the desired integral may be estimated by the weighted average

$$\hat{I}(h) = \sum_{i=1}^{n} g(\boldsymbol{\psi}_i) w_i^{(h)}. \tag{2.6}$$

The significance of this development is that it shows that with a single Markov chain run, we can estimate the entire family of posterior expectations $\{I(h),\, h \in \mathcal{H}\}$. As was the case for the estimate on the left side of (2.1), the estimate (2.6) is remarkable in its simplicity. To compute it, we need to know only the ratio of the *priors*, and not the posteriors.

2.3 Serial Tempering

Unfortunately, (2.6) suffers a serious defect: unless h is close to h_*, ν_h can be nearly singular with respect to ν_{h_*} over the region where the $\boldsymbol{\psi}_i$'s are likely to be, resulting in a very unstable estimate. A similar remark applies to the estimate on the left side of (2.1). In other words, there is effectively a "radius" around h_* within which one can safely move. To state the problem more explicitly: there does not exist a single h_* for which the ratios $\nu_h(\boldsymbol{\psi})/\nu_{h_*}(\boldsymbol{\psi})$ have small variance simultaneously 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 $h \in \mathcal{H}$, ν_h is "close to" at least one of $\nu_{h_1}, \ldots, \nu_{h_J}$. We then replace ν_{h_*} in the denominator by $(1/J)\sum_{j=1}^{J} b_j \nu_{h_j}$, 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 ν_h is close to ν_{h_j}, the variance of $\nu_h(\boldsymbol{\psi})/[(1/J)\sum_{j=1}^J b_j \nu_{h_j}(\boldsymbol{\psi})]$ is small; hence the variance of $\nu_h(\boldsymbol{\psi})/[(1/J)\sum_{j=1}^J b_j \nu_{h_j}(\boldsymbol{\psi})]$ is small simultaneously for all $h \in \mathcal{H}$. Whereas for the estimates (2.1) and (2.6) we need a Markov chain with invariant distribution is $\nu_{h*,\boldsymbol{w}}$, in the present situation we need a Markov chain whose invariant distribution is the mixture $(1/J)\sum_{j=1}^J \nu_{h_j,\boldsymbol{w}}$. This approach may be implemented by a methodology called serial tempering (Marinari and Parisi (1992); Geyer and Thompson (1995)), originally developed for the purpose of improving mixing rates of certain Markov chains that are used to simulate physical systems in statistical mechanics. However, it can be used for a very different purpose, namely to increase the range of values over which importance sampling estimates have small variance. We now summarize this methodology, in the present context, and show how it can be used to produce estimates that are stable over a wide range of h values. Our explanations are detailed, because the material is not trivial and because we wish to deal with estimates of both marginal likelihood and posterior expectations. The reader who is not interested in the detailed explanations can skip the rest of this subsection with no loss regarding understanding the rest of the material in this paper, and simply regard serial tempering as a black box that produces estimates of the marginal likelihood (up to a constant) and of posterior expectations (cf. (2.1) and (2.6)) that are stable over a wide h-region.

To simplify the discussion, suppose that in line 2 of the LDA model we take $\boldsymbol{\alpha} = (\alpha, \ldots, \alpha)$, i.e. $\mathrm{Dir}_K(\boldsymbol{\alpha})$ is a symmetric Dirichlet, so that $\mathcal{H}$ is effectively two-dimensional, and suppose that we take $\mathcal{H}$ to be a bounded set of the form $\mathcal{H} = [\eta_L, \eta_U] \times [\alpha_L, \alpha_U]$. Our goal is to generate a Markov chain with invariant distribution $(1/J)\sum_{j=1}^J \nu_{h_j,\boldsymbol{w}}$. The updates will sample different components of this mixture, with jumps from one component to another. We now describe this carefully. Let Ψ denote the state space for $\boldsymbol{\psi}$. Recall that $\boldsymbol{\psi}$ has some continuous components and some discrete components. To proceed rigorously, we will take ν_h and $\nu_{h,\boldsymbol{w}}$ to all be densities with respect to a measure μ on Ψ. Define $\mathcal{L} = \{1, \ldots, J\}$, and for $j \in \mathcal{L}$, suppose that Φ_j is a Markov transition function on Ψ with invariant distribution equal to the posterior $\nu_{h_j,\boldsymbol{w}}$. On occasion we will write ν_j instead of ν_{h_j}. This notation is somewhat inconsistent, but we use it in order to avoid having double and triple subscripts. We have $\nu_{h,\boldsymbol{w}} = \ell_{\boldsymbol{w}}\,\nu_h/m(h)$ and $\nu_{h_j,\boldsymbol{w}} = \ell_{\boldsymbol{w}}\,\nu_j/m(h_j)$, $j = 1, \ldots, J$.

Serial tempering involves considering the state space $\mathcal{L} \times \Psi$, and forming the family of distributions $\{P_\zeta,\, \zeta \in \mathbb{R}^J\}$ on $\mathcal{L} \times \Psi$ with densities

$$p_\zeta(j, \boldsymbol{\psi}) \propto \ell_{\boldsymbol{w}}(\boldsymbol{\psi})\nu_j(\boldsymbol{\psi})/\zeta_j. \tag{2.7}$$

(To be pedantic, these are densities with respect to $\mu \times \sigma$, where σ is counting measure on $\mathcal{L}$.) The vector ζ is a tuning parameter, which we discuss shortly. For any value of ζ, by standard methods involving the Metropolis-Hastings algorithm, we can generate a Markov chain having invariant distribution equal to (2.7). If we take $\zeta_j = am(h_j)$ for $j = 1, \ldots, J$, where a is an arbitrary constant, then the $\boldsymbol{\psi}$-marginal of p_ζ is exactly $(1/J)\sum_{j=1}^J \nu_{h_j,\boldsymbol{w}}$, so we can generate a Markov chain with the desired invariant distribution. Unfortunately, the values $m(h_1), \ldots, m(h_J)$ are unknown (our objective is precisely to estimate them). It will turn out that for any value of ζ, a Markov chain with invariant distribution (2.7) enables us to estimate the vector $(m(h_1), \ldots, m(h_j))$ up to a constant, and the closer ζ is to a constant multiple of $(m(h_1), \dots, m(h_j))$, the better is our estimate. This gives rise to a natural iterative procedure for estimating $(m(h_1), \dots, m(h_j))$. We now give the details.

Let $\Gamma(j, \cdot)$ be a Markov transition function on $\mathcal{L}$. In our context, we would typically take $\Gamma(j, \cdot)$ to be the uniform distribution on $\mathcal{N}_j$, where $\mathcal{N}_j$ is a set consisting of the indices of the h_l's which are close to h_j. Serial tempering is a Markov chain on $\mathcal{L} \times \Psi$ which can be viewed as a two-block Metropolis-Hastings (i.e. Metropolis-within-Gibbs) algorithm, and is run as follows. Suppose that the current state of the chain is $(L_{i-1}, \boldsymbol{\psi}_{i-1})$.

- A new value $j \sim \Gamma(L_{i-1}, \cdot)$ is proposed. We set $L_i = j$ with the Metropolis probability

$$\rho = \min\left\{1, \frac{\Gamma(j, L_{i-1})}{\Gamma(L_{i-1}, j)} \frac{\nu_j(\boldsymbol{\psi}_{i-1})/\zeta_j}{\nu_{L_{i-1}}(\boldsymbol{\psi}_{i-1})/\zeta_{L_{i-1}}}\right\}, \tag{2.8}$$

  and with the remaining probability we set $L_i = L_{i-1}$.

- Generate $\boldsymbol{\psi}_i \sim \Phi_{L_i}(\boldsymbol{\psi}_{i-1}, \cdot)$.

By standard arguments, the density (2.7) is an invariant density for the serial tempering chain. A key observation is that the $\boldsymbol{\psi}$-marginal density of p_ζ is

$$f_\zeta(\boldsymbol{\psi}) = (1/c_\zeta) \sum_{j=1}^J \ell_{\boldsymbol{w}}(\boldsymbol{\psi}) \nu_j(\boldsymbol{\psi})/\zeta_j, \qquad \text{where} \qquad c_\zeta = \sum_{j=1}^J m(h_j)/\zeta_j. \tag{2.9}$$

Suppose that $(L_1, \boldsymbol{\psi}_1), (L_2, \boldsymbol{\psi}_2), \dots$ is a serial tempering chain. To estimate $m(h)$, consider

$$\widehat{M}_\zeta(h) = \frac{1}{n} \sum_{i=1}^n \frac{\nu_h(\boldsymbol{\psi}_i)}{(1/J) \sum_{j=1}^J \nu_j(\boldsymbol{\psi}_i)/\zeta_j}. \tag{2.10}$$

Note that this estimate depends only on the $\boldsymbol{\psi}$-part of the chain. Assuming that we have established that the chain is ergodic, we have

$$\begin{aligned} \widehat{M}_\zeta(h) &\xrightarrow{\text{a.s.}} \int \frac{\nu_h(\boldsymbol{\psi})}{(1/J) \sum_{j=1}^J \nu_j(\boldsymbol{\psi})/\zeta_j} \frac{\sum_{j=1}^J \ell_{\boldsymbol{w}}(\boldsymbol{\psi}) \nu_j(\boldsymbol{\psi})/\zeta_j}{c_\zeta} \, d\mu(\boldsymbol{\psi}) \\ &= \int \frac{\ell_{\boldsymbol{w}}(\boldsymbol{\psi}) \nu_h(\boldsymbol{\psi})}{c_\zeta/J} \, d\mu(\boldsymbol{\psi}) \\ &= \frac{m(h)}{c_\zeta/J}. \end{aligned} \tag{2.11}$$

This means that for any ζ, the family $\{\widehat{M}_\zeta(h),\, h \in \mathcal{H}\}$ can be used to estimate the family $\{m(h),\, h \in \mathcal{H}\}$, up to a single multiplicative constant.

To estimate the family of integrals $\left\{\int g(\boldsymbol{\psi}) \, d\nu_{h,\boldsymbol{w}}(\boldsymbol{\psi}),\, h \in \mathcal{H}\right\}$, we proceed as follows. Let

$$\widehat{U}_\zeta(h) = \frac{1}{n} \sum_{i=1}^n \frac{g(\boldsymbol{\psi}_i) \nu_h(\boldsymbol{\psi}_i)}{(1/J) \sum_{j=1}^J \nu_j(\boldsymbol{\psi}_i)/\zeta_j}. \tag{2.12}$$

By ergodicity we have

$$\begin{aligned}
\widehat{U}_\zeta(h) &\xrightarrow{\text{a.s.}} \int \frac{g(\boldsymbol{\psi})\nu_h(\boldsymbol{\psi})}{(1/J)\sum_{j=1}^J \nu_j(\boldsymbol{\psi})/\zeta_j} \frac{\sum_{j=1}^J \ell_{\boldsymbol{w}}(\boldsymbol{\psi})\nu_j(\boldsymbol{\psi})/\zeta_j}{c_\zeta}\, d\mu(\boldsymbol{\psi}) \\
&= \int \frac{\ell_{\boldsymbol{w}}(\boldsymbol{\psi})g(\boldsymbol{\psi})\nu_h(\boldsymbol{\psi})}{c_\zeta/J}\, d\mu(\boldsymbol{\psi}) \qquad (2.13) \\
&= \frac{m(h)}{c_\zeta/J} \int g(\boldsymbol{\psi})\, d\nu_{h,\boldsymbol{w}}(\boldsymbol{\psi}).
\end{aligned}$$

Combining the convergence statements (2.13) and (2.11), we see that

$$\hat{I}^{\text{st}}_\zeta(h) := \frac{\widehat{U}_\zeta(h)}{\widehat{M}_\zeta(h)} \xrightarrow{\text{a.s.}} \int g(\boldsymbol{\psi})\, d\nu_{h,\boldsymbol{w}}(\boldsymbol{\psi}). \qquad (2.14)$$

Suppose that for some constant a, we have

$$(\zeta_1, \dots, \zeta_J) = a(m(h_1), \dots, m(h_J)). \qquad (2.15)$$

Then $c_\zeta = J/a$, and as noted earlier, $f_\zeta(\boldsymbol{\psi}) = (1/J)\sum_{j=1}^J \nu_{h_j,\boldsymbol{w}}(\boldsymbol{\psi})$, i.e. the ψ-marginal of p_ζ (see (2.9)) gives equal weight to each of the component distributions in the mixture. (Expressing this slightly differently, if (2.15) is true, then the invariant density (2.7) becomes $p_\zeta(j, \boldsymbol{\psi}) = (1/J)\nu_{h_j,\boldsymbol{w}}(\boldsymbol{\psi})$, so the L-marginal distribution of p_ζ gives mass $(1/J)$ to each point in $\mathcal{L}$.) Therefore, for large n, the proportions of time spent in the J components of the mixture are about the same, a feature which is essential if serial tempering is to work well. In practice, we cannot arrange for (2.15) to be true, because $m(h_1), \dots, m(h_J)$ are unknown. However, the vector $(m(h_1), \dots, m(h_J))$ may be estimated (up to a multiplicative constant) iteratively as follows. If the current value is $\zeta^{(t)}$, then set

$$\big(\zeta_1^{(t+1)}, \dots, \zeta_J^{(t+1)}\big) = \big(\widehat{M}_{\zeta^{(t)}}(h_1), \dots, \widehat{M}_{\zeta^{(t)}}(h_J)\big). \qquad (2.16)$$

From the convergence result (2.11), we get $\widehat{M}_{\zeta^{(t)}}(h_j) \xrightarrow{\text{a.s.}} m(h_j)/a_{\zeta^{(t)}}$, where $a_{\zeta^{(t)}}$ is a constant, i.e. (2.15) is nearly satisfied by $\big(\zeta_1^{(t+1)}, \dots, \zeta_J^{(t+1)}\big)$. To determine the number of iterations needed, at each iteration we record the proportions of time spent in the J different components of the mixture, i.e. the vector $\big((1/n)\sum_{i=1}^n I(L_i = 1), \dots, (1/n)\sum_{i=1}^n I(L_i = J)\big)$, and we stop the iteration when this vector is nearly uniform. In all our examples, three or four iterations were sufficient. Pseudocode is given in Algorithm 1.

To sum up, we estimate the family of marginal likelihoods (up to a constant) and the family of posterior expectations as follows. First, we obtain the vector of tuning parameters ζ via the iterative scheme given by (2.16). To estimate the family of marginal likelihoods (up to a constant) we use $\widehat{M}_\zeta(h)$ defined in (2.10), and to estimate the family of posterior expectations we use $\hat{I}^{\text{st}}_\zeta(h) = \widehat{U}_\zeta(h)/\widehat{M}_\zeta(h)$ (see (2.12) and (2.10)).

We point out that it is possible to estimate the family of marginal likelihoods (up to a constant) by

$$\widetilde{M}_\zeta(h) = \frac{1}{n}\sum_{i=1}^n \frac{\nu_h(\boldsymbol{\psi}_i)}{\nu_{L_i}(\boldsymbol{\psi}_i)/\zeta_{L_i}}. \qquad (2.17)$$

Algorithm 1: Serial tempering. See the discussion in Sec. 2.3 and Appendices A.1 and 4.

Data: Observed words $\boldsymbol{w}$
Result: A Markov chain on $\mathcal{L} \times \Psi$

specify $h_1, \ldots, h_J \in \mathcal{H}$;
initialize $\zeta_1^{(1)}, \ldots, \zeta_J^{(1)}$;
initialize $\boldsymbol{\psi}_0 = (\boldsymbol{\beta}^{(0)}, \boldsymbol{\theta}^{(0)}, \boldsymbol{z}^{(0)})$, L_0;
compute count statistics n_{dk} and m_{dkv}, for $d = 1, \ldots, D$, $k = 1, \ldots, K$, $v = 1, \ldots, V$;
for *tuning iteration* $t = 1, \ldots$ **do**
    for *MCMC iteration* $i = 1, \ldots$ **do**
        `// The Metropolis-Hastings update`
        `// Set` L_i `via the probability` ρ `given by (2.8)`
        propose index $j \sim \Gamma(L_{i-1}, \cdot)$;
        sample $U \sim \text{Uniform}(0, 1)$;
        if $U < \rho$ **then**
            set $L_i = j$;
        else
            set $L_i = L_{i-1}$;
        `// Generate` $\boldsymbol{\psi}_i = (\boldsymbol{\beta}^{(i)}, \boldsymbol{\theta}^{(i)}, \boldsymbol{z}^{(i)}) \sim \Phi_{L_i}(\boldsymbol{\psi}_{i-1}, \cdot)$
        for *document* $d = 1, \ldots, D$ **do**
            for *word* w_{dr}, $r = 1, \ldots, n_d$ **do**
                sample topic index $z_{dr}^{(i)}$ via the CGS (Griffiths and Steyvers, 2004);
                update count statistics n_{dk} and m_{dkv} according to $z_{dr}^{(i)}$ and w_{dr};
        for *topic* $k = 1, \ldots, K$ **do**
            sample topic $\beta_k^{(i)}$ via (4.5);
        for *document* $d = 1, \ldots, D$ **do**
            sample the distribution on topics $\theta_d^{(i)}$ via (4.5);
    `// Update tuning parameters` $\zeta_1, \ldots, \zeta_J$
    compute the estimates $\widehat{M}_{\zeta^{(t)}}(h_1), \ldots, \widehat{M}_{\zeta^{(t)}}(h_J)$ via (2.10) using $\boldsymbol{\psi}_i$ and $\zeta_1^{(t)}, \ldots, \zeta_J^{(t)}$;
    set $\big(\zeta_1^{(t+1)}, \ldots, \zeta_J^{(t+1)}\big) = \big(\widehat{M}_{\zeta^{(t)}}(h_1), \ldots, \widehat{M}_{\zeta^{(t)}}(h_J)\big)$;

Note that $\widetilde{M}_\zeta(h)$ uses the sequence of pairs $(L_1, \boldsymbol{\psi}_1), (L_2, \boldsymbol{\psi}_2), \ldots$, and not just the sequence $\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \ldots$. To see why (2.17) is a valid estimator, observe that by ergodicity we have

$$\begin{aligned}
\widetilde{M}_\zeta(h) &\xrightarrow{\text{a.s.}} \iint \frac{\nu_h(\boldsymbol{\psi})}{\nu_L(\boldsymbol{\psi})/\zeta_L} \cdot \left[\frac{1}{c_\zeta} \ell_{\boldsymbol{w}}(\boldsymbol{\psi}) \nu_L(\boldsymbol{\psi})/\zeta_L\right] d\mu(\boldsymbol{\psi})\, d\sigma(L) \\
&= \iint \frac{m(h)}{c_\zeta} \nu_{h,\boldsymbol{w}}(\boldsymbol{\psi})\, d\mu(\boldsymbol{\psi})\, d\sigma(L) \\
&= J \frac{m(h)}{c_\zeta}.
\end{aligned} \tag{2.18}$$

(Note that the limit in (2.18) is the same as the limit in (2.11).) Similarly, we may estimate the integral $\int g(\boldsymbol{\psi})\, d\nu_{h,\boldsymbol{w}}(\boldsymbol{\psi})$ by the ratio

$$\tilde{I}^{\text{st}}_{\zeta}(h) = \sum_{i=1}^{n} \frac{g(\boldsymbol{\psi}_i)\nu_h(\boldsymbol{\psi}_i)}{\nu_{L_i}(\boldsymbol{\psi}_i)/\zeta_{L_i}} \Bigg/ \sum_{i=1}^{n} \frac{\nu_h(\boldsymbol{\psi}_i)}{\nu_{L_i}(\boldsymbol{\psi}_i)/\zeta_{L_i}}.$$

The estimate $\tilde{I}^{\text{st}}_{\zeta}(h)$ is also based on the pairs $(L_1, \boldsymbol{\psi}_1), (L_2, \boldsymbol{\psi}_2), \ldots$, and it is easy to show that $\tilde{I}^{\text{st}}_{\zeta}(h) \xrightarrow{\text{a.s.}} \int g(\boldsymbol{\psi})\, d\nu_{h,\boldsymbol{w}}(\boldsymbol{\psi})$.

The estimates $\widetilde{M}_{\zeta}(h)$ and $\tilde{I}^{\text{st}}_{\zeta}(h)$ are the ones that are used by Marinari and Parisi (1992) and Geyer and Thompson (1995), but $\widehat{M}_{\zeta}(h)$ and $\hat{I}^{\text{st}}_{\zeta}(h)$ appear to significantly outperform $\widetilde{M}_{\zeta}(h)$ and $\tilde{I}^{\text{st}}_{\zeta}(h)$ in terms of accuracy. We demonstrate this in Section 2.4.

Remark 3 *Theorem 1 continues to be true when we use the serial tempering chain, as opposed to the simple ACGS. The needed changes are that in the statement of the theorem B_n is replaced with $\widehat{M}_{\zeta}$, and Condition A5 is replaced by the following. If $h^{(1)}, \ldots, h^{(J)}$ are the grid points used in running the serial tempering chain, then the stipulation on h_* given by* (2.3) *is satisfied by $h^{(j)}$ for at least one index j. See the proof of Theorem 1 in the Appendix.*

Globally-Valid Confidence Bands for $\{I(h),\, h \in \mathcal{H}\}$ Based on Serial Tempering Here we explain how to form confidence bands for the family $\{I(h),\, h \in \mathcal{H}\}$ based on $\{\hat{I}^{\text{st}}_{\zeta}(h),\, h \in \mathcal{H}\}$. Our arguments are informal, and we focus primarily on the algorithm for constructing the bands. The proof that the method works is given in Section A.3 of the Appendix. We will write $\hat{I}$ instead of $\hat{I}^{\text{st}}_{\zeta}$ to lighten the notation. Suppose that $\sup_{h\in\mathcal{H}} n^{1/2}|\hat{I}(h) - I(h)|$ has a limiting distribution as $n \to \infty$ (in the Appendix we explain why such a result is true), and suppose that we know the .95 quantile of this distribution, i.e. we know the value $c_{.95}$ such that

$$P\Bigl(\sup_{h\in\mathcal{H}} n^{1/2}|\hat{I}(h) - I(h)| \le c_{.95}\Bigr) = .95. \tag{2.19}$$

In this case we may rewrite (2.19) as

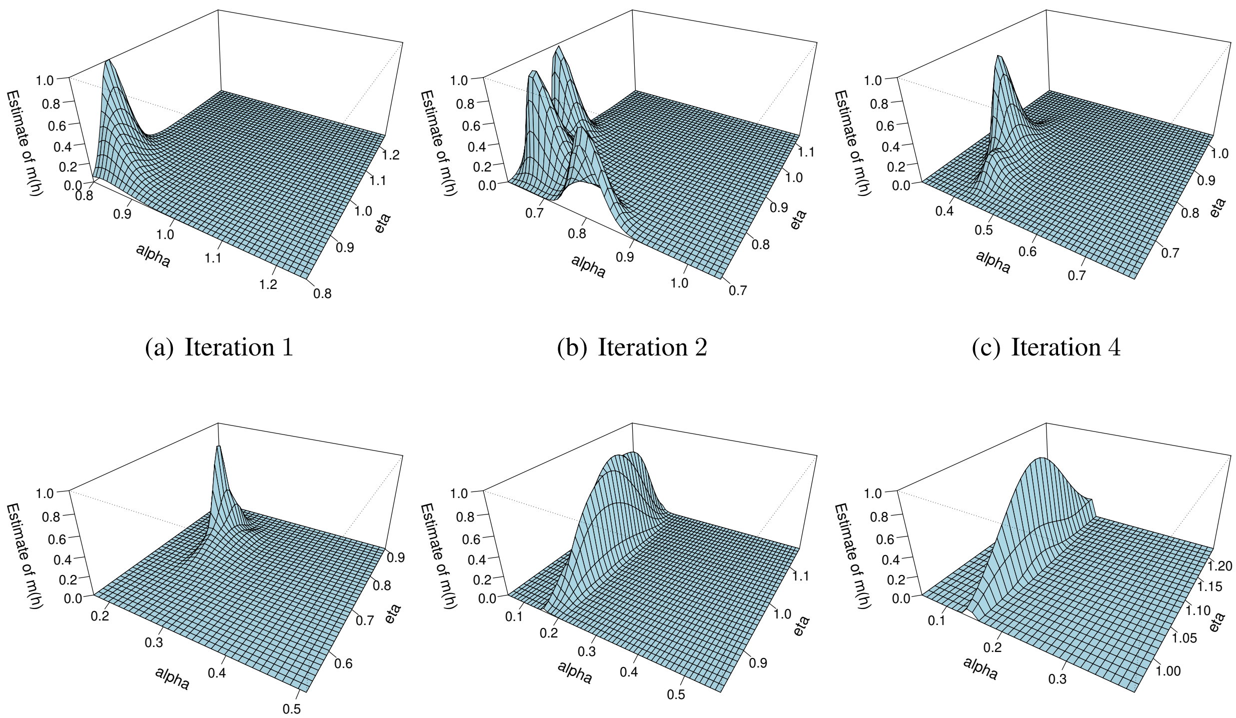
$$P\Bigl(\hat{I}(h) - c_{.95}/n^{1/2} \le I(h) \le \hat{I}(h) + c_{.95}/n^{1/2} \text{ for all } h \in \mathcal{H}\Bigr) = .95,$$

meaning that the band $\hat{I}(h) \pm c_{.95}/n^{1/2}$ is a globally-valid confidence band for $\{I(h),\, h \in \mathcal{H}\}$. (In contrast, for a pointwise band $(L(h), U(h)),\ h \in \mathcal{H}$, we can only make the statement $P\bigl(L(h) \le I(h) \le U(h)\bigr) = .95$ for each $h \in \mathcal{H}$, and we cannot make any statement regarding simultaneous coverage.)

The difficulty is in obtaining $c_{.95}$, and we now show how this quantity can be estimated through the method of batching, which is described as follows. The sequence $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_n$ is broken up into J consecutive pieces of equal lengths called batches. For $j = 1, \ldots, J$, let $\hat{I}_j(h)$ be the estimate of $I(h)$ produced by batch j. Now the $\hat{I}_j(h)$'s are each formed from a sample of size n/J. Informally, if n is large and n/J is also large, then for $j = 1, \ldots, J$, $\sup_{h\in\mathcal{H}} (n/J)^{1/2}|\hat{I}_j(h) - I(h)|$ and $\sup_{h\in\mathcal{H}} n^{1/2}|\hat{I}(h) - I(h)|$ have approximately the same distribution. Therefore, to estimate $c_{.95}$, we let $S_j = \sup_{h\in\mathcal{H}} (n/J)^{1/2}|\hat{I}_j(h) - I(h)|$, and as our estimate of $c_{.95}$ we use the 95th percentile of the sequence $S_1, \dots, S_J$. Unfortunately, the S_j's are not available, because they involve $I(h)$, which is unknown. So instead we use $\mathcal{S}_j = \sup_{h \in \mathcal{H}} (n/J)^{1/2} |\hat{I}_j(h) - \hat{I}(h)|$, in which we have substituted $\hat{I}(h)$ for $I(h)$. To conclude, let $\mathcal{S}_{[1]} \leq \mathcal{S}_{[2]} \leq \cdots \leq \mathcal{S}_{[J]}$ denote the ordered values of the sequence $\mathcal{S}_1, \dots, \mathcal{S}_J$. We estimate $c_{.95}$ via $\mathcal{S}_{[.95J]}$, and our 95% globally-valid confidence band for $\{I(h),\, h \in \mathcal{H}\}$ is $\{\hat{I}(h) \pm \mathcal{S}_{[.95J]}/n^{1/2},\, h \in \mathcal{H}\}$. In the Appendix we show that the probability that the entire function $\{I(h),\, h \in \mathcal{H}\}$ lies inside the band converges to .95 as $n \to \infty$. There are conditions on J: we need that $J \to \infty$ and $n/J \to \infty$; a good choice is $J = n^{1/2}$. The Markov chain length n should be chosen such that the band is acceptably narrow.

Iterative Scheme for Choosing the Grid The performance of serial tempering depends crucially on the choice of grid points $h_1, \dots, h_J$, and it is essential that $\arg\max_h m(h)$ be close to at least one of the grid points, for the reason discussed at the beginning of this section. This creates a circular problem: the ideal grid is one that is centered or nearly centered at $\arg\max_h m(h)$, but $\arg\max_h m(h)$ is unknown. The problem is compounded by the fact that the grid has to be "tight," i.e. the points $h_1, \dots, h_J$ need to be close together. This is because when the corpus is large, if h_j and $h_{j'}$ are not close, then for $j \neq j'$, ν_{h_j} and $\nu_{h_{j'}}$ are nearly singular (each is a product of a large number of terms—see (2.2)). In the serial tempering chain, this near singularity causes the proposal $j \sim \Gamma(L_{i-1}, \cdot)$ (see (2.8)) to have high probability of being rejected, and the chain does not mix well. To deal with this problem, we use an iterative scheme which proceeds as follows. We initialize the experiment with a fixed $h^{(0)}$ (for example $h^{(0)} = (1, 1)$) and a subgrid that "covers" $h^{(0)}$ (for example a subgrid with convex hull equal to $[1/2, 2] \times [1/2, 2]$). We then subsample a small set of documents from the corpus and run the serial tempering chain to find the estimate of the maximizer of the marginal likelihood for the subsampled corpus, using the current grid setting. We iterate: at iteration t, we set $h^{(t)}$ to be the estimate of the maximizer obtained from the previous iteration, and select a subgrid that covers $h^{(t)}$. As the iteration number t increases, the grid is made more narrow, and the number of subsampled documents is increased. This scheme works because in the early iterations the number of documents is small, so the near-singularity problem does not arise, and we can use a wide grid. In our experience, decreasing the dimensions of the α- and η-grids by 10% and increasing the number of subsampled documents by 10% at each iteration works well. It is very interesting to note that convergence may occur before the subsample size is equal to the number of documents in the corpus, in which case there is no need to ever deal with the entire corpus, and in fact this is typically what happens, unless the corpus is small. (By "convergence" we mean that $h^{(t)}$ is nearly the same as the values from the previous iterations.) Of course, for small corpora the near-singularity problem does not arise, and the iterative scheme can be skipped entirely.

To illustrate the scheme, we generated a corpus according to the LDA model with $D = 10^5$, $K = 50$, $V = 500$, $n_d = 80$ for all d, and $h_{\text{true}} = (\eta, \alpha) = (.8, .2)$, and ran the scheme using Markov chains of length $n = 50{,}000$ and grids of size $J = 100$. As will be clear shortly, our results would have been identical if D had been *any* number bigger than 10^5. Figure 2 shows the marginal likelihood surfaces as the iterations progress. At iteration 1, the α-value of the maximizer is outside the convex hull of the grid, and at the second iteration, the grid is centered at that point. Figure 3 gives precise information on the number of subsampled documents (left panel), and the lower and upper endpoints of the α- and η-values used in the grids, as the iterations progress (right panel). The right panel also gives α- and η-values of the estimate of the argmax as the iterations progress. As can be seen from Figure 3, the scheme has effectively converged after about 18 iterations, and at convergence the number of subsampled documents is only 200.



(a) Iteration 1 (b) Iteration 2 (c) Iteration 4

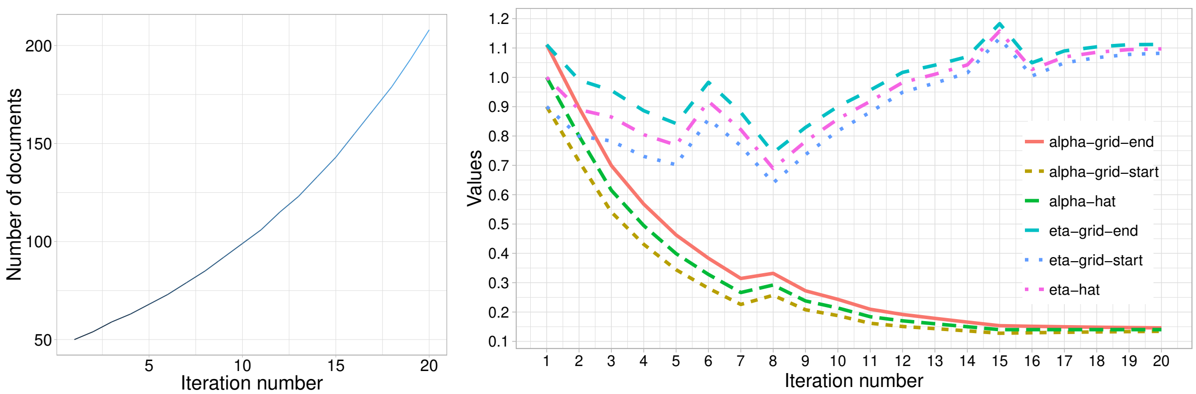
(d) Iteration 8 (e) Iteration 12 (f) Iteration 18

Figure 2: Values of $\widehat{M}(h)$ for iterations 1, 2, 4, 8, 12, 18 using a synthetic corpus generated according to the LDA model with $K = 20$, $n_d = 100$ for each d, $V = 100$, and $h_{\text{true}} = (.8, .2)$.

Serial tempering is a method for enhancing the simple estimator (2.1) which works well when $\dim(h)$ is low. The method does not scale well when $\dim(h)$ increases. In Section 6 we discuss this issue and present an idea on a different way to enhance (2.1) when h is high dimensional.

2.4 Illustration on a Wikipedia Corpus

In Section 1 we mentioned that the hyperparameter h has a strong effect on the prior distribution of the parameters in the model. Here we show empirically that it has a strong impact on the posterior distribution, and hence on inference based on this posterior distribution. To this end, we considered a corpus of articles from Wikipedia, constructed as follows. When a Wikipedia article is created, it is typically tagged to one or more categories, one of which is the "primary category." The corpus consists of 8 documents from the category *Leopardus*, 8 from the category *Lynx*, and 7 from *Prionailurus*. There are 303 words in the vocabulary, and the total number of words in the corpus is 7788. We took $K = 3$, so implicitly we envisage a topic being induced by each of the three categories. The corpus is quite small, but it is challenging to analyze because the topics



(a) Number of documents (b) Grid configurations and α- and η-values of estimate of the argmax

Figure 3: Iterations in the serial tempering scheme used on the synthetic corpus in Figure 2: left panel gives the number of documents subsampled at each iteration; right panel gives the specifications for the grid at each iteration.

are very close to each other, so in the posterior distribution there is a great deal of uncertainty regarding the latent topic indicator variables, and this is why we chose this data set. (In our analysis of this corpus, we treat the articles as unlabeled, i.e. we act as if for each article we don't know the category from which the article is taken.) As mentioned in Section 1, two quantities of interest are the posterior probability that the topic indicator variables for documents i and j are close, i.e. $\nu_{h,\boldsymbol{w}}(\|\theta_i - \theta_j\| \leq \epsilon)$, and the posterior expectation of the distance between topics i and j, which is given by the integral $\int \|\beta_i - \beta_j\| \, d\nu_{h,\boldsymbol{w}}(\boldsymbol{\psi})$. Figure 4 gives plots of estimates of these posterior probabilities and expectations, as h varies, together with 95% globally-valid confidence sets. The plots clearly show that these posterior probabilities and expectations vary considerably with h.

Each plot was constructed from a serial tempering chain, using the methodology described in Section 2.3. Details regarding the chain and the plots are as follows. We took the sequence $h_1, \ldots, h_J$ to consist of an 11×20 grid of 220 evenly-spaced values over the region $(\eta, \alpha) \in [.6, 1.1] \times [.15, 1.1]$. For each hyperparameter value h_j ($j = 1, \ldots, 220$), we took Φ_j to be the Markov transition function of the Augmented Collapsed Gibbs Sampler alluded to earlier and described in detail in Section 4 (in all our experiments we used the Augmented Collapsed Gibbs Sampler, but the Grouped Gibbs Sampler gives results which are very similar). We took the Markov transition function $K(j, \cdot)$ on $\mathcal{L} = \{1, \ldots, 220\}$ to be the uniform distribution on $\mathcal{N}_j$ where $\mathcal{N}_j$ is the subset of $\mathcal{L}$ consisting of the indices of the h_l's that are neighbors of the point h_j. (An interior point has eight neighbors, an edge point has five, and a corner point has three.)[1]

In Section 2.3, we stated that $\widehat{M}_\zeta(h)$ and $\hat{I}^{\text{st}}_\zeta(h)$ appear to significantly outperform $\widetilde{M}_\zeta(h)$ and $\tilde{I}^{\text{st}}_\zeta(h)$ in terms of accuracy. We now provide some evidence for this, and we will deal with the estimates of $I(h)$ (a comparison of $\widehat{M}_\zeta(h)$ and $\widetilde{M}_\zeta(h)$ is given in George (2015)). We considered the Wikipedia Cats corpus described above, and we took $I(h) = \nu_{h,\boldsymbol{w}}(\|\theta_7 - \theta_8\| \leq .07)$. We

[1]Software for implementation of our algorithms as well as datasets we use are available as an R package at `https://github.com/clintpgeorge/ldamcmc`

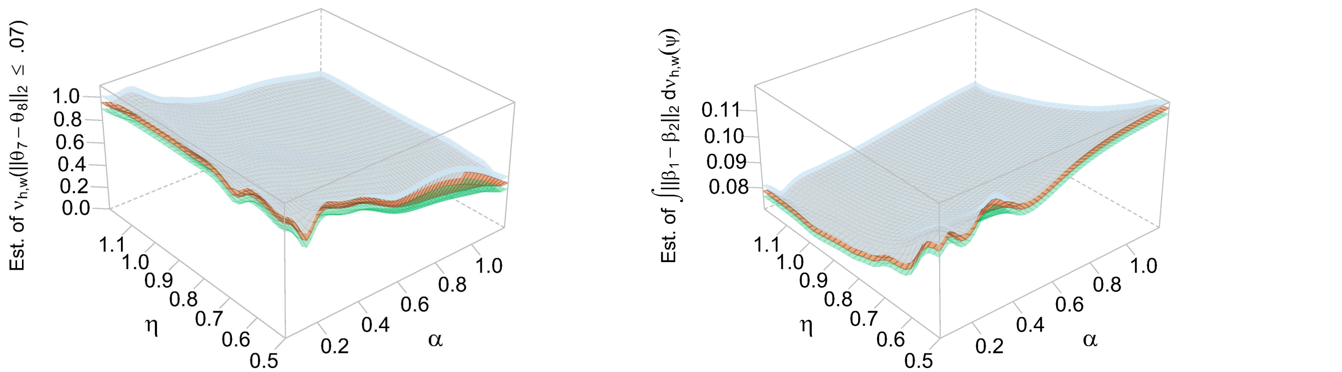


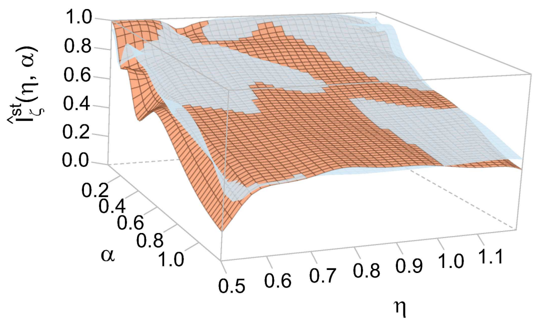
Figure 4: Variability of posterior probabilities and expectations for the Cats corpus from Wikipedia. Left panel: estimate of the posterior probability that documents 7 and 8 have essentially the same topics, in the sense that $\|\theta_7 - \theta_8\| \le .07$, as h varies. Right panel: estimate of the posterior expectation of the (Euclidean, i.e. L_2) distance between topics 1 and 2 as h varies.

calculated $\hat{I}_\zeta^{\text{st}}(h)$ twice, using two different seeds, and also calculated $\tilde{I}_\zeta^{\text{st}}(h)$ twice, using two different seeds, in every case using the same h-range that was used in Figure 4. The four surfaces were constructed via four independent serial tempering experiments, each involving two iterations (each of length $50{,}000$ after a short burn-in period) to form the tuning parameter ζ, which was given initial value $\zeta^{(0)} = \big(\zeta_1^{(0)}, \ldots, \zeta_{220}^{(0)}\big) = (1, \ldots, 1)$, and one final iteration (of length $100{,}000$) to form the estimate of $I(h)$. Figure 5(a) shows the two estimates $\hat{I}_\zeta^{\text{st}}(h)$, and Figure 5(b) shows the two estimates $\tilde{I}_\zeta^{\text{st}}(h)$. The figures show that the two independent estimates $\hat{I}_\zeta^{\text{st}}(h)$ are close to each other, whereas the two independent estimates $\tilde{I}_\zeta^{\text{st}}(h)$ are not.

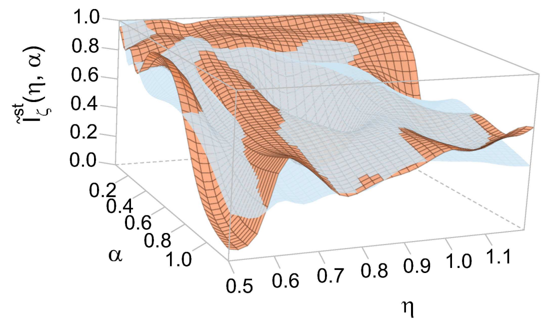
Although the variability of $\hat{I}_\zeta^{\text{st}}(h)$ is significantly smaller than that of $\tilde{I}_\zeta^{\text{st}}(h)$, the figures perhaps don't show this very clearly because a visual comparison of two surfaces is not easy. Therefore, we extracted two one-dimensional slices from each panel in Figure 5, which we used to create Figure 6. The figure shows the values of the two versions of $\hat{I}_\zeta^{\text{st}}(\eta, \alpha)$ and the two versions of $\tilde{I}_\zeta^{\text{st}}(\eta, \alpha)$ when η is fixed at .70 (two left panels); and it shows these plots when η is fixed at 1.00 (two right panels). The superiority of $\hat{I}_\zeta^{\text{st}}$ over $\tilde{I}_\zeta^{\text{st}}$ is striking. We mention that, ostensibly, $\widehat{M_\zeta}(h)$ and $\hat{I}_\zeta^{\text{st}}(h)$ require more computation, but the quantities $(1/J) \sum_{j=1}^J \nu_j(\boldsymbol{\psi}_i)/\zeta_j$, $i = 1, \ldots, n$ are calculated once, and stored. Doing this essentially eliminates the increased computing cost.

3 Empirical Assessment of the Estimator of the Argmax

Consider the LDA model with a given hyperparameter value, which we will denote by h_{true}, and suppose we carry out steps 1–4 of the model, where in the final step we generate the corpus $\boldsymbol{w}$. The maximum likelihood estimate of h is $\hat{h} = \arg\max_h m(h)$ and, as we mentioned earlier, for any constant a, known or unknown, $\arg\max_h m(h) = \arg\max_h am(h)$. As noted earlier, the

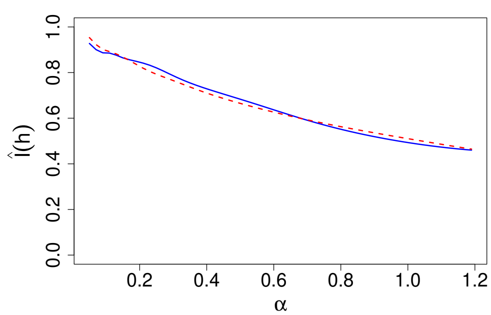


(a) $\hat{I}^{\text{st}}_{\zeta}(\eta, \alpha)$

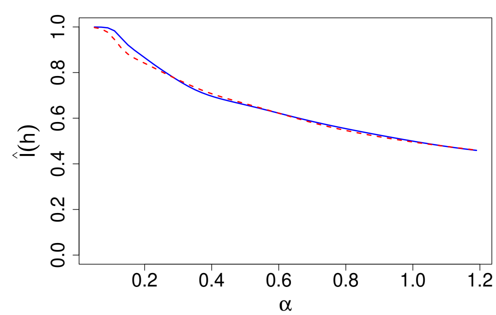


(b) $\tilde{I}^{\text{st}}_{\zeta}(\eta, \alpha)$

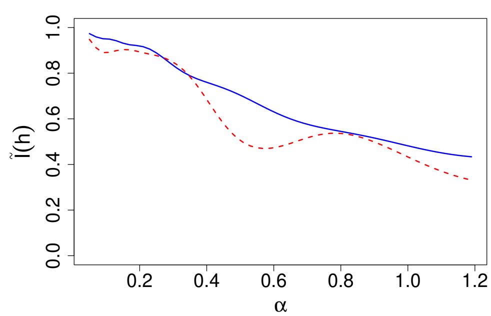
Figure 5: Comparison of the variability of $\hat{I}^{\text{st}}_{\zeta}$ and $\tilde{I}^{\text{st}}_{\zeta}$. Left panel shows two independent estimates of $I(\eta, \alpha) = \nu_{h,\boldsymbol{w}}(\|\theta_7 - \theta_8\| \leq .07)$ using $\hat{I}^{\text{st}}_{\zeta}(\eta, \alpha)$. Right panel uses $\tilde{I}^{\text{st}}_{\zeta}$ instead of $\hat{I}^{\text{st}}_{\zeta}$.



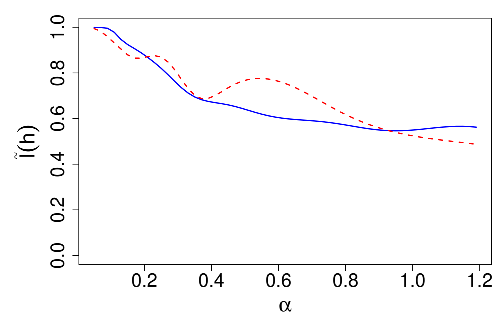
(a) $\hat{I}^{\text{st}}_{\zeta}(.70, \alpha)$



(b) $\hat{I}^{\text{st}}_{\zeta}(1.0, \alpha)$



(c) $\tilde{I}^{\text{st}}_{\zeta}(.70, \alpha)$



(d) $\tilde{I}^{\text{st}}_{\zeta}(1.0, \alpha)$

Figure 6: Two one-dimensional views of the plots in Figure 5. Each of the top two panels shows two independent estimates of $I(\eta, \alpha)$, using $\hat{I}^{\text{st}}_{\zeta}(\eta, \alpha)$. For the left panel, $\eta = 0.70$, and for the right panel, $\eta = 1.00$. The bottom two panels use $\tilde{I}^{\text{st}}_{\zeta}$ instead of $\hat{I}^{\text{st}}_{\zeta}$. The plots show that the variability of $\hat{I}^{\text{st}}_{\zeta}$ is much smaller than that of $\tilde{I}^{\text{st}}_{\zeta}$.

family $\{\widehat{M_\zeta}(h),\, h \in \mathcal{H}\}$, where $\widehat{M_\zeta}(h)$ is given by (2.10), may be used to estimate the family $\{m(h),\, h \in \mathcal{H}\}$ up to a multiplicative constant. So we may use $\arg\max_h \widehat{M_\zeta}(h)$ to estimate $\hat{h}$.

Recall that $B_n(h)$ is the estimate of $m(h)/m(h_*)$ given by the left side of equation (2.1). In theory, $\arg\max_h B_n(h)$ can also be used. However, as we pointed out earlier, $B_n(h)$ is stable only for h close to h_*—a similar remark applies to $\hat{I}(h)$—and unless the region of hyperparameter values of interest is small, we would not use $B_n(h)$ and $\hat{I}(h)$, and we would use estimates based on serial tempering instead. We have included the derivations of $B_n(h)$ and $\hat{I}(h)$ primarily for motivation, as these makes it easier to understand the development of the serial tempering estimates. In Section 2.4 we presented an experiment which strongly suggested that $\hat{I}_\zeta^{\text{st}}(h)$ is significantly better than $\tilde{I}_\zeta^{\text{st}}(h)$ in terms of variance. George (2015) gives experimental evidence that, analogously, $\widehat{M_\zeta}(h)$ is significantly better than $\widetilde{M_\zeta}(h)$. Therefore, for the rest of this paper, we use only $\widehat{M_\zeta}(h)$ and $\hat{I}_\zeta^{\text{st}}(h)$.

Here we present the results of some experiments which demonstrate good performance of $\hat{\hat{h}} := \arg\max_h \widehat{M_\zeta}(h)$ as an estimate of h_{true}. We took $\boldsymbol{\alpha} = (\alpha, \dots, \alpha)$, i.e. $\text{Dir}_K(\boldsymbol{\alpha})$ is a symmetric Dirichlet, so that the hyperparameter in the model reduces to $h = (\eta, \alpha) \in (0, \infty)^2$. Our experiment is set up as follows: the vocabulary size is $V = 40$, the number of documents is $D = 400$, the document lengths are $n_d = 80$, $d = 1, \dots, D$, and the number of topics is $K = 8$. We used four settings for the hyperparameter under which we generate the model: h_{true} is taken to be $(.25, .25)$, $(.25, 4)$, $(4, .25)$, and $(4, 4)$. We estimated the marginal likelihood surfaces (up to a constant) on an evenly-spaced 50×50 grid of 2500 values using $\widehat{M_\zeta}(h)$ calculated from a serial tempering chain implemented as follows. The size of the subgrid was taken to be $11 \times 11 = 121$, and we used ten iterations of the iterative scheme described in Section 2.3 to form the final subgrid. The subgrid for each of the four corpora is shown in the first section of the supplementary document George and Doss (2017). For each hyperparameter value h_j ($j = 1, \dots, 121$), we took Φ_j to be the Markov transition function of the Augmented Collapsed Gibbs sampler. We took the Markov transition function $K(j, \cdot)$ on $\mathcal{L} = \{1, \dots, 121\}$ to be the uniform distribution on $\mathcal{N}_j$ where $\mathcal{N}_j$ is the subset of $\mathcal{L}$ consisting of the indices of the h_l's that are neighbors of the point h_j. We obtained the value ζ^{final} via three iterations of the scheme given by (2.16), in which we ran the serial tempering chain in each tuning iteration for $100{,}000$ iterations after a short burn-in period, and we initialized $\zeta^{(0)} = \big(\zeta_1^{(0)}, \dots, \zeta_{121}^{(0)}\big) = (1, \dots, 1)$. Using ζ^{final}, we ran the final serial tempering chain for the same number of iterations as in the tuning stage.

Figure 7 gives plots of the estimates $\widehat{M_\zeta}(h)$ and also of their Monte Carlo standard errors (MCSE) for the four specifications of h_{true}. We computed these standard error estimates using the method of batch means, which is implemented in the R package `mcmcse` (Flegal et al., 2016); they are valid pointwise, as opposed to globally, over the h-region of interest. They indicate that the accuracy of $\widehat{M_\zeta}(\cdot)$ is adequate over the entire h-range for each of the four cases of h_{true}. (We produced error margins that are valid locally, as opposed to globally, because it is of interest to see the regions where the variability is high.) In the supplementary document George and Doss (2017) we show plots of the occupancy times for the 121 components of the mixture distribution. For each of the four values of h_{true}, these occupancy times are close to uniform, indicating adequate mixing. We note that $\arg\max_h \widehat{M_\zeta}(h)$ can be obtained through a grid search from the plots in Figure 7, which is what we did in this particular illustration, but in practice these plots don't need to be generated, and $\arg\max_h \widehat{M_\zeta}(h)$ can be found very quickly through standard optimization algorithms such as those that work through gradient-based approaches (which are very easy to implement here, since $\dim(h)$ is only 2). These algorithms take very little time because they require calculation of $\widehat{M_\zeta}(\cdot)$ for only a few values of h. For the case where $\dim(h)$ is large, we mention in particular Bergstra and Bengio (2012), who argue that random search is more efficient than grid search when only a few components of h matter. As can be seen from the figure, $\arg\max_h \widehat{M_\zeta}(h)$ provides fairly good estimates of h_{true}. This experiment involves modest sample sizes; when we increase the number of documents, the surfaces become more peaked, and $\hat{\hat{h}}$ is closer to h_{true} (experiments not shown).

George (2015) shows that estimates based on $\widetilde{M_\zeta}$ also provide good estimates of h_{true}, and he compares the $\widetilde{M_\zeta}$ and the $\widehat{M_\zeta}$ estimates. From his comparison, we can conclude that the extent of the superiority of the estimates based on $\widehat{M_\zeta}$ is about the same on the synthetic corpora of the present section as in the real data illustration of Section 2.4.

4 Construction of Two Markov Chains with Invariant Distribution $\nu_{h_*,w}$

In order to develop Markov chains on $\boldsymbol{\psi} = (\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{z})$ whose invariant distribution is the posterior $\nu_{h,\boldsymbol{w}}$, we first express the posterior in a convenient form. We start with the familiar formula

$$\nu_{h,\boldsymbol{w}}(\boldsymbol{\psi}) \propto \ell_{\boldsymbol{w}}(\boldsymbol{\psi}) \nu_h(\boldsymbol{\psi}), \tag{4.1}$$

where the likelihood $\ell_{\boldsymbol{w}}(\boldsymbol{\psi}) = p^{(h)}_{\boldsymbol{w}\,|\,\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{\beta}}(\boldsymbol{w} \mid \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\beta})$ is given by line 4 of the LDA model statement. For $d = 1, \dots, D$ and $j = 1, \dots, K$, let $S_{dj} = \{i : 1 \le i \le n_d \text{ and } z_{dij} = 1\}$, which is the set of indices of all words in document d whose latent topic variable is j. With this notation, from line 4 of the model statement we have

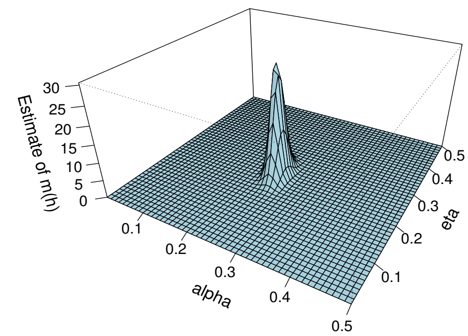
$$\begin{aligned} p^{(h)}_{\boldsymbol{w}\,|\,\boldsymbol{z},\boldsymbol{\theta},\boldsymbol{\beta}}(\boldsymbol{w} \mid \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{\beta}) &= \prod_{d=1}^{D} \prod_{i=1}^{n_d} \prod_{j: z_{dij}=1} \prod_{t=1}^{V} \beta_{jt}^{w_{dit}} = \prod_{d=1}^{D} \prod_{j=1}^{K} \prod_{t=1}^{V} \prod_{i \in S_{dj}} \beta_{jt}^{w_{dit}} \\ &= \prod_{d=1}^{D} \prod_{j=1}^{K} \prod_{t=1}^{V} \beta_{jt}^{\sum_{i \in S_{dj}} w_{dit}} = \prod_{d=1}^{D} \prod_{j=1}^{K} \prod_{t=1}^{V} \beta_{jt}^{m_{djt}}, \end{aligned} \tag{4.2}$$

where $m_{djt} = \sum_{i \in S_{dj}} w_{dit}$ counts the number of words in document d for which the latent topic is j and the index of the word in the vocabulary is t. Recalling the definition of n_{dj} given just before (A.1), and noting that $\sum_{i \in S_{dj}} w_{dit} = \sum_{i=1}^{n_d} z_{dij} w_{dit}$, we see that

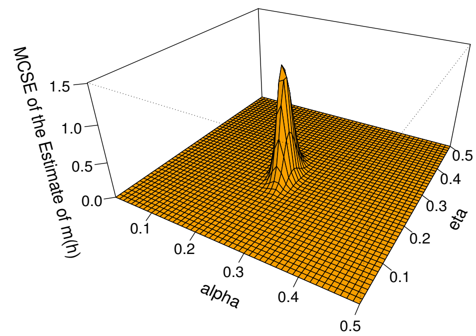
$$m_{djt} = \sum_{i=1}^{n_d} z_{dij} w_{dit} \quad \text{and} \quad \sum_{t=1}^{V} m_{djt} = n_{dj}. \tag{4.3}$$

Plugging the likelihood (4.2) and the prior (A.1) into (4.1), and absorbing Dirichlet normalizing constants into an overall constant of proportionality, we have

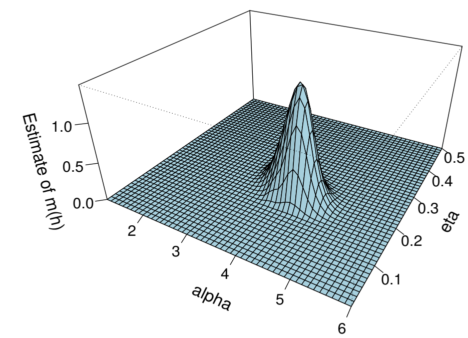
$$\nu_{h,\boldsymbol{w}}(\boldsymbol{\psi}) \propto \left[\prod_{d=1}^{D} \prod_{j=1}^{K} \prod_{t=1}^{V} \beta_{jt}^{m_{djt}} \right] \left[\prod_{d=1}^{D} \prod_{j=1}^{K} \theta_{dj}^{n_{dj}} \right] \left[\prod_{d=1}^{D} \prod_{j=1}^{K} \theta_{dj}^{\alpha_j - 1} \right] \left[\prod_{j=1}^{K} \prod_{t=1}^{V} \beta_{jt}^{\eta - 1} \right]. \tag{4.4}$$



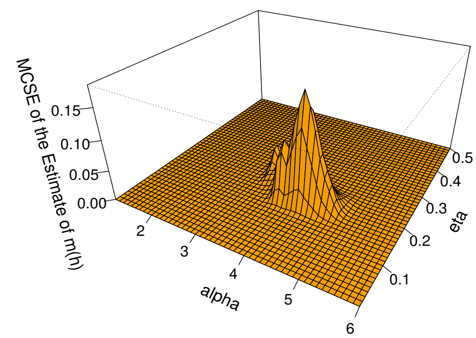
(a) $\widehat{M}_\zeta(h)$: $h_{\text{true}} = (.25, .25)$, $\hat{\hat{h}} = (.24, .24)$



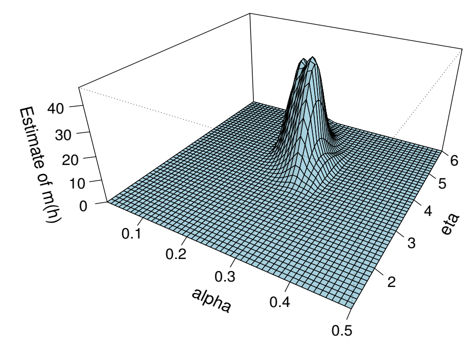
(b) MCSE of $\widehat{M}_\zeta(h)$: $h_{\text{true}} = (.25, .25)$



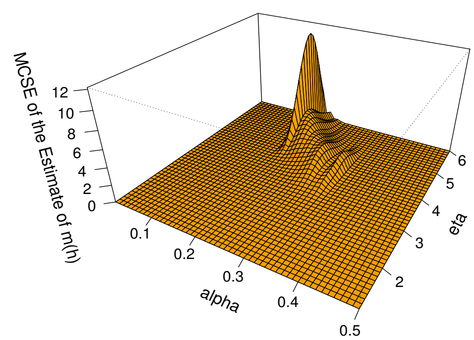
(c) $\widehat{M}_\zeta(h)$: $h_{\text{true}} = (.25, 4)$, $\hat{\hat{h}} = (.19, 4.2)$



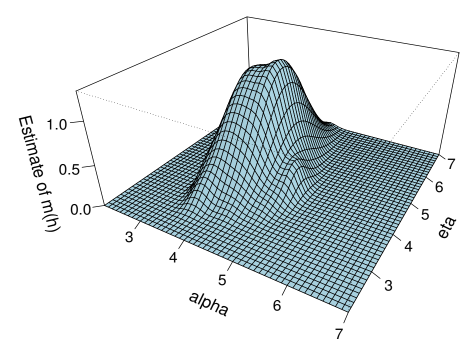
(d) MCSE of $\widehat{M}_\zeta(h)$: $h_{\text{true}} = (.25, 4)$



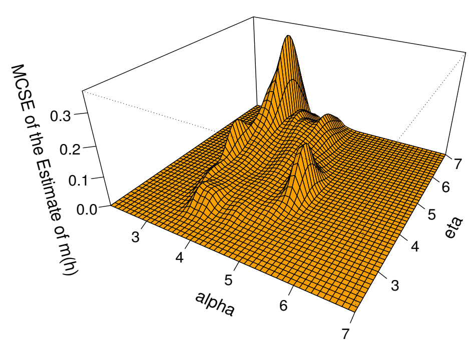
(e) $\widehat{M}_\zeta(h)$: $h_{\text{true}} = (4, .25)$, $\hat{\hat{h}} = (4.2, .27)$



(f) MCSE of $\widehat{M}_\zeta(h)$: $h_{\text{true}} = (4, .25)$



(g) $\widehat{M}_\zeta(h)$: $h_{\text{true}} = (4, 4)$, $\hat{\hat{h}} = (4.9, 4.2)$



(h) MCSE of $\widehat{M}_\zeta(h)$: $h_{\text{true}} = (4, 4)$

Figure 7: $\widehat{M}_\zeta(h)$ and MCSE of $\widehat{M}_\zeta(h)$ for four values of h_{true}. In each case, $\hat{\hat{h}}$ is close to h_{true}.

The expression for $\nu_{h,\boldsymbol{w}}(\boldsymbol{\psi})$ above also appears in the unpublished report Fuentes et al. (2011).

The Conditional Distributions of $(\boldsymbol{\beta}, \boldsymbol{\theta})$ Given $\boldsymbol{z}$ and of $\boldsymbol{z}$ Given $(\boldsymbol{\beta}, \boldsymbol{\theta})$

All distributions below are conditional distributions given $\boldsymbol{w}$, which is fixed, and henceforth this conditioning is suppressed in the notation. Note that in (4.4), the terms m_{djt} and n_{dj} depend on $\boldsymbol{z}$. By inspection of (4.4), we see that given $\boldsymbol{z}$,

$$\begin{aligned}
&\theta_1, \ldots, \theta_D \text{ and } \beta_1, \ldots, \beta_K \text{ are all independent,} \\
&\theta_d \sim \mathrm{Dir}_K\big(n_{d1} + \alpha_1, \ldots, n_{dK} + \alpha_K\big), \\
&\beta_j \sim \mathrm{Dir}_V\big(\textstyle\sum_{d=1}^D m_{dj1} + \eta, \ldots, \sum_{d=1}^D m_{djV} + \eta\big).
\end{aligned} \tag{4.5}$$

From (4.4) we also see that

$$\begin{aligned}
p^{(h)}_{\boldsymbol{z}\,|\,\boldsymbol{\theta},\boldsymbol{\beta}}(\boldsymbol{z}\,|\,\boldsymbol{\theta},\boldsymbol{\beta}) &\propto \prod_{d=1}^{D}\prod_{j=1}^{K}\left(\left[\prod_{t=1}^{V}\beta_{jt}^{m_{djt}}\right]\theta_{dj}^{n_{dj}}\right) \\
&= \prod_{d=1}^{D}\prod_{i=1}^{n_d}\prod_{j=1}^{K}\left[\prod_{t=1}^{V}\beta_{jt}^{z_{dij}w_{dit}}\theta_{dj}^{z_{dij}w_{dit}}\right] \qquad (4.6) \\
&= \prod_{d=1}^{D}\prod_{i=1}^{n_d}\prod_{j=1}^{K}\left[\prod_{t=1}^{V}\big(\beta_{jt}\theta_{dj}\big)^{w_{dit}}\right]^{z_{dij}}, \qquad (4.7)
\end{aligned}$$

where (4.6) follows from (4.3). Let $p_{dij} = \prod_{t=1}^{V}\big(\beta_{jt}\theta_{dj}\big)^{w_{dit}}$. By inspection of (4.7) we see immediately that given $(\boldsymbol{\theta}, \boldsymbol{\beta})$,

$$\begin{aligned}
&z_{11}, \ldots, z_{1n_1}, z_{21}, \ldots, z_{2n_2}, \ldots, z_{D1}, \ldots, z_{Dn_D} \text{ are all independent,} \\
&z_{di} \sim \mathrm{Mult}_K(p_{di1}, \ldots, p_{diK}).
\end{aligned} \tag{4.8}$$

The conditional distribution of $(\boldsymbol{\beta}, \boldsymbol{\theta})$ given by (4.5) can be used, in conjunction with the CGS of Griffiths and Steyvers (2004), to create a Markov chain on $\boldsymbol{\psi}$ whose invariant distribution is $\nu_{h,\boldsymbol{w}}$: if $\boldsymbol{z}^{(1)}, \boldsymbol{z}^{(2)}, \ldots$ is the CGS, then for $l = 1, 2, \ldots$, we generate $(\boldsymbol{\beta}^{(l)}, \boldsymbol{\theta}^{(l)})$ from $p^{(h)}_{\boldsymbol{\theta},\boldsymbol{\beta}\,|\,\boldsymbol{z}}(\cdot\,|\,\boldsymbol{z}^{(l)})$ given by (4.5) and form $(\boldsymbol{z}^{(l)}, \boldsymbol{\beta}^{(l)}, \boldsymbol{\theta}^{(l)})$—this is what we have called the Augmented CGS. The CGS is uniformly ergodic (Theorem 1 of Chen and Doss (2017)) and an easy argument shows that the resulting ACGS is therefore also uniformly ergodic (and in fact, the rate of convergence of the ACGS is exactly the same as that of the CGS; see Diaconis et al. (2008, Lemma 2.4)).

The two conditionals (4.5) and (4.8) also enable a direct construction of a two-cycle Gibbs sampler that runs on the pair $(\boldsymbol{z}, (\boldsymbol{\beta}, \boldsymbol{\theta}))$—this is what we have called the Grouped Gibbs Sampler. This Gibbs sampler has the very attractive feature that it can be parallelized: From (4.5), we see that given $\boldsymbol{z}$ and $\boldsymbol{w}$, the θ_d's and β_t's are all independent, so can be updated simultaneously by different processors; and from (4.8), we see that given $(\boldsymbol{\beta}, \boldsymbol{\theta})$ and $\boldsymbol{w}$, all the components of $\boldsymbol{z}$ are independent, so can also be updated simultaneously by different processors. This scheme was noted earlier by Newman et al. (2009), who dismissed it on the grounds that the Collapsed Gibbs Sampler has superior mixing properties because, according to Liu et al. (1994), collapsing improves the mixing rate. However, the theorem from Liu et al. (1994) that Newman et al. (2009) are citing does not apply to the present situation. To be specific, Liu et al. (1994) consider a Gibbs sampling situation involving three variables X, Y, and Z. They show that a Gibbs sampler on the pair (X, Y) (with Z integrated out), which they call a collapsed Gibbs sampler, is superior to a Gibbs sampler on the triple (X, Y, Z). But for the LDA model, the CGS on $\boldsymbol{z} = (z_{11}, \ldots, z_{1n_1}, \ldots, z_{D1}, \ldots, z_{Dn_D})$ is not a collapsed version of the Gibbs sampler that runs on the pair $\big(\boldsymbol{z}, (\boldsymbol{\beta}, \boldsymbol{\theta})\big)$ in any sense, so which of the two Gibbs samplers is superior in terms of mixing rate is an open question. George (2015) compared the mixing rates for various parameters empirically, and found that the mixing rate for the CGS is faster, but not much faster. A paper based on George (2015) that studies this Grouped Gibbs Sampler, including its mixing rate and computational complexity, is under preparation (Doss and George, 2017).

5 Evaluation: Choice of Estimator of $\arg\max_h m(h)$ and Resulting Model Fit

The maximizer of the marginal likelihood, $\hat{h} = \arg\max_h m(h)$, may be estimated via the MCMC scheme described in the present paper, or by some version of the EM algorithm (VI-EM or Gibbs-EM). Our main goal in this section is two-fold. (1) We show empirically that neither the VI-EM nor the Gibbs-EM method provides estimates of $\hat{h}$ that are as accurate as ours, and we briefly discuss why theoretically neither VI-EM nor Gibbs-EM, at least in its current implementation, can be expected to work correctly. We also compare VI-EM to Gibbs-EM in terms of accuracy, which to the best of our knowledge has not been done before, and compare VI-EM, Gibbs-EM, and our estimator in terms of speed. This is done in Section 5.1. (2) We consider some of the default choices of h used in the literature that use ad-hoc (i.e. non-principled) criteria. We look at model fit and show empirically that when we use any of the three estimates of $\hat{h}$ (VI-EM, Gibbs-EM, or our serial tempering method), model fit is better than if we use any of the ad-hoc choices. This is done in Section 5.2.

5.1 Comparison of Methods for Estimating $\arg\max_h m(h)$

For uniformity of notation, let $\hat{\hat{h}}_{\text{ST}}$, $\hat{\hat{h}}_{\text{VEM}}$, and $\hat{\hat{h}}_{\text{GEM}}$ be the estimates of $\hat{h}$ formed from serial tempering MCMC, VI-EM, and Gibbs-EM, respectively, and recall that $\hat{\hat{h}}_{\text{ST}} = \hat{\hat{h}} = \arg\max_h \widehat{M_\zeta}(h)$.

VI-EM The estimate $\hat{\hat{h}}_{\text{VEM}}$ proposed by Blei et al. (2003) 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 $\nu_{h^{(k)}, \boldsymbol{w}}$. This step is infeasible because $\nu_{h^{(k)}, \boldsymbol{w}}$ 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_{ϕ_*}, which is "closest" to $\nu_{h^{(k)}, \boldsymbol{w}}$. Let $Q(h)$ be the expected value of $\log(p_h(\boldsymbol{\psi}, \boldsymbol{w}))$ with respect to q_{ϕ_*}. 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.

The implementation of the EM algorithm through variational inference methods outlined above describes what Blei et al. (2003) do *conceptually*, but not exactly. Actually, Blei et al. (2003) apply VI-EM to a model that is different from ours. In that model, $\boldsymbol{\beta}$ is viewed as a fixed but unknown parameter, to be estimated, and the latent variable is $\boldsymbol{\vartheta} = (\boldsymbol{\theta}, \boldsymbol{z})$. Thus, the observed and missing data are, respectively, $\boldsymbol{w}$ and $\boldsymbol{\vartheta}$, and the marginal likelihood is a function of two variables, h and $\boldsymbol{\beta}$. Abstractly speaking, the description of VI-EM given above is exactly the same. We implemented VI-EM to the version of the LDA model considered in this paper, by modifying the Blei et al. (2003) code. While VI-EM can handle very large corpora with many topics, there are no theoretical results regarding convergence of the sequence $h^{(k)}$ to $\arg\max_h m(h)$, and VI-EM has the following problems: it may have poor performance if the approximation of $\nu_{h^{(k)},\boldsymbol{w}}$ by q_{ϕ_*} is not good; and if the likelihood surface is multimodal, as in Figure 7(e), then it can fail to find the global maximum (as is the case for all EM-type algorithms and also gradient-based approaches).

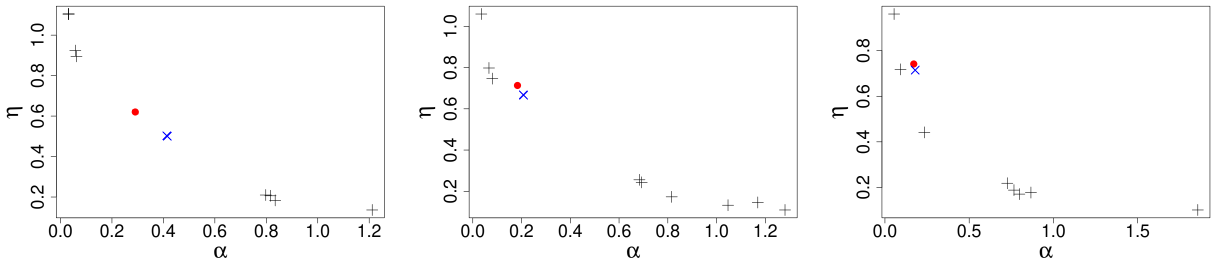
Gibbs-EM Monte Carlo EM (MC-EM), in which the E-step is replaced by a Monte Carlo estimate, dates back to Wei and Tanner (1990), and was introduced to the machine learning community in Andrieu et al. (2003). As mentioned earlier, since an error is introduced at every iteration, there is no reason to expect that the algorithm will converge at all, let alone to the true maximizer of the likelihood. In fact, Wei and Tanner (1990) recognized this problem and suggested that the Markov chain length be increased at every iteration of the EM algorithm. We will let m_k denote the MC length at the k^{th} iteration. Convergence of MC-EM (of which the Gibbs-EM algorithm of Wallach (2008) is a special case) is a nontrivial issue. It was studied by Fort and Moulines (2003), who showed that a minimal condition is that $m_k \to \infty$ at the rate of k^a, for some $a > 1$. However, they do not give guidelines for choosing a. Other conditions imposed in Fort and Moulines (2003) are fairly stringent, and it is not clear whether they are satisfied in the LDA model. In the current implementation of Gibbs-EM (Wallach, 2006), the latent variable is taken to be $\boldsymbol{z}$ (because the standard Markov chain used to estimate posterior distributions in this model is the CGS). At the k^{th} iteration, a Markov chain $\boldsymbol{z}_1, \dots, \boldsymbol{z}_{m_k}$ with invariant distribution equal to the posterior distribution of $\boldsymbol{z}$ given $\boldsymbol{w}$ is generated, and the function $G(h) = (1/m_k)\sum_{i=1}^{m_k} \log(p_h(\boldsymbol{z}_i, \boldsymbol{w}))$ must be maximized. This is done by solving the equation $\nabla G(h) = 0$ using fixed-point iteration, and because $\nabla G(h)$ is computationally intractable, an approximation (Minka, 2003) is used (in effect, a lower bound to $G(h)$ is found, and the lower bound is what is maximized). This approximation introduces a second potential problem for Gibbs-EM. A third potential problem is that, as for VI-EM, the iterations may get stuck near a local maximum when the likelihood surface is multimodal.

To evaluate the performance of the VI-EM, Gibbs-EM, and serial tempering MCMC methods of estimating $\hat{h}$, we generated small synthetic corpora according to the LDA model with the following specifications: the true hyperparameter value is $h = (\eta, \alpha) = (.8, .2)$, the vocabulary size is $V = 20$, the number of words in each document is $n_d = 80$, the number of topics is $K = 4$ and 8, and the number of documents is $D = 20$, 40, and 100, for a total of 6 specifications. For each specification, we formed $\hat{\hat{h}}_{\text{ST}}$, $\hat{\hat{h}}_{\text{VEM}}$, and $\hat{\hat{h}}_{\text{GEM}}$. For $\hat{\hat{h}}_{\text{GEM}}$, we used the algorithm given in Wallach (2006), in which the Markov chain is the CGS. We took the number of cycles of the Gibbs sampler to be 10,000—this is considerably greater than the default value of 20 in the MALLET package (McCallum, 2002); and we formed 10 independent estimates, using 10 different initial values. Likewise, for VI-EM we formed 10 estimates using 10 different initial values. For the serial tempering estimate, our principal goal was to form a confidence set for $\hat{h}$, and we did this as follows. We ran 10 independent serial tempering chains, for which the sequence $h_1, \ldots, h_J$ consisted of a 7×9 grid of 63 values over the region $(\eta, \alpha) \in [.6, .9] \times [.1, .3]$ (this region was obtained from a small number of iterations of the iterative scheme described in Section 2.3), and the rest of the specifications were the same as those described in the experiments of Section 2.4; each chain was run for 100,000 iterations. Let $\hat{\hat{h}}_{\text{ST}}^{[\ell]}$ be the estimate of $\hat{h}$ formed from serial tempering chain ℓ, for $\ell = 1, \ldots, 10$. According to Theorem 1 in Section 2.1 and Remark 3 in Section 2.3, the independent variables $\hat{\hat{h}}_{\text{ST}}^{[1]}, \ldots, \hat{\hat{h}}_{\text{ST}}^{[10]}$ are approximately bivariate normally distributed with mean vector $\hat{h}$. Therefore, they can be used to form a 95% confidence ellipse for $\hat{h}$, 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). The confidence set could also have been formed from a single long chain, using the method described in Theorem 1; the two methods use about the same computational resources. Figure 8 shows the results, and we make two general observations.

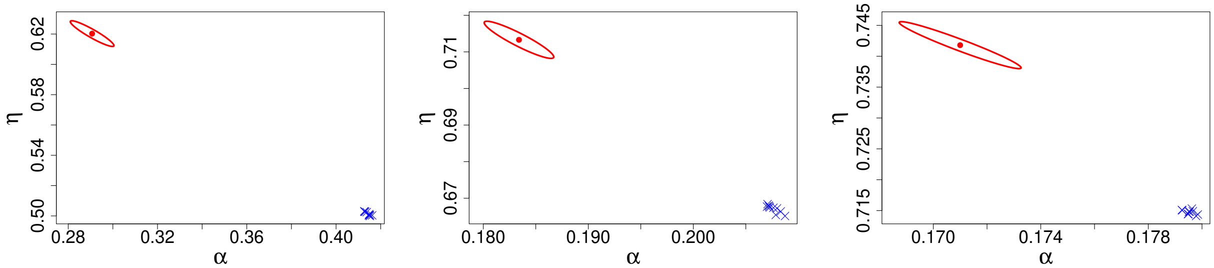
1. From the plots in rows 1 and 3 (plots (a), (b), (c), (g) (h), and (i)), we see that the VI-EM method does not perform well: in each of the 6 cases, the estimates are far from the true value, $\arg\max_h m(h)$, and also strongly depend on the starting values. We created plots (d), (e), (f), (j), (k), and (l), which are zoomed-in versions of plots (a), (b), (c), (g) (h), (i), respectively; these magnify a region which contains the serial tempering estimate and associated confidence ellipse. We see that while the plots in rows 1 and 3 show that the Gibbs-EM estimates greatly outperform the VI-EM estimates (they are both closer to the true value and less dependent on the starting value), the zoomed-in plots in rows 2 and 4 show that the Gibbs-EM points are far from being inside the 95% confidence ellipse. We carried out some experiments in which we followed the recommendations in Fort and Moulines (2003) to increase the number of cycles in the Gibbs sampling inner loop. Specifically, we took $m_1 = 2^7$ and doubled the length of the Gibbs sampler run with every iteration, i.e. we took $m_n = 2^{(6+n)}$, $n = 1, \ldots, 20$. Unfortunately, this did not give significant improvement. The Gibbs-EM estimates were never close to being inside the ellipse, The problem could be with our rate of increase, or that Gibbs-EM simply does not produce consistent estimates, or with the implementation of the maximization step (which uses an approximation).

2. Both Gibbs-EM and VI-EM improve as the number of documents, D, increases. A possible explanation of this is that as D increases, generally speaking the EM algorithm converges faster because the likelihood surface becomes more peaked. Of course, the larger the value of D, the weaker is the effect of the choice of h—this is the Bernstein-von Mises Theorem (see Freedman (1999) and the references therein), which loosely speaking states that as $D \to \infty$, the data swamp the prior.

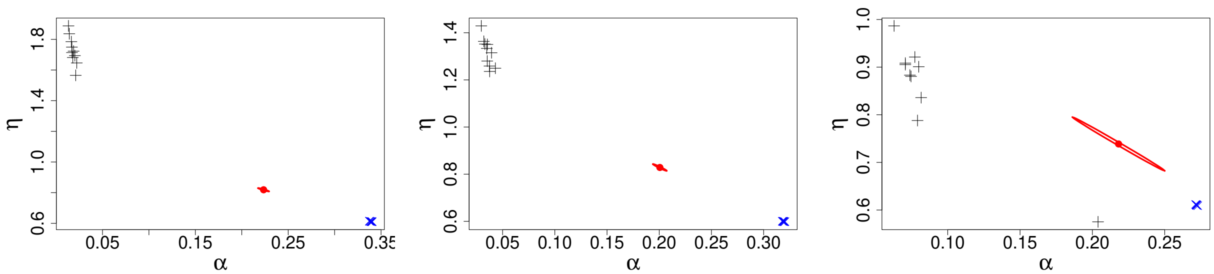
To assess the computational burden, we computed $\hat{\hat{h}}_{\text{VEM}}$, $\hat{\hat{h}}_{\text{GEM}}$, and $\hat{\hat{h}}_{\text{ST}}$ for the six corpora



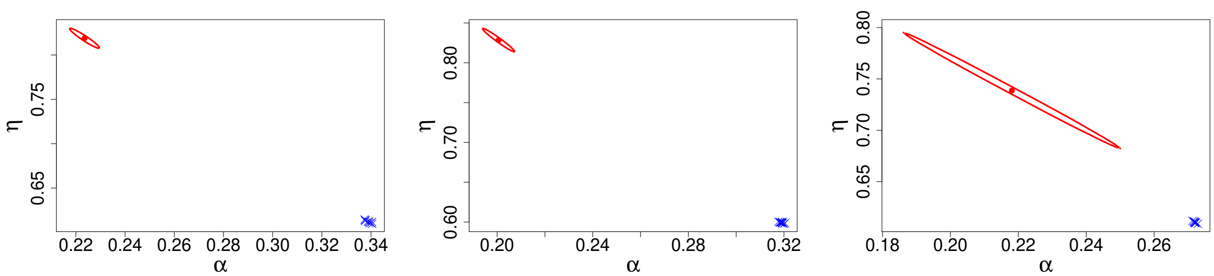
(a) $K = 4, D = 20$ (b) $K = 4, D = 40$ (c) $K = 4, D = 100$



(d) $K = 4, D = 20$ (e) $K = 4, D = 40$ (f) $K = 4, D = 100$



(g) $K = 8, D = 20$ (h) $K = 8, D = 40$ (i) $K = 8, D = 100$



(j) $K = 8, D = 20$ (k) $K = 8, D = 40$ (l) $K = 8, D = 100$

Figure 8: Plots of estimates of $\hat{h}$ for the 6 corpora described in text. Points marked $\times$ are estimates formed by Gibbs-EM, points marked $+$ are estimates formed by VI-EM. A point marked $\bullet$ is the average of 10 independent estimates of $\hat{h}$ formed via ST chains, and the ellipse is a 95% confidence set for $\hat{h}$ formed from the 10 estimates. The three plots in row 2 are zoomed-in versions of the three plots in row 1, magnifying a region which contains the ST estimate, so the ellipse becomes visible. Similarly, the plots in row 4 are zoomed-in versions of the plots in row 3.

we considered. For $\hat{\hat{h}}_{\text{VEM}}$, each run consisted of 100 EM iterations and in each EM iteration there were 100 variational inference iterations. For $\hat{\hat{h}}_{\text{GEM}}$, each run consisted of 50 EM iterations and in each EM iteration the CGS was run for $11{,}000$ cycles, of which the first 1000 were deleted as burn-in. For $\hat{\hat{h}}_{\text{ST}}$, each run consisted of 2 tuning iterations with a chain length of $51{,}000$ cycles, and a final iteration with a chain length of $101{,}000$ cycles; and in each case, the first 1000 cycles were deleted as burn-in. Our experiments were conducted through the R programming language, using Rcpp, on a 3.70GHz quad core Intel Xeon Processor E5-1630V3. Table 1 gives the results. From the table, we see that the time for $\hat{\hat{h}}_{\text{ST}}$ is about seven times the time for $\hat{\hat{h}}_{\text{GEM}}$, and the time for $\hat{\hat{h}}_{\text{GEM}}$ is about 55 times the time for $\hat{\hat{h}}_{\text{VEM}}$. These numbers are not as extreme as they look, because for both $\hat{\hat{h}}_{\text{GEM}}$ and $\hat{\hat{h}}_{\text{ST}}$ we could have gotten comparable results with much smaller chain lengths.

K	D	Time for $\hat{\hat{h}}_{\text{VEM}}$	Time for $\hat{\hat{h}}_{\text{GEM}}$	Time for $\hat{\hat{h}}_{\text{ST}}$
8	100	.34	11.64	90.55
8	40	.12	6.08	45.87
8	20	.04	2.96	31.35
4	100	.19	10.73	49.75
4	40	.08	4.88	25.71
4	20	.04	2.23	16.72

Table 1: Length of time, in minutes, it takes to compute the VI-EM, Gibbs-EM, and serial tempering estimates of $\hat{h}$ for six corpora.

5.2 Comparison of Model Fit: Empirical Bayes Choice vs. Ad-Hoc Choices of the Hyperparameter

In the literature, the following choices for $h = (\eta, \alpha)$ have been presented: $h_{\text{DG}} = (0.1, 50/K)$, used in Griffiths and Steyvers (2004); $h_{\text{DA}} = (0.1, 0.1)$, used in Asuncion et al. (2009); and $h_{\text{DR}} = (1/K, 1/K)$, used in the `Gensim` topic modelling package (Řehůřek and Sojka, 2010), a well-known package used in the topic modelling community. These choices are ad-hoc, and not based on any particular principle.

Criterion for Model Fit The criterion we use is a score that is inversely related to the so-called "perplexity" score which is sometimes used in the machine learning literature. When applied to the LDA context, the score is obtained as follows. For $d = 1, \dots, D$, 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 $d = 1, \dots, D$, 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

$$L_d(h) = \int \ell_{\boldsymbol{w}_d}(\boldsymbol{\psi})\, d\nu_{h,\boldsymbol{w}_{(-d)}}(\boldsymbol{\psi}), \tag{5.1}$$

where $\ell_{\boldsymbol{w}_d}(\boldsymbol{\psi})$ is the likelihood of $\boldsymbol{\psi}$ for the held-out document d, and $\nu_{h,\boldsymbol{w}_{(-d)}}$ is the posterior distribution of $\boldsymbol{\psi}$ given $\boldsymbol{w}_{(-d)}$. We form the score $S(h) = \left[\prod_{d=1}^{D} L_d(h)\right]^{1/D}$. Two different values of hyperparameter h are compared via their scores. Conceptually, it is easy to estimate $L_d(h)$ by direct Monte Carlo: let $\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \ldots$ be an ergodic Markov chain with invariant distribution $\nu_{h,\boldsymbol{w}_{(-d)}}$. We then approximate the integral by $(1/n)\sum_{i=1}^{n} \ell_{\boldsymbol{w}_d}(\boldsymbol{\psi}_i)$. Care needs to be exercised, however, because in (5.1), 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 (5.1).

Real Datasets Here we compare the fit of LDA models based on various choices of the hyperparameter, on several corpora of real documents. We created two sets of document corpora, one from the 20Newsgroups dataset[2], and the other from the English Wikipedia. The 20Newsgroups dataset is commonly used in the machine learning literature for experiments on applications of text classification and clustering algorithms. It contains approximately 20,000 articles that are partitioned relatively evenly across 20 different newsgroups or categories. We created the second set of corpora from web articles downloaded from the English Wikipedia, with the help of the MediaWiki API[3].

We created the 20Newsgroups corpora as follows. We formed five subsets of the 20Newsgroups dataset, which we call C-1–C-5, with the feature that the articles within the subsets are increasingly difficult to distinguish: for corpus C-1 the topics for the different articles are very different, and for corpus C-5 the topics for the different articles are similar. For each article, we took its true topic label to be the newsgroup to which the article is assigned. Thus, for corpora C-1–C-5, it becomes increasingly difficult to place the articles into the correct newsgroup. We built corpus C-1 from a random subset of articles from the 20Newsgroups categories Medicine, Christianity, and Baseball; these three categories are highly unrelated and easily recognizable from article texts. We built corpus C-2 from a random subset of articles from the categories Automobiles, Motorcycles, Baseball, and Hockey (all four of these categories are classified under the super-category Recreation in the 20Newsgroups dataset), and we built corpus C-3 from a random subset of articles from the categories Cryptography, Electronics, Medicine, and Space (all four of these categories are classified under the super-category Science in the 20Newsgroups dataset). Compared to the categories in corpus C-1, the categories in corpora C-2 and C-3 are moderately related. Lastly, we created corpus C-4 using articles under the categories Autos and Motorcycles, and corpus C-5 using articles under the categories PC Hardware and Mac Hardware. In corpora C-4 and C-5, the corresponding categories are closely related to each other and hard to distinguish from article texts.

We created the Wikipedia corpora as follows. When a Wikipedia article is created, it is typically tagged to one or more categories, one of which is the "primary category." For each article, we took its true topic label to be the primary category label for the article. We created corpus C-6 from a subset of the Wikipedia articles under the categories *Leopardus*, *Lynx*, and *Prionailurus* and corpus C-7 from a subset of the Wikipedia articles under the categories *Acinonyx*, *Leopar-*

[2] http://qwone.com/~jason/20Newsgroups
[3] http://www.mediawiki.org/wiki/API:Query

dus, *Prionailurus*, and *Puma*. All the categories of corpora C-6 and C-7 are part of the Wikipedia super-category *Felines*. We created corpus C-8 from a subset of the Wikipedia articles under the categories *Coyotes*, *Jackals*, and *Wolves*. All three categories of corpus C-8 are under the Wikipedia super-category *Canis*. Finally, we created corpus C-9 from a subset of the Wikipedia articles under the categories *Eagles*, *Falco (genus)*, *Falconry*, *Falcons*, *Harriers*, *Hawks*, *Kites*, and *Owls*. All eight categories of corpus C-9 are subcategories of the Wikipedia category *Birds of Prey*. For each of the four Wikipedia corpora that we created, the categories of the articles are closely related to each other, and fairly hard to distinguish from article texts.

Table 2 gives some information on the nine corpora we created. In the table, the column labeled V gives the vocabulary size for each corpus, the column labeled N gives the total number of words for each corpus, and the column labeled Categories gives newsgroup categories for each 20Newsgroup corpus, and Wikipedia categories for each Wikipedia corpus. The numbers shown in parentheses next to the category names are the number of documents associated with the corresponding categories. For each corpus, we took the number of topics K to be equal to the number of categories for the corpus.

Corpus	Categories	V	N
C-1	sci.med (50), soc.religion.christian (50), rec.sport.baseball (50)	807	12,092
C-2	rec.autos (50), rec.motorcycles (50), rec.sport.baseball (50), rec.sport.hockey (50)	1,061	16,579
C-3	sci.crypt (50), sci.electronics (50), sci.med (50), sci.space (50)	1,033	15,828
C-4	rec.autos (50), rec.motorcycles (50)	488	6,602
C-5	comp.sys.ibm.pc.hardware (50), comp.sys.mac.hardware (50)	502	7,454
C-6	Leopardus (8), Lynx (8), Prionailurus (7)	303	7,788
C-7	Acinonyx (6), Leopardus (8), Prionailurus (7), Puma (8)	622	12,831
C-8	Coyotes (7), Jackals (7), Wolves (8)	447	9,212
C-9	Eagles (62), Falco (genus) (45), Falconry (52), Falcons (10), Harriers (21), Hawks (16), Kites (22), Owls (76)	1,369	116,135

Table 2: Corpora created from the 20Newsgroups dataset and the Wikipedia pages.

Comparison of Model Fit We now compare the performance of the LDA models indexed by $\hat{\hat{h}}_{\text{ST}}$, $\hat{\hat{h}}_{\text{GEM}}$, $\hat{\hat{h}}_{\text{VEM}}$, h_{DR}, h_{DA}, and h_{DG} for corpora C-1–C-9, using the estimate of the score $S(h)$, which we denote by $\widehat{S}(h)$, described in the beginning of this subsection. Details regarding how $\hat{\hat{h}}_{\text{ST}}$ was computed and regarding its accuracy are given in the supplementary document George and Doss (2017). The actual values of $\hat{\hat{h}}_{\text{ST}}$, $\hat{\hat{h}}_{\text{GEM}}$, and $\hat{\hat{h}}_{\text{VEM}}$, are also given in George and Doss (2017).

To compute $\widehat{S}(h)$ for a corpus, for every held-out document, we used Chen's (2015) method with a full Gibbs sampling chain of length $2{,}000$, after discarding a short burn-in period. Table 3 gives the ratios $\widehat{S}(h)/\widehat{S}(\hat{\hat{h}}_{\text{ST}})$, where h is $\hat{\hat{h}}_{\text{GEM}}$, $\hat{\hat{h}}_{\text{VEM}}$, h_{DR}, h_{DA}, and h_{DG}, for all nine corpora. From the table, we make three main observations: (1) Any of the estimates of $\hat{h}$ are better than any of the ad-hoc choices, uniformly, and by wide margins. (2) Within the estimates of $\hat{h}$, ST does better than either GEM or VEM on the whole, although not in every case, and when it is outperformed, it is not by much. (3) As a general pattern, the lack of fit of the models indexed by the ad-hoc choices of h is worse for the Wikipedia corpora than for the 20Newsgroups corpora. The Wikipedia corpora may be considered "difficult," in the sense that for these corpora the articles are very similar, and thus hard to distinguish from article texts. On the other hand, within the group of estimates of $\hat{h}$, it is not clear what are the characteristics of a corpus which affect the fit—there may be factors, beyond similarity of the documents, that are relevant.

Corpus	$\hat{\hat{h}}_{\text{GEM}}$	$\hat{\hat{h}}_{\text{VEM}}$	h_{DR}	h_{DA}	h_{DG}
C-1	6.78×10^{-01}	4.83×10^{-01}	3.54×10^{-01}	$1.11 \times 10^{+00}$	8.24×10^{-04}
C-2	5.11×10^{-01}	8.19×10^{-01}	5.23×10^{-01}	2.52×10^{-02}	7.21×10^{-05}
C-3	9.86×10^{-01}	5.58×10^{-01}	2.98×10^{-01}	1.41×10^{-01}	1.33×10^{-02}
C-4	8.21×10^{-01}	7.71×10^{-01}	3.48×10^{-01}	1.22×10^{-01}	6.66×10^{-02}
C-5	9.98×10^{-01}	$1.62 \times 10^{+00}$	4.58×10^{-01}	1.61×10^{-01}	9.36×10^{-02}
C-6	$2.48 \times 10^{+00}$	$1.12 \times 10^{+01}$	7.31×10^{-03}	5.71×10^{-06}	6.57×10^{-08}
C-7	4.39×10^{-01}	$7.82 \times 10^{+00}$	5.34×10^{-03}	1.51×10^{-10}	1.89×10^{-14}
C-8	$2.04 \times 10^{+00}$	6.40×10^{-01}	9.90×10^{-04}	1.77×10^{-09}	3.29×10^{-12}
C-9	$1.04 \times 10^{+00}$	1.75×10^{-02}	2.17×10^{-02}	7.04×10^{-03}	5.56×10^{-09}

Table 3: Ratios of the estimates of the fit criterion $S(h)$ to estimate of $S(\hat{\hat{h}}_{\text{ST}})$ for five choices of h, for all nine corpora. A small number indicates a lack of fit, thus a poor choice of h, and by this criterion, all ad-hoc choices perform poorly.

Implementation Details To compute $\widehat{M_\zeta}(h)$, we implemented the serial tempering scheme described in Section 2 as follows. The size of the subgrid was taken to be $7 \times 13 = 91$, and we used six iterations of the iterative scheme described in Section 2.3 to form the final subgrid, using Markov chains of length $10{,}000$. For the run using the final subgrid, we used three iterations of the scheme given by (2.16) to obtain ζ^{final}, with a Markov chain length of $50{,}000$ per iteration (after a short burn-in period). The final run, using ζ^{final}, also used a Markov chain length of $50{,}000$. To estimate the standard error of $\widehat{M_\zeta}(h)$, we used the method of batch means, which is implemented by the R package `mcmcse` in Flegal et al. (2016). Diagnostics that establish that the serial tempering chain mixes adequately are given in the supplementary document George and Doss (2017). Table 4 gives the time it took to compute $\hat{\hat{h}}_{\text{VEM}}$, $\hat{\hat{h}}_{\text{GEM}}$, and $\hat{\hat{h}}_{\text{ST}}$, for three of the real corpora used in this section.

It is natural to ask why it has not been noted before that VI-EM and Gibbs-EM sometimes perform poorly. Evaluations have been typically done through a model fit criterion such as the one we used in this subsection, and to the best of our knowledge the literature has not given an assessment of how close $\hat{\hat{h}}_{\text{VEM}}$ and $\hat{\hat{h}}_{\text{GEM}}$ are to h_{true} for corpora generated from an LDA model indexed by h_{true}, as is done in Section 5.1.

Corpus	K	$\hat{\hat{h}}_{\text{VEM}}$	$\hat{\hat{h}}_{\text{GEM}}$	$\hat{\hat{h}}_{\text{ST}}$
C-2	4	0.18	7.72	409.75
C-4	2	0.10	3.78	195.90
C-9	8	1.20	59.12	287.97

Table 4: Execution times, in minutes, for three corpora, on a 3.70GHz quad core Intel Xeon Processor E5-1630V3.

6 Discussion

Inference from LDA depends heavily on the choice of hyperparameters used to fit the model. To estimate the hyperparameters, we view the analytically intractable $\hat{h} = \arg\max_h m(h)$, which is a function of the document corpus itself, as the gold standard, and we have developed a methodology for estimating $\hat{h}$. The basis for our approach is a stable method, based on a single serial tempering Markov chain, for estimating the entire marginal likelihood function $m(h)$ (up to a constant). For a given function of the parameters of the model, essentially the same method enables us to estimate the entire family of posterior expectations of the parameters as the hyperparameter varies, and this feature enables us to carry out an analysis of sensitivity of our inference with respect to the hyperparameters.

Hyperparameter selection is a simple form of model selection and we note that, generally speaking, in carrying out model selection there are two competing goals. One goal is to select the correct model, and the other goal is to select the model that "provides the best inference." These two goals are not the same. The second goal is particularly relevant when the document corpus is a real data set, i.e. the corpus is not necessarily generated from the LDA model, and we use LDA as a convenient model through which to make inference. Selection of the hyperparameter via maximization of the marginal likelihood is akin to maximum likelihood estimation and, as such, should have the standard properties of maximum likelihood estimates. We will avoid giving a technical explanation of this last fact, and instead state it informally as follows: for a corpus generated according to the LDA model indexed by h_{true}, if the corpus is large, then $\hat{h}$ is close to h_{true}. So the empirical Bayes method achieves the first goal by its very nature, and we have verified this empirically in Section 3. The evaluation in Section 5 shows (at least empirically) that the empirical Bayes method also accomplishes the second goal.

A Fully Bayes Approach to Empirical Bayes Inference For serial tempering to work, it is necessary for the grid points $h_1, \dots, h_J$ to cover $\mathcal{H}$. Unfortunately, when $\dim(\mathcal{H})$ is large, the value of J that is needed is huge, and the approach breaks down. Here we discuss an entirely different method. Although there is no inherent limitation on $\dim(\mathcal{H})$ for the method to work, we view it as useful for the case where $\dim(\mathcal{H})$ is moderate: we re-iterate our caution stated in Remark 2 of Section 2.1 that it is not advisable to use a high-dimensional h.

Suppose that $\mathcal{H}$ is a bounded hyper-rectangle. We put a uniform distribution on $\mathcal{H}$, denoted $u(h)$, and in this fully-Bayes situation the parameter is now $(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{z}, h)$. The marginal posterior distribution of h is then $\pi(h) \propto m_{\boldsymbol{w}}(h)u(h) \propto m_{\boldsymbol{w}}(h)$, and we see that $\arg\max_h m_{\boldsymbol{w}}(h) = \arg\max_h \pi(h)$. Suppose that $(\boldsymbol{\beta}^{(1)}, \boldsymbol{\theta}^{(1)}, \boldsymbol{z}^{(1)}, h^{(1)}), \dots, (\boldsymbol{\beta}^{(n)}, \boldsymbol{\theta}^{(n)}, \boldsymbol{z}^{(n)}, h^{(n)})$ is a Markov chain whose invariant distribution is the posterior distribution of $(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{z}, h)$ given $\boldsymbol{w}$ (see Wallach (2008)). From the marginal sequence $h^{(1)}, \ldots, h^{(n)}$ we may estimate $\pi(h)$ via a multivariate density estimator, and hence $\arg\max_h \pi(h)$. Call this estimate $\bar{h}$. We then use (2.1), with $\bar{h}$ as the value of h_*, to estimate $m_{\boldsymbol{w}}(h)$ in a small neighborhood of $\bar{h}$, which is all that we need in order to estimate $\arg\max_h m_{\boldsymbol{w}}(h)$. In effect, $\bar{h}$ is an initial coarse estimate of $\arg\max_h m_{\boldsymbol{w}}(h)$, and (2.1) is then used to fine-tune it. We hope to develop this idea fully in future work.

Appendix

A.1 A Likelihood Ratio Formula for the Parameters in the LDA Model

To obtain the ratio of densities formula (2.2), we note that from the hierarchical nature of the LDA model we have

$$\nu_h(\boldsymbol{\psi}) = \nu_h(\boldsymbol{\beta}, \boldsymbol{\theta}, \boldsymbol{z}) = p^{(h)}_{\boldsymbol{z}\,|\,\boldsymbol{\theta},\boldsymbol{\beta}}(\boldsymbol{z} \,|\, \boldsymbol{\theta}, \boldsymbol{\beta})\; p^{(h)}_{\boldsymbol{\theta}}(\boldsymbol{\theta})\; p^{(h)}_{\boldsymbol{\beta}}(\boldsymbol{\beta})$$

in self-explanatory notation, where $p^{(h)}_{\boldsymbol{z}\,|\,\boldsymbol{\theta},\boldsymbol{\beta}}$, $p^{(h)}_{\boldsymbol{\theta}}$, and $p^{(h)}_{\boldsymbol{\beta}}$ are given by lines 3, 2, and 1, respectively, of the LDA model. Let $n_{dj} = \sum_{i=1}^{n_d} z_{dij}$, i.e. n_{dj} is the number of words in document d that are assigned to topic j. Using the Dirichlet and multinomial distributions specified in lines 1–3 of the model, we obtain

$$\nu_h(\boldsymbol{\psi}) = \left[\prod_{d=1}^{D}\prod_{j=1}^{K} \theta_{dj}^{n_{dj}}\right]\left[\prod_{d=1}^{D}\left(\frac{\Gamma\big(\sum_{j=1}^{K}\alpha_j\big)}{\prod_{j=1}^{K}\Gamma(\alpha_j)}\prod_{j=1}^{K}\theta_{dj}^{\alpha_j-1}\right)\right]\left[\prod_{j=1}^{K}\left(\frac{\Gamma(V\eta)}{\Gamma(\eta)^V}\prod_{t=1}^{V}\beta_{jt}^{\eta-1}\right)\right]. \qquad \text{(A.1)}$$

We now apply (A.1) to ν_h and ν_{h_*} and obtain (2.2).

A.2 Proof of Theorem 1

The convergence in (2.1) holds for each fixed h; however, $\arg\max_h B_n(h)$ depends on the function $B_n(\cdot)$. Before proving Theorem 1 we provide an example to show that if f_n and f are real-valued functions, then convergence of f_n to f pointwise does not imply convergence of $\arg\max_h f_n(h)$ to $\arg\max_h f(h)$. In our example, the domain of the functions is the interval $[0, 1]$, and the functions are displayed in Figure 9. The functions f_n and f are identical on the interval $[2/n, 1]$. Clearly $f_n(h) \to f(h)$ for each $h \in [0, 1]$, but $\arg\max_h f_n(h) = 1/n$ while $\arg\max_h f(h) = .9$.

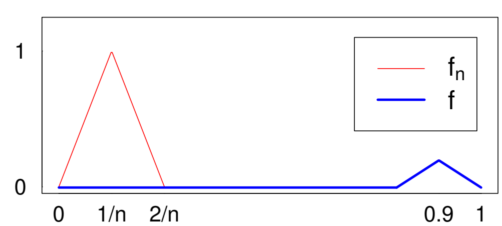


Figure 9: Non-convergence of the argmax.

Theorem 1 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 For each n, the maximizer of $B_n(\cdot)$ is unique (thus we can talk about $\arg\max_h B_n(\cdot)$).

A5 The point $h_* = (\eta^*, \boldsymbol{\alpha}^*)$ satisfies $2\eta - \eta^* > 0$ and $2\alpha_j - \alpha_j^* > 0$, $j = 1, \dots, K$ for all $h = (\eta, \boldsymbol{\alpha}) \in \mathcal{H}$.

A6 The marginal likelihood 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.

Proof of Part 1 Note the following:

- In Section 4 we showed that the ACGS is uniformly ergodic. So in particular, it is Harris ergodic.
- The LDA model is an exponential family with parameter (η, α) (Section 3.3 of Wainwright and Jordan (2008)).
- In their Remark 1, Doss and Park (2016) show that if $\{\nu_h,\ h \in \Omega\}$ is an exponential family, where Ω is the natural parameter space, and if $\mathcal{H}$ is a compact subset of the interior of Ω, then $\int \sup_{h \in \mathcal{H}} (\nu_h / \nu_{h_*}) \, d\nu_{h_*, \boldsymbol{w}} < \infty$. (In empirical process theory, finiteness of this integral is the main condition that is needed to obtain uniformity in the Law of Large Numbers.)
- In the context of the present situation, Theorem 3 of Doss and Park (2016) states that under Harris ergodicity of the sequence $\boldsymbol{\psi}_1, \boldsymbol{\psi}_2, \dots$ and finiteness of $\int \sup_{h \in \mathcal{H}} (\nu_h / \nu_{h_*}) \, d\nu_{h_*, \boldsymbol{w}}$, the convergence in (2.1) is uniform on $\mathcal{H}$, i.e. Condition C3 of Section 2.1 holds.
- Suppose that $f_n,\ n = 1, 2, \dots$ and f are real-valued functions defined on a compact subset X of Euclidean space. Suppose further that f is continuous and that each of $f_n,\ n = 1, 2, \dots$ and f has a unique maximizer. Under these conditions, uniform convergence of f_n to f on X implies $\arg\max_{x \in \mathsf{X}} f_n(x) \to \arg\max_{x \in \mathsf{X}} f(x)$. Verification of this fact is routine. A detailed proof is given in Lemma 1 of Doss and Park (2016).

Combining these facts, we see that under A1–A4, $\arg\max_h B_n(h) \to \arg\max_h m(h)$ with probability one (Assumptions A5 and A6 are not needed for Part 1 of the theorem.) $\square$

Proof of Part 2 Theorem 4 of Doss and Park (2016) asserts the asymptotic normality stated in Part 2 of Theorem 1 under A1–A4 and A6, the condition

$$\text{for every } h \in \mathcal{H} \text{ there exists } \epsilon > 0 \text{ such that } \int \|\nabla_h (\nu_h / \nu_{h_*})\|^{2+\epsilon} \, d\nu_{h_*, \boldsymbol{w}} < \infty, \tag{A.2}$$

where $\|\cdot\|$ is the Euclidean norm in $\mathbb{R}^p$, and the condition that the Markov chain used is geometrically ergodic. Using standard calculus, we can check that if $2\eta - \eta^* > 0$ and $2\alpha_j - \alpha_j^* > 0$, $j = 1, \dots, K$, then for sufficiently small ϵ the integral in (A.2) is finite. Thus, Condition A5 implies (A.2). As mentioned in the proof of Part 1 of the theorem, the ACGS is uniformly ergodic; so in particular, it is geometrically ergodic. Thus, we have established the asymptotic normality stated in Part 2 of the theorem under A1–A6. $\square$

Proof of Part 3 The variance matrix Σ in Part 2 is analytically intractable, but fortunately is easy to estimate via the method of batching, as follows. For $j = 1, \dots, J$, let $h^{[j]}$ be the estimate of the argmax produced from batch j, and let $h^{[]}$ be the estimate of the argmax produced from the entire sequence. The batch-based estimate is $\widehat{\Sigma}_n = (n/J)\bigl\{[1/(J-1)] \sum_{j=1}^J \bigl(h^{[j]} - h^{[]}\bigr)\bigl(h^{[j]} - h^{[]}\bigr)^\top\bigr\}$. (The quantity inside the braces is essentially the sample covariance matrix of $h^{[1]}, \dots, h^{[J]}$, except that we use $h^{[]}$ instead of the average of $h^{[1]}, \dots, h^{[J]}$ as the centering value; and the term n/J is the number of samples per batch.) 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). Invertibility of $\widehat{\Sigma}_n$ for large n follows from positive definiteness of Σ and the convergence $\widehat{\Sigma}_n \xrightarrow{\text{a.s.}} \Sigma$; in fact we have $\widehat{\Sigma}_n^{-1} \xrightarrow{\text{a.s.}} \Sigma^{-1}$. Therefore, applying Part 2, we get

$$n^{1/2}\bigl(\arg\max_h B_n(h) - \arg\max_h m(h)\bigr)\widehat{\Sigma}_n^{-1} n^{1/2}\bigl(\arg\max_h B_n(h) - \arg\max_h m(h)\bigr)^\top \xrightarrow{d} \chi_p^2,$$

which establishes the statement regarding the ellipse. $\square$

Proof of Theorem 1 for the Serial Tempering Chain The main change in the proof is that the requirement (A.2) is replaced with

$$\text{for every } h \in \mathcal{H} \text{ there exists } \epsilon > 0 \text{ such that } \int \|\nabla_h(\nu_h/\nu_\zeta)\|^{2+\epsilon}\, df_\zeta < \infty, \tag{A.3}$$

where $\nu_\zeta = (1/J)\sum_{j=1}^J \nu_j(\boldsymbol{\psi}_i)/\zeta_j$, and f_ζ is given by (2.9). It is easy to see that (A.3) is satisfied if the stipulation on h_* given by (2.3) holds for $h^{(j)}$ for at least one index j. $\square$

A.3 Proof of Validity of the Confidence Band for $\{I(h),\, h \in \mathcal{H}\}$

In addition to assuming that $J \to \infty$ and $n/J \to \infty$, we will need the following conditions:

A7 The stipulation on h_* given by (2.3) holds for $h^{(j)}$ for at least one index j.

A8 The function g satisfies the moment condition

$$\text{for every } h \in \mathcal{H} \text{ there exists } \epsilon > 0 \text{ such that } \int \biggl(g\frac{\nu_h}{\nu_\zeta}\biggr)^{2+\epsilon} df_\zeta < \infty.$$

Note that A8 is automatically satisfied if A7 holds and g is bounded (for example if g is an indicator function, as in Section 2.4). In the following, we will assume Conditions A1, A7, and A8. The heart of the proof is the assertion that $\sup_{h\in\mathcal{H}} n^{1/2}\bigl|\hat{I}_\zeta^{\text{st}}(h) - I(h)\bigr|$ has a limiting distribution as $n \to \infty$, and we show this in three steps:

1. We observe that for each h, $n^{1/2}\bigl(\hat{I}_\zeta^{\text{st}}(h) - I(h)\bigr)$ has an asymptotic normal distribution.
2. We show that more can be said, and that the stochastic process $\bigl\{n^{1/2}\bigl(\hat{I}_\zeta^{\text{st}}(h) - I(h)\bigr),\, h \in \mathcal{H}\bigr\}$ converges in distribution to a mean-zero Gaussian process indexed by h.
3. We conclude from Step 2 that $\sup_{h\in\mathcal{H}} n^{1/2}\bigl|\hat{I}_\zeta^{\text{st}}(h) - I(h)\bigr|$ has a limiting distribution as $n \to \infty$.

We now provide the details.

1. Note that $\hat{I}_\zeta^{\text{st}}(h)$, defined in (2.14), is a ratio of $\widehat{M_\zeta}(h)$ and $\widehat{U}_\zeta(h)$, which are given by (2.10) and (2.12), respectively. Each of these is an average of a function of $\boldsymbol{\psi}_1, \ldots, \boldsymbol{\psi}_n$, so we have a bivariate central limit theorem, as follows. For economy of notation, let $U_i^{(h)}$ be the summands in (2.12) and let $M_i^{(h)}$ be the summands in (2.10). We have

$$n^{1/2} \begin{pmatrix} \dfrac{1}{n} \displaystyle\sum_{i=1}^n U_i^{(h)} - \dfrac{m(h)}{c_\zeta/J} \int g \, d\nu_{h,\boldsymbol{w}} \\ \dfrac{1}{n} \displaystyle\sum_{i=1}^n M_i^{(h)} - \dfrac{m(h)}{c_\zeta/J} \end{pmatrix} \xrightarrow{d} \mathcal{N}_2(0, \Sigma_h),$$

where Σ_h is a covariance matrix. (If the $\boldsymbol{\psi}$'s were an iid sequence, then Σ_h would be simply the covariance matrix of the pair $(U_1^{(h)}, M_1^{(h)})$; however, in the present situation, Σ_h is the more complicated covariance matrix that arises in the Markov chain central limit theorem.) Therefore, by the delta method applied to the function $\varphi\colon \mathbb{R}^2 \to \mathbb{R}$ defined by $\varphi(u, m) = u/m$, we have

$$n^{1/2} \left(\frac{\sum_{i=1}^n U_i^{(h)}}{\sum_{i=1}^n M_i^{(h)}} - \int g \, d\nu_{h,\boldsymbol{w}} \right) \xrightarrow{d} \mathcal{N}\bigl(0, (\nabla\varphi)^\top \Sigma_h \nabla\varphi\bigr), \tag{A.4}$$

where the gradient $\nabla\varphi$ is evaluated at $\bigl((1/n)\sum_{i=1}^n U_i^{(h)}, (1/n)\sum_{i=1}^n M_i^{(h)}\bigr)$. Now note that the quantity to the left of the "$\xrightarrow{d}$" sign in (A.4) is precisely $n^{1/2}\bigl(\hat{I}_\zeta^{\text{st}}(h) - I(h)\bigr)$.

2. To extend convergence in distribution for each fixed h to convergence as a stochastic process, we use Part 4 of Theorem 6 of Doss and Park (2016). Because we assume Conditions A1, A7, and A8 and because the distributions of the latent parameters in the LDA model form an exponential family, the regularity conditions for that theorem are satisfied, and we conclude that $n^{1/2}\bigl(\hat{I}_\zeta^{\text{st}}(\cdot) - I(\cdot)\bigr) \xrightarrow{d} G(\cdot)$, where $G(\cdot)$ is a mean-zero Gaussian process indexed by h. Here, convergence in distribution takes place in $C(\mathcal{H})$, the space of continuous real-valued functions defined on $\mathcal{H}$, endowed with the sup-norm topology.

3. The map $T\colon C(\mathcal{H}) \to [0, 1]$ defined by $T(f) = \sup_{h\in\mathcal{H}} |f(h)|$ is continuous, so from Step 2 we conclude that $\sup_{h\in\mathcal{H}} n^{1/2}\bigl|\hat{I}_\zeta^{\text{st}}(h) - I(h)\bigr| \xrightarrow{d} \sup_{h\in\mathcal{H}} |G(h)|$.

Substitution of the $\mathcal{S}_j$'s for the S_j's is valid under the assumption that $J \to \infty$, convergence in probability of $\mathcal{S}_{[.95J]}$ to $c_{.95}$ is a consequence of the condition $n/J \to \infty$, and the validity of the bands now follows. The literature's recommendation of $J = n^{1/2}$ is made in the different context of estimating the variance of an average, not for forming globally-valid confidence bands; nevertheless, in our experience this choice works well also in the present situation.

Acknowledgments

We are grateful to three referees for their very helpful constructive criticism.

References

Andrieu, C., De Freitas, N., Doucet, A. and Jordan, M. I. (2003). An introduction to MCMC for machine learning. *Machine Learning* **50** 5–43.

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.

Bergstra, J. and Bengio, Y. (2012). Random search for hyper-parameter optimization. *Journal of Machine Learning Research* **13** 281–305.

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

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.

Chen, Z. and Doss, H. (2017). Inference for the number of topics in the latent Dirichlet allocation model via Bayesian mixture modelling. (Under revision for *Journal of Computational and Graphical Statistics*) .

Diaconis, P., Khare, K. and Saloff-Coste, L. (2008). Gibbs sampling, exponential families and orthogonal polynomials (with discussion). *Statistical Science* **23** 151–178.

Doss, H. and George, C. P. (2017). A grouped Gibbs sampler for parallel computing in the latent Dirichlet allocation model. Tech. rep., Department of Statistics, University of Florida.

Doss, H. and Park, Y. (2016). An MCMC approach to empirical Bayes inference and Bayesian sensitivity analysis via empirical processes. *Annals of Statistics* (to appear) .

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.

Fort, G. and Moulines, E. (2003). Convergence of the Monte Carlo expectation maximization for curved exponential families. *The Annals of Statistics* **31** 1220–1259.

Freedman, D. (1999). Wald Lecture: On the Bernstein-von Mises theorem with infinite-dimensional parameters. *The Annals of Statistics* **27** 1119–1141.

Fuentes, C., Gopal, V., Casella, G., George, C. P., Glenn, T. C., Wilson, J. N. and Gader, P. D. (2011). Product partition models for Dirichlet allocation. Tech. rep., Department of Computer and Information Science and Engineering, University of Florida.

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. (2017). Supplement to "Principled selection of hyperparameters in the latent Dirichlet allocation model".

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.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika* **57** 97–109.

Hobert, J. P. and Casella, G. (1996). The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *Journal of the American Statistical Association* **91** 1461–1473.

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.

Liu, J. S., Wong, W. H. and Kong, A. (1994). Covariance structure of the Gibbs sampler with applications to the comparisons of estimators and augmentation schemes. *Biometrika* **81** 27–40.

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

McCallum, A. K. (2002). MALLET: A machine learning for language toolkit. `http://mallet.cs.umass.edu`.

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

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

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.

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.

Wainwright, M. J. and Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning* **1** 1–305.

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.

Wei, G. C. G. and Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association* **85** 699–704.