
On Priors in Latent Dirichlet Allocation for Variational Bayes

Jean-Marc Andreoli
Xerox Research Centre Europe
Grenoble, France

Adrien Dulac
Université Grenoble Alpes
Laboratoire d'Informatique de Grenoble

Eric Gaussier

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 one being Variational Bayes. 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 propose here a new variant of LDA, concerning the choice of its prior, simply applying the principle of Bayesian conjugacy which underlies the rest of its design. This leads to new, efficient update rules for the variational Bayes estimation algorithm and we explore that new design on a few experimental 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 systematic application of its underlying design principle, which can be informally summarised as: conjugate priors work best.

2 RELATED WORK

essai git

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** distribution¹, 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, τ where $m \in \mathbb{R}$ is a scalar and $\tau \in \mathbb{R}^N$ 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 $\mathbf{x} \in \mathbb{R}_+^N$ by

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

where \mathcal{B} denotes the multivariate beta function $\mathcal{B}(\mathbf{x}) \triangleq \frac{\prod_n \Gamma(x_n)}{\Gamma(\sum_n x_n)}$ which is also the normalising constant of the

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

¹a.k.a. the Snark distribution, because the Snark is a Boojum, you see.

Dirichlet distribution with parameter \mathbf{x} , and $Z(m, \boldsymbol{\tau})$ is the normalising constant of the **Boojum** distribution itself, defined by

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

The expression $\boldsymbol{\tau} \mathbf{x}$ 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 **Boojum**($m, \boldsymbol{\tau}$) is proper, i.e. $Z(m, \boldsymbol{\tau}) < \infty$, if and only if²*

$$\forall n \in N \, \tau_n > 0 \text{ and } m < 1 \text{ and } (m \geq 0 \text{ or } \sum_{n \in N} \exp -\frac{\tau_n}{|m|} < 1)$$

The proof of this result is far from trivial, and can be provided on demand by the authors. On the other hand, conjugacy with the Dirichlet distribution, as expressed by the following property, is quite straightforward.

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

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

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 variables³: (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.

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

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

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

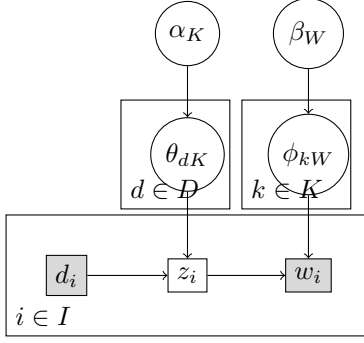
The VB method solves the following optimisation problem:

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

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 \hat{q} of q^* in \mathcal{C} . It proceeds by simple coordinate descent, along the components \hat{q}_X , one for each independent variable X in (2). The update, at each step of the coordinate descent focussing

²For any whole number N , we use the shorthand $n \in N$ to mean $n \in \{1 \dots N\}$

³In 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.



$$\begin{aligned}
 p(\alpha_K \beta_W \theta_{DK} \phi_{KW} z_I w_I | d_I) &= \\
 & p(\alpha_K) p(\beta_W) \\
 & \prod_{d \in D} p(\theta_{dK} | \alpha_K) \prod_{k \in K} p(\phi_{kW} | \beta_W) \\
 & \prod_{i \in I} p(z_i | d_i \theta_{DK}) p(w_i | z_i \phi_{KW}) \\
 \alpha_K &\sim p^{(K)} \quad \beta_W \sim p^{(W)} \\
 \forall d \in D \quad \theta_{dK} | \alpha_K &\sim \text{Dirichlet}(\alpha_K) \\
 \forall k \in K \quad \phi_{kW} | \beta_W &\sim \text{Dirichlet}(\beta_W) \\
 \forall i \in I \quad z_i | d_i \theta_{DK} &\sim \text{Cat}(\theta_{d_i K}) \\
 \forall i \in I \quad w_i | z_i \phi_{KW} &\sim \text{Cat}(\phi_{z_i W})
 \end{aligned}$$

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

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

on variable X , is given by

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

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

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

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 forms⁴:

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

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

⁴To avoid an inflation in Greek letters, we denote the variational parameters with the same letter as the corresponding model variables, decorated with a dot.

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:

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

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

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

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

Figure 2: Generic variational updates for the LDA model.

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

 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.

where **Boojum** is the conjugate distribution of the Dirichlet in the exponential family introduced in Section 3. To proceed further, we need to choose the parameters $p^{(K)}, p^{(W)}$ (the model priors) so as to make the updates of $\dot{q}^{(K)}, \dot{q}^{(W)}$ concrete. 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 **Boojum**(0, 0). Reporting in $[\mathbf{d}]$, this *naturally* constrains the distribution $\dot{q}^{(K)}$ to be equal to **Boojum**($-D, \sum_d \bar{\alpha}_{dK}$). However, update $[\mathbf{D}]$ 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):

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

The arg max expression above, computing the mode of the **Boojum** distribution, can be simplified by introducing function Φ defined for any vector u_N by

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

The resulting update rules $[\mathbf{D}]$ and $[\mathbf{d}]$ are given in Figure 3. By constraining $\dot{q}^{(K)}$ to be Dirac, the computation of its expectation becomes trivial in update $[\mathbf{D}]$, but at the price of its own update $[\mathbf{d}]$, which now requires the computation of the mode of the approximated distribution. 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, 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 **Boojum**($\dot{m}, \dot{\zeta}_W$) distribution becomes a simple Exponential distribution⁵ with rate $\dot{\zeta}_W$, the expectation of which is trivial to compute.

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

The resulting update rules $[\mathbf{W}]$ and $[\mathbf{w}]$ 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 $[\mathbf{W}]$.

⁵We mean here a multivariate Exponential, product of independent scalar Exponentials.

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 in our case 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. $\hat{q}^{(W)}$ is an Exponential, and obviously proper: it is easy to show that its parameter $\zeta^{(W)}$ always remain strictly within the positive orthant. As for $\hat{q}^{(K)}$, which is a Dirac, we should rather consider the distribution which it approximates, namely **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 $p^{(W)}$ on β_W is very different in shape from $p^{(K)}$ on α_K . 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. 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 profile if we choose our prior on that side).

The document topic side and the topic word side of the LDA model are strictly symmetric, just as dimensions are treated identically 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 variant could not be applied to the document topic side of the model, since it would make its prior 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.

6 EXPERIMENTS

7 DISCUSSION

References