Latent IBP compound Dirichlet Allocation

Cédric Archambeau, Balaji Lakshminarayanan, Guillaume Bouchard

Abstract—We introduce the four-parameter IBP compound Dirichlet process (ICDP), a stochastic process that generates sparse non-negative vectors with potentially an unbounded number of entries. If we repeatedly sample from the ICDP we can generate sparse matrices with an infinite number of columns and power-law characteristics. We apply the four-parameter ICDP to sparse nonparametric *topic modelling* to account for the very large number of topics present in large text corpora and the power-law distribution of the vocabulary of natural languages. The model, which we call latent IBP compound Dirichlet allocation (LIDA), allows for power-law distributions, both, in the number of topics summarising the documents and in the number of words defining each topic. It can be interpreted as a sparse variant of the hierarchical Pitman-Yor process when applied to topic modelling. We derive an efficient and simple collapsed Gibbs sampler closely related to the collapsed Gibbs sampler of latent Dirichlet allocation (LDA), making the model applicable in a wide range of domains. Our nonparametric Bayesian topic model compares favourably to the widely used hierarchical Dirichlet process and its heavy tailed version, the hierarchical Pitman-Yor process, on benchmark corpora. Experiments demonstrate that accounting for the power-distribution of real data is beneficial and that sparsity provides more interpretable results.

Index Terms—Bayesian nonparametrics, power-law distribution, sparse modelling, topic modelling, clustering, bag-of-words representation, Gibbs sampling

1 Introduction

Probabilistic topic models such as *latent Dirichlet allocation* (LDA) [1], [2] are widespread tools to analyse and explore large text corpora. LDA models the documents in the corpus as a mixture of K discrete distributions over the vocabulary, which are called topics. The key simplifying assumption made in LDA is that the sequential structure of text can be ignored to capture the semantic structure of the corpus. As a result, LDA considers a bag-of-words representation of the documents. Among the many, notable extensions include the modelling of topical trends over time [3], their particularisation to the discovery of topics in conjunction with the underlying social network [4] or the joint representation of topics and sentiment [5]. In recent years, topic models have been used in numerous applications, not only in text analysis, but also to model huge image databases [6], [7], software bugs [8] or regulatory networks in systems biology [9], and they proved to give state-of-the-art results in the unsupervised extraction of human intelligible topics from a wide variety of documents [10].

While LDA can be viewed as a hierarchical Bayesian extension of probabilistic latent semantic analysis (PLSA) [11] and can be interpreted as a multinomial PCA model [12], its tremendous success (over 4800 citations according to Google Scholar at the time of writing) can be attributed to its simplicity and its natural interpretation. The model not only proposes an appealing generative model of documents, but it also enjoys a relatively simple inference procedure (i.e. a collapsed Gibbs sampler [13]) based on simple word counts, which is able to handle millions of documents in a couple of minutes. A practical issue with LDA, however, is model selection (i.e., the identification of the number of topics capturing the underlying semantics). When modelling real data, the number of topics are expected to grow logarithmically with the size of the corpus. When the number of documents in the corpus increases, it is reasonable to assume that new topics will appear, but that the increase will not be linear in the number of documents; there will be a saturation effect. Model selection can be dealt with in a principled way by considering the *hierarchical Dirichlet process* (HDP), which can be interpreted as the nonparametric extension of LDA [14].

Despite the success and appealing generative construct of HDP-LDA for topic modelling, it is interesting to note that the distributions it postulates are inappropriate for modelling real corpora. Data sampled from HDP-LDA show typically significant departure from real observation counts. For example, it is well-known that the ordered frequencies of the vocabulary words observed in most real corpora follow Zipf's law [15]: the frequency of a specific word is proportional to the inverse of its rank. This is illustrated in Figure 1, where the ordered word frequencies of the four corpora we will consider in the experiments are shown. A more realistic nonparametric topic model would be based on the Poisson-Dirichlet process [16], also known as the *Pitman-Yor process* (PYP) [17]. This stochastic process is a generalisation of the *Dirichlet process* (DP) [18]; it has one additional parameter, called the discount parameter, which enables it to account for power-law characteristics in the data. Its hierarchical extension, the *hierarchical Pitman-Yor process* (HPYP) was used successfully for language modelling

- *C. Archambeau is with Amazon Berlin, Germany. E-mail: cedric.p.archambeau@gmail.com*
- *G. Bouchard is with Xerox Research Centre Europe, France. E-mail: guillaume.bouchard@xrce.xerox.com*
- *B. Lakshminarayanan is with the Gatsby Computational Neuroscience Unit, CSML, University College London. E-mail: balaji@gatsby.ucl.ac.uk*

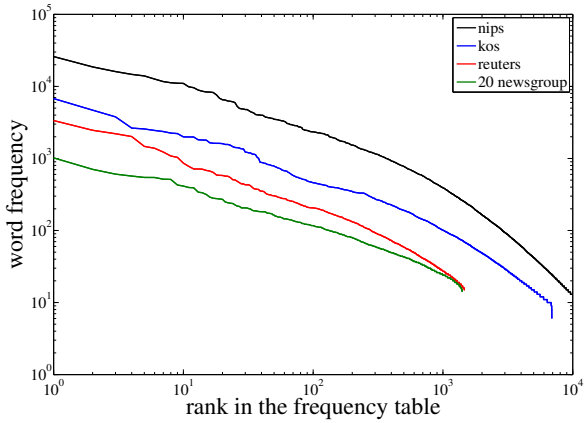


Fig. 1. Ordered word frequencies of the four benchmark corpora that will be considered in the experiments (see Section 5 for a detailed description). Let f_w be the frequency of word w in the corpus. It can be observed that the ranked word frequencies follow Zipf's law, which is an example of a power-law distribution: $p(f_w) \propto f_w^{-c}$, where c is a positive constant. Like many natural phenomena, human languages including English exhibit this property. Intuitively, this means that human languages have a very heavy tail: few words are extremely frequent, while many words are very infrequent.

in [19] and for more general sequence modelling in [20]. It was also shown to have a remarkable connection to interpolated Kneser-Ney, which is currently still one of the most effective language models since it was first proposed more than a decade ago [21], [22]. HPYP-LDA was proposed for topic modelling to account for the power-law distribution of text [23] and it was shown to outperform HDP-LDA in terms of predictive likelihood on several benchmark data sets such as the Reuters and the New York Journal corpora.

Recently, sparsity-enforcing priors have been proposed to enable topics to be defined by a small subset of the vocabulary. Sparsity enforcing priors lead to compression as well as an easier interpretation of the topics. A suitable candidate in the Bayesian nonparametric domain is the *IBP compound-Dirichlet* process (ICDP) [24]. On top of the simple sparsity-promoting advantage, the ICDP enables to decouple the topic inter-document frequency and intra-document frequency [25]. An HDP-LDA model assumes implicitly that an infrequent topic should also be infrequent in every document where it appears. Hence, unlike HDP-LDA, the ICDP can lead to very specific topics that might be very rare in a document corpus overall, but relate to a lot of words in the few documents that include this topic. The ICDP assumes that a random infinite binary matrix generated by an *Indian Buffet Process* (IBP) [26], [27] prior "selects" a subset of the components before applying a Dirichlet prior on the subset of activated components. The ICDP based on the one-parameter IBP has been applied as a prior for the document-topic distribution in a model called the *Focused Topic Model* (FTM) to enable a small number of topics allocated per document [24]; it has also been applied as prior for the topic-word matrix in the *Sparse-Smooth Topic Model* (SSTM) to obtain topics with fewer words describing them [25]. While the posterior Dirichlet associated with each topic in HDP-LDA is peaked for a small value of its hyperparameters, it puts non-zero probability mass on all words of the vocabulary. SSTM removes this constraint and it enables topics to be expressed by words that might be very discriminative, but do not necessarily appear in the same proportion in all documents associated with these topics.

The primary goal of FTM and SSTM is to render topic models more expressive, either by allowing more diverse topic distributions, or by favouring more specialised topic definitions. FTM decouples the prevalence of a topic in the corpus from its prevalence in individual documents, while SSTM decouples the prevalence of a word occurring in the corpus and its prevalence in the individual topics. However, neither of these models address the fundamental weakness of HDP-LDA regarding the power-law distribution of natural language and possibly of topics. Both, FTM and SSTM, are based on the one-parameter IBP, which typically generates very tall binary matrices. As a result, most of the features are shared by most topics, which is undesirable.

In this work, we introduce the four-parameter IBP compound Dirichlet process (ICDP), which is based on the three-parameter IBP [28]. As illustrated in Figure 2, the ranked frequencies of features generated from a one-parameter IBP are not power-law distributed, while they are for a three-parameter IBP. Hence, it is natural to consider the four-parameter ICDP to model real text corpora. We also propose a unified framework for the power-law extensions of FTM and SSTM, which contains them as special cases. The power-law extension of FTM can be viewed as a sparse variant of HPYP-LDA [23]. Moreover, unlike previous methods we derive a very simple collapsed Gibbs sampler in the same vein as the collapsed Gibbs sampler for LDA. In the experimental section, we apply the proposed models on several text corpora; the predictive likelihood compares favourably with respect to the widely used HDP-LDA. A detailed analysis of the results show that the topic models based on the four-parameter ICDP are more expressive and result in a higher number of topics, many of them being infrequent. The most common topics tend to be easier to interpret than HDP-based topics. The infrequent topics are often associated with a subset of documents and are more difficult to interpret, but specialised.

The paper is organised as follows. First, we introduce the four-parameter IBP compound Dirichlet process. Next, we present *latent IBP compound Dirichlet allocation* (LIDA), a sparse nonparametric Bayesian topic model with power-law characteristics. Subsequently, we discuss a simple collapsed Gibbs sampler. Finally, we validate the model on several toy and benchmark corpora.

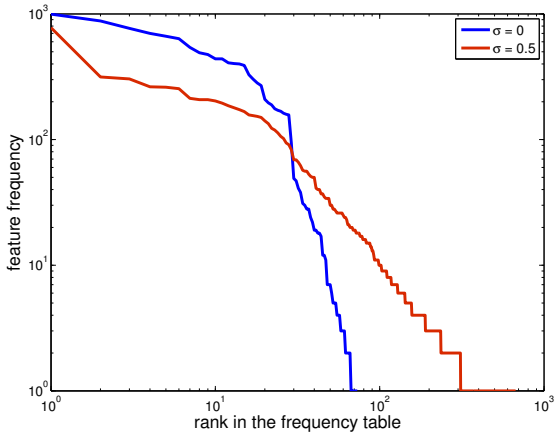


Fig. 2. Ordered feature frequencies drawn from a three-parameter IBP with parameters $\eta = 10$ and $\delta = 1$. The features are power-law distributed when $\sigma > 0$.

2 IBP COMPOUND DIRICHLET PROCESS

The *Dirichlet process* (DP) [18] has been extensively studied in the literature as a prior for the number of components in a mixture model [29], [30]. Draws from a DP are completely random measures, which means that they are discrete with probability one [31]:

$$H \sim \mathrm{DP}(\alpha_0 H_0) \quad \Rightarrow \quad H = \sum_{k=1}^{\infty} \tau_k \delta(\phi_k), \tag{1}$$

where $\alpha_0 > 0$ is the mass parameter, $\delta(\cdot)$ is Dirac's delta function and H_0 is the base measure. The atoms $\{\phi_k\}_{k=1}^{\infty}$ are drawn from the base measure H_0. The weights $\{\tau_k\}_{k=1}^{\infty}$ associated with the atoms depend on α_0 and $\sum_{k=1}^{\infty} \tau_k = 1$. If we think of H_0 as a prior on the mean and the covariance of a multivariate Gaussian distribution, then a draw from H_0 would generate an atom ϕ_k, which would correspond to the mean vector and the covariance matrix of component k. Its mixture weight would then be τ_k. Hence, it is natural to use the Dirichlet process as a prior for the weights in an infinite mixture, e.g. of Gaussians.

The HDP is a two-level hierarchical extension of the Dirichlet process. It is still assumed that the data are generated from an underlying mixture model, but every data point now corresponds to a group of observations, each of which has been generated from one of the mixture components. Hence, we can think of every data points as a mixture model with a specific weighting of the mixture components. Let $H \sim \mathrm{DP}(\alpha H_0)$ be a prior for measure G_d, such that

$$G_d \sim \mathrm{DP}(\alpha H) \quad \Rightarrow \quad G_d = \sum_{k=1}^{\infty} \theta_{kd} \delta(\phi_k). \tag{2}$$

The first point to note from this expression is that G_d is again a mixture model and that it shares all its atoms with H. The second point to note is that each draw G_d is characterised by a set of weights $\{\theta_{kd}\}_{k=1}^{\infty}$, which satisfy $\sum_{k=1}^{\infty} \theta_{kd} = 1$. As shown in [14], the set $\{\tau_k\}_{k=1}^{\infty}$ defines a prior on every set $\{\theta_{kd}\}_{k=1}^{\infty}$:

$$\boldsymbol{\theta}_d \sim \mathrm{DP}(\alpha \boldsymbol{\tau}), \tag{3}$$

where $\boldsymbol{\theta}_d$ and $\boldsymbol{\tau}$ can be thought of as infinite dimensional vectors. One of the most popular applications of the HDP is topic modelling, which we have denoted HDP-LDA. Each data point corresponds to a document, which is summarised by a bag of words and each word in the document is assumed to be generated from an underlying topic. A topic is just a discrete distribution over the vocabulary and documents are modelled as a mixture of topics, the mixture weights being document dependent. Hence, the topics correspond to the atoms, while the topic proportions are the mixture weights.

Unfortunately, the main issue comes from (3): the distribution over topics (i.e., components) depends only on $\alpha\boldsymbol{\tau}$, meaning that the importance of each topic (i.e. component) in the whole corpus is linked to its probability of being associated with any document (i.e., data point). This is undesirable as it might well be that a specific topic does not occur often, but is very important to one specific document. The one-parameter ICDP was proposed in [24] to address this weakness. However, this process relies on the one-parameter IBP [26], [27], in which the expected number of active topics is coupled to the number of topics that are shared among documents (see Figure 3 top left corner). Hence, the one-parameter ICDP only partially addresses the issue with HDP-LDA as relatively few topics are shared by many documents. We will address this issue by considering the four-parameter ICDP extension. Our model relies on the three-parameter IBP [28], which exhibits a power-law behaviour as we will illustrate in the next section.

2.1 Three-parameter Indian Buffet process

The *Indian Buffet process* (IBP) is a nonparametric Bayesian model typically used to generate an unbounded number of latent features when we can assume the data are exchangeable. However, for a finite number of observations, let say D, the number of features is finite with probability one. In the IBP metaphor, the observations are called customers and the features dishes [27]. Let $\eta > 0$, $\delta > -\sigma$ and $\sigma \in [0, 1)$ be the three parameters of the IBP. The generative process of the features is as follows [28]:

1) The first customer tries $\mathrm{Poisson}(\eta)$ dishes;
2) Next, customer d tries dish k with probability $\frac{m_k - \sigma}{d - 1 + \delta}$ and $\mathrm{Poisson}\left(\eta \frac{\Gamma(1+\delta)\Gamma(d-1+\delta+\sigma)}{\Gamma(d+\delta)\Gamma(\delta+\sigma)}\right)$ new dishes,

where $\Gamma(\cdot)$ is the Gamma function and m_k is the number of customers having tried dish k.

While usually defined by one parameter [26] or two parameters [32], the IBP is used as a generative model for binary matrices. Each row is obtained by repeatedly

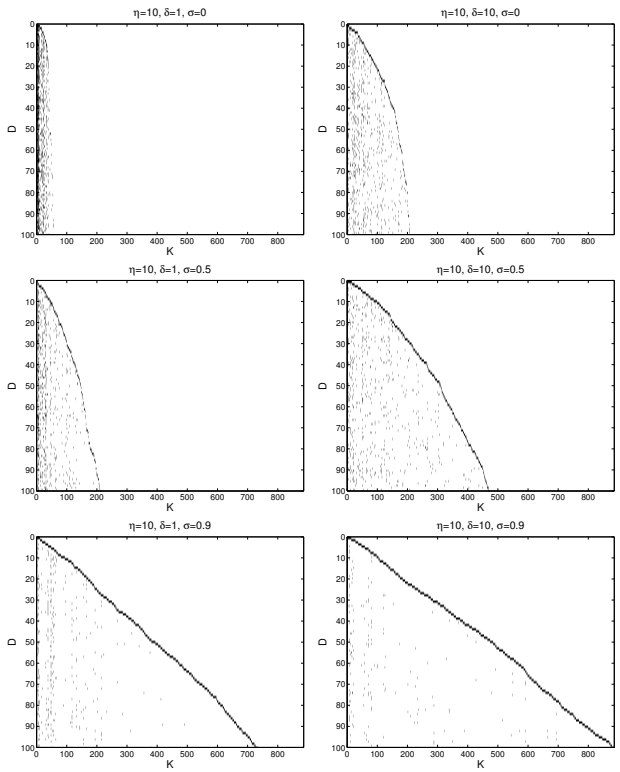


Fig. 3. Binary matrices sampled from a three-parameter IBP. The mass parameter η is constant, the concentration parameter δ increases across columns, while the discount parameter σ increases across rows. In all cases the expected number of non-zero entries is ηD. The amount of features that is shared decreases when the concentration parameter δ increases. The discount parameter σ is responsible for the power-law characteristic. The top left matrix is drawn from a one-parameter IBP, while the matrix below is drawn from a two-parameter IBP with the same mass parameter, but a larger concentration parameter.

sampling from a beta-Bernoulli process where the beta process is integrated out [33]:

$$G_d|H \sim \mathrm{BeP}(H), \qquad H \sim \mathrm{BP}(\delta, \eta H_0), \tag{4}$$

where $\eta > 0$ is the mass parameter and H_0 is a smooth base distribution. The concentration parameter δ is positive and it is set to 1 in the one-parameter case. Both, the Bernoulli and the beta process are examples of completely random measures [34]. In particular, the beta process is a completely random measure with an unnormalised beta distribution as rate measure. It is defined on the product space $[0,1] \times \mathbb{F}$:

$$\Lambda_0(d\pi \times d\phi) = \eta\delta\pi^{-1}(1-\pi)^{\delta-1}d\pi H_0(\phi)d\phi. \tag{5}$$

Draws from the beta process correspond to draws from the Poisson process with rate measure Λ_0, which integrates to infinity over its entire domain. This means that the random measure H has a countably infinite number of atoms, each independently and identically distributed according to H_0:

$$H = \sum_{k=1}^{\infty} \pi_k \delta(\phi_k). \tag{6}$$

Since all π_k lie in $[0,1]$, we can interpret H as an infinite collection of coin-tossing probabilities. Hence, any random measure G_d drawn from the Bernoulli process with base measure H is of the form

$$G_d = \sum_{k=1}^{\infty} \bar{\theta}_{kd} \delta(\phi_k), \tag{7}$$

where $\bar{\theta}_{kd} \sim \mathrm{Bernoulli}(\pi_k)$ and the atoms $\{\phi_k\}_{k=1}^{\infty}$ are the same as the ones of H.

The mass parameter η regulates the expected number of active features per observation G_d and thus the total number of features active in the random matrix formed by stacking the observations $\{G_d\}_{d=1}^{D}$. The concentration parameter δ can be interpreted as a repulsion parameter: when it increases, the number of different features will increase for a given number of expected active features. This is illustrated in Figure 3. The first column compares draws from the one-parameter and a two-parameter IBP. When $\delta > 1$ it can be observed that less features are shared.

The three-parameter IBP is based on a generalisation of the beta process [35], [28]:

$$G_d|H \sim \mathrm{BeP}(H), \qquad H \sim \mathrm{CRM}(\Lambda_0). \tag{8}$$

The rate measure $\Lambda_0(d\pi \times d\phi)$ is now given by

$$\eta \frac{\Gamma(1+\delta)}{\Gamma(1-\sigma)\Gamma(\delta+\sigma)} \pi^{-\sigma-1}(1-\pi)^{\delta+\sigma-1} d\pi H_0(\phi) d\phi, \tag{9}$$

where $\sigma \in [0,1)$ is the discount parameter and $\delta > -\sigma$. Again, the three-parameter IBP is obtained by integrating out the completely random measure H and draws G_d are of the form (7).

Like the two-parameter IBP, the number of active features depends on η and the number of non-zero entries is decoupled from the number of shared entries thanks to δ. However, the three-parameter IBP also exhibits a power-law in the number of unique features thanks to the additional discount parameter [36], which has a similar role as the discount parameter in the PYP [17]. The effect of the discount parameter is shown in Figure 3.

Another, less formal way to understand the three-parameter IBP is by taking the limit of the finite-dimensional case. Let $\bar{\mathbf{\Theta}}$ be a random binary matrix of size $K \times D$ where K is finite, with rows $\{\bar{\boldsymbol{\theta}}_k\}_{k=1}^{K}$. We define the intensity ε as the expected number of rows with at least one activated feature:

$$\varepsilon = \eta \sum_{j=0}^{D-1} \frac{\Gamma(1+\delta)\Gamma(j+\delta+\sigma)}{\Gamma(j+1+\delta)\Gamma(\delta+\sigma)}. \tag{10}$$

We assume that the total number of rows K is greater than ε so that the fraction of activated rows is ε/K in

expectation. The matrix $\bar{\mathbf{\Theta}}$ is generated according to the following process:

$$
\begin{aligned}
&a_k \sim \text{Bernoulli}\left(\frac{\varepsilon}{K}\right)\\
&\text{if } a_k = 0 \text{ (row } k \text{ is not activated)}\\
&\quad \bar{\theta}_{kd} = 0 \text{ for all } d \in \{1, 2, \cdots, D\}\\
&\text{else (row } k \text{ is activated)}\\
&\quad d \sim \text{Uniform}\left(\{1, 2, \cdots, D\}\right)\\
&\quad\quad \bar{\theta}_{kd} = 1\\
&\quad\quad \pi_k|\bar{\theta}_{kd} = 1 \sim \text{Beta}\left(1 + \frac{\eta\delta}{K} - \sigma, \delta + \sigma\right)\\
&\quad \text{for } d' \neq d\\
&\quad\quad \bar{\theta}_{kd'} \sim \text{Bernoulli}(\pi_k)
\end{aligned}
\tag{11}
$$

Hence, when the latent activation variable a_k is equal to one, row k is activated. The posterior probability of π_k given $\bar{\boldsymbol{\theta}}_k$ is obtained by multiplying the Bernoulli likelihood (11) with the improper prior $p(\pi_k) \propto \pi^{\frac{\eta\delta}{K}-\sigma-1}(1-\pi)^{\delta+\sigma-1}$ and normalising. This leads to

$$
\pi_k|\bar{\boldsymbol{\theta}}_{k\cdot} \sim \text{Beta}\left(\frac{\eta\delta}{K} + \bar{\theta}_{k\cdot} - \sigma, D - \bar{\theta}_{k\cdot} + \delta + \sigma\right), \tag{12}
$$

where $\bar{\theta}_{k\cdot} \geqslant 1$. The notation "$\cdot$" indicates a summation over the index. When more than one column is activated, i.e. $\bar{\theta}_{k\cdot}^{\setminus kd} \geqslant 1$, we can further derive the conditional by integrating out π_k:

$$
\bar{\theta}_{kd}|\bar{\mathbf{\Theta}}^{\setminus kd}; \bar{\theta}_{k\cdot}^{\setminus kd} \geqslant 1 \sim \text{Bernoulli}\left(\frac{\frac{\eta\delta}{K} + \bar{\theta}_{k\cdot}^{\setminus kd} - \sigma}{\frac{\eta\delta}{K} + D - 1 + \delta}\right), \tag{13}
$$

where the notation "$\setminus kd$" indicates the contribution of $\bar{\theta}_{kd}$ was removed. We recover the expression derived in [28] when identifying $\bar{\theta}_{k\cdot}^{\setminus kd}$ with m_k and letting K tend to infinity. When $\bar{\theta}_{kd}$ is the only active entry in row k, i.e. $\bar{\theta}_{k\cdot}^{\setminus kd} = 0$, setting $\bar{\theta}_{kd}$ equal to 1 is equivalent to creating a new row. Hence, we get

$$
\bar{\theta}_{kd}|\bar{\mathbf{\Theta}}^{\setminus kd}; \bar{\theta}_{k\cdot}^{\setminus kd} = 0 \sim \text{Bernoulli}\left(\frac{\epsilon}{K}\right). \tag{14}
$$

More generally, for a given number of columns D, the number K^* of activated rows is $\text{Binomial}\left(\frac{\varepsilon}{K}, K\right)$. If we invoke the law of rare events, we find that when K tends to infinity, K^* tends in distribution to a Poisson distribution with mean parameter ε, again recovering the result of [28]. The expected number of new features also tends to a Poisson distribution with parameter equal to the difference between the two intensities, i.e. $\text{Poisson}\left(\eta\frac{\Gamma(1+\delta)\Gamma(D+\delta+\sigma)}{\Gamma(D+1+\delta)\Gamma(\delta+\sigma)}\right)$, again recovering the original formulation of the three-parameters IBP.

2.2 Four-parameter ICD process

The name IBP compound Dirichlet process (ICDP) was first coined in [24], where its one-parameter version was proposed as an alternative Bayesian nonparametric prior to the DP. In contrast to the DP, which can be highly peaked and thus quasi sparse for small values of its hyperparameter α, the ICDP is truly sparse. This has not only advantages in terms of storage, but also in terms of representation capabilities. In this section, we introduce the four-parameter extension of the ICDP. On top of being more flexible, it exhibits power-law characteristics. We start our discussion with finite dimensional matrices and then generalise to the infinite case.

Let $\bar{\mathbf{\Theta}}$ be a binary matrix of size $K \times D$, where K is finite. We assume $\bar{\mathbf{\Theta}}$ is a binary entry selection mask for $\mathbf{\Theta} \in \mathbb{R}^{K\times D}$, such that they share the same non-zero entries. The prior on the columns of $\mathbf{\Theta}$ can be formalised as follows:

$$
\boldsymbol{\theta}_d|\bar{\boldsymbol{\theta}}_d \sim \text{Dirichlet}(\alpha\bar{\boldsymbol{\theta}}_d) = \frac{\Gamma(\bar{\theta}_{\cdot d}\alpha)}{\prod_{k:\bar{\theta}_{kd}\neq 0}\Gamma(\bar{\theta}_{kd}\alpha)} \prod_{k:\bar{\theta}_{kd}\neq 0}^{K} \theta_{kd}^{\alpha-1}, \tag{15}
$$

where it is assumed $\bar{\theta}_{\cdot d} > 0$. The Dirichlet distribution is degenerate: it is defined over the simplex of dimension $\sum_k \bar{\theta}_{kd} - 1$. By convention, we force θ_{kd} to be equal to 0 if it does not belong to the support (i.e., if $\bar{\theta}_{kd} = 0$). This distribution was proposed in [37] for modelling large discrete domains such as in language modelling.

A finite sparsity inducing prior for $\mathbf{\Theta}$ can be constructed based on the truncated IBP:

$$
\bar{\boldsymbol{\theta}}_d \sim \prod_k \text{Bernoulli}(\pi_k), \quad \mathbf{\Theta}|\bar{\mathbf{\Theta}} \sim \prod_d \text{Dirichlet}(\alpha\bar{\boldsymbol{\theta}}_d), \tag{16}
$$

where $p(\pi_k) \propto \pi^{\frac{\eta\delta}{K}-\sigma-1}(1-\pi)^{\delta+\sigma-1}$. Each column $\bar{\boldsymbol{\theta}}_d$ of $\bar{\mathbf{\Theta}}$ is a binary vector, its k^{th} entry being equal to one with probability π_k. The random variable $\boldsymbol{\pi}$ is a K-dimensional vector containing the Beta random variables $\{\pi_k\}_{k=1}^K$.The infinite extension is obtained by letting K tend to infinity and integrating out $\boldsymbol{\pi}$:

$$
\bar{\mathbf{\Theta}} \sim \text{IBP}(\eta, \delta, \sigma), \qquad \mathbf{\Theta}|\bar{\mathbf{\Theta}} \sim \prod_d \text{DP}(\alpha\bar{\boldsymbol{\theta}}_d). \tag{17}
$$

The last equation yields $\boldsymbol{\theta}_d \sim \text{DP}(\alpha\bar{\boldsymbol{\theta}}_d)$, which should be compared to (3), where $\boldsymbol{\theta}_d \sim \text{DP}(\alpha\boldsymbol{\tau})$. Here, each $\bar{\boldsymbol{\theta}}_d$ is independently and identically distributed according to the marginal beta-Bernoulli process inducing (17). In other words, the $\{\boldsymbol{\theta}_d\}$ do not share the same prior as in the case of the DP, as desired.

The four-parameter ICDP is obtained by integrating out the latent binary mask $\bar{\mathbf{\Theta}}$:

$$
\mathbf{\Theta} \sim \text{ICDP}(\alpha, \eta, \delta, \sigma) = \sum_{\bar{\mathbf{\Theta}}} p(\mathbf{\Theta}|\bar{\mathbf{\Theta}})p(\bar{\mathbf{\Theta}}). \tag{18}
$$

Since the number of observed features K is finite when the number of columns D is finite, the four-parameter ICDP can be understood as a mixture of degenerate Dirichlet distributions over simplices of different dimensions. A similar type of degenerate Dirichlet priors was used in [37], [25], [24]. All considered the special case where $\delta = 1$ and $\sigma = 0$. Only [24] considered the ICDP, but the weights associated with each column of $\bar{\mathbf{\Theta}}$ were independently drawn from a Gamma distribution. A Dirichlet distribution was then recovered by normalising

the weights. We do not follow this route as it leads to a complicated sampler partially based on Hybrid Monte Carlo (see e.g. [38]). Instead, we draw the weights from a degenerate Dirichlet with mass parameter α shared by all columns. Thanks to the binary mask, α does not need to be small to ensure that the individual Dirichlet priors are peaked on a small set of features. Importantly, this construction enables us to derive a relatively simple collapsed Gibbs sampler as discussed in Section 4.

Draws from several ICDPs are shown in Figure 4 and they are compared to DPs with a similar mass parameter α. It can be observed that the columns of the matrices drawn from the ICDP have a variable number of active elements, even for large values of the mass parameter α. This is not the case for matrices drawn from the DP. In the context of topic modelling, the ICDP can be used as a prior on the topic proportion matrix as well as the topic distribution. This means that some topics (words) might only occur in few documents (topics) or, conversely, some document (topics) could only have few topics (words) associated with them. As we will show in the experimental section, this is a more realistic assumption than the ones made in conventional topic model like HDP-LDA or even HPY-LDA.

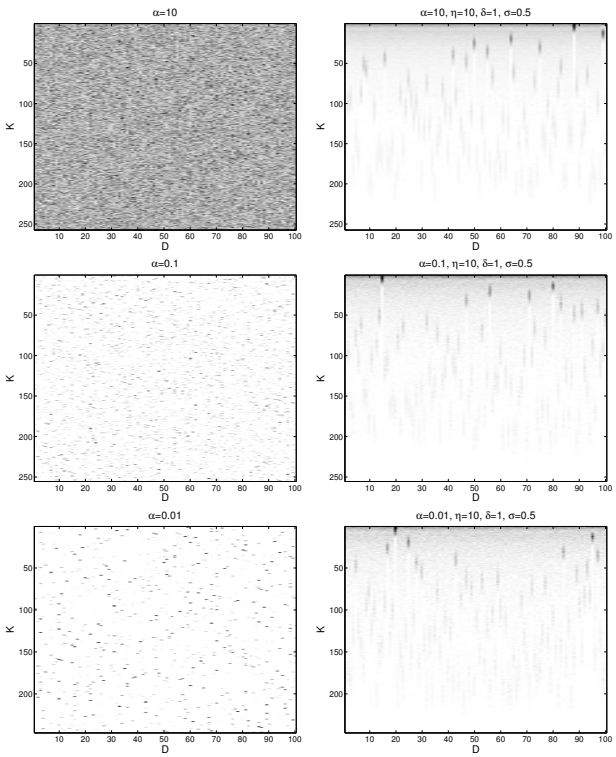


Fig. 4. Left column matrices are generated from a Dirichlet distribution (like in HDP-LDA) with mass parameter $\alpha \in \{10, 0.1, 0.01\}$. It can be observed that the amount of sparsity for each column of the matrix is similar; if one is interested in sparse topics, then all of them show a similar level of sparsity. The matrices in right column are generated from an ICDP (like in LIDA). It can be observed that the amount of sparsity varies across the columns of the matrix.

3 Latent IBP compound Dirichlet Allocation

As LDA and its nonparametric extension HDP-LDA, *latent IBP compound Dirichlet allocation* (LIDA) is a generative model of documents that is based on their bag-of-words representation. At first sight, this might be perceived as a crude assumption, but it has been proven to be valid for topic modelling. However, one unsatisfactory aspect about LDA and HDP-LDA is that they assume the vocabulary is known in advance. Hence, they could be considered to be incomplete generative models as they are incapable of incorporating new words when new documents are observed. LIDA does not suffer from this weakness as explained below. Moreover, LIDA imposes ICDP priors on the topic and the word proportion matrices. As a result, LIDA can account for power-laws in the topics and the words. The latter is especially appealing as the vocabulary of real corpora exhibits power-law characteristics as discussed in Section 1.

The generative process of documents based on LIDA can be summarised as follows:

1) Topic generation:
   - The first topic picks $\text{Poisson}(\gamma)$ words;
   - Next, topic k picks a previously used word v with probability $\frac{m_v - \zeta}{k - 1 + \xi}$ and enriches the topic with $\text{Poisson}\left(\gamma \frac{\Gamma(1+\xi)\Gamma(k-1+\xi+\zeta)}{\Gamma(k+\xi)\Gamma(\xi+\zeta)}\right)$ new words;
   - Topic k is then defined by drawing a discrete distribution over the subset of V_k words defining it from a $\text{Dirichlet}(\beta \mathbf{1}_{V_k})$,

   where $\beta > 0$, $\gamma > 0$, $\xi > -\zeta$ and $\zeta \in [0, 1)$. The count variable m_v indicates the number of times word v appeared in previously observed topics.
2) Document generation:
   - The first document picks $\text{Poisson}(\eta)$ topics;
   - Next, document d picks a previously used topic k with probability $\frac{m_k - \sigma}{d - 1 + \delta}$ and draws $\text{Poisson}\left(\eta \frac{\Gamma(1+\delta)\Gamma(d-1+\delta+\sigma)}{\Gamma(d+\delta)\Gamma(\delta+\sigma)}\right)$ new topics;
   - The topic proportions associated with document d are then obtained by drawing a discrete distribution over the subset of K_d topics from a $\text{Dirichlet}(\alpha \mathbf{1}_{K_d})$;
   - Word w_i in document d is generated by first drawing a topic z_i from the topic proportion distribution of document d and then drawing w_i from the word distribution of topic z_i,

   where $\alpha > 0$, $\eta > 0$, $\delta > -\sigma$ and $\sigma \in [0, 1)$. The count variable m_k indicates the number of times topic k appeared in previous documents.

Hence, the main differences with the generative model of HDP-LDA are:

- For every newly generated document, a subset of the previously observed topics is selected (according to their importance) and potentially a small set of new topics are generated. Hence, words will be

generated not only from previously observed topics, but also from new topics.
- Similarly, every time a word is generated it is selected from a subset of the previously observed vocabulary words or a small set of new candidate words. Hence, every topic definition will take into account the fact that the size of the vocabulary increases when the corpus increases.

More formally, let $\boldsymbol{\Theta} \in \mathbb{R}^{K\times D}$ be the topic proportions matrix and $\bar{\boldsymbol{\Theta}} \in \mathbb{R}^{K\times D}$ its associated binary mask. Further, let $\boldsymbol{\Phi} \in \mathbb{R}^{V\times K}$ be the word proportions matrix and $\bar{\boldsymbol{\Phi}} \in \mathbb{R}^{V\times K}$ its binary mask. Both, K and V are potentially infinite, but given a finite number of documents D, they have a finite number of non-zero entries. The probabilistic model of LIDA is defined as follows:

$$\boldsymbol{\Theta} \sim \mathrm{ICDP}(\alpha, \eta, \delta, \sigma), \tag{19}$$

$$\boldsymbol{\Phi} \sim \mathrm{ICDP}(\beta, \gamma, \xi, \zeta), \tag{20}$$

$$z_i|\boldsymbol{\theta}_d \sim \mathrm{Discrete}(\boldsymbol{\theta}_d), \tag{21}$$

$$w_i|z_i, \{\boldsymbol{\phi}_k\}_{k=1}^{\infty} \sim \mathrm{Discrete}(\boldsymbol{\phi}_{z_i}), \tag{22}$$

where $i = \{1, \ldots, N_d\}$ and $d = \{1, \ldots, D\}$. The total number of words in the corpus is given by $N = \sum_d N_d$. The priors (19) and (20) can be decomposed according to (17). While $\boldsymbol{\theta}_d$ and $\boldsymbol{\phi}_k$ are infinite dimensional vectors, they have a finite number of non-zero entries, which sum up to one. Hence, (21) and (22) are proper discrete distributions. However, it should be noted that LIDA is a nonparametric model where both the number of topics and the number of vocabulary words are unbounded. The graphical model is depicted in Figure 5.

Next, we derive a collapsed Gibbs sampler to infer the latent topic assignments and the latent binary masks. It is relatively simple to implement and is closely related to the collapsed Gibbs samplers for LDA [13], the Chinese restaurant franchise process [14] and the IBP [26].

4 Inference

The sampler that we present is derived from the truncated version of LIDA. The nonparametric version described in the subsequent subsections is obtained trivially from the posteriors by passing to the infinite limit.

First, we integrate out the latent weights $\{\boldsymbol{\theta}_d\}$ and $\{\boldsymbol{\phi}_k\}$ associated respectively with the topics and the words. This leads to the following marginal likelihoods:

$$\boldsymbol{z}|\bar{\boldsymbol{\Theta}} \sim \prod_d \frac{\Gamma(\bar{\theta}_{\cdot d}\alpha)}{\Gamma(\bar{\theta}_{\cdot d}\alpha + n_{\cdot\cdot d})} \prod_{k:\bar{\theta}_{kd}\neq 0} \frac{\Gamma(\bar{\theta}_{kd}\alpha + n_{\cdot kd})}{\Gamma(\bar{\theta}_{kd}\alpha)}, \tag{23}$$

$$\boldsymbol{w}|\boldsymbol{z}, \bar{\boldsymbol{\Phi}} \sim \prod_{k:\bar{\theta}_{k\cdot}\neq 0} \frac{\Gamma(\bar{\phi}_{\cdot k}\beta)}{\Gamma(\bar{\phi}_{\cdot k}\beta + n_{\cdot k\cdot})} \prod_{v:\bar{\phi}_{vk}\neq 0} \frac{\Gamma(\bar{\phi}_{vk}\beta + n_{vk\cdot})}{\Gamma(\bar{\phi}_{vk}\beta)}, \tag{24}$$

where $\boldsymbol{w} = \{w_i\}_{i=1}^N$, $\boldsymbol{z} = \{z_i\}_{i=1}^N$ and n_{vkd} is the number of times word v was assigned to topic k in document d. The notation $\cdot$ means we sum over the corresponding index. Note that $n_{\cdot kd} = 0$ if $\bar{\theta}_{kd} = 0$ (as $\theta_{kd} = 0$) and $n_{vk\cdot} = 0$ if $\bar{\phi}_{vk} = 0$ (as $\phi_{vk} = 0$). Based on the above conditionals, we can write the collapsed Gibbs updates:

$$p(\bar{\theta}_{kd}|\boldsymbol{z}, \bar{\boldsymbol{\Theta}}^{\setminus kd}) \propto p(\boldsymbol{z}|\bar{\boldsymbol{\Theta}}) p(\bar{\theta}_{kd}|\bar{\boldsymbol{\Theta}}^{\setminus kd}), \tag{25}$$

$$p(\bar{\phi}_{vk}|\boldsymbol{w}, \boldsymbol{z}, \bar{\boldsymbol{\Phi}}^{\setminus vk}) \propto p(\boldsymbol{w}|\boldsymbol{z}, \bar{\boldsymbol{\Phi}}) p(\bar{\phi}_{vk}|\bar{\boldsymbol{\Phi}}^{\setminus vk}), \tag{26}$$

$$p(z_i|\boldsymbol{w}, \boldsymbol{z}^{\setminus i}, \bar{\boldsymbol{\Theta}}, \bar{\boldsymbol{\Phi}}) \propto p(\boldsymbol{w}|\boldsymbol{z}, \bar{\boldsymbol{\Phi}}) p(\boldsymbol{z}|\bar{\boldsymbol{\Theta}}). \tag{27}$$

In the finite case, $p(\bar{\theta}_{kd}|\bar{\boldsymbol{\Theta}}^{\setminus kd})$ is given by (13) if $\bar{\theta}_{k\cdot}^{\setminus kd} \geqslant 1$ and by (14) otherwise. The conditional $p(\bar{\phi}_{vk}|\bar{\boldsymbol{\Phi}}^{\setminus vk})$ has the same functional form as $p(\bar{\theta}_{kd}|\bar{\boldsymbol{\Theta}}^{\setminus kd})$.

4.1 Sampling the topic activations

The topic activations per document are sampled according to (25). If there is at least one word allocated to topic k in document d, the document-topic indicator $\bar{\theta}_{kd}$ is automatically set to 1 (i.e., if $n_{\cdot kd} > 0$, we have $p(\bar{\theta}_{kd} = 1|\boldsymbol{z}, \bar{\boldsymbol{\Theta}}^{\setminus kd}) = 1$); otherwise, when $n_{\cdot kd} = 0$, the probability that $\bar{\theta}_{kd}$ is activated is given by

$$\bar{\theta}_{kd}|\boldsymbol{z}, \bar{\boldsymbol{\Theta}}^{\setminus kd} \sim \mathrm{Bernoulli}\left(\pi_{kd}\right), \tag{28}$$

where $\mathcal{B}(\cdot,\cdot)$ is the Beta function and π_{kd} is defined as

$$\pi_{kd} = \frac{1}{1 + \frac{\mathcal{B}(\bar{\theta}_{\cdot d}^{\setminus kd}\alpha, \alpha)(D - 1 - \bar{\theta}_{k\cdot}^{\setminus kd} + \delta + \sigma)}{\mathcal{B}(\bar{\theta}_{\cdot d}^{\setminus kd}\alpha + n_{\cdot\cdot d}, \alpha)(\bar{\theta}_{k\cdot}^{\setminus kd} - \sigma)}}. \tag{29}$$

This formula is obtained using (13) and letting K tend to infinity. However, it is not valid when the topic is a "new" topic. This is the case when, for a given topic k, the binary mask is only active in the current document d, that is, when $\bar{\theta}_{k\cdot} = \bar{\theta}_{kd}$. In this case, the probability of creating the topic k given by (14) is involved, leading to

$$\tilde{\pi}_{kd} = \frac{\eta}{\eta + \frac{\mathcal{B}(\bar{\theta}_{\cdot d}^{\setminus kd}\alpha, \alpha)\Gamma(D + \delta)\Gamma(\delta + \sigma)}{\mathcal{B}(\bar{\theta}_{\cdot d}^{\setminus kd}\alpha + n_{\cdot\cdot d}, \alpha)\Gamma(1 + \delta)\Gamma(D - 1 + \delta + \sigma)}}. \tag{30}$$

Algorithm 1 Collapsed Gibbs Sampler Pseudocode

Initialize $\{\bar{\theta}_{kd}\}_{k=1,d=1}^{K,D}$, $\{\bar{\phi}_{vk}\}_{v=1,k=1}^{V,K}$ and $\{z_i\}_{i=1}^N$.
do
  for $d = 1, \ldots, D$
    for $k = 1, \ldots, K$
      if $n_{\cdot kd} = 0$
        if $\bar{\theta}_{k\cdot} = \bar{\theta}_{kd}$
          Sample $\bar{\theta}_{kd}$ according to (30).
        else
          Sample $\bar{\theta}_{kd}$ according to (29).
    Sample $\mathrm{Poisson}(\tilde{\pi}_{kd})$ new topics using (30).
  for $k = 1, \ldots, K$
    for $v = 1, \ldots, V$
      if $n_{vk\cdot} = 0$
        Sample $\bar{\phi}_{vk}$ according to (31).
    Sample $\mathrm{Poisson}(\tilde{\kappa}_{vk})$ new words using (33).
  for $i = 1, \ldots, N$
    Sample z_i according to (35).
  Sample $\alpha, \eta, \delta, \sigma, \beta, \gamma, \xi, \zeta$ (see Section 4.5).
while not converged

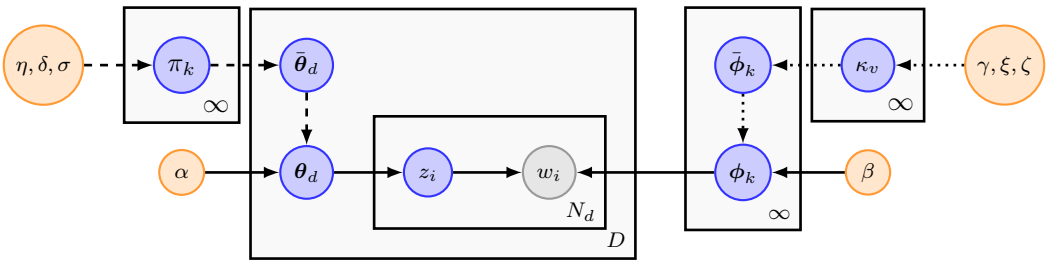


Fig. 5. Graphical models for the different configurations: HDP-LDA: solid arrows only; four-parameter FTM: solid + dashed arrows; LIDA: solid + dashed + dotted arrows. Nodes indicate random variables (gray: observed variables; blue: latent variables; orange: hyperparameters). Rectangle plates correspond to repetitions.

The number of topics K is infinite, so this corresponds to activating $\text{Poisson}(\tilde{\pi}_{kd})$ topics.

4.2 Sampling the word activations

The word activations per topic are sampled according to (26). Similar to the topic activations, when $n_{vk\cdot} > 0$, the corresponding topic-word indicator $\bar{\phi}_{vk}$ is automatically set to 1; otherwise, when $n_{vk\cdot} = 0$, the probability that $\bar{\phi}_{vk}$ is activated is given by

$$\bar{\phi}_{vk} | \boldsymbol{w}, \boldsymbol{z}, \bar{\boldsymbol{\Phi}}^{\setminus vk} \sim \text{Bernoulli}\left(\kappa_{vk}\right), \tag{31}$$

where

$$\kappa_{vk} = \frac{1}{1 + \frac{\mathcal{B}(\bar{\phi}_{\cdot k}^{\setminus vk}\beta, \beta)(K^* - 1 - \bar{\phi}_{v\cdot}^{\setminus vk} + \xi + \zeta)}{\mathcal{B}(\bar{\phi}_{\cdot k}^{\setminus vk}\beta + n_{\cdot k \cdot}, \beta)(\bar{\phi}_{v\cdot}^{\setminus vk} - \zeta)}}, \tag{32}$$

where K^* is the total number of activated topics. For the observed words, this expression is always valid. However, LIDA accounts for potential unobserved words in every topic. Again, in this case the formula is not valid when the unobserved word is a "new" one, that is, it does not belong to the current vocabulary (observed or unobserved). This leads to

$$\tilde{\kappa}_{vk} = \frac{\gamma}{\gamma + \frac{\mathcal{B}(\bar{\phi}_{\cdot k}^{\setminus vk}\beta, \beta)\Gamma(K^* + \xi)\Gamma(\xi + \zeta)}{\mathcal{B}(\bar{\phi}_{\cdot k}^{\setminus vk}\beta + n_{\cdot k \cdot}, \beta)\Gamma(1 + \xi)\Gamma(K^* - 1 + \xi + \zeta)}}. \tag{33}$$

The number of words V is potentially infinite, so this corresponds to activating $\text{Poisson}(\tilde{\kappa}_{vk})$ new words.

However, if one desires to consider a finite vocabulary, one should assign at least one vocabulary word to every topic and sample word activations using the finite version of the Gibbs update, which uses the finite version of the activation probability:

$$\kappa_{vk} = \frac{1}{1 + \frac{\mathcal{B}(\bar{\phi}_{\cdot k}^{\setminus vk}\beta, \beta)(K^* - 1 - \frac{\gamma\xi}{V} - \bar{\phi}_{v\cdot}^{\setminus vk} + \xi + \zeta)}{\mathcal{B}(\bar{\phi}_{\cdot k}^{\setminus vk}\beta + n_{\cdot k \cdot}, \beta)(\frac{\gamma\xi}{V} + \bar{\phi}_{v\cdot}^{\setminus vk} - \zeta)}}. \tag{34}$$

4.3 Sampling the topic assignments

The topic assignments are sampled according to (27). The variable z_i indicates the topic assigned to word w_i in document d. The posterior is given by

$$p(z_i = k | \boldsymbol{w}, \boldsymbol{z}^{\setminus i}, \bar{\boldsymbol{\Theta}}, \bar{\boldsymbol{\Phi}}) \propto \frac{(\alpha + n_{\cdot kd}^{\setminus i})(\beta + n_{vk\cdot}^{\setminus i})}{\bar{\phi}_{\cdot k}\beta + n_{\cdot k\cdot}^{\setminus i}} \bar{\phi}_{vk}\bar{\theta}_{kd}. \tag{35}$$

The product $\bar{\phi}_{vk}\bar{\theta}_{kd}$ equals one only if the topic-document and topic-word binary masks are activated. The inference algorithm is a Gibbs sampler, which alternates between the sampling of the discrete variables $\bar{\theta}_{kd}$, $\bar{\phi}_{vk}$ and z_{id} conditionally to the others. The sampler is summarised in Algorithm 1.

4.4 Special cases

The four-parameter *Focussed Topic Model* (FTM) is obtained if $\bar{\boldsymbol{\phi}}_k = \mathbf{1}_V$. In this case, there is no need to sample the topic-word activation variables $\{\bar{\phi}_{vk}\}$. The topic assignments are then sampled as follows:

$$p(z_i | \boldsymbol{w}, \boldsymbol{z}^{\setminus i}, \bar{\boldsymbol{\Theta}}) \propto \frac{(\alpha + n_{\cdot kd}^{\setminus i})(\beta + n_{vk\cdot}^{\setminus i})}{V\beta + n_{\cdot k\cdot}^{\setminus i}} \bar{\theta}_{kd}. \tag{36}$$

The two-parameter FTM proposed in [24] is recovered when setting $\delta = 1$ and $\sigma = 0$ in (29).

Similarly, the infinite version of *Sparse-Smooth Topic Model* (SSTM) is obtained if $\bar{\boldsymbol{\theta}}_d = \mathbf{1}_K$ and there is no need to sample the topic-document activation variables $\{\bar{\theta}_{kd}\}$. The topic assignments are sampled as follows:

$$p(z_{id} = k | \boldsymbol{w}, \boldsymbol{z}^{\setminus i}, \bar{\boldsymbol{\Phi}}) \propto \frac{(\alpha + n_{\cdot kd}^{\setminus i})(\beta + n_{vk\cdot}^{\setminus i})}{\bar{\phi}_{\cdot k}\beta + n_{\cdot k\cdot}^{\setminus i}} \bar{\phi}_{vk}. \tag{37}$$

It should be noted that SSTM is infinite in the size of the vocabulary, unlike the version proposed in [25] where a DP prior was used to account for an infinite number of topics.

Finally, standard LDA is recovered by letting $\bar{\boldsymbol{\phi}}_k = \mathbf{1}_V$ and $\bar{\boldsymbol{\theta}}_d = \mathbf{1}_K$. This leads to the well-known collapsed Gibbs sampler [13]:

$$p(z_i = k | \boldsymbol{w}, \boldsymbol{z}^{\setminus i}) \propto \frac{(\alpha + n_{\cdot kd}^{\setminus i})(\beta + n_{vk\cdot}^{\setminus i})}{V\beta + n_{\cdot k\cdot}^{\setminus i}}. \tag{38}$$

4.5 Hyperparameter sampling

In order to infer the hyperparameter values we also use Markov Chain Monte Carlo. When we cannot derive a Gibbs sampler, we use Metropolis-Hastings [38] to sample hyperparameter values.

TABLE 1
Data characteristics and average log-perplexity (with standard errors) on held-out data. LIDA outperforms all other methods, except on 20 newsgroups, where the four-parameter FTM performs best.

	20 newsgroup	Reuters	KOS	NIPS
#docs	1000	2000	3430	1500
#unique words	1407	1472	6906	12419
HDP-LDA	6.568±0.033	6.188±0.009	NA	NA
FTM ($\delta = 1$, $\sigma = 0$)	6.342±0.020	5.623±0.015	7.262±0.007	6.901±0.005
HPY-LDA	6.572±0.029	6.164±0.011	NA	NA
FTM	**6.332±0.020**	5.622±0.013	7.266±0.009	6.883±0.008
LIDA	6.376±0.026	**5.592±0.024**	**7.257±0.010**	**6.795±0.007**

For α and β, we obtain closed form Gibbs updates:

$$p(\alpha|\boldsymbol{z}, \bar{\boldsymbol{\Theta}}) \propto p(\boldsymbol{z}|\bar{\boldsymbol{\Theta}}, \alpha)p(\alpha), \tag{39}$$

$$p(\beta|\boldsymbol{w}, \boldsymbol{z}, \bar{\boldsymbol{\Phi}}) \propto p(\boldsymbol{w}|\boldsymbol{z}, \bar{\boldsymbol{\Phi}}, \beta)p(\beta), \tag{40}$$

where $p(\boldsymbol{z}|\bar{\boldsymbol{\Theta}}, \alpha)$ is given by (23) and $p(\alpha) \propto \frac{1}{\alpha}$. Similarly, $p(\boldsymbol{w}|\boldsymbol{z}, \bar{\boldsymbol{\Phi}}, \beta)$ is given by (24) and $p(\beta) \propto \frac{1}{\beta}$.

In order to sample η and δ, we use the joint likelihood of the three-parameter IBP. This leads to

$$p(\eta|\bar{\boldsymbol{\Theta}}, \delta, \sigma) \propto p(\bar{\boldsymbol{\Theta}}|\eta, \delta, \sigma)p(\eta), \tag{41}$$

$$p(\delta|\bar{\boldsymbol{\Theta}}, \eta, \sigma) \propto p(\bar{\boldsymbol{\Theta}}|\eta, \delta, \sigma)p(\delta), \tag{42}$$

where $p(\eta) = \text{Gamma}(a, b)$ and $p(\delta) \propto \frac{1}{\delta+\sigma} - \sigma$. The joint marginal likelihood of the document-topic indicator matrix was derived in [28]. It is given by

$$P(\bar{\boldsymbol{\Theta}}|\eta, \delta, \sigma) = \exp\left(-\eta \sum_{j=0}^{D-1} \frac{\Gamma(1+\delta)\Gamma(j+\delta+\sigma)}{\Gamma(j+1+\delta)\Gamma(\delta+\sigma)}\right) \eta^{K^*} \times \prod_{k \leqslant K^*} \frac{\Gamma(1+\delta)\Gamma(D-\bar{\theta}_{k\cdot}+\delta+\sigma)\Gamma(\bar{\theta}_{k\cdot}-\sigma)}{\Gamma(1-\sigma)\Gamma(\delta+\sigma)\Gamma(D+\delta)}, \tag{43}$$

where K^* is the number of activated features. For completeness, we provide an alternate derivation in Appendix A to the one proposed in [28], which is derived from the truncated IBP.

Setting $\sigma = 0$ in (43) leads to the marginal likelihood for the two-parameter IBP [32]. Further setting $\delta = 1$ leads to the marginal likelihood for the one-parameter IBP as originally derived in [26]. Hence, we can derive a closed form Gibbs update for η as the Gamma distribution is conjugate to the joint marginal likelihood (43) and we use a Metropolis-Hastings step for δ. We use a similar procedure to sample γ and ξ. We did not sample σ or ζ, but this is possible in principle.

5 Experiments

This section is divided into three parts. First, we detail how we approximate the log-predictive likelihood of the words of a test corpus. Next, we study the properties of the the Gibbs sampler and the convergence of the parameters on a toy example. Finally, we evaluate the four-parameter FTM and LIDA on standard benchmarks data sets.

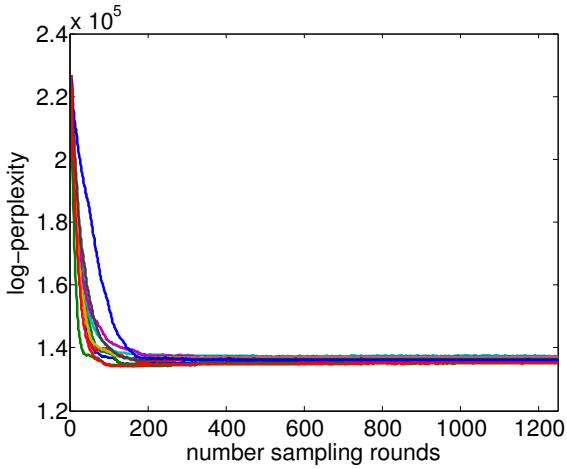


Fig. 6. Training log-perplexity for 10 randomly generated corpora. It can be observed that the sampler converges relatively quickly.

5.1 Evaluation

Let $\boldsymbol{w}_*$ denote the test corpus of size N_* . We use perplexity as a performance metric. Lower perplexity is better. It corresponds to the harmonic mean of the inverse test likelihood:

$$\text{Perplexity}(\boldsymbol{w}_*) = \exp\left(-\frac{\ln P(\boldsymbol{w}_*|\boldsymbol{w})}{N_*}\right). \tag{44}$$

The test log-likelihood $\ln P(\boldsymbol{w}_*|\boldsymbol{w})$ is approximated by empirical averages based on samples after burn-in:

$$\ln P(\boldsymbol{w}_*|\boldsymbol{w}) \approx \sum_{d_*} \sum_{v} n_{v\cdot d_*} \ln \mathbb{E}\left[\boldsymbol{\phi}_v^\top|\boldsymbol{z}_*\right] \; \mathbb{E}\left[\boldsymbol{\theta}_{d_*}|\boldsymbol{z}_*\right], \tag{45}$$

where the posterior expectation of the topic proportions associated with the test documents are computed as follows:

$$\mathbb{E}\left[\theta_{kd_*}|\boldsymbol{z}_*\right] \approx \frac{1}{S}\sum_{s=1}^{S} \frac{\bar{\theta}_{kd_*}\alpha + n_{\cdot kd_*}}{\bar{\theta}_{\cdot d_*}\alpha + n_{\cdot\cdot d_*}}. \tag{46}$$

The posterior expectation of the word proportions is computed in the same fashion:

$$\mathbb{E}\left[\phi_{vk}|\boldsymbol{w}, \boldsymbol{z}\right] \approx \frac{1}{S}\sum_{s=1}^{S} \frac{\bar{\phi}_{vk}\beta + n_{vk\cdot}}{\bar{\phi}_{\cdot k}\beta + n_{\cdot k\cdot}}. \tag{47}$$

In practice, we sample the topics of the test documents on half of the corpus and evaluate perplexity on the other half. In Section 5.3, we split the data sets randomly into five folds and report mean and standard error of the log-perplexities to assess the significance of the results.

5.2 Toy data set

We generated an artificial corpus consisting of 1000 documents, each having an expected number of 150 words. We focussed on the convergence of the sampler of the ICDP. Hence, we restricted the analysis to the

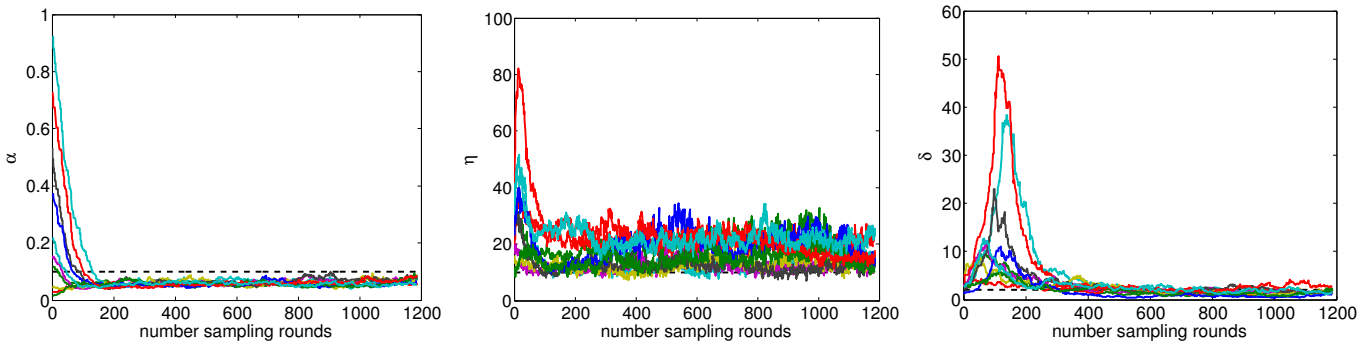


Fig. 7. Toy example – Convergence of the hyperparameters based on ten random initialisations and fixed σ. The correct value is indicated by the constant dashed line. In all cases, the hyperparameter samples converge to a value close to the ground truth. However, α and η show a small bias.

case where topics are sampled from finite dimensional Dirichlet distributions with parameter $\beta = 0.01$. We randomly initialise α, η and δ. We set σ to its true value, 0.1. While in principle we could sample the discount parameter, we noticed that for values greater than 0.25, a very large number of topics could be generated during burn-in, slowing down the sampler significantly. We thus recommend constructing models based on a grid of values that lie in $[0, 1)$ in practice.

The log-perplexity for 10 randomly generated corpora is shown in Figure 6. We observe a modest burn-in of approximately 250 sampling rounds. The parameter samples for 10 random initialisations is shown in Figure 7 for one of these corpora. While the parameters converge to reasonably similar values in all cases, it can be observed that α is underestimated, while η is overestimated. This essentially means that a larger number of sparse topics is preferred compared to the ground truth. We observed experimentally, that this bias reduces when the number of documents increases, but that that this number needs to be relatively large for the reduction to be significant.

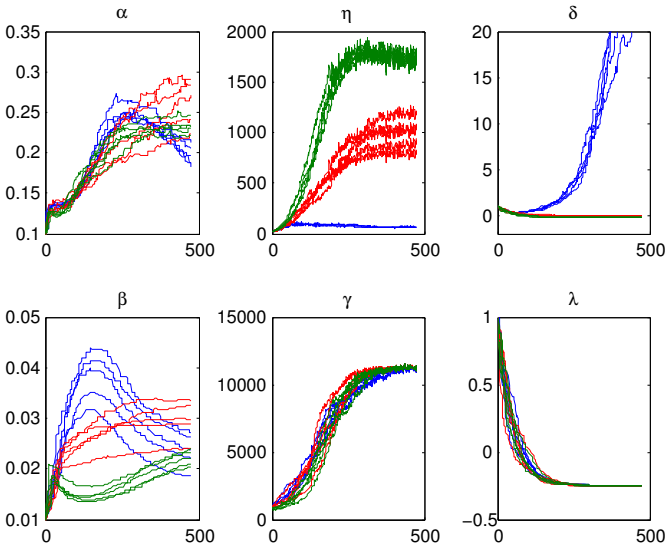


Fig. 8. Convergence of the hyperparameters for NIPS when $\zeta = 0.25$.The blue, red and green curves correspond respectively to $\sigma = 0$, $\sigma = 0.10$ and $\sigma = 0.25$.

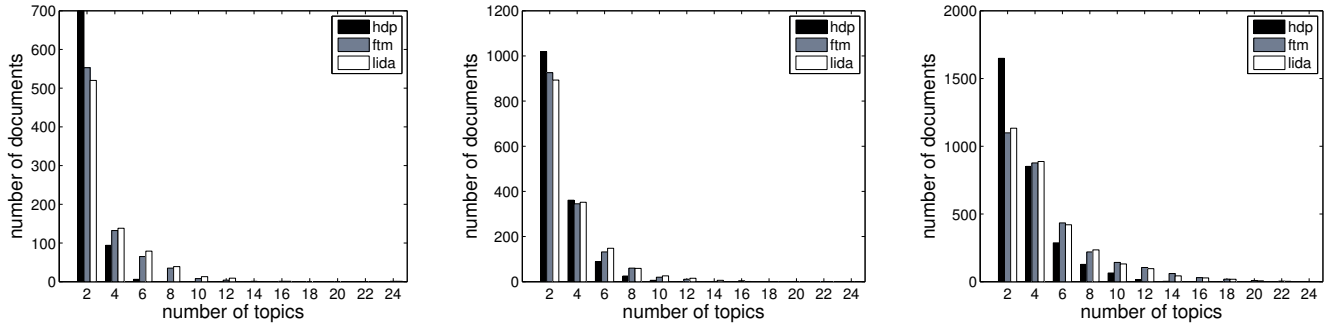
5.3 Benchmark data sets

We consider four benchmark corpora: 20 newsgroup, Reuters, KOS and NIPS. The characteristics of these data sets are reported in Table 1. The KOS and NIPS data sets are from the UCI Machine Learning Repository[1]; we did not perform any further processing, such as removal of stopwords. For the 20 newsgroup and Reuters-21758 data sets, we used the preprocessed version from [24]. Further details about the pre-processing steps is available in [24]. Figure 1 shows the power-law distribution of their word occurrences.

As a baseline, we used the Matlab implementation of the HDP-LDA topic model by Teh.[2] The mass parameters of the DP were set to 1, while the mass parameter of the Dirichlet prior on topic distribution was set to 0.10. We also compared our results to the HPY-LDA topic model; further details are available in Appendix B. We used the Chinese restaurant franchise sampler for both HDP-LDA and HPY-LDA (with discount equal to 0.25).

The results of test log-perplexity are shown in Table 1. Both, FTM and LIDA significantly outperform HDP-LDA and HPY-LDA. The results reported for FTM and LIDA are for the optimal hyperparameters. While the optimal value for σ is typically close to 0, the optimal value for ζ is typically moderate, around 0.25. This indicates that the power-law is more prominent for the word occurrences than for the topic occurrences. However, the performance gain between FTM and LIDA is modest. The concentration δ is in all cases different from 1, meaning that the two-parameter FTM similar to the one proposed in [24] is always suboptimal, as shown in the table. While its performance is close to the four-parameter FTM, it should be noted that the number of topics created in the former is typically three

1. http://archive.ics.uci.edu/ml.
2. www.stats.ox.ac.uk/~teh/research/npbayes/npbayes-r1.tgz



(a) 20 newsgroup. (b) Reuters. (c) KOS.

Fig. 9. Histogram of the number of topics per document. The FTM and LIDA assign more topics to the documents compared to HDP-LDA. We do not note a significant difference between FTM and LIDA, even though LIDA is consistently outperforming FTM in terms of test log-perplexity.

times larger or more than in the latter. Thus, accounting for the power-law enables us to learn not only a better performing model, but also a model with a lower model complexity which is much faster to learn.

Table 1 shows that FTM actually slightly outperforms LIDA on the 20 newsgroup data set. The preprocessing carried out by [24] consists among others to filter out low-frequency words, which is in favour of FTM. To assess the effect of the preprocessing, we ran additional experiments on the unprocessed version of the 1000 documents in the 20 newsgroup data, which contains 20659 unique words in the vocabulary. The perplexities are 7.017 ± 0.040, 6.999 ± 0.048, and 6.759 ± 0.036 for FTM ($\delta = 1, \sigma = 0$), unconstrained FTM and LIDA respectively. LIDA now outperforms FTM as it is able to better account for the power law in word distribution. This supports the fact that LIDA is more suitable for modelling real world data with power law characteristic.

We only report the performance of HDP-LDA and HPY-LDA on 20 newsgroup and Reuters as they did not converge in a reasonable amount of time on KOS and NIPS. Sampling the topics was extremely slow due to the dense vector of topic proportions. For example, it took more than 120 hours to run 30 training iterations for HPY-LDA on KOS or NIPS data sets. We also experimented with the C++ implementation of HDP-LDA by Wang,[3] but found that the sampler was again very slow. It can be observed that HPY-LDA outperforms HDP-LDA on Reuters data set, which is larger than 20 newsgroup, suggesting that accounting for the power-law characteristic is beneficial in more realistic settings. It is worth noting that HPY-LDA does not account for the power-law distribution of the word occurrences. Also, it should be noted that HPY-LDA performs worse than FTM, indicating that sparsity is more important than the power-law distribution of the topic proportions.

Figure 8 shows the Markov chains of the hyperparameters of LIDA for $\sigma \in \{0, 0.10, 0.25\}$. The discount parameter ζ is fixed to 0.25, which is the value that led to the test log-perplexity reported in Table 1. It can be observed that most chains stabilise after approximately 250 to 500 sampling rounds. The mass parameter α converges to a relatively large value when $\sigma \neq 0$ compared to optimal values for HDP-LDA, where it is typically equal to 0.1 or smaller to allow for a sparse topic assignment. Similarly, the mass parameter η converges to a very large value. The concentration parameter δ is slightly negative. The model behaves very differently when $\sigma = 0$: the mass parameter α is relatively small like in HDP-LDA, favouring a peaked degenerate Dirichlet posterior, while the concentration δ is large, favouring a large number of topics. In other words, the model attempts to compensate for the absence of power-law characteristics by creating many, quasi-sparse vectors of topic proportions. In all cases, the mass parameter β is relatively large compared to typical values used in HDP-LDA or HPY-LDA. The large number of very sparse topics that are created (over 4000) authorise β to be large as it is not necessary to enforce spiked word distributions.

Figure 9 compares the histograms of the number of topics per document. It can be observed that the sparse models assign a larger number of topics to each document, indicating that the documents are more accurately characterised by the topics and less topics are shared by many documents. This is confirmed when computing the average number of words per topic and the average number of documents per topic. For example, in the case of KOS, the average number of words per topic for HDP-LDA, FTM and LIDA is respectively 158, 43, and 47, while the average number of documents per topic is respectively 9, 4, and 4 The histograms of the number of topics per word (See Figure 10 for an example) indicate that the sparse models tend to learn more diverse topics.

We end our discussion by showing and comparing topics inferred by HDP-LDA, FTM and LIDA. Examples of topics extracted from Reuters are shown in Table 2 and the ones extracted from NIPS in Table 3. We selected

3. www.cs.cmu.edu/~chongw/software/hdp.tar.gz

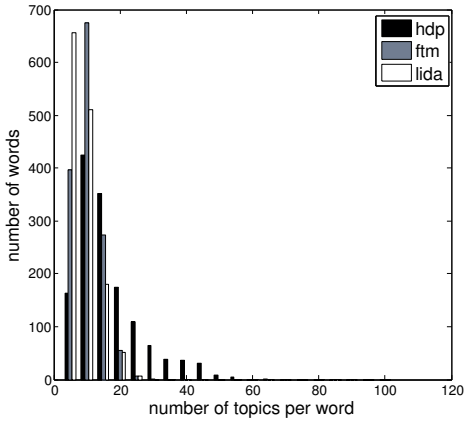


Fig. 10. Histogram of the number of topics per word for 20 newsgroup. Most words are assigned to a relatively small number of topics in FTM and LIDA, which increases the diversity of the topics.

random topics from FTM and then identified the closest topics inferred by HDP-LDA and LIDA. As distance measure between topics, we used the minimum mean absolute distance:

$$k_{\text{win}} = \underset{k}{\text{argmin}} \sum_{v} |\phi_{vk} - \phi^*_{vl}|, \tag{48}$$

where ϕ^*_l is the reference topic. The words are ordered by decreasing weight. While all models return relatively clean topics, the ranked list of words provided by the sparse models appear more coherent (e.g. third, fourth or sixth topic extracted from Reuters). When comparing the topics inferred by FTM and LIDA, the latter appear more descriptive. This is more apparent when comparing the topics extracted from NIPS (e.g. first, third or last topic).

6 Conclusion

In this work, we studied a family of partially exchangeable arrays [39] exhibiting sparse and power-law characteristics. We introduced the four-parameter IBP compound Dirichlet process (ICDP) which is a sparse alternative to the hierarchical Pitman-Yor process (PYP). The sparsity in the ICDP is controlled by a latent IBP. It was shown that the three-parameter IBP is more suitable than the one-parameter IBP when modelling real textual corpora and we expect this to apply to a wide variety of non-textual corpora. The new type of sparse nonparametric topic models we propose better fit real data in terms of predictive likelihood compared to the widely used HDP-based topic models (HDP-LDA) or its heavy-tailed counterpart (HPY-LDA). Besides the fact that the resulting topic-document and topic-word matrices are sparser and thus easier to handle in downstream applications, the advantage is that it decouples the word and/or topic occurrences in single documents and their occurrences in the corpus.

The main contributions of the paper are the introduction of a unified framework to encode sparsity both in the topic-document and topic-word matrices, and the fact that the generative model accounts for the power-law distributions of the word and the topic frequencies. We also propose a simple collapsed Gibbs sampler that scales better in terms of speed and memory requirements compared to the popular samplers currently used in HDP-LDA or in HPY-LDA.

As a concluding remark, recent advances showed that variational techniques can be used to obtain scalable topic model algorithms able to handle millions of documents [40]. These algorithms are based on deterministic approximations of the posterior distributions [2]. They rely on the stick-breaking construction of HDP-LDA and HPY-LDA. Similarly, we anticipate that the stick breaking construction of the the three-parameter IBP [28], [36] could be easily extended to scale up the three-parameter FTM and LIDA to the same data sizes.

Acknowledgments

The authors would like to thank the reviewers for constructive comments. The authors would like to thank Chong Wang and Sinead Williamson for sharing the pre-processed datasets, and Yee Whye Teh for helpful discussions.

References

[1] T. L. Griffiths and M. Steyvers, "Prediction and semantic association," in *Advances in Neural Information Processing Systems (NIPS)*, 2002.

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

[3] C. Wang, D. Blei, and D. Heckerman, "Continuous time dynamic topic models," in *International Conference on Uncertainty in Artificial Intelligence (UAI)*, 2008.

[4] A. McCallum, A. Corrada-Emmanuel, and X. Wang, "Topic and role discovery in social networks," in *International Joint Conference on Artificial Intelligence (IJCAI)*, 2005.

[5] I. Titov and R. McDonald, "A joint model of text and aspect ratings for sentiment summarization," in *International Conference on Computational Linguistics and Annual Meeting of the Association for Computational Linguistics (ACL)*, 2008.

[6] C. Wang, D. M. Blei, and L. Fei-Fei, "Simultaneous image classification and annotation," in *International Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009.

[7] Y. Wang and G. Mori, "Human action recognition by semilatent topic models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 10, pp. 1762–1774, 2009.

[8] S. K. Lukins, N. A. Kraft, and L. H. Etzkorn, "Source code retrieval for bug localization using Latent Dirichlet Allocation," in *Working Conference on Reverse Engineering (WCRE)*, 2008, pp. 155–164.

[9] B. Liu, L. Liu, A. Tsykin, G. J. Goodall, J. E. Green, M. Zhu, C. H. Kim, and J. Li, "Identifying functional miRNAmRNA regulatory modules with correspondence latent Dirichlet allocation," *Bioinformatics*, vol. 26, no. 24, pp. 3105–3111, 2010.

[10] J. Chang, J. Boyd-Graber, S. Gerrish, C. Wang, and D. Blei, "Reading tea leaves: How humans interpret topic models," in *Advances in Neural Information Processing Systems (NIPS)*, 2009.

[11] T. Hofmann, "Probabilistic latent semantic analysis," in *International Conference on Uncertainty in Artificial Intelligence (UAI)*, 1999, pp. 289–296.

[12] W. Buntine, "Variational extensions to EM and multinomial PCA," in *European Conference on Machine Learning (ECML)*, 2002, pp. 23–34.

[13] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences*, vol. 101, pp. 5228–5235, 2004.

TABLE 2
Examples of random topics associated with Reuters. The topics were matched using the minimum mean absolute error with respect to the topics extracted by FTM.

FTM

trade	shares	shares	board	company	japan	yen	ec	interest
japan	pct	common	committee	chief	japanese	japan	european	payments
reagan	stake	mln	special	executive	pact	dollar	community	loans
states	stock	pct	directors	officer	ministry	japanese	ministers	income
goods	group	share	proposal	president	semiconductor	bank	system	brazil
japanese	investment	outstanding	chairman	inc	agreement	paris	meeting	debt
united	securities	stock	inc	chairman	makers	currency	told	first
president	corp	company	acquisition	december	industry	nations	member	received
agreement	exchange	inc	offered	act	reduce	agreed	market	net
tariffs	commission	dlrs	share	suit	international	february	minister	manufacturers

HDP-LDA

trade	pct	shares	special	chairman	makers	yen	ec	payments
bill	stake	offer	committee	inc	quarter	dollar	west	interest
house	investment	dlrs	board	president	japanese	bank	monetary	income
billion	group	company	inc	executive	japan	japan	finance	brazil
foreign	shares	share	share	december	pact	dealers	meeting	loans
year	exchange	inc	directors	sale	output	dollars	need	first
imports	securities	stock	acquisition	talks	second	tokyo	countries	debt
reagan	total	mln	dlrs	company	production	currency	rates	longterm
legislation	commission	pct	companys	chief	semiconductor	bought	currencies	medium
congress	stock	corp	restructuring	officer	ministry	trading	germany	net

LIDA

trade	securities	shares	special	chief	japan	paris	system	payments
foreign	stock	common	committee	executive	japanese	currency	ec	interest
told	investment	stock	restructuring	dlrs	trade	last	agreed	income
surplus	shares	company	come	president	last	dollar	countries	brazil
last	exchange	mln	proposals	company	agreement	nations	monetary	manufacturers
products	commission	pct	interest	chairman	states	yen	central	received
imports	stake	share	acquisition	officer	action	japan	belgian	loans
president	group	outstanding	offered	inc	japans	exchange	finance	brazilian
minister	pct	shareholders	effort	corp	pact	six	market	year
goods	inc	three	firms	vice	semiconductor	stability	member	reduced

TABLE 3
Examples of random topics associated with NIPS. The topics were matched using the minimum mean absolute error with respect to the topics extracted by FTM.

FTM

model	control	classifier	chip	gradient	network	learning	optimization	direction
bayesian	controller	classification	neural	learning	system	robot	constraint	field
data	model	training	analog	descent	point	field	problem	motion
parameter	learning	pattern	weight	rate	dynamic	arm	annealing	receptive
estimator	system	error	network	stochastic	attractor	model	method	unit
variables	task	set	neuron	momentum	delay	control	objective	layer
method	critic	class	implementation	convergence	neural	dynamic	solution	visual
variance	forward	data	circuit	error	fixed	motor	energy	model
criterion	actor	mlp	digital	adaptive	stability	task	neural	selectivity
selection	architecture	decision	vlsi	parameter	connection	space	point	moving

LIDA

model	control	classifier	chip	algorithm	network	model	point	motion
data	model	classification	neural	gradient	system	movement	problem	direction
estimation	controller	training	weight	error	dynamic	field	function	point
parameter	robot	problem	bit	function	point	arm	optimization	moving
cross	learning	class	digital	descent	attractor	trajectory	objective	field
bayesian	task	decision	implementation	problem	neural	control	algorithm	model
posterior	system	set	analog	method	equation	dynamic	method	trajectory
prediction	forward	performance	hardware	convergence	dynamical	motor	annealing	unit
validation	action	error	synapse	learning	delay	point	constraint	flow
estimate	space	data	vlsi	local	fixed	hand	neural	velocity

[14] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical Dirichlet processes," *Journal of the American Statistical Association*, vol. 101, no. 476, pp. 1566–1581, 2006.

[15] G. K. Zipf, *The Psychobiology of Language.* Houghton-Mifflin, 1935.

[16] J. Pitman and M. Yor, "The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator," *Annals of Probability*, vol. 25, pp. 855–900, 1997.

[17] H. Ishwaran and L. F. James, "Gibbs sampling methods for stick-breaking priors," *Journal of the Americal Statistical Association*, vol. 96, p. 453, 2001.

[18] T. Ferguson, "A Bayesian analysis of some nonparametric problems," *Annals of Statistics*, vol. 1, no. 2, pp. 209–230, 1973.

[19] Y. W. Teh, "A hierarchical Bayesian language model based on Pitman-Yor processes," in *International Conference on Computational Linguistics and Annual Meeting of the Association for Computational Linguistics (ACL)*, 2006, pp. 985–992.

[20] F. Wood, J. Gasthaus, C. Archambeau, L. James, and Y. W. Teh, "The sequence memoizer," in *Communications of the ACM*, vol. 54, no. 2, 2011, pp. 91–98.

[21] R. Kneser and H. Ney, "Improved backing-off for m-gram language modeling," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1995.

[22] S. F. Chen and J. Goodman, "An empirical study of smoothing techniques for language modeling," *Computer Speech and Language*, vol. 13, p. 359394, 1999.

[23] I. Sato and H. Nakagawa, "Topic models with power-law using Pitman-Yor process," in *International conference on Knowledge Discovery and Data Mining (KDD)*. ACM, 2010.

[24] S. Williamson, C. Wang, K. A. Heller, and D. M. Blei, "The IBP-compound Dirichlet process and its application to focused topic modeling," in *International Conference on Machine Learning (ICML)*, 2010.

[25] C. Wang and D. M. Blei, "Decoupling sparsity and smoothness in the discrete hierarchical Dirichlet process," in *Advances in Neural Information Processing Systems (NIPS)*, 2010.

[26] T. L. Griffiths and Z. Ghahramani, "Inifinite latent feature models and the Indian buffet process," in *Advances in Neural Information Processing Systems (NIPS)*, L. Saul, Y. Weiss, and L. Bottou, Eds., 2005.

[27] T. Griffiths and Z. Ghahramani, "The Indian buffet process: An introduction and review," *Journal of Machine Learning Research*, vol. 12, pp. 1185–1224, 2011.

[28] Y. W. Teh and D. Görür, "Indian buffet processes with power-law behavior," in *Advances in Neural Information Processing Systems (NIPS)*, 2009.

[29] C. Antoniak, "Mixtures of Dirichlet processes with applications to Bayesian nonparametric," *Annals of Statistics*, vol. 2, no. 6, p. 11521174, 1974.

[30] M. Escobar and M. West, "Bayesian density estimation and inference using mixtures," *Annals of Statistics*, vol. 2, p. 577588, 1995.

[31] J. Sethuraman, "A constructive definition of Dirichlet priors," *Statistica Sinica*, vol. 4, pp. 639–650, 1994.

[32] Z. Ghahramani, T. Griffiths, and P. Sollich, "Bayesian nonparametric latent feature models (with discussion)," in *Bayesian Statistics 8*, 2007, pp. 201–226.

[33] R. Thibaux and M. I. Jordan, "Hierarchical beta processes and the Indian buffet process," in *International Conference on Artificial Intelligence and Statistics (UAI)*, 2007.

[34] J. F. C. Kingman, "Completely random measures," *Pacific Journal of Mathematics*, vol. 21, no. 1, pp. 59–78, 1967.

[35] Y. Kim and J. Lee, "On posterior consistency of survival models," *Annals of Statistics*, vol. 666, pp. 666–686, 2001.

[36] T. Broderick, M. Jordan, and J. Pitman, "Beta processes, stick-breaking, and power laws," *Bayesian Analysis*, vol. 7, no. 1, pp. 1–38, 2012.

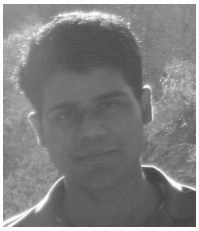
[37] N. Friedman and Y. Singer, "Efficient Bayesian parameter estimation in large discrete domains," in *Advances in Neural Information Processing Systems (NIPS)*, S. A. Solla, T. K. Leen, and K.-R. Müller, Eds., 1999.

[38] C. Robert and G. Casella, *Monte Carlo Statistical Methods.* Springer, 2004.

[39] D. J. Aldous, "More uses of exchangeability: representations of complex random structures," arXiv:0909.4339v2, Tech. Rep., 2010.

[40] C. Wang, J. Paisley, and D. Blei, "Online variational inference for the hierarchical Dirichlet process," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2011.

Cédric Archambeau received the Ph. D. in Applied Sciences from the Université catholique de Louvain, Belgium, in 2005. In October 2009, he joined the Services Innovation Laboratory at Xerox Research Centre Europe, France, where he led the Machine Learning group until October 2013. Currently, he is with Amazon, Berlin, and holds an Honorary Senior Research Associate position in the Centre for Computational Statistics and Machine Learning at University College London. His research interests include probabilistic machine learning and data science, with applications in natural language processing, relational learning, personalised content creation and data assimilation. He has published more than 40 papers in international journals and conferences.



Balaji Lakshminarayanan is a Ph. D. student at the Gatsby Computational Neuroscience Unit, University College London. He is interested in machine learning (specifically, probabilistic machine learning and efficient Bayesian inference methods) and its real world applications.

Guillaume Bouchard is a Senior Research Scientist at Xerox Research Centre Europe. He graduated from INSA de Rouen in 2001 and received his Ph. D. in Statistics from INRIA research centre in 2004 before joining Xerox. He has more than 10 years of research experience in machine learning, with a specific focus on tractable probabilistic methods, including Bayesian inference. His research impacted several real world applications in the domain of user modelling, relational data modelling, transportation as well as text understanding.