A Left-to-right Algorithm for Likelihood Estimation in Gamma-Poisson Factor Analysis

Joan Capdevila[1,2], Jesús Cerquides[3], Jordi Torres[1,2], François Petitjean[4], and Wray Buntine[4]

[1] Universitat Politècnica de Catalunya (UPC), Barcelona, Spain
[2] Barcelona Supercomputing Center (BSC), Barcelona, Spain
{jc,torres}@ac.upc.edu
[3] Institut d'Investigació en Intel.ligència Artificial (IIIA-CSIC), Bellaterra, Spain
{cerquide}@iiia.csic.es
[4] Monash University, Victoria, Australia
{francois.petitjean, wray.buntine}@monash.edu

Abstract. Computing the probability of unseen documents is a natural evaluation task in topic modeling. Previous work has addressed this problem for the well-known Latent Dirichlet Allocation (LDA) model. However, the same problem for a more general class of topic models, referred here to as Gamma-Poisson Factor Analysis (GaP-FA), remains unexplored, which hampers a fair comparison between models. Recent findings on the exact marginal likelihood of GaP-FA enable the derivation of a closed-form expression. In this paper, we show that its exact computation grows exponentially with the number of topics and non-zero words in a document, thus being only solvable for relatively small models and short documents. Experimentation in various corpus also indicates that existing methods in the literature are unlikely to accurately estimate this probability. With that in mind, we propose L2R, a left-to-right sequential sampler that decomposes the document probability into a product of conditionals and estimates them separately. We then proceed by confirming that our estimator converges and is unbiased for both small and large collections.

Keywords: Topic models · Gamma-Poisson · Factor Analysis · Left-to-right · Importance Sampling · Estimation methods.

1 Introduction

Probabilistic topic models [1] have enabled the thematic exploration of document collections at a scale which would have been unfeasible for unassisted humans. Despite the growing interest in these models, there is some disagreement on the methodologies to evaluate and compare them. Their unsupervised nature makes it difficult to propose a silver-bullet metric and this has led to a myriad of application-specific methods for evaluation ranging from document classification to word prediction. However, their probabilistic nature also suggests that computing the likelihood of a collection of held-out documents is a legitimate

measure of the generalization capabilities of these models, independent of their final application.

Although previous work [15, 4] has looked at this problem for the well-known Latent Dirichlet Allocation (LDA) [2], similar studies have not yet been conducted for related models like Gamma-Poisson (GaP) [5] and Poisson matrix factorisation (PMF) [7], and related forms of non-negative matrix factorisation [9]. GaP, PMF and their extensions, referred here to as GaP Factor Analysis (GaP-FA), represent a more general and expressive class of models [3] which explicitly take into account the document length. Because of this, GaP-FA have been successfully applied in many other domains beyond topic modeling [5, 17]. For instance, they have been used to include implicit feedback in recommendation systems [7] or to perform statistical relational learning in sparse networks [16]. Therefore, the intrinsic evaluation of GaP-FA in terms of the likelihood of held-out data becomes relevant for much broader domains, even though in this paper we only consider applications of text analysis.

Computing the probability of a single document requires integrating out all document-level latent variables. This marginal distribution has no analytical solution in the original GaP model, but recent progress in the field has enabled the derivation of a closed-form expression by means of an augmented GaP model [6]. Nonetheless, we will show that the complexity of the exact solution grows exponentially with the number of topics and the number of non-zero words in the document. Moreover, the base of the exponential depends on the maximum count of any word in the document. This means that the exact marginal is only tractable in reasonably small scenarios. Thus, approximation methods to the marginal document likelihood are essential for evaluating GaP-FA under more realistic conditions.

Simple approximation methods, such as Direct Sampling (DS) or the Harmonic Mean (HM) method [12], are known to produce inaccurate estimations, particularly in high-dimensional setups. Despite this, their ease of implementation and low computational cost have promoted their use in LDA-like models [8, 14]. As a result of this misuse, there is a need for more accurate and computationally efficient estimation methods. One approach which has been reported to output state-of-the-art results in LDA is the Left-to-right Sequential Sampler [4]. By leveraging the chain rule of probability, the algorithm decomposes the joint document probability into a product of conditionals, one conditional per word. Then, unbiased estimates can be built for each conditional given the posterior samples on the left-hand topics. However, three issues arise in GaP-FA due to the Gamma-Poisson construction: (1) The posterior distribution over the left-hand topic assignments is not tractable. (2) The computational cost of each exact conditional is exponential with the number of topics. (3) The time complexity grows quadratically with the number of non-zero words.

In this paper, we propose L2R, a left-to-right sequential sampler for GaP-FA that addresses (1) by means of Gibbs sampling the augmented model, (2) via Importance Sampling with proposal distributions conditioned on the left-hand samples (3) through a mathematical simplification that enables computing the

conditional probability for all zero words at once. Moreover, we compare the accuracy of L2R to that of existing estimation methods in two different setups:

- in reasonably small scenarios, where the exact marginal can be assessed in moderate time and hence, conclusions about their accuracy can be drawn;
- in realistic scenarios, where the exact marginal and the vanilla left-to-right are computationally unfeasible and hence, only their convergence can be studied.

In the rest of this paper, we introduce some preliminary concepts about GaP-FA in Sec. 2. In Sec. 3, we formulate the problem in terms of computing the marginal document likelihood. We present an overview of existing estimation methods in Sec. 4. In Sec. 5, we describe the L2R algorithm. Finally, Sec. 6 contains the experimental work carried out in both scenarios.

2 Background

2.1 Gamma-Poisson Factor Analysis (GaP-FA)

Poisson Factor Analysis (PFA) is a type of discrete component or factor analysis [3] with Poisson likelihoods. This means that PFA assumes that the full count matrix $\boldsymbol{Y} \in \mathbb{N}_0^{N \times W}$, where N refers to the number of documents and W to the vocabulary size, can be generated from a multivariate Poisson distribution parametrized through the product of two smaller matrices,

$$\boldsymbol{Y} \sim \text{Pois}(\boldsymbol{\Theta\Phi}) \tag{1}$$

where $\boldsymbol{\Theta} \in \mathbb{R}_+^{N \times K}$ is the factor score matrix and $\boldsymbol{\Phi} \in \mathbb{R}_+^{K \times W}$, the factor loading matrix. K refers to the dimension of the latent factors or topics in topic modeling. This method can be augmented with latent factor/topics counts x_{nwk} and express each count y_{nw} as the sum of the K independent counts,

$$y_{nw} = \sum_{k=1}^{K} x_{nwk},\ x_{nwk} \sim \text{Pois}(\theta_{nk}\phi_{kw}) \tag{2}$$

where y_{nw} are the observed word counts of the w-th word in the n-th document, and x_{nwk} corresponds to the hidden or latent counts in the k-th topic for the same document and word. θ_{nk} and ϕ_{kw} refer to the corresponding row/column entries in matrices $\boldsymbol{\Theta}$, $\boldsymbol{\Phi}$, respectively.

Several models have been developed from this by placing different types of priors over the factor score or loading matrices [17]. In this work, we restrict attention to methods that assume a Gamma distribution for each factor score. We refer to these models as Gamma-Poisson Factor Analysis (GaP-FA). This group includes Non-negative Matrix Factorization (NMF) [9], gamma-Poisson (GaP) [5] and the three hierarchical models Γ-PFA, $\beta\Gamma$-PFA, $\beta\gamma\Gamma$-PFA presented in [17], among many others. For mathematical convenience, we consider the shape-scale parameterization of the Gamma distribution,

$$\theta_{i,k} \sim \text{Ga}\left(r_k, \frac{p_k}{1-p_k}\right) \; k = 1...K \tag{3}$$

where r_k corresponds to the shape and $\frac{p_k}{1-p_k}$ to the scale of the k-th factor. To satisfy the constraints of the Gamma, we must ensure that $r > 0$ and $0 < p < 1$.

Next, we review two compound probability distributions that are the result of assuming that the rate of a univariate and multivariate Poisson distribution is controlled by a Gamma random variable. These distributions will be useful in deriving marginal and conditional likelihoods for GaP-FA.

2.2 Negative Binomial (NB)

The Negative Binomial (NB) distribution is a discrete distribution for the number of successes in a sequence of i.i.d Bernoulli trials with probability p after observing a given number of r failures. The NB can be constructed by marginalizing a Poisson distribution whose rate θ is controlled by a gamma random variable parameterized as in Eq. (3) above. In other words,

$$\text{NB}(x; r, p) = \int \text{Pois}(x|\theta)\, \text{Ga}\left(\theta; r, \frac{p}{1-p}\right) \text{d}\theta. \tag{4}$$

2.3 Negative Multinomial (NM)

The Negative Multinomial (NM) distribution [13] is the multivariate generalization of the NB distribution to W outcomes ($W > 1$), each occurring with probability q_w and for a given number of failures r.

As shown in [6], the NM can be built by marginalizing W independent Poisson distributions whose rate is controlled by a gamma random variable θ that is scaled by a vector $\phi_{:}$ of length W. This can be expressed mathematically as,

$$\text{NM}\left(x_{:}; r, q_{:} = \frac{p\phi_{:}}{1-p+p\sum_w \phi_w}\right) = \int \prod_w \text{Pois}(x_w|\theta\phi_w)\, \text{Ga}\left(\theta; r, \frac{p}{1-p}\right) \text{d}\theta \tag{5}$$

where r are the number of failures and $q_{:} = \frac{p\phi_{:}}{1-p+p\sum_w \phi_w}$ is the vector of W success probabilities. When $\phi_{:}$ is a probability vector, which sums up to 1, the success probabilities of the NM become $q_{:} = p\phi_{:}$.

3 Problem Statement

A common and reasonable strategy to compute the probability of unseen documents in topic models is to use point estimates for the set of global parameters, instead of a fully Bayesian approach which would marginalize across all parameters [15, 4]. This enables factorizing the held-out probability across documents in GaP-FA since documents are conditionally independent given the global parameters. As a result of this, the problem then reduces to calculating the *marginal*

document likelihood for each held-out document independently by integrating out the document-level latent variables. We can express this for the GaP-FA model in Eq. (1) as,

$$p(\boldsymbol{Y};\Phi,p,r)=\prod_n p(y_{n:};\Phi,p,r)=\prod_n\int p(y_{n:},\theta_{n:};\Phi,p,r)\,\mathrm{d}\theta_{n:} \tag{6}$$

where the first equality expresses that documents $y_{n:}$ are independent given the set of global parameters $\Omega=\{\Phi,p,r\}$, and the second equality says that this probability is equal to the product across all marginal document likelihoods.

Next, we focus on deriving a closed-from expression for the *marginal document likelihood* in GaP-FA, $p(y_{n:};\Phi,p,r)$, and show the computational issues that arise in its evaluation. It is important to note that the derivation and approximation of this marginal in the rest of this paper is equivalent for testing and training documents, so we will use $y_{n:}$ indistinctly to refer to both.

3.1 Exact Marginal Document Likelihood in GaP-FA

Following [6], the marginal likelihood of the n-th document in GaP-FA can be written from the augmented model in Eq. (2). Note that we can write the marginal as the sum of the marginal on $x_{n::}$ over all possible topic counts, which must add up to the observed counts $y_{n:}$ in the n-th document. This can be expressed formally as,

$$p(y_{n:};\Phi,p,r)=\sum_{x_{n::}\in\mathbb{X}_{y_{n:}}}\prod_k p(x_{n:k};r_k,p_k,\phi_{k:}) \tag{7}$$

where the summation set $\mathbb{X}_{y_{n:}}=\{x_{n::}\in(\mathbb{N}_0)^{W\times K}\mid y_{n:}=\sum_{k=1}^K x_{n:k}\}$ corresponds to all the possible partitions of the topic counts in the n-th document into K parts. Factorization across the marginals on the topic counts is due to independence across these counts, as in Eq. (2).

Then, deriving a closed-form expression for the marginal document likelihood boils down to finding an analytical expression for $p(x_{n:k};r_k,p_k,\phi_{k:})$. Following [6], this probability can be calculated by marginalizing out θ_{nk} in the augmented model. Moreover, a parametric distribution can be derived by noting that this marginal matches the NM distribution definition introduced in Eq. (5). In other words, the marginal distribution on the counts of the k-th topic can be written as,

$$\begin{aligned}p(x_{n:k};\phi_{k:},p_k,r_k)&=\int\prod_w\mathrm{Pois}(x_{nwk}|\theta_{nk}\phi_{kw})\mathrm{Ga}\left(\theta_{nk};r_k,\frac{p_k}{1-p_k}\right)\mathrm{d}\theta_{nk}\\&=\mathrm{NM}\left(x_{n:k};r_k,q_{k:}=\frac{p_k\phi_{k:}}{1-p_k+p_k\sum_w\phi_{kw}}\right)\end{aligned} \tag{8}$$

where $\phi_{k:},p_k,r_k$ are topic-dependent and $x_{n:k},\theta_{nk}$ document-dependent too.

3.2 On the Time Complexity of the Exact Marginal

Evaluating Eq. (7) means summing the independent marginals on $x_{n:k}$ over all elements in the set $\mathbb{X}_{y_{n:}}$. As shown in Eq. (8), each marginal consists of a NM distribution which has a cost linear with the number words W in an unoptimized implementation of NM, or linear with the number of non-zero word counts W_c when all zero words are evaluated together. Therefore, the cost of each summand is linear with both the number of topics K and the number of non-zeros W_c, since K marginals need to be computed for each summand.

The number of sums in Eq. (7) equals the cardinality of the set $|\mathbb{X}_{y_{n:}}|$. As shown in [6], the cardinality is given by the product of the partitions in each word w. The latter consist of the number of partitions of a natural number, i.e. y_{nw}, into K parts, which is the combinatorial term of selecting $K-1$ objects from a collection of $y_{nw}+K-1$. Therefore, the overall number of partitions for document n is $\prod_{\{w|y_{nw}\neq 0\}} \binom{y_{nw}+K-1}{K-1}$, where $\{w|y_{nw} \neq 0\}$ corresponds to the W_c non-zeros.

In the limit, one can show that this set grows exponentially with both the number of topics and the number of non-zeros $\mathcal{O}((y_{nmax})^{KW_c})$. We note that the base of the exponent is the maximum word count in the n-th document y_{nmax}. Therefore, the cost of summing over the set $|\mathbb{X}_{y_{n:}}|$ dominates the complexity of evaluating the exact marginal document likelihood.

As a result, the exact evaluation of the marginal document likelihood for GaP-FA is only tractable for reasonably small problems, such as in models with 5 topics, documents with 10 non-zero words and all words having 1 or 2 counts. However, the existence of this closed-from expression motivates the development of tailored estimation methods and to calibrate their outputs with the exact.

4 Related Work

Wallach et al. [15] presented several estimation methods for evaluating LDA in terms of held-out likelihood. Buntine [4] also compared the performance of these methods against the exact calculation for the same LDA model. The conclusion of both studies was that simple and commonly-used estimation methods fail to accurately estimate the document likelihood, specially in high-dimensional scenarios. But Wallach's Left-to-right algorithm was modified to a Sequential Sampler scheme and proven to be unbiased by Buntine. Given the quick convergences and unbiasedness properties of the Left-to-right Sequential Sampler, it can now be used as a gold standard for estimation in LDA with large number of samples.

To the best of our knowledge, no prior work exists for document likelihood estimation in GaP-FA. However, it is natural to wonder whether LDA methods can be directly applied in GaP-FA. As we have seen previously, the Gamma-Poisson construction differs from that of LDA and the time complexity of its marginal document likelihood is far more complex. The number of sums in LDA grows exponentially with the document length. Therefore, existing estimation

methods [15, 4] for LDA have to be amended accordingly. Next, we discuss the amendments and limitations imposed by GaP-FA.

In contrast with LDA, *Direct Sampling (DS)* or Importance Sampling with the prior as proposal cannot be formulated over the discrete variables of the augmented model $x_{n::}$, because the observed counts $y_{n:}$ follow a deterministic relationship with the topic counts $x_{n::}$. Therefore, DS has to be formulated over the continuous variables $\theta_{n:}$ as the Monte Carlo sampling of Eq. (6),

$$p(y_{n:};\Phi,p,r) \approx \frac{1}{S}\sum_{s=1}^{S} p(y_{n:}|\theta_{n:}^{(s)};\Phi,p,r) \qquad \text{where } \theta_{n:}^{(s)} \sim p(\theta_{n:};p,r)\ , \tag{9}$$

where the likelihood $p(y_{n:}|\theta_{n,:};\Phi,p,r)$ is W-variate Poisson with rates given by the vector $\theta_{n:}\Phi$ and $p(\theta_{n:};p,r)$ is given by Eq. (3) for each topic $k < K$. Although this estimator is unbiased, the main caveat is that the proposal distribution ignores the observed counts and too many samples might be needed when the prior is far from the joint distribution.

An alternative to this formulation is to use samples from the posterior distribution and build an unbiased estimator through the *Harmonic Mean (HM)* method [12]. To sample the posterior, one needs to consider the augmented GaP-FA and perform Gibbs Sampling on the locally conjugated complete conditionals as in [17]. Although this method has been used in LDA-like topic models [8, 14], the same authors expressed some reservations when introducing it due to the non-stable convergence and high variance. Note that this estimator cannot either be built on the discrete variables of the augmented model $x_{n::}$.

In fact, the deterministic relationship between the observed counts and the latent topic counts is what causes difficulties to tune other methods such Annealed Importance Sampling (AIS) [11], which transitions between the prior over the topic assignments and its posterior through a series of tempered distributions, or Chib-style estimators [10].

Related work in Poisson Factorization for topic modeling computes perplexity scores by holding out some random words in the document-term matrix instead of the full document [17]. A similar approach in LDA-like models consists of holding out the second half of a document, while the first half is added to the training data. The evaluation task, known as document completion [15], consists of computing the probability of the second half given the first. Although this task is known to be well correlated but biased for LDA, rigorous studies have not yet been conducted for GaP-FA. This work also paves the way for calibrating a document completion style method against the exact calculation and to develop specialized and unbiased sampling methods that approximate word prediction or document completion.

5 L2R: a Left-to-right Algorithm for GaP-FA

In this section we present L2R, a tailored left-to-right sequential sampler [4] for GaP-FA. L2R builds on the general product rule of probability, in which any

joint distribution can be decomposed into the product of several conditionals. By considering a left-to-right order of words, the joint probability of a document is decomposed by the product of W conditional probabilities where each is conditioned on the preceding left words. We can express this decomposition for GaP-FA as,

$$p(y_{n:}; \Phi, p, r) = \prod_{w=1}^{W} p(y_{nw}|y_{n<w}; \Phi, p, r) \tag{10}$$

where "$< w$" refers to words on the left side of w. Nonetheless, the exact calculation of these conditionals is still as intractable as the previous marginal likelihood. We now introduce the left topic counts $x_{n<w:}$ and marginalize them out as follows,

$$p(y_{n:}; \Phi, p, r) = \prod_{w=1}^{W} \sum_{x_{n<w:}} p(y_{nw}, x_{n<w:}|y_{n<w}; \Phi, p, r). \tag{11}$$

Given that the w-th word counts, y_{nw}, are conditionally independent from the left-hand side counts $y_{n<w}$ given their topic counts $x_{n<w:}$, the joint expression above can be split into two factors as,

$$p(y_{n:}; \Phi, p, r) = \prod_{w=1}^{W} \sum_{x_{n<w:}} p(y_{nw}|x_{n<w:}; \Phi, p, r) p(x_{n<w:}|y_{n<w}; \Phi, p, r). \tag{12}$$

This expression uncovers a sampling structure which suggests to draw samples from the posterior over the topic counts on the left-hand side of w and to evaluate the conditional probability of the current word count given these left samples. In other words, the two step process can be summarized as follows

$$x^{(s)}_{n<w:} \sim p(x_{n<w:}|y_{n<w}; \Phi, p, r) \tag{13}$$

$$p(y_{n:}; \Phi, p, r) \approx \prod_{w=1}^{W} \frac{1}{S} \sum_{s=1}^{S} p(y_{nw}|x^{(s)}_{n<w:}; \Phi, p, r) \tag{14}$$

Next, we present a method for drawing samples from the posterior over the topic counts in Eq. (13) and a strategy to approximate the inner conditionals in Eq. (14). This will enable us to address the first two issues mentioned in the Introduction. Then, we show that if we re-order documents in a particular way, we can avoid computing the product in Eq. (14) across all words in the vocabulary W, which addresses the third issue. Finally, we summarize all these contributions in the pseudo-code for the L2R algorithm and discuss its computational complexity.

5.1 Sampling the Left-hand Topics

The posterior distribution in Eq. (13) does not have a closed-form expression due to the intractable normalizing constant. Therefore, a common thing to do is

to build a Gibbs sampler to draw samples from it. However, the complete conditionals $p(x_{nw':}|x_{n<w:}^{\neg w'}, y_{n<w}; \Phi, p, r)\ \forall w' < w$ do not admit a computationally feasible sampler due to the conditioning on the observed counts $y_{nw'}$.

One way to sample from this posterior is to consider the augmented model in Eq. (2), but only over the left-hand side of w. This makes the model locally conjugate and it enables the derivation of the complete conditionals as,

$$p(\theta_{nk}|-) = \mathrm{Ga}\left(\theta_{nk}; r_k + \sum_{w'<w} x_{n<w'k}, \frac{p_k}{1 - p_k + p_k \sum_{w'<w} \phi_{kw'}}\right) \quad \forall k \leq K \tag{15}$$

$$p(x_{nw':}|-) = \mathrm{Mult}\left(x_{nw':}; y_{nw'}, \frac{\phi_{:w'}\theta_{n:}}{\sum_k \phi_{kw'}\theta_{nk}}\right) \quad \forall w' < w \tag{16}$$

where "$|-$)" refers to all variables except the conditioned. These expressions can be integrated in a Gibbs sampling scheme in which we first sample Eq. (15) and then each of the left word counts as in Eq. (16), or vice-versa. However, only samples from the left-hand topics need to be recorded for the L2R algorithm.

5.2 Approximating the Conditional Probability

The inner conditional probability in Eq. (14) can be expressed as the sum of the marginal on $x_{nw:}$ over all possible topic counts, which must add up to the w-th word count y_{nw}. Given that topic counts are independent among them, the marginal also factorizes. We can write this as,

$$p(y_{nw}|x_{n<w:}^{(s)}; \Phi, p, r) = \sum_{x_{nwk} \in \mathbb{X}_{y_{nw}}} \prod_{k=1}^{K} p(x_{nwk}|x_{n<wk}^{(s)}; \phi_{k:}, p_k, r_k). \tag{17}$$

where the summation set $\mathbb{X}_{y_{nw}} = \{x_{nw:} \in (\mathbb{N} \cup 0)^K \mid y_{n:} = \sum_{k=1}^{K} x_{n:k}\}$ has cardinality $|\mathbb{X}_{y_{nw}}| = \binom{y_{nw}+K-1}{K-1}$.

The marginal above, which is conditioned to the left samples, can be derived by leveraging on the augmented model. By introducing θ_{nk}, the probability of the actual count x_{nwk} becomes conditionally independent of the left samples $x_{n<wk}^{(s)}$ given the introduced θ_{nk}. Therefore, the left samples influence the probability over θ_{nk}, but not that over x_{nwk} as shown,

$$p(x_{nwk}|-) = \int p(x_{nwk}|\theta_{nk}; \phi_{kw}) p(\theta_{nk}|x_{n<wk}^{(s)}; \phi_{k:}, p_k, r_k)\, \mathrm{d}\theta_{nk} \tag{18}$$

where "$-$)" refers to the set $\{x_{n<wk}^{(s)}, \phi_{k:}, p_k, r_k\}$.

In the integral above, we substitute the probability over θ_{nk} for the Poisson distribution in Eq. (2) and that over θ_{nk} for the complete conditional in Eq. (15). The resulting integral corresponds to the compound probability distribution in Eq. (4), which is a Negative Binomial (NB) parameterized as follows,

$$p(x_{nwk}|-) = \text{NB}\left(x_{nwk}; r_k + \sum_{w'<w} x^{(s)}_{nw'k}, \frac{\phi_{wk}p_k}{1 - p_k + p_k \sum_{w' \leq w} \phi_{w'k}}\right). \quad (19)$$

Although it is possible to compute the exact conditional probability through the closed-form expression given by Eq. (17), its computational cost still grows exponentially with the number of topics (note that the exponential growth is now independent of the number of non-zeros) and hence it is only tractable for a small number of topics or word counts y_{nw}.

Therefore, our alternative to the exact calculation consists in replacing the complicated sum in Eq. (17) with a Monte Carlo estimate. To do that, we propose to perform Importance Sampling with a proposal distribution which is conditioned on the left samples as follows,

$$q(x_{nw:}|x^{(s)}_{n<w:}; \phi_{:w}, p, r) = \text{Mult}(x_{nw:}; y_{nw}, \propto \phi_{:w}\mathbb{E}_{p(\theta_{n:}|x^{(s)}_{n<w:}, \Phi, r, p)}[\theta_{n:}]) \quad (20)$$

where expectation over $\theta_{n:}$ is computed w.r.t the complete conditional in Eq. (15). Given that this proposal is built taking into account the left-hand samples, the proposal will be close to the marginal $x_{nw:}$ as long as the left counts are good predictors of the target.

Finally, we estimate the conditional probability as,

$$\begin{aligned} x^{(s')}_{nw:} &\sim q(x_{nw:}|x^{(s)}_{n<w:}; \phi_{:w}, p, r) \\ p(y_{nw}|x^{(s)}_{n<w:}; \Phi, p, r) &\approx \frac{1}{S'} \sum_{s'} \frac{p(x^{(s')}_{nw:}|x^{(s)}_{n<w:}; \Phi, p, r)}{q(x^{(s')}_{nw:}|x^{(s)}_{n<w:}; \phi_{:w}, p, r)} \end{aligned} \quad (21)$$

where S' corresponds to another set of samples which replace the intractable sum in Eq. (17). However, we will show in the experiments that with one single sample $S' = 1$, we can accurately approximate the exact in situations where the topics for the w-th word are likely to be predicted from the preceding topics, which is often the case if some thematic structure exists in the corpus.

5.3 Dealing with zero words

The left-to-right decomposition rule in Eq. (10) does not impose any specific word order to be valid. Besides, the inspection of the exact conditional formula from Eqs. (17) (19) reveals that words without counts contribute with a tractable term which only depends on the left-hand counts.

This suggests that if we re-order documents in such a way that all non-zero words precede zeros, we can re-use the posterior samples drawn for non-zero words to calculate the probability of zeros. Note that zeros do not contribute to the posterior sampling over the left-hand topics. This allows one to build a conditional probability for all words without counts $n \geq w_z$ that occur after the non-zeros $< w_z$. A closed-form expression can be derived for this probability which can be computed in linear time with the number of topics as,

$$p(y_{n\geq w_z}|x^{(s)}_{n<w_z:};\Phi,p,r)=\prod_k\left(\frac{1-p_k+p_k\sum_{w'<w_z}\phi_{w'k}}{1-p_k+p_k\sum_{w'\leq W}\phi_{w'k}}\right)^{r_k+\sum_{w'<w_z}x^{(s)}_{nw'k}}. \tag{22}$$

By re-ordering the document, reusing the posterior samples and the mathematical simplification shown above, we can speed up the algorithm from computing the conditional probability across all words in the vocabulary W to only those with non-zero counts W_c. Given that for most corpora, the vocabulary size is larger than the non-zero words per document ($W \geq W_c$), this makes a critical enhancement to the time-complexity of this algorithm as we show later.

5.4 Algorithm Pseudocode

In Algorithm 1, we present the pseudocode of L2R, summarizing the developments from the previous sections. The input data consists of the number of samples S used to approximate each of the factors in the left-to-right decomposition, the number of samples S' to draw from the proposal distribution in the case of sampled conditionals, the n-th document $y_{n:}$ sorted as in Section 5.3 and the point estimates for the global parameters $\Omega = \{\Phi, p, r\}$. The algorithm outputs the approximate marginal document likelihood $p(y_{n:};\Phi,p,r)$.

Algorithm 1: L2R algorithm

input : $S, S', y_{n:}, \Omega = \{\Phi, p, r\}$
output: $p(y_{n:};\Omega)$

1 **for** $w \leftarrow 1$ **to** W_c **do**
2 **for** $s \leftarrow 1$ **to** S **do**
3 $x^{(s)}_{n<w:} \leftarrow$ `PostSamp`$(x^{(s)}_{n<w:}, \Omega)$; *Eqs.* (15) (16)
4 $p(y_{nw}|x^{(s)}_{n<w:};\Omega) \leftarrow$ `CondProb` $(x^{(s)}_{n<w:}, \Omega, S')$; *Eq.* (21)
5 $p(y_{nw}|y_{n<w};\Omega) = \frac{1}{S}\sum_s p(y_{nw}|x^{(s)}_{n<w:};\Omega)$
6 $w_z \leftarrow W_c + 1$
7 **for** $s \leftarrow 1$ **to** S **do**
8 $x^{(s)}_{n<w_z:} \leftarrow$ `PostSamp`$(x^{(s)}_{n<w_z:}, \Omega)$; *Eqs.* (15) (16)
9 $p(y_{n\geq w_z}|x^{(s)}_{n<w_z:};\Omega) \leftarrow$ `CondProbZeros` $(x^{(s)}_{n<w_z:}, \Omega)$; *Eq.* (22)
10 $p(y_{nw_z}|y_{n<w_z};\Omega) = \frac{1}{S}\sum_s p(y_{n\geq w_z}|x^{(s)}_{n<w_z:};\Omega)$
11 $p(y_{n:};\Omega) \approx \prod_{w\leq w_z} p(y_{nw}|y_{n<w};\Omega)$

From line 1 to 5, the algorithm approximates the conditionals distributions for non-zero words by computing the averaged probability across S samples for each word. To approximate this conditional probability, the algorithm uses the Importance Sampling scheme defined in Eq. (21).

From line 6 to 10, the algorithm approximates the conditionals for all words without counts following the same procedure as for non-zeros, except that the conditional for all non-zeros is computed at once in line 9 through its exact form given by Eq. (22).

The final estimate for marginal document likelihood is build from the product of the $W_c + 1$ probabilities in line 11.

5.5 On the Time Complexity of the L2R algorithm

The time complexity of the L2R algorithm can be derived from the cost of the subprocesses in Algorithm 1. We first note that the cost of computing the conditionals for all non-zero words dominates over that of zeros because line 4 is linear with both the number of samples S' and the number of topics K, whereas line 9 is only linear with the latter. The cost of the posterior sampling process in line 3 and 8 is also linear with the number of topics K and non-zeros W_c. Therefore, the overall cost is given by $\mathcal{O}(W_c S(W_c + K + S'))$ which is quadratic in the number of non-zero words. Note also that without the optimization of zeros it would have been quadratic with the vocabulary size and without the approximate conditionals, exponential with the number of topics.

6 Empirical results

In this section, we present the comparison results of the L2R algorithm against the exact marginal likelihood, Direct Sampling (DS) and the Harmonic Mean (HM) method. The code for L2R, DS and HM methods, as well as the processed corpora and trained GaP-FA have been made public[5].

6.1 Experiment setup

We follow the setup in [4] which first compares methods against the exact in tractable scenarios and then looks at convergence in more realistic cases. In addition, we introduce a proper comparison metric, new document collections and a model which also infers the number of topics.

Experiments We define two sets of experiments. The first consists of comparing the output probabilities of each estimator against the exact marginal likelihood. Given that the computational complexity of the marginal likelihood is only tractable in *reasonably small scenarios* and it is dominated by the number of sums in Eq. (7), we restrict to scenarios in which the cardinality of the summation set is less than 10^9. To do this, we choose a maximum of 1000 documents from a downsized corpus whose word counts do not exceed this limit in a GaP-FA with a maximum of 5 topics. The second set of experiments consists of assessing the estimator's convergence in *more realistic conditions* for which the marginal likelihood is not tractable. In this setup, we use 1,000 evaluation documents for each collection in a GaP-FA with a maximum of 100 topics.

[5] https://github.com/jcapde/L2R https://doi.org/10.7910/DVN/GDTAAC

Dataset	Vocabulary	Num. Docs	Doc. Length
NIPS	11,463	5,811	$1,899 \pm 513$
AP	10,473	2,246	194 ± 111
20NGs	11,928	18,846	123 ± 247
Reuters	8,843	19,043	79 ± 75
Twitter	6,344	10,523	25 ± 4
WS	4,679	12,309	9 ± 3

Table 1: Document collections

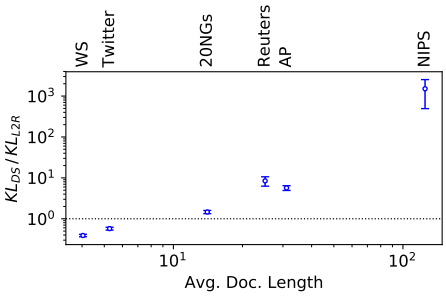

Fig. 1: KL Comparison of L2R vs. DS

Comparison measures To compare several document probabilities to their true marginal, we propose to use the *Kullback-Leibler (KL)* or relative entropy, which is a proper divergence measure for probability distributions. We can interpret it as the number of extra bits required from using the estimated probabilities instead of the exact in decoding a codebook of length the number of evaluated documents. To study the convergence in realistic scenarios, we plot the *log-likelihood* for all evaluated documents as function of the number of samples.

Document collections Table 1 contains the 6 collections used in the experimentation. All datasets, except NIPS which was used as it is published, were pre-processed by removing stopwords, non-letters and words with two or less characters. We have also applied Porter Stemming and filtered out words that appeared less than 5 times or in more than 50% of documents. Then, vocabularies were cropped to the 100 most frequent words for experiments with the exact marginal and they were used as in Table 1 for experiments in realistic conditions. Note that collections in Table 1 are ordered decreasingly on the average document length, being datasets at the top commonly used as long-text corpus, while those at the bottom used in short-text studies.

Model hyperparameters, training and samplers parameters Among all possible GaP-PFA models, we have chosen to train the $\beta\Gamma$-PFA model [17]. This model also corresponds to the GaP model [5] with inference on the number of topics by placing a Beta Process over the p hyperparameter in the Gamma-distributed factors from Eq. (3). This allows us to avoid model selection on a critical parameter such the number of topics. But, several other model hyperparameters need to be specified, such as the maximum number of topics which was set as described above for the different experiments, the Dirichlet prior over Φ which was set $\alpha = 0.1$, the scale for gamma $r = 1$ and the Beta hyperparameters $c = 1$ and $\epsilon = 1/K_{max}$, as in [17]. The training of the global parameters Φ, p, r is performed with the complete collection following the Gibbs Sampling scheme in [17] which runs for 1000 iterations and discards a burn-in period of 500. Regarding the samplers, we varied the number of samples up to $S = 10,000$ for

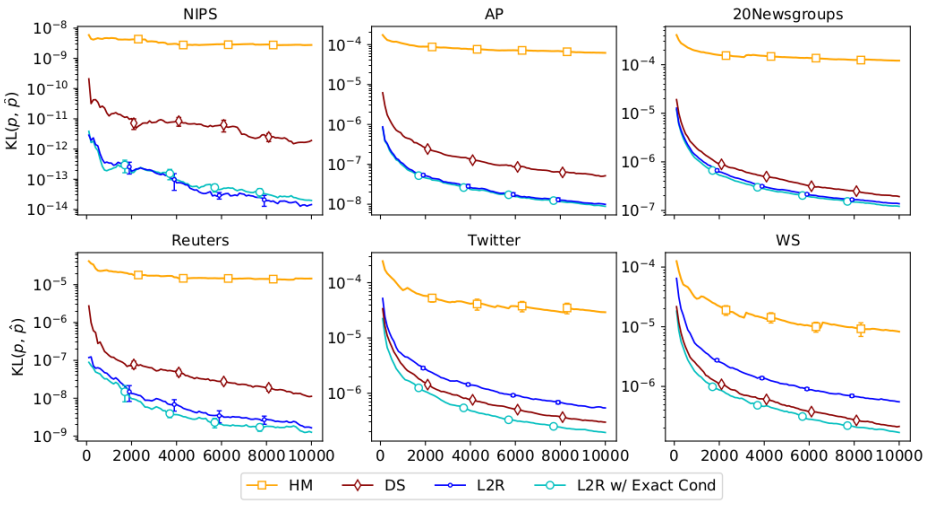

Fig. 2: Relative Entropy or KL between the estimated document probabilities and the exacts as a function of samples used. (Lower KL is better)

all methods, and we used $S' = 1$ for L2R to keep the same overall number of samples for all estimators.

6.2 Experiments in dimensioned document collections

Fig. 2 shows the KL divergence between the exact and estimated probabilities as a function of the number of samples used by each estimation method. In this experiment, we have included the L2R with exact conditionals given by Eqs. (17)-(19) to compare against the proposed sampling. We have calculated the KL for all 4 estimation methods in the 6 collections with 1000 documents, except in NIPS and AP which only contained 1 and 460 documents with a tractable marginal, respectively. Each experiment was repeated 10 times and we plotted their mean and standard error.

Results show that L2R with exact conditionals achieves the lowest KL across all 6 datasets, followed very closely by the proposed L2R algorithm with $S' = 1$ which obtains the second lowest KL in 4 datasets. We note that L2R performs worse than DS in Twitter and WS datasets, which both are the shortest text datasets. This poor performance in short-text can also be explained by the fact that vanilla topic models struggle to learn predictive topic structure due to few word co-occurrence in a document, and hence the proposal in Eq. (20) is not close enough to the target to accurately estimate the conditionals with a single sample. Unreported experiments confirm that a larger S' makes the L2R estimates closer to those of the L2R with exact conditionals. In Fig. 1, we have compared the quality of L2R vs DS, as per the results obtained in the last sample of Fig. 2, as a function of the average document length of the downsized corpora. We observe

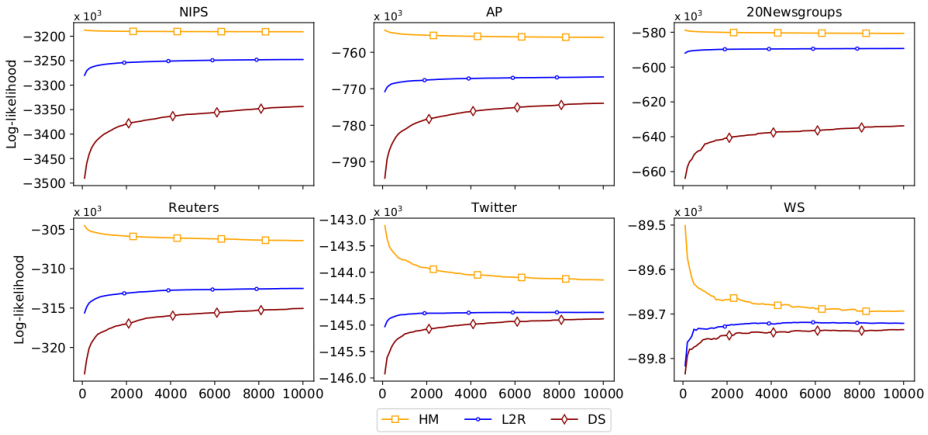

Fig. 3: Log-likelihood of the evaluated documents as a function of the number of samples.

a favorable tendency for L2R with longer documents which motivates its use in more realistic scenarios.

6.3 Experiments in realistic document collections

In Fig. 3, we plot the log-likelihood of 1,000 documents as a function of samples for the three methods that scale to the realistic scenario described above. Results show that L2R converges faster than DS in all 6 collections. The HM method also has a good convergence rate in the 4 datasets with longest documents, although the inaccuracy reported previously suggests that the method might be over-estimating the document likelihood like in LDA [15, 4]. Therefore, the fast convergence and the fact that its estimates are sandwiched by estimators that tend to under- and over- estimate, validates L2R's use for document likelihood estimation in GaP-FA with just a few hundred samples.

7 Conclusions

In this paper, we have proposed L2R, a left-to-right algorithm for estimating the marginal document likelihood in GaP-FA. The accurate estimation in reasonably small scenarios and the quick convergence in realistic scenarios encourages its use for evaluating and comparing GaP-FA topic models in terms of unseen document likelihood.

Future work should explore new estimation methods capable of reducing the time complexity of L2R, which is quadratic in the number of non-zero words. Exploring the use of these methods and the exact calculation for other evaluation tasks like document completion or word prediction is another interesting avenue for future work.

Acknowledgements This work was supported in part by Obra Social "La-Caixa", by the Australian Research Council under award DE170100037, by the SGR programs of the Catalan Government (2014-SGR-1051, 2014-SGR-118), by the Severo Ochoa Program SEV2015-0493 and by the the Spanish Ministry of Economy and Competitivity (MINECO) and the European Regional Development Fund (ERDF) under contracts TIN2015-65316 and Collectiveware TIN2015-66863-C2-1-R (MINECO/FEDER).

References

1. Blei, D.M.: Probabilistic topic models. Communications of the ACM **55**(4), 77–84 (2012)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent Dirichlet allocation. the Journal of machine Learning research **3**, 993–1022 (2003)
3. Buntine, W., Jakulin, A.: Discrete component analysis. In: Subspace, Latent Structure and Feature Selection, pp. 1–33. Springer (2006)
4. Buntine, W.L.: Estimating likelihoods for topic models. ACML **9**, 51–64 (2009)
5. Canny, J.: GaP: a factor model for discrete data. In: Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. pp. 122–129. ACM (2004)
6. Filstroff, L., Lumbreras, A., Févotte, C.: Closed-form marginal likelihood in gamma-Poisson factorization. arXiv preprint arXiv:1801.01799 (2018)
7. Gopalan, P., Ruiz, F.J., Ranganath, R., Blei, D.: Bayesian nonparametric Poisson factorization for recommendation systems. In: Artificial Intelligence and Statistics. pp. 275–283 (2014)
8. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proceedings of the National Academy of Sciences **101**(suppl 1), 5228–5235 (2004)
9. Lee, D.D., Seung, H.S.: Algorithms for non-negative matrix factorization. In: Advances in neural information processing systems. pp. 556–562 (2001)
10. Murray, I., Salakhutdinov, R.R.: Evaluating probabilities under high-dimensional latent variable models. In: Advances in neural information processing systems. pp. 1137–1144 (2009)
11. Neal, R.M.: Annealed importance sampling. Statistics and computing **11**(2), 125–139 (2001)
12. Newton, M.A., Raftery, A.E.: Approximate Bayesian inference with the weighted likelihood bootstrap. Journal of the Royal Statistical Society. Series B (Methodological) pp. 3–48 (1994)
13. Sibuya, M., Yoshimura, I., Shimizu, R.: Negative multinomial distribution. Annals of the Institute of Statistical Mathematics **16**(1), 409–426 (1964)
14. Wallach, H.M.: Topic modeling: beyond bag-of-words. In: Proceedings of the 23rd international conference on Machine learning. pp. 977–984. ACM (2006)
15. Wallach, H.M., Murray, I., Salakhutdinov, R., Mimno, D.: Evaluation methods for topic models. In: Proceedings of the 26th Annual International Conference on Machine Learning. pp. 1105–1112. ACM (2009)
16. Zhao, H., Du, L., Buntine, W.: Leveraging node attributes for incomplete relational data. In: International Conference on Machine Learning. pp. 4072–4081 (2017)
17. Zhou, M., Hannah, L., Dunson, D.B., Carin, L.: Beta-negative binomial process and Poisson factor analysis. In: International Conference on Artificial Intelligence and Statistics. pp. 1462–1471 (2012)