On Latent Dirichlet Allocation Priors and Variational Bayes

Abstract

Latent Dirichlet Allocation (LDA) is a successful model capable of explaining large document collections using a small number of topics. Several methods exist to estimate it, one of the earliest ones being Variational Bayes (VB). While the core of the model has been stable since its inception ca. 2003, its "border", namely the treatment of the hyperparameters, has been widely debated, since its importance in practice has been acknowledged. We study here in a systematic way the impact of the choice of LDA model priors on the Variational Bayes algorithm, then propose a new "natural" prior obtained by simply applying the principle of Bayesian conjugacy which underlies the design of the rest of the model. This leads to new, efficient VB update rules, which we discuss and experiment on a few document collections.

1 INTRODUCTION

One of the earliest applications of the Variational Bayes method in machine learning was in the domain of topic modelling, with the LDA model, where potentially large document collections are summarised by a small set of topics, each document being seen as a sparse mixture of topics and each topic as a sparse mixture of words. Part of the success of the method originates in the simplicity of its algorithm, which relies on a small set of well founded update rules. While the rules associated to the core of the LDA model are stable and consensual, those dealing with the "border" of the model have often been debated, and no clear consensus has emerged. In the Bayesian network terminology, the problem is that of the choice of priors, the importance of which has often been acknowledged. We propose here a systematic study of this problem, as well as yet another variant of LDA obtained by simple application of its underlying design principle, which can be informally summarised as: in the absence of better criterion, try conjugate priors.

Preliminary work. Under review by AISTATS 2016. Do not distribute.

2 RELATED WORK

The LDA model [3] has been extended in many different ways to address different problems. Information about document authorship has, for example, been added to the model in [9], whereas the integration of correlations between topics in the model has been explored in [2]; more recently, [4] describes an extension to model multilingual unaligned collections. Other extensions have focused on streaming or online versions of the model: [13] focuses on efficient methods for inference in streaming collections, whereas [12] introduces a new model for text streams based on transition probabilities between topics of successive documents and [7] proposes an online variational Bayes algorithm for LDA based on mini-batches.

Inference in LDA is usually performed through variational Bayes (as proposed in the original LDA paper [3]) or Gibbs sampling (as proposed in [6]). Both methods have been extensively studied and collapsed versions, resulting in faster inference, have been proposed [10, 8]. This last study, besides comparing the different approaches to inference in LDA, introduces gamma priors on the hyperparameters[1] which yield more stable models (in the sense that the difference between inference methods disappears when the priors are used). [11] goes one step further by assessing the importance of priors on both the document-topic and word-topic distributions. The priors considered are asymmetric Dirichlet priors and the authors show that the use of such asymmetric priors has substantial advantages over the use of symmetric priors, especially for the document-topic distribution.

We follow here a similar approach but consider a complete family of priors, namely the conjugate prior to the Dirichlet distribution, which we introduce in the next section. As for asymmetric Dirichlet priors, the use of a conjugate prior leads to efficient inference methods; it also yields a broader family of distributions that provides a complete Bayesian treatment (and re-interpretation) of the variational Bayes inference in the LDA model.

[1] A similar prior is used in [5] for Indian Buffet Processes.

3 A CONJUGATE PRIOR FOR THE DIRICHLET DISTRIBUTION

It is well known that the exponential family of distributions satisfies important stability properties, which lead to elegant, analytical expressions in Bayesian inference, hence its massive use in Bayesian models. One property is that any distribution in that family has a natural conjugate also in that family, and there is a systematic procedure to compute it. For example, the Dirichlet distribution can be derived in this way as a conjugate to the multinomial distribution, and they are both in the exponential family. In the sequel, we are interested in a conjugate to the Dirichlet itself, which is easy to derive by application of the same systematic procedure. We name it here the **Boojum** distribution[2], for lack of a better name. It is a distribution over the positive orthant $\mathbb{R}_+^N$ (the parameter space of the N-dimensional Dirichlet), and has two parameters $m, \boldsymbol{\tau}$ where $m \in \mathbb{R}$ is a scalar and $\boldsymbol{\tau} \in \mathbb{R}^N$ is a vector (as a general rule, the parameter space of the conjugate has dimension one plus the dimension of the parameter space of the original distribution). It is defined, for $\boldsymbol{x} \in \mathbb{R}_+^N$ by

$$\textbf{Boojum}(\boldsymbol{x}; m, \boldsymbol{\tau}) \triangleq \frac{1}{Z(m,\boldsymbol{\tau})} \mathcal{B}(\boldsymbol{x})^m \exp -\boldsymbol{\tau}\boldsymbol{x}$$

where $\mathcal{B}$ denotes the multivariate beta function $\mathcal{B}(\boldsymbol{x}) \triangleq \frac{\prod_n \Gamma(x_n)}{\Gamma(\sum_n x_n)}$ which is also the normalising constant of the Dirichlet distribution with parameter $\boldsymbol{x}$, and $Z(m, \boldsymbol{\tau})$ is the normalising constant of the **Boojum** distribution itself, defined by

$$Z(m, \boldsymbol{\tau}) \triangleq \int_{\boldsymbol{x} \in \mathbb{R}_+^N} \mathcal{B}(\boldsymbol{x})^m \exp -\boldsymbol{\tau}\boldsymbol{x} \, \mathbf{d}\boldsymbol{x}$$

The expression $\boldsymbol{\tau}\boldsymbol{x}$ in the definition denotes the scalar product of the two vectors. Of course, for the distribution to be proper, the normalising constant must be finite, which is not always the case. Although no analytical formula is known for $Z(m, \boldsymbol{\tau})$, one exists for its finiteness:

Proposition 1. *The distribution* $\textbf{Boojum}(m, \boldsymbol{\tau})$ *is proper, i.e.* $Z(m, \boldsymbol{\tau}) < \infty$*, if and only if*[3]

$$\begin{array}{l} \forall n \in N \ \tau_n > 0 \quad \textit{and} \quad m < 1 \quad \textit{and} \\ (m \geq 0 \quad \textit{or} \quad \sum_{n \in N} \exp -\frac{\tau_n}{|m|} < 1) \end{array}$$

The proof of this result is not trivial, and is available on demand from the authors. On the other hand, conjugacy with the Dirichlet distribution, as expressed by the following property, is quite straightforward.

[2] a.k.a. the Snark distribution, because the Snark is a Boojum, you see.

[3] For any whole number N, we use the shorthand $n \in N$ to mean $n \in \{1 \dots N\}$

Proposition 2. *Let* $(\boldsymbol{y}_p)_{p \in P}$ *be a family of random variables over the simplex of* $\mathbb{R}_+^N$*, assumed mutually independent given a random variable* $\boldsymbol{x}$ *over* $\mathbb{R}_+^N$*.*

$$\begin{array}{ll} \textit{Prior:} & \boldsymbol{x} \sim \textbf{Boojum}(m, \boldsymbol{\tau}) \\ \textit{Observation:} & \forall p \in P \ \boldsymbol{y}_p | \boldsymbol{x} \sim \textbf{Dirichlet}(\boldsymbol{x}) \\ \implies \textit{Posterior:} & \boldsymbol{x} | (\boldsymbol{y}_p)_{p \in P} \sim \textbf{Boojum}(m', \boldsymbol{\tau}') \end{array}$$
$$\textit{where } m' = m - |P| \textit{ and } \boldsymbol{\tau}' = \boldsymbol{\tau} - \textstyle\sum_{p \in P} \log \boldsymbol{y}_p$$

This holds whenever the prior is proper, in which case so is the posterior.

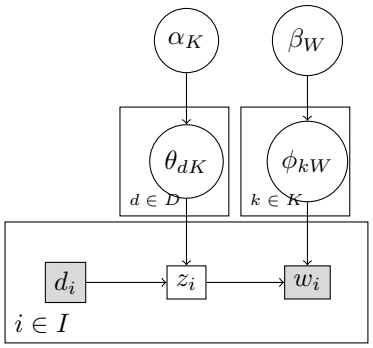
4 LDA REVISITED

Given a number D of documents, W of words (or terms), K of topics and I of occurrences, a realisation consists of the following random variables[4]: (i) a tuple $(d_i, z_i, w_i)_{i \in I}$, where for each $i \in I$, the discrete objects $d_i \in D, z_i \in K, w_i \in W$ denote, respectively, the document, the topic and the word associated with occurrence i; (ii) two stochastic matrices θ_{DK} (of dimension $D \times K$) and ϕ_{KW} (of dimension $K \times W$) which characterise each document d as the distribution of topics θ_{dK} and each topic k as the distribution of words ϕ_{kW}; (iii) two positive vectors α_K and β_W which characterise the topics and words of the whole collection. The full graphical representation of the LDA model is given in Figure 1.

The space of realisations is that of complete collections of documents: this explains why the collection-level variables α_K and β_W are random, and not parameters. Besides the known size parameters D, W, K, I, the model has two possibly unknown parameters $p^{(K)}$ and $p^{(W)}$, which are distributions for α_K and β_W, respectively. For the sake of symmetry, the occurrence-document vector d_I is considered a random variable, but it is assumed always observed and independent of all the rest, so it could as well have been treated as a known parameter. All probability expressions are conditioned upon it. Conditioned to the observation of the occurrence-word vector w_I, the log probability is given, up to some additive constant depending only on d_I, w_I, by (1).

Note that, at this point, we make no assumption on the distributions $p^{(K)}$ and $p^{(W)}$. Our main contribution is precisely in studying the impact of the choice of these distributions. We show that an appropriate choice leads to a new formulation of the variational approximation of LDA, which we investigate.

[4] In the sequel, we take the convention of not distinguishing the vectors by typesetting them in boldface, but rather by systematically recalling their index ranges as subscripts.



$$
\begin{array}{l}
p(\alpha_K \beta_W \theta_{DK} \phi_{KW} z_I w_I | d_I) = \\
\qquad p(\alpha_K) p(\beta_W) \\
\qquad \prod_{d \in D} p(\theta_{dK} | \alpha_K) \prod_{k \in K} p(\theta_{kW} | \beta_W) \\
\qquad \prod_{i \in I} p(z_i | d_i \theta_{DK}) p(w_i | z_i \phi_{KW})
\end{array}
$$

$$
\begin{array}{rrcl}
\alpha_K & \sim & p^{(K)} & \beta_W \sim p^{(W)} \\
\forall d \in D & \theta_{dK} | \alpha_K & \sim & \mathbf{Dirichlet}(\alpha_K) \\
\forall k \in K & \phi_{kW} | \beta_W & \sim & \mathbf{Dirichlet}(\beta_W) \\
\forall i \in I & z_i | d_i \theta_{DK} & \sim & \mathbf{Cat}(\theta_{d_i K}) \\
\forall i \in I & w_i | z_i \phi_{KW} & \sim & \mathbf{Cat}(\phi_{z_i W})
\end{array}
$$

$$
\begin{aligned}
\log p(\alpha_K \beta_W \theta_{DK} \phi_{KW} z_I | w_I d_I) &= \log p^{(K)}(\alpha_K) + \log p^{(W)}(\beta_W) - D \log \mathcal{B}(\alpha_K) - K \log \mathcal{B}(\beta_W) \qquad (1) \\
&+ \sum_{dk} (\alpha_k - 1) \log \theta_{dk} + \sum_{kw} (\beta_w - 1) \log \phi_{kw} + \sum_{ik} \mathbb{I}[z_i = k] (\log \theta_{d_i k} + \log \phi_{k w_i})
\end{aligned}
$$

Figure 1: The full Bayesian network of the LDA model and the decomposition of its joint distribution q.

5 PRIORS FOR VARIATIONAL BAYES LDA

The posterior probability given by (1) does not admit an analytical expression. The VB method tries to approximate it, by projecting it onto a simpler space $\mathcal{C}$ of probability distributions, namely that of distributions q of the form

$$
\begin{array}{l}
q(\alpha_K \beta_W \theta_{DK} \phi_{KW} z_I) = q^{(K)}(\alpha_K) q^{(W)}(\beta_W) \\
\qquad \prod_{d \in D} q_d(\theta_{dK}) \prod_{k \in K} q_k(\phi_{kW}) \prod_{i \in I} q_i(z_i)
\end{array} \qquad (2)
$$

Hence, in the variational model, the variables $\alpha_K, \beta_W, \{\theta_{dK}\}_{d \in D}, \{\phi_{kW}\}_{k \in K}, \{z_i\}_{i \in I}$ are assumed independent.

5.1 The VB method

The VB method solves the following optimisation problem:

$$
q^* \quad = \quad \arg\min_{q \in \mathcal{C}} \mathbb{K}(q, \tilde{p}) \quad \text{where} \quad \tilde{p} = p | w_I d_I \qquad (3)
$$

Hence q^* is the distribution of class $\mathcal{C}$, i.e. decomposable according to (2), which is closest, for the KL-divergence, to the posterior distribution $\tilde{p}$ given by (1). VB computes an estimate $\dot{q}$ of q^* in $\mathcal{C}$. It proceeds by simple coordinate descent, along the components $\dot{q}_X$, one for each independent variable X in (2). The update, at each step of the coordinate descent focussing on variable X, is given by

$$
\dot{q}_X(x) \quad \leftarrow\propto \quad \exp \mathbb{E}_{y \sim \dot{q}_{\neg X}} [\log \tilde{p}(x, y)] \qquad (4)
$$

where the expectation is taken over the set $\neg X$ of all the independent variables in (2) other than X. Hence, the estimate $\dot{q}$ converges to q^*, or at least to a local minimum of $\mathbb{K}(q, \tilde{p})$, taken to be an approximation of $\tilde{p}$. When the right-hand side in (4) does not have an analytical expression, VB uses an additional approximation, by *externally* constraining the estimate $\dot{q}_X$ to be in a specific class of distributions. Typically, one may want to constrain $\dot{q}_X$ to be a Dirac distribution with some parameter $\dot{\pi}_X$ and the update becomes

$$
\dot{\pi}_X \quad \leftarrow \quad \arg\max_x \mathbb{E}_{y \sim \dot{q}_{\neg X}} [\log \tilde{p}(x, y)] \qquad (5)
$$

The right-hand side of (5) is the mode of the right-hand side distribution in (4), since the closest Dirac approximation of any distribution is the Dirac at its mode.

In the case of the LDA model, it is well known that the choice of conjugate (Dirichlet) priors for variables θ_{DK} and ϕ_{KW} lead to update rules (4) which *naturally* constrain the variational distributions to the following forms[5]:

$$
\begin{array}{ll}
\forall d \in D & \dot{q}_d(\theta_K) = \mathbf{Dirichlet}(\theta_K; \dot{\alpha}_{dK}) \\
\forall k \in K & \dot{q}_k(\phi_W) = \mathbf{Dirichlet}(\phi_W; \dot{\beta}_{kW})
\end{array}
$$

Furthermore, $\dot{q}_i$ for $i \in I$ is by construction a categorical distribution, with some parameter μ_{iK}. However, the resulting updates do not differentiate between the distributions $\dot{q}_i$ where d_i, w_i are identical, i.e. multiple occurrences of the same word in the same document. Hence, the variational parameter μ_{IK} can be replaced by a parameter μ_{DWK} such that $\mu_{iK} = \mu_{d_i w_i K}$ for all $i \in I$. Likewise, the observation d_I, w_I can be summarised by the sufficient statistics n_{DW}, which is the document-word count matrix:

$$
n_{dw} \quad \triangleq \quad |\{i \in I | d_i = d, w_i = w\}|
$$

[5] To avoid a proliferation of Greek letters, we denote the variational parameters with the same letter as the corresponding model variables, decorated with a dot.

Finally, the updates given by (4) applied to the LDA model are summarised in Figure 2, where the different updates are given labels (in brackets under the arrow) for reference purpose. For clarity sake, we have introduced the intermediate variational quantities $\bar{\alpha}_{DK}$ and $\bar{\beta}_{KW}$ defined by

$$\bar{\alpha}_{dk} \triangleq \mathbb{E}_{\theta_K \sim \dot{q}_d}[-\log \theta_k] \quad \bar{\beta}_{kw} \triangleq \mathbb{E}_{\phi_W \sim \dot{q}_k}[-\log \phi_w]$$

which have a simple analytical expression. One recognises in Figure 2 the standard updates of the LDA model, except maybe for the bottom two **[d]** and **[w]**, discussed below.

5.2 Treatment of the parameters $p^{(K)}$ and $p^{(W)}$

Let's first justify the update **[d]** of $\dot{q}^{(K)}$ (the same applies to update **[w]** of $\dot{q}^{(W)}$). By eliminating from (1) all the terms which do not involve α_K, hence contribute only a multiplicative constant, and taking expectations on the others, Equation (4) becomes

$$\begin{aligned}
\dot{q}^{(K)}(\alpha_K) \quad &\leftarrow\propto \quad \exp \log p^{(K)}(\alpha_K) - D \log \mathcal{B}(\alpha_K) \\
&+ \quad \sum_d \mathbb{E}_{\theta_{dK} \sim \dot{q}_d}[\sum_k \alpha_k \log \theta_{dk}] \\
&= \quad p^{(K)}(\alpha_K) \mathcal{B}(\alpha_K)^{-D} \exp \sum_{dk} -\alpha_k \bar{\alpha}_{dk} \\
&\propto \quad p^{(K)}(\alpha_K) \mathbf{Boojum}(\alpha_K; -D, \sum_d \bar{\alpha}_{dK})
\end{aligned}$$

where **Boojum** is the conjugate distribution of the Dirichlet in the exponential family introduced in Section 3. Note that this formula, involving the **Boojum** distribution, holds in general, whatever the choice of priors $p^{(K)}, p^{(W)}$. To make the updates of $\dot{q}^{(K)}, \dot{q}^{(W)}$ concrete, we now need to proceed with that choice. Since the two variables α_K and β_W are parameters of Dirichlet distributions, on θ_{DK} and ϕ_{KW} respectively, we naturally choose their modelling distributions $p^{(K)}, p^{(W)}$ in the conjugate class of Dirichlet, namely **Boojum**. This *naturally* ensures that the corresponding variational distributions $\dot{q}^{(K)}, \dot{q}^{(W)}$ are also in that class.

We first show that the so called EM type 2 hyperparameter estimation, which has been proposed for VB LDA, is in fact a special case of this approach. Indeed, EM in general is known to be a special case of VB, and what we give here is just the pure VB presentation of the method, leading to the same updates. In VB, the method amounts to choosing an "uninformative" $p^{(K)}$, i.e. $p^{(K)} \propto 1$, which is also the improper distribution $\mathbf{Boojum}(0, 0)$. Reporting in **[d]**, this *naturally* constrains the distribution $\dot{q}^{(K)}$ to be equal to $\mathbf{Boojum}(-D, \sum_d \bar{\alpha}_{dK})$. However, update **[D]** requires the computation of its expectation, which is intractable. Instead, $\dot{q}^{(K)}$ is *externally* constrained to be a Dirac distribution, hence we can apply Equation (5):

$$\begin{aligned}
p^{(K)} \propto 1 \qquad & \dot{q}^{(K)} = \mathbf{Dirac}(\dot{\eta}_K) \\
\dot{\alpha}_{dk} \quad \leftarrow \quad & \dot{\eta}_k + \textstyle\sum_w n_{dw} \mu_{dwk} \\
\dot{\eta}_K \quad \leftarrow \quad & \arg\max \mathbf{Boojum}(-D, \textstyle\sum_d \bar{\alpha}_{dK})
\end{aligned}$$

The arg max expression above, computing the mode of the **Boojum** distribution, can be simplified by introducing function $\mathbf{\Phi}$ defined for any vector u_N by

$$\mathbf{\Phi}(u_N) \quad \triangleq \quad \arg\min_{x_N} \log \mathcal{B}(x_N) + u_N x_N$$

The resulting update rules **[D]** and **[d]** are given in Figure 3. By constraining $\dot{q}^{(K)}$ to be Dirac, the computation of its expectation becomes trivial in update **[D]**, but at the price of its own update **[d]**, which now requires the computation of the mode of the distribution thus approximated by the Dirac. In other words, we have traded a complex integration for a complex optimisation, but the latter is still more tractable than the former.

Let's now detail an alternative approach, which we call here "full conjugacy", on the topic-word side for the sake of presentation (but it works just as well on the document-topic side). We still choose $p^{(W)}$ in the **Boojum** class as before, but we don't force its parameter (m, ζ_W) to be 0 as above. Hence, $\dot{q}^{(W)}$ is also in the same class, and let $(\dot{m}, \dot{\zeta}_W)$ be its parameter. The updates are then straightforward to derive. The key observation here is that parameter $\dot{m}$ is assigned the expression $m - K$, which never changes in subsequent updates. Furthermore, if we choose $m = K$, that expression is null, i.e. $\dot{m} = 0$. This is particularly helpful, because, then, the intractable $\mathbf{Boojum}(\dot{m}, \dot{\zeta}_W)$ distribution becomes a simple Exponential distribution[6] with rate $\dot{\zeta}_W$, the expectation of which is trivial to compute.

$$\begin{aligned}
p^{(W)} = \mathbf{Boojum}(K, \zeta_W) \;\; & \dot{q}^{(W)} = \mathbf{Expon}(\dot{\zeta}_W) \\
\dot{\beta}_{kw} \quad \leftarrow \quad & \dot{\zeta}_w^{-1} + \textstyle\sum_d n_{dw} \mu_{dwk} \\
\dot{\zeta}_w \quad \leftarrow \quad & \zeta_w + \textstyle\sum_k \dot{\beta}_{kw}
\end{aligned}$$

The resulting update rules **[W]** and **[w]** are given in Figure 3. Note that parameter ζ_W is still free: it could be set to 0, or preferably to some small machine value (the same for all components) to avoid numerical instability in the inversion of the rate defining the expectation of the Exponential in rule **[W]**.

5.3 Discussion

In both variants of LDA, the priors are improper: in the EM type 2 estimation, because it is "uniform" on a space of infinite measure (the positive orthant), and

[6] We mean here a multivariate Exponential, product of independent scalar Exponentials.

$$\mu_{dwk} \underset{[\mathbf{K}]}{\leftarrow\propto} \exp -(\bar{\alpha}_{dk} + \bar{\beta}_{kw})$$

$$\begin{array}{llll} \bar{\alpha}_{dk} & \leftarrow & \Psi(\sum_{k'} \dot{\alpha}_{dk'}) - \Psi(\dot{\alpha}_{dk}) & \bar{\beta}_{kw} \leftarrow \Psi(\sum_{w'} \dot{\beta}_{kw'}) - \Psi(\dot{\beta}_{kw}) \\ \dot{\alpha}_{dk} & \underset{[\mathbf{D}]}{\leftarrow} & \mathbb{E}[\dot{q}^{(K)}]_k + \sum_w n_{dw}\mu_{dwk} & \dot{\beta}_{kw} \underset{[\mathbf{W}]}{\leftarrow} \mathbb{E}[\dot{q}^{(W)}]_w + \sum_d n_{dw}\mu_{dwk} \\ \dot{q}^{(K)} & \underset{[\mathbf{d}]}{\leftarrow\propto} & p^{(K)}\mathbf{Boojum}(-D, \sum_d \bar{\alpha}_{dK}) & \dot{q}^{(W)} \underset{[\mathbf{w}]}{\leftarrow\propto} p^{(W)}\mathbf{Boojum}(-K, \sum_k \bar{\beta}_{kW}) \end{array}$$

Figure 2: Generic variational updates for the LDA model.

$$\begin{array}{llll} \dot{\alpha}_{dk} & \underset{[\mathbf{D}]}{\leftarrow} & \dot{\eta}_k + \sum_w n_{dw}\mu_{dwk} & \dot{\beta}_{kw} \underset{[\mathbf{W}]}{\leftarrow} \dot{\zeta}_w^{-1} + \sum_d n_{dw}\mu_{dwk} \\ \dot{\eta}_K & \underset{[\mathbf{d}]}{\leftarrow} & \mathbf{\Phi}(\frac{1}{D}\sum_d \bar{\alpha}_{dK}) & \dot{\zeta}_w \underset{[\mathbf{w}]}{\leftarrow} \zeta_w + \sum_k \bar{\beta}_{kw} \end{array}$$

Figure 3: Updates $[\mathbf{D}], [\mathbf{W}], [\mathbf{d}], [\mathbf{w}]$ in our variants: one is equivalent to the EM type 2 estimation method and the other is original. For the sake of presentation, we apply the former to the document-topic side and the latter to the topic-word side, but they are interchangeable, or the same variant could be applied to both sides.

in the full conjugacy method, by Proposition 1, since K is obviously not less than 1.

Using an improper prior is not a problem so long as the posterior is guaranteed to remain proper. While we cannot check that on the true posterior, we can at least check that its approximation computed by the VB method is proper. For the full conjugacy method, $\dot{q}^{(W)}$ is an Exponential, and obviously proper: it is easy to show that its parameter $\dot{\zeta}^{(W)}$ always remain strictly within the positive orthant. As for the EM type 2 method, $\dot{q}^{(K)}$ is a Dirac, but we should rather consider the distribution which it approximates, namely $\mathbf{Boojum}(-D, \sum_d \bar{\alpha}_{dK})$, and show that the latter is proper. By Proposition 1, and after a few transformations (essentially replacing $\bar{\alpha}_{DK}$ by its definition in terms of $\dot{\alpha}_{DK}$) we have to show

$$\log \sum_k \exp \frac{1}{D} \sum_d \Psi(\dot{\alpha}_{dk}) < \frac{1}{D} \sum_d \Psi(\sum_k \dot{\alpha}_{dk})$$

And indeed, this is a direct consequence of the convexity of $\log \sum \exp$, together with some known property of function Ψ.

The prior of the full conjugacy method is very different in shape from that of the EM type 2 method. While the latter has a uniform (improper) density, the former is strongly peaked at 0 and sharply decreasing away from it. It tends to favour values of β_W close to 0. But β_W is the parameter of the Dirichlet for ϕ_{kW}, and a Dirichlet with parameter close to 0 tends to favour values towards the borders, and even more the corners, of the simplex. This is indeed confirmed by looking at the shape of the (improper) prior distribution of the whole matrix ϕ_{KW}, which can be computed analytically by collapsing β_W. Since ϕ_{kW} is the distribution of words of topic k, being close to the borders and corners of the simplex essentially means being sparse. Hence, our choice of prior favours sparsity in the topic word profiles (the same applies to the document topic profiles if we apply full conjugacy on that side).

The document-topic side and the topic-word side of the LDA model are strictly symmetric, just as dimensions are treated symmetrically in a matrix factorisation. However, when considering the asymptotic behaviour of the model, that symmetry is broken, as the vocabulary is often seen as fixed, while the number of documents can grow to infinity. In that case, our full conjugacy variant could not be applied to the document topic side of the model, since it would make its prior $p^{(K)} = \mathbf{Boojum}(D, \eta_K)$ dependent on the size D of the document collection, and it would not asymptotically vanish as is usually expected. On the other hand, the topic word side is not affected, as the corresponding prior depends only on the number of topics K.

6 Experiments

In order to evaluate our proposal, we ran experiments over two standard corpora, namely: 20 Newsgroup[7] and Reuters50[8]. The first one consists of 13000 messages taken from 20 newsgroups, whereas the second one contains 5000 documents authored by 50 different authors. Table **??** summarizes the characteristics of the two corpora. Each corpus is divided in a training and test set with a ratio 80-20 percent. Models are evaluated according to their perplexity on the test set and we follow here the approach of [1] based on the following fold-in procedure: the word-topic distribution

[7] http://qwone.com/~jason/20Newsgroups

[8] https://archive.ics.uci.edu/ml/datasets/Reuter_50_50

is learned on the training set and considered fixed, whereas the document-topic distribution is learned for each document of the test set, by using the first half of the document only; the perplexity is then computed on each test document, then averaged over all test documents, using:

$$\log p(x^{test}) = \sum_{dw} n_{dw} \log \sum_k \hat{\theta}_{kd} \hat{\phi}_{wk}$$
$$perplexity(x^{test}) = \exp(-\frac{\log p(x^{test})}{\sum_d n_d})$$

with $n_d = \sum_w n_{dw}$.

We have compared several variants of the LDA model according to the previous development:

- With a symmetric prior, fixed or estimated with the EM type 2 procedure, for α and/or β;
- With a fixed asymmetric prior, for α and/or β;
- With a Boojum prior, estimated with the procedure described in Section 5, for α and/or β.

The best combination is obtained with a fixed asymmetric prior on α and a Boojum on β. In the remainder, we refer to this model as *conjugate LDA* (*lda-conjugate* for short). Both the Boojum prior and the asymmetric prior on α yield better results than the symmetric one, in acordance with the results obtained in [11] which illustrate the imporatnce of an asymmetric prior on the document-topic distribution. Furthermore, the Boojum prior on β helps here to improve the models.

Figure 6 illustrates the behavior of the conjugate LDA model wrt to the standard LDA model (with symmetric priors estimated via the EM type 2 procedure). The perplexity is averaged over 10 runs to assess the influence of the random initilization of the parameters. The number of topics is set to 6 and the number of documents to 1000. As one can note, the perplexity of the conjugate model is lower than the one of the standard model, and decreases with the number of iterations, as expected. The Boojum prior on the word-topic distribution leads to a higher decrease in perplexity.

Figure 6 shows how the conjugate LDA model behaves, compared to the standard LDA model, with respect to the size of the collection. Here again, the number of topics K is first set to 6 and the curves correspond to the ratio of perplexity of the conjudate LDA model wrt to the standard LDA model. As one can note, the ratio of perplexity is below 1 when the size of the collections is small (roughly below 3000 documents), and above 1 after that. This indicates that the conjugate LDA model compares favorably to the standard model for small collections, which can be explained by the additional information brought by the prior on such collections (the influence of the prior then decreases with the size of the collection). Figure 6 also displays the ratio of perplexity for the two collections with the number of topics set to the actual number of classes in the collections: $K = 20$ for 20 Newsgroup, as the data originates from 20 different news groups, and $K = 50$ for Reuters50 as the texts are authored by 50 different persons. As one can note, we observe the same tendency as the one mentioned above, even though the difference between the two models is less marked.

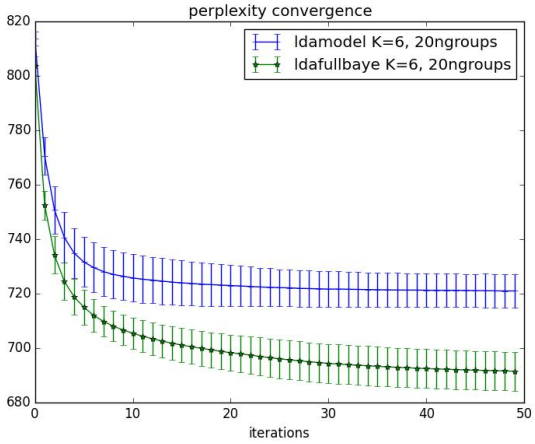


Figure 4: Evolution of perplexity on the test sets according to the number of iterations of the variationnal Bayes methods for both conjugate and standard LDA mdoels, for 20 Newsgroup

To be kept and completed if we have a nice illustration! Figure ?? illustrates the qualitative impact of the Boojum prior on the word-topic distribution.

Lastly, Figure 6 compares the execution time of the inference procedures for the two models for $K = 20$. Similar plots are obtained for different values of K. As one can note, there is no significant difference between the inference in the conjugate and the standard models.

7 Discussion

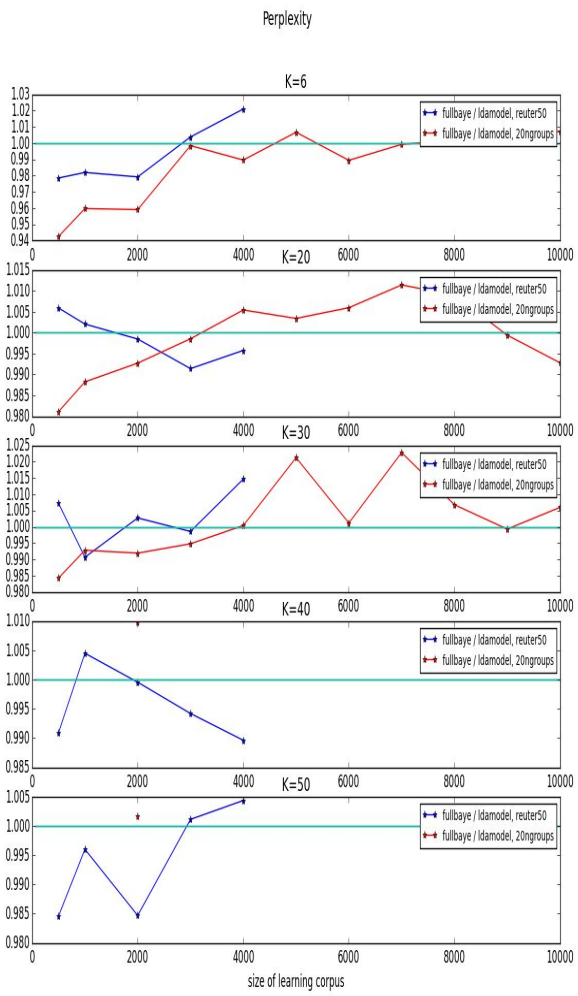


Figure 5: Perplexity ratio for the conjugate LDA model and the standard LDA model according to the size of the collection ($K = 6, 20$ for 20 Newsgroup and $K = 6, 50$ for Reuters50)

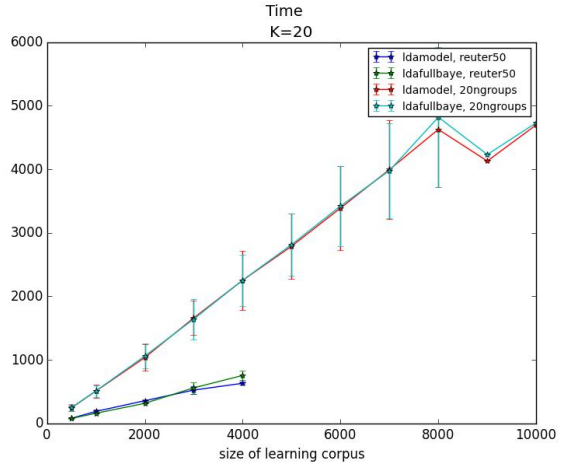


Figure 6: Time of inference for 50 iterations of variational bayes.

References

[1] A. Asuncion, M. Welling, P. Smyth, and Y. W. Teh. On smoothing and inference for topic models. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, pages 27–34. AUAI Press, 2009.

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

[3] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022, 2003.

[4] J. Boyd-Graber and D. M. Blei. Multilingual topic models for unaligned text. In *Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, UAI '09, pages 75–82, Arlington, Virginia, United States, 2009. AUAI Press.

[5] D. Görür, F. Jäkel, and C. E. Rasmussen. A choice model with infinitely many latent features. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 361–368, New York, NY, USA, 2006. ACM.

[6] T. L. Griffiths and M. Steyvers. Finding scientific topics. *PNAS*, 101(suppl. 1):5228–5235, 2004.

[7] M. Hoffman, F. R. Bach, and D. M. Blei. Online learning for latent dirichlet allocation. In *advances in neural information processing systems*, pages 856–864, 2010.

[8] I. Porteous, D. Newman, A. Ihler, A. Asuncion, P. Smyth, and M. Welling. Fast collapsed gibbs sampling for latent dirichlet allocation. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 569–577. ACM, 2008.

[9] M. Rosen-Zvi, T. Griffiths, M. Steyvers, and P. Smyth. The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, UAI '04, pages 487–494, Arlington, Virginia, United States, 2004. AUAI Press.

[10] Y. W. Teh, D. Newman, and M. Welling. A collapsed variational Bayesian inference algorithm

for latent Dirichlet allocation. In *Advances in neural information processing systems*, pages 1353–1360, 2006.

[11] H. M. Wallach, D. M. Mimno, and A. McCallum. Rethinking LDA: Why Priors Matter. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, editors, *Advances in Neural Information Processing Systems 22*, pages 1973–1981. Curran Associates, Inc., 2009.

[12] Y. Wang, E. Agichtein, and M. Benzi. Tm-lda: Efficient online modeling of latent topic transitions in social media. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '12, pages 123–131, New York, NY, USA, 2012. ACM.

[13] L. Yao, D. Mimno, and A. McCallum. Efficient methods for topic model inference on streaming document collections. In *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '09, pages 937–946, New York, NY, USA, 2009. ACM.