Submitted to the Annals of Statistics

CONSISTENCY IN LATENT ALLOCATION MODELS

By Andrew Womack, Elias Moreno and George Casella

University of Florida and Universidad de Granada

A probabilistic formulation for latent allocation models was introduced in the machine learning literature by Blei et al. (2003) in the study of a corpora of documents. This article addresses the consistency properties of various posterior probabilities on the space of latent allocations, focusing on the "bag of words" model. It is shown that the Latent Dirichlet Allocation and Ewens-Pitman priors provide inconsistent Bayesian inference for the number of latent contexts generating a collection of samples. Hierarchical priors are derived for the latent allocation problem and it is shown that they provide consistent posterior inferences in a collection of samples under appropriate conditions.

1. Introduction. A probabilistic formulation for latent allocation models was introduced in the machine learning literature in the early 2000s (Blei et al., 2003). The original motivating example was in language processing, where the goal was to model corpora of documents. After this, the adoption and modification of these models moved quickly.

Latent allocation models arise in the analysis of discrete data where each of n observations in a sample can take one of V different values, called features. In contrast to traditional models where each observation arises from the same data generating process, latent allocation models assume that each observation arises from one of T different possible data generating processes, referred to as contexts. Thus, the set of observations that take a particular value, say v, is modeled as a product of models for these T contexts. For the language processing application, the features that can be observed belong to a list of vocabulary words, the contexts are the topics that are being discussed in a document, and the document represents a sample.

The process giving rise to the latent allocation model is equivalent to an urn process. For a sample (document) there are T urns (topics) filled with varying numbers of balls. Each urn has balls of V different colors (features), and the proportion of each color varies among the urns. For each observation in the sample, an urn is chosen with probability proportional to the number of balls within the urn. Then a ball is drawn at random from the chosen urn, its color is recorded, and the ball is returned to the urn. Only the color of the ball is recorded, and not the urn from which it was drawn. A sample is generated by performing n draws from the T urns.

AMS 2000 subject classifications: Primary 62F15, 62C10

Keywords and phrases: Latent Dirichlet Allocation, Bayes' Factors, Clustering, Consistency

Thus, the latent allocation model is described by two processes. The first process generates contexts (topics or urns), and the second generates features (words or balls) within a context. Each observation in a sample, indexed by i = 1,...,n, gives rise to two key variables. The latent random variable z_i describes what context observation i is drawn from and can take values 1,...,T. The random variable w_i records the observed feature and can take values 1,...,V . In the simplest formulation of the model, called the "bag of words" model, the ordering of the observations in a sample is unimportant. For the language processing example, this implies that there is no syntactic structure.

For the "bag of words" model, the ordering of the observed features is irrelevant and the contexts are not observed. Hence, the data can be summarized by the proportion of observations attributed to each element of {1,. . . ,V}. Call these proportions $\{\gamma_v\}^V_{v=1}$. As the number of observations n grows, the uncertainty in the point estimate γ_v diminishes. Asymptotically, the data provide consistent estimators of the true proportions, γ^*_v . However, a key limitation is that the data provide no other information about the model, specifically about the unobserved contexts z.

Latent allocation models have moved far beyond their original formulation and application. As a tool in supervised and unsupervised learning, they have proved quite popular for classification (Li and Fei-Fei, 2007; Xing and Girolami, 2007; Li et al., 2009; Zhang et al., 2009). Issues in implementation have been addressed through variational and MCMC methods (Mariote et al., 2007; Newman et al., 2007; Teh et al., 2007; Porteous et al., 2008; Canini et al., 2009). Finally, additional structure has been built into the models to account for more complicated data generating processes (Teh et al., 2006; Wang and Grimson, 2007; Doyle and Elkan, 2009; Endres et al., 2009).

The data generating process for the sample $w = (w_1,...,w_n)$ is described conditional on the unobserved $z = (z_1,...,z_n)$ by a product of multinomial distributions. The goal is to carry out an asymptotic analysis of the posterior distribution of z, which will allows for the determination of an appropriate prior distribution. Here the focus is on the "bag of words" formulation of the latent allocation model for two reasons. First, consistency properties of the model have not been discussed in the literature although the prior utilized in the Latent Dirichlet Allocation model produces an inconsistent Bayesian procedure. Second, this analysis can provide insights into how to address

Page 3

consistency in the numerous extensions of the original model.

In Section 2, in addition to the original Latent Dirichlet Allocation (LDA) prior for z, consideration is given to three alternative priors: a variation of the Ewens-Pitman (EP) (Pitman, 1996), a variation of the Hierarchical Uniform Prior (HUP) (Casella et al., 2011), and a new Non-Uniform Hierarchical Prior (NUHP). The asymptotic posterior behavior under each of these priors for a single sample is explored in Section 3. In section 4, inferences about the behavior of the HUP and NUHP priors in a collection of samples and the potential clustering of samples are considered. The paper concludes with a brief discussion of open questions.

2. Likelihood and Prior Specifications for a Single Sample.

2.1. Likelihood Specification. The likelihood of z for the data w is derived using the assumption that the set of observations that have a common context $j \in \{1,...,T\}$ are exchangeable. DeFiniti's Theorem implies that these features are conditionally independent with a multinomial sampling distribution with parameter β_j; this is the conditional sampling distribution $p(w_i|z_i = j,\beta_j)$. The distribution of these observed features is completed by a prior distribution for β_j, which is assumed to be a Dirichlet distribution with parameter $\alpha_j = (\alpha_{j,1},...,\alpha_{j,V})$.

Thus, the sampling distribution of w given z is given by

$$w_i|z_i,\beta, n, T \sim \text{Multinomial}(1,\beta_{z_i}) \tag{1}$$
$$\beta_j|\alpha, n, T \sim \text{Dirichlet}(\alpha_j) \text{ for } j = 1,\ldots,T$$

Because z allocates the observations into T latent contexts, the likelihood factors as a product over the T contexts as

$$p(w|z,\alpha, n, T) = \prod_{j=1}^{T} \left[\left(\prod_{k=1}^{V} \frac{\Gamma(n_{j,k}(z,w) + \alpha_{j,k})}{\Gamma(\alpha_{j,k})} \right) \frac{\Gamma(\sum_{k=1}^{V} \alpha_{j,k})}{\Gamma(n_j(z) + \sum_{k=1}^{V} \alpha_{j,k})} \right] \tag{2}$$

where $n_{j,k}(z,w)=\#\{i : w_i = k \;\&\; z_i = j\}$ and $n_j(z)=\#\{i : z_i = j\}$. Further, it is assumed that each $\alpha_{j,k} > 0$.

This definition encapsulates two distinct models. If it is assumed that the β_j have a common prior ($\alpha_j = \alpha$ for all j), then this sampling distribution fits into the framework of product partition modeling (Hartigan, 1990; Barry and Hartigan, 1992). In this paper the condition is relaxed, following Doyle and Elkan (2009), to allow context specific priors for the context specific multinomial distributions. In supervised learning, this allows context specific feature distributions. In a single sample, the assumption that $\alpha_t = \alpha$ for all

t introduces a multiplicity of $\frac{T!}{(T-p)!}$ if p contexts are used in z. Since this multiplicity is finite, any asymptotic results that are obtained in the model with differing α_t will hold for the model with a single α.

The effect of different priors on z is explored in two ways. First, the asymptotic distribution of $F = \left(\frac{n_1(z)}{n}, \cdots, \frac{n_T(z)}{n}\right) = (f_1,\ldots,f_T)$ is explored. In particular, if p contexts are used in z but only t of them are used infinitely often, then the differences in the penalization as a function of n, p, and t are of interest. For these discussions, z is fixed with the properties that: 1) $n_{j_1}(z),\ldots,n_{j_t}(z)$ are increasing such that $\frac{n_{j_k}(z)}{n} \to p_{j_k} > 0$ as $n \to \infty$; 2) $n_j(z)$ is finite as $n \to \infty$ for $j \in \{j_1,\ldots,j_p\}\backslash\{j_1,\ldots,j_t\}$; and 3) $n_j(z)=0$ for $j \,/\!\in \{j_1,\ldots,j_p\}$.

2.2. Latent Dirichlet Allocation (LDA). To develop the prior, the z_i are assumed to be exchangeable, and once again the deFiniti Theorem is invoked. The z_i are given a multinomial prior with parameter $\theta = (\theta_1,\ldots,\theta_T)$ and θ is given a Dirichlet distribution with parameter $a = (a_1,\ldots,a_T)$. This hierarchical formulation,

$$z_i|\theta, n, T \sim \text{Multinomal}(1,\theta)$$
$$\theta|a, n, T \sim \text{Dirichlet}(a),$$

gives rise to a Multinomial-Dirichlet prior,

$$\pi_{LDA}(z|a, n, T) = \left(\prod_{j=1}^{T} \frac{\Gamma(n_j(z) + a_j)}{\Gamma(a_j)} \right) \frac{\Gamma\left(\sum_{j=1}^{T} a_j\right)}{\Gamma\left(n + \sum_{j=1}^{T} a_j\right)}, \tag{3}$$

called the Latent Dirichlet Allocation model because the latent states z, which allocate the contexts, are given a Multinomial-Dirichlet sampling distribution.

The random variable $F = \left(\frac{n_1(z)}{n}, \cdots, \frac{n_T(z)}{n}\right)$ follows a Dirichlet(a) distribution asymptotically, and thus places no mass on the sub-faces of the T-simplex asymptotically. Observing F with any $f_i = 0$ has probability 0 asymptotically. In order to obtain a posterior where at least one of the f_i is 0, the likelihood of data must overcome the fact that the prior only places mass on the interior of the simplex asymptotically. As discussed in Section 3, this does not occur with the latent allocation likelihood. However, if restricted to a sub-face, the LDA converges to a proper measure. In particular, restricting to the z using contexts $j_1,\dots,j_p$ produces an asymptotic measure for $(f_{j_1},\dots,f_{j_p})$ that is a Dirichlet distribution with parameter $(a_{j_1},\dots,a_{j_p})$.

The LDA prior has poor behavior in terms of penalization as a function of the number of used contexts p as n increases. Applying the Stirling approximation for the Γ function provides

$$\pi_{LDA}(z|n, T,\alpha,a) \propto n^{\frac{1-t-2\sum_{j\notin\{j1,\dots,jt\}} a_j}{2}} \left(\prod_{k=1}^{t} p_{jk}^{np_{jk}} \right) \tag{4}$$

as the asymptotic form of the LDA prior. As can be seen in (4), the LDA prior penalizes asymptotically based on the number of topics used infinitely often, t, as opposed to the actual number of topics used, p. This produces behavior that does not properly penalize z and leads to inconsistent posterior inference for p.

2.3. Ewens-Pitman (EP). The latent allocation problem can be viewed as a clustering problem. A prior is first defined on the space of clusters and then, conditioned on a particular partition of $[n] = \{1,...,n\}$, a conditional prior is defined on the topics assigned to the clusters. The Ewens-Pitman prior arises as the marginal prior over partitions in a Dirichlet Process prior and thus has a single parameter λ. $\varrho(z)$ is a fixed partition assignment defined by a particular z and $n(\varrho)=(n_1(\varrho),n_2(\varrho),...,n_p(\varrho))$ is defined as the vector of cluster sizes for the partition ϱ (here the subscripts have no relationship to contexts). The EP prior for ϱ is given by

$$(5) \qquad \pi_{EP}(\varrho l n,T,\lambda) = c_{n,T} \frac{\Gamma(\lambda)}{\Gamma(n+\lambda)} \lambda_p \prod_{j=1}^{p} \Gamma(n_j(\varrho))$$

where $c_{n,T}$ is a normalizing constant which accounts for the fact that $1 \le p \le T$. The EP prior is completed for the latent allocation problem by placing a prior on the contexts used by z given ϱ. The uniform prior is a natural choice, providing $\frac{(T-p)!}{T!}$ if ϱ is a partition consisting of p clusters. This provides an unconditional EP prior for z given by

$$(6) \qquad \pi_{EP}(z l n,T,\lambda) = c_{n,T} \frac{(T-p)!}{T!} \frac{\Gamma(\lambda)}{\Gamma(n+\lambda)} \lambda_p \prod_{k=1}^{p} \Gamma(n_{j_k}(z))$$

where $n_{j_k}(z) > 0$ for $k = 1,...,p$ and $n_j(z) = 0$ for all $j /\in \{j_1,...,j_p\}$.

The random vector F is understood on the simplex by considering its behavior when restricted to sub-faces and the probability of various sub-faces. The set of z that use the contexts $j_1,...,j_p$ provides information about the behavior of F for $f_i = 0$ for $i /\in \{j_i,...,j_p\}$. In order to determine the distribution of F, one has to count the number of z with $n_{j_k}(z) =$

$n_{j_k}(z)$ for $k = 1,...,p$, of which there are $\binom{n}{n_{j_1}(z),\cdots,n_{j_p}(z)}$. Thus, the conditional probability of observing $n_{j_k}(z) = n_{j_k}$ is given (up to proportion) by $(n_{j_1}\times\cdots\times n_{j_p})_{-1}$. This quantity is maximized when $p - 1$ of the n_{j_k} are 1 and the remaining n_{j_k} is $n - p + 1$. Thus, when restricted to a face of the T-simplex, the prior pushes probability towards the corners of the face. In fact, the distribution of F converges in the weak sense to the Haldane measure, $(f_{j_1}\times\cdots\times f_{j_p})_{-1}$, asymptotically.

The behavior of the EP prior on the simplex is completed by determining the probability of various faces, which is given by

(7)

$$\pi_{EP}(\text{contexts } j_1,\dots,j_p|n,T,\lambda) \propto \frac{(T-p)!}{T!}\lambda_p \sum_{\substack{n_1+\cdots+n_p=n\\ n_1>0,\dots,n_p>0}} \frac{n}{n_1\times\cdots\times n_p}$$

Calculation of (7) requires the burdensome computation of the sum in (7). Define this sum to be $H_p(n)$. It is not difficult to show that $\frac{H_{p+1}(n)}{H_p(n)} \to \infty$ as $n \to \infty$. In fact $H_p(n)$ increases at a rate of $(\log(n))^{p-1}$. Thus, though the prior when restricted to a face converges to the Haldane measure, the prior asymptotically places all of its mass on the full T simplex and not a sub-face.

Alternatively, the EP prior can be viewed in terms of penalization. Applying the Stirling approximation for the Γ function provides

$$\pi_{EP}(z|n,T,\lambda) \propto \lambda_p n^{\frac{1-t-2\lambda}{2}} \left(\prod_{k=1}^{t} p_{j_k}^{n p_{j_k}} \right) \tag{8}$$

as the asymptotic form of the EP prior. As with the LDA prior, (8) shows that the EP prior penalizes (as a function of n) asymptotically based on the number of topics used infinitely often, t, as opposed to the actual number of topics used, p. Instead of using the a_j as in (4), the penalization in (8) uses λ in the power of n. Thus the asymptotic posterior behavior of the EP posterior should be similar to that of LDA posterior.

2.4. Hierarchical Uniform Prior (HUP). The EP prior places mass 1 on the full T-simplex asymptotically. and when restricted to a face converges weakly to a non-integrable measure. An alternative prior is the Hierarchical Uniform Prior, which places a prior on the T-simplex providing better behavior on both counts. The HUP for a partition ϱ is defined by utilizing uniform priors hierarchically. A given partition ϱ has the following related quantities: p(ϱ) is defined to be the number of clusters in the partition

ϱ; and class(ϱ) = $(n_1(\varrho),\dots,n_p(\varrho))$ is defined to be the partition class of ϱ. Further, b(n, p) is defined as the number of partition classes of [n] into p clusters, which are in a one-to-one correspondence with the number of

Young's diagrams containing n boxes in p rows. For a given class $(n_1,\dots,n_p)$, $N(n_1,\dots,n_p)$ is defined as the number of partitions of that class, which are in one-to-one correspondence with filling a particular Young's diagram with integers 1,...,n, up to permutation of rows that have the same length. The HUP is defined by

$$\pi_{HUP}(p \text{ contexts}|n, T) = \frac{1}{T} \tag{9}$$

$$\pi_{HUP}(\text{class}(n_1,\dots,n_p)|p \text{ contexts}, n, T) = \frac{1}{b(n, p)} \tag{10}$$

$$\pi_{HUP}(\varrho|\text{class}(n_1,\dots,n_p), p \text{ contexts}, n, T) = \frac{1}{N(n_1,\dots,n_p)} \tag{11}$$

In order to complete the HUP for the latent allocation problem, a uniform prior is placed on the particular contexts used in z, providing

$$\pi_{HUP}(z|n, T) = \frac{1}{T}\,\frac{1}{b(n, p)}\,\frac{1}{N(n_{j_1}(z),\dots,n_{j_p}(z))}\,\frac{(T-p)!}{T!} \tag{12}$$

as the prior for z that uses contexts $j_1,\dots,j_p$.

The first important fact about the HUP is that it gives a finite prior weight (independent of n) on each choice of contexts that could generate the data, $\pi_{HUP}(\text{contexts } j_1,\dots,j_p|n, T) = \frac{(T-p)!p!}{T!T}$, which is in stark contrast to the LDA and EP priors. Additionally, it is easy to see that when restricted to a given set of contexts $(j_1,\dots,j_p)$, the distribution of F is asymptotically uniform. Thus, the HUP both places a finite mass on each sub-face of the T-simplex and is asymptotically given by a proper measure when restricted to each sub-face.

The HUP also produces a penalization that depends directly on p as a function of n. Formulas for b(n, p) and $N(n_1,\dots,n_p)$ are needed for this penalization to be made explicit. The latter is easier to determine. It is just a multinomial coefficient reduced by a redundancy factor that accounts for rows of the Young's diagram with the same length, and is given by

$$N(n_1,\dots,n_p) = \binom{n}{n_1,\dots,n_p}\frac{1}{R(n_1,\dots,n_p)} \tag{13}$$

$$R(n_1,\dots,n_p) = \prod_{i=1}^{n}\left(\sum_{k=1}^{p} I[n_p = i]\right)! \tag{14}$$

Lemma 3 in the Supplementary Materials establishes that the asymptotic limit of $\frac{b(n,p)}{n^{p-1}}$ is $(p!(p-1)!)^{-1}$ as $n \to \infty$ for fixed p. Thus, the asymptotic form of the HUP for z is

$$\pi_{HUP}(z|n, T, \alpha) \propto n^{\frac{1+t-2p}{2}} \left(\prod_{k=1}^{t} p_{j_k}^{np_{j_k}} \right) . \tag{15}$$

The penalization as n increases depends not only on t, the number of contexts used infinitely often, but also on p, the total number of contexts used. In particular, there is a penalization of $n^{\frac{-p}{2}}$ for using p contexts and an additional penalization of $n^{\frac{t-p}{2}}$ for using only using t of the contexts infinitely often.

2.5. Nonuniform Hierarchical Prior (NUHP). Although the HUP does have the desirable properties that it asymptotically places finite mass on each sub-face of the T-simplex and that F is asymptotically a proper measure when restricted to each sub-face, it is undesirable to need this measure to be uniform. From a clustering point of view, it is not obvious how to force this measure to be anything other than uniform. However, one can take a clue from the LDA prior itself. Though the LDA places all of its mass on the interior of the T-simplex, it produces an appropriate Dirichlet measure when restricted to a particular sub-face. As an alternative to the HUP and the LDA prior, the NUHP first places a uniform prior on the choice of contexts and then an appropriate Multinomial-Dirichlet prior on the z which arise from these contexts. The form of the NUHP for z using contexts $j_1, \ldots, j_p$ is

$$\begin{aligned} \pi_{NUHP}(z|n, T, a) = {} & \frac{(T-p)!p!}{T!T} c(j_1, \ldots, j_p) \\ & \times \left(\prod_{k=1}^{p} \frac{\Gamma(n_{j_k}(z) + a_{j_k})}{\Gamma(a_{j_k})} \right) \frac{\Gamma(\sum_{k=1}^{p} a_{j_k})}{\Gamma(n + \sum_{k=1}^{p} a_{j_k})} \end{aligned} \tag{16}$$

where $c(j_1, \ldots, j_p)$ is a constant that accounts for the fact that the conditional prior of z given $j_1, \ldots, j_p$ is restricted to the set of z with $n_{j_k}(z) > 0$ for $k = 1, \ldots, p$ and $n_j(z) = 0$ for $j \not\in \{j_1, \ldots, j_p\}$

The NUHP produces appropriate clustering and LDA behavior asymptotically on the T-simplex due to its construction. Viewing the NUHP in terms of a penalization for fixed z provides the asymptotic approximation

$$\pi_{NUHP}(z|n, T, a) \propto n^{\frac{1-t-2\sum_{k=t+1}^{p} a_{jk}}{2}} \prod_{k=1}^{t} p_{j_k}^{np_{j_k}} \tag{17}$$

When $a_j = 1$ for all j, this provides the same asymptotic penalization as the HUP and, in contrast to the LDA, the penalization depends directly on the p contexts that are used in z. Thus, though the NUHP produces a different prior than the HUP, it provides a means of obtaining an equivalent asymptotic analysis that can be generalized to provide LDA like behavior on the sub-faces of the T-simplex.

2.6. Small Sample Prior Comparison. The small sample behavior of the priors provides an intuition for the posterior properties of the models. Of particular interest is the prior probability of using p contexts out of a possible T contexts in the prior. Figure 1 shows the prior probability of using p = T contexts for small values of n. In these figures, it is assumed that a = 1 and $\lambda = 1$. As expected, there is a large contrast between the hierarchical priors and the LDA or EP priors. While the former provide constant values for the prior of using p = T contexts, the latter have this probability increasing with n. Though there is a quantitative difference between the LDA and EP models (the LDA increases $\pi(p = T)$ more quickly as a function of n than the EP), both models produce degenerate prior distributions asymptotically.

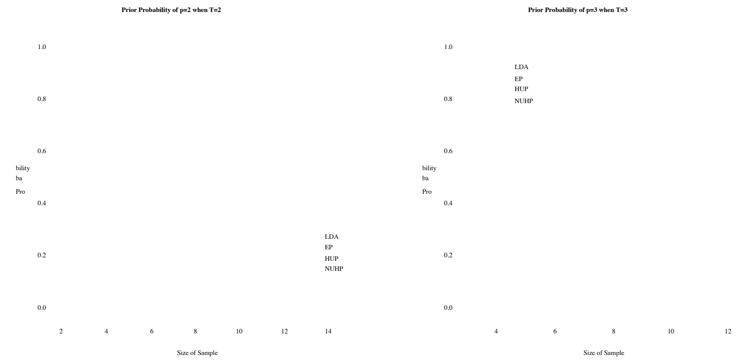


Fig 1. Small Sample Prior Behavior

The plots show the prior probability for using the maximum number of available contexts for T = 2 (left) and T = 3 (right). In this plot the hyperparameters of the priors are fixed at a = 1 and $\lambda = 1$.

3. Posterior Behavior for a Single Sample. This section addresses the asymptotic probability assigned to sets of contexts as the number of observed features grows in a single sample. Regardless of the true sampling

distribution, the LDA and EP models place probability 1 on the set of all available contexts. The HUP and NUHP produce non-zero posterior probabilities for each collection of contexts. While one might desire consistency from the HUP and NHUP models, there is simply not enough information in the data to provide such consistency. However, this behavior is not unexpected as the data can only provide an estimate of γ^*_v.

3.1. Inconsistency for LDA and EP Priors. The LDA and EP models are shown to produce inconsistent posterior distributions through introducing a construct called a "lift" of an allocation vector.

Definition 1 (Lift of an Allocation Vector). Let z be a context allocation vector using p<T contexts, $I \subset \{1,...,n\}$ be an index set, and t be a context not used in z. The lift $l(z,I,t)$ is defined as the context allocation vector $z(I,t)$ where z_i is replaced by t for all $i \in I$. For a given index set I, observed data w, and allocation vector z, define $I_k(w) \subset I$ to be the set of indices $i \in I$ such that $w_i = k$ and further define $I_{j,k}(w,z) \subset I_k$ to be the set of indices $i \in I$ such that $z_i = j$ and $w_i = k$.

The lift $l(z,I,t)$ maps the allocation vector z to an allocation vector using at most p + 1 contexts (less than p + 1 contexts if and only if I contains all of the i such that z_i is a particular context). Two lifts for different z can coincide. Trivially, if $z_1 = 1_n$, $z_2 = 2_n$, $I_1 = I_2 = \{1,...,n\}$, and $t_1 = t_2 = 3$, then $l(z_1,I_1,t_1) = l(z_2,I_2,t_2)$. Here attention is focused on a particular set of lifts as $n \to \infty$ such that there is a finite multiplicity for the number of choices of z, I, and t giving rise to the same lifted allocation vector.

Proposition 1 (Multiplicity of Lifts). Fix a sequence of $(z_n)_{n=1}^{\infty}$ such that z_m consists of the first m elements of z_{m+1}. Define $J_\infty \subset N$ to be an index set of infinite size satisfying: for a given k, $J^n_{j,k} = \{i \in J : w_i = k \,\&\, z^n_i = j\}$ is such that $|J^n_{j,k}| = n_{j,k}(z_n,w) \to \infty$ for some $j = j_k$ and $J^n_{j,k} = \varnothing$ for $j = j_k$. Let I be a finite subset of J_∞. There are at most pv such choices of $(z_n)_{n=1}^{\infty}$ and J_∞ satisfying these conditions which map to the same lifted allocation vector.

In this setting, a lift can be viewed as an injective map from the p-simplex into the p + 1-simplex, moving points on the boundary of the p + 1-simplex into its interior. The set of lifts can then be can then be viewed as a fiber bundle of the map from the p+1-simplex onto its p-simplex boundaries. Essentially, lifts satisfying the conditions of Proposition 1 provide any desired number of “copies” of p contexts contained in p + 1 contexts. The LDA and EP models place too much mass on these lifts, producing posterior inconsistency.

3.1.1. Inconsistency in the LDA Model. For each z, fix an index set J satisfying the conditions of Proposition 1. Inconsistency of the LDA prior is established by showing that the sum of probabilities over all lifts associated with index sets I of size L contained in J is greater than a constant times L^{-1} times the probability of z. This provides the “copies” of z with p contexts contained in the set of p+1 contexts that drive the ordering of the posterior probabilities of R_p, the class of partitions with p clusters. The proof of inconsistency for the LDA model is proven through the use of Proposition 1 and (21), which is an asymptotic lower bound of the posterior probability of the lifts of a given allocation vector z. Throughout the rest of the paper it is assumed that, regardless of the sampling model, the set {w} with $\#\{i : w_i = k\} < \infty$ for some i has probability zero as $n = |w| \to \infty$. (21) is established through a series of propositions.

Proposition 2 (Probability of a Single Lift). Let z be an allocation vector using p contexts and choose an index set I as in Proposition 1 and a context t which is not used in z. Then

$$
\begin{aligned}
(18)\qquad p_{LDA}(l(z,I,t),w|n,T,\alpha,a) &= p_{LDA}(z,w|n,T,\alpha,a) \\
&\times \frac{(L-1+a_t)\cdots(a_t)}{(L-1+\sum_k \alpha_{t,k})\cdots(\sum_k \alpha_{t,k})} \\
&\times \left(\prod_k \frac{(L_{j_k,k}-1+\alpha_{t,k})\cdots(\alpha_{t,k})}{(n_{j_k,k}(z,w)-L_{j_k,k}+\alpha_{j_k,k})\cdots(n_{j_k,k}(z,w))}\right) \\
&\times \left(\prod_j \frac{(n_j(z)-L_j+\sum_k \alpha_{j,k})\cdots(n_j(z)-L_j+\sum_k \alpha_{j,k})}{(n_j(z)-L_j+a_j)\cdots(n_j(z)-1+a_j))}\right)
\end{aligned}
$$

where j_k is the context for I_k, $L_{j_k,k}$ is the size of I_k, and $L_j = \sum_{k:j_k=j} L_{j_k,k}$.

Proposition 3 (Probability of a Sum of Similar Lifts). Summing (18) over all I with the same $L_{j_k,k}$ and taking the limit as $n \to \infty$ provides

$$
(19)\qquad \sum_{\substack{I \\ L\ =L_{j_k,k}}} p_{LDA}(l(z,I\ ,t),w|n,T,\alpha,a) \approx p_{LDA}(z,w|n,T,\alpha,a)
$$

jk,k

$$\times \frac{(L-1+a_t)\cdots(a_t)}{(L-1+\sum_k \alpha_{t,k})\cdots(\sum_k \alpha_{t,k})} \left(\prod_k \frac{(L_{jk,k}-1+\alpha_{t,k})\cdots(\alpha_{t,k})}{L_{jk,k}!} \right)$$

Proof. This follows because there are

$$\prod_j \binom{n_{jk,k}(z,w)}{L_{jk,k}}$$

choices of such I and the rational functions in $n_j(z)$ and $n_{jk,k}(z,w)$ are fractions of polynomials of the the same finite degree with leading coefficient 1 in the numerator and denominator and are thus asymptotically 1.

Lemma 1. The sum

$$\sum \frac{1}{(L-1+\sum_k \alpha_{t,k})\cdots(\sum_k \alpha_{t,k})} \left(\prod_k \frac{(L_{jk,k}-1+\alpha_{t,k})\cdots(\alpha_{t,k})}{L_{jk,k}!} \right)$$

over all choices for the $L_{jk,k}$ that sum to L is $\frac{1}{L!}$.

Proposition 4 (Probability of Sum of Lifts). Summing (19) over all choices for $L_{jk,k}$ provides

(20)
$$\sum_I p_{LDA}(l(z,I,t),w|n,T,\alpha,a) = p_{LDA}(z,w|n,T,\alpha,a) \frac{(L-1+a_t)\cdots(a_t)}{L!}$$

For any z and choice of t, (20) provides a lower bound of

$$\sum_I p_{LDA}(l(z,I,t),w|n,T,\alpha,a) \geq p_{LDA}(z,w|n,T,\alpha,a) \frac{\min_j\{a_j\}}{L} \tag{21}$$

This lower bound provides the proof for Theorem 1.

Theorem 1 (Inconsistency of LDA Model). Regardless of the true sampling distribution, the posterior probabilities of the R_p are asymptotically given by

$$\{ 0 \quad p<T$$

$$(22) \qquad \lim_{n\to\infty} p_{LDA}(R_p|w,n,T,\alpha,a) = \qquad 1 \quad p = T$$

and thus the LDA model is inconsistent.

Proof. Due to (21) and Proposition 1, one can choose a set of lifts that change L indices and obtain

$$\frac{p_{LDA}(R_{p+1}|w,n,T,\alpha,a)}{p_{LDA}(R_p|w,n,T,\alpha,a)} \geq \frac{\min_j\{a_j\}}{p v L}$$

Given any $M > 0$, the divergence of the series $\sum \frac{1}{L}$ allows one to choose L_* such that $\sum_{L=1}^{L_*} \frac{1}{L} \geq \frac{p v M}{\min_j\{a_j\}}$. Applying Proposition 1 and summing over $L = 1,\ldots,L_*$, one obtains

$$\frac{p_{LDA}(R_{p+1}|w,n,T,\alpha,a)}{p_{LDA}(R_p|w,n,T,\alpha,a)} \geq M$$

Thus, the LDA model is inconsistent.

In Figure 2, the inconsistency of the LDA prior is shown for $T = 2$. In particular, $\alpha_1 = (2,4)$ and $\alpha_2 = (4,2)$ are chosen to provide separation between the sampling distributions of the two models containing a single context. The choice of $a = (0.05,0.05)$ is made so that the prior probability of two contexts is relatively small. The first plot shows the probability of two contexts as a function of n where m_1, the number of observed features taking value 1, is the modal value from a Beta-Binomial distribution with parameter $\alpha = c(2,4)$ (and is thus the maximally supported observation under the assumption that just context 1 generated the data). From this plot, one can easily see that the posterior probability of two contexts increases to 1 quite rapidly with n. The second plot shows the cumulative distribution function of the posterior probability of two contexts when the samples are generated by context 1 for $n = 50,100$, and 300. The CDF is clearly showing that the distribution of the posterior probability of two contexts is converging to 1 asymptotically.

Corollary 1. The latent allocation model is inconsistent under the following priors

1. Uniform prior on the space of z.
2. $\pi(z_i = j|n, T,\theta) = \theta_j > 0$.

3.1.2. Inconsistency in the EP Model. Because the asymptotic penalization is similar between the EP and LDA models, inconsistency for the EP model can be proven by following arguments similar to those for the LDA model.

Proposition 5 (Probability of a Single Lift). Let z be an allocation vector using p contexts and choose an index set I as in Proposition 1 and a

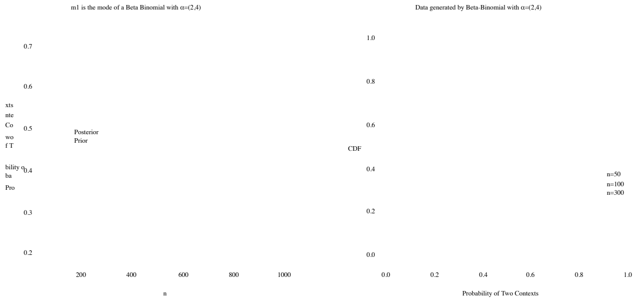


Fig 2. Plots of the posterior probability of two contexts for the LDA prior. The true data generating process in each plot is 1 context with $\alpha_1 = (2, 4)$. The left plot shows the prior and posterior probabilities of two contexts as a function of n. The right plot shows the CDF of the probability of two contexts; the vertical line shows the prior probability of two contexts.

context t which is not used in z. Then

$$p_{EP}(l(z,I,t),w|n, T,\alpha,\lambda) = p_{EP}(z,w|n, T,\alpha,\lambda) \times \frac{\lambda}{(T-p)} \frac{(L-1)!}{(L-1+\sum_k \alpha_{t,k})\cdots(\sum_k \alpha_{t,k})} \tag{23}$$

$$\times \left(\prod_k \frac{(L_{j_k,k} - 1 + \alpha_{t,k})\cdots(\alpha_{t,k})}{(n_{j_k,k}(z,w) - L_{j_k,k} + \alpha_{j_k,k})\cdots(n_{j_k,k}(z,w))} \right)$$

$$\times \left(\prod_j \frac{(n_j(z) - L_j + \sum_k \alpha_{j,k})\cdots(n_j(z) - L_j + \sum_k \alpha_{j,k})}{(n_j(z) - L_j)\cdots(n_j(z) - 1))} \right)$$

where j_k is the context for I_k, $L_{j_k,k}$ is the size of I_k, and $L_j = \sum_{k:j_k=j} L_{j_k,k}$.

Theorem 2 (Inconsistency of EP Model). Regardless of the true sampling distribution, the posterior probabilities of the R_p are asymptotically given by

$$\lim_{n\to\infty} p_{EP}(R_p|w,n,T,\alpha,\lambda) = \begin{cases} 0 & p<T \\ 1 & p = T \end{cases} \tag{24}$$

and thus the EP model is inconsistent.

Proof. Following the arguments of Propositions 3,4 and Lemma 1, one

can choose a set of lifts that change L indices and obtain

$$\frac{p_{EP}(R_{p+1}|w,n,T,\alpha,\lambda)}{p_{EP}(R_p|w,n,T,\alpha,\lambda)} \geq \frac{\lambda}{(T-p)p_v L}$$

Given any $M > 0$, one can use the divergence of the series $\sum \frac{1}{L}$ to choose L_* such that $\sum_{L=1}^{L_*} \frac{1}{L} \geq \frac{(T-p)p_v M}{\lambda}$. Applying Proposition 1, and summing over $L = 1,\ldots,L_*$ provides

$$\frac{p_{EP}(R_{p+1}|w,n,T,\alpha,\lambda)}{p_{EP}(R_p|w,n,T,\alpha,\lambda)} \geq M$$

Thus, the EP model is inconsistent.

Figure 3 shows the inconsistency of the EP prior when $T = 2$. As with Figure 2, $\alpha_1 = (2,4)$ and $\alpha_2 = (4,2)$. The choice of $\lambda = 0.1$ is made in order for the prior probability of two contexts to be relatively small. The first plot shows the probability of two contexts as a function of n where m_1, the number of observed features taking value 1, is the modal value from a Beta-Binomial distribution with parameter α = c(2,4) (and is thus the maximally supported observation under the assumption that just context 1 generated the data). From this plot, it is clear that the posterior probability of two contexts increases to 1 as n → ∞. The second plot shows the cumulative distribution function of the posterior probability of two contexts when the samples are generated by context 1 for n = 100. The CDF is clearly showing that the distribution of the posterior probability of two contexts is converging to 1 asymptotically.

3.2. Behavior of HUP and NUHP Models. The HUP and NUHP models produce posterior distributions on the sets of topics which are asymptotically positive as n → ∞. First, an integral representation for the asymptotic posterior probability of a set of contexts is derived for the HUP. A few modifications of the machinery accommodates the additional parameter, a, of the NUHP and a similar integral is derived for the NUHP.

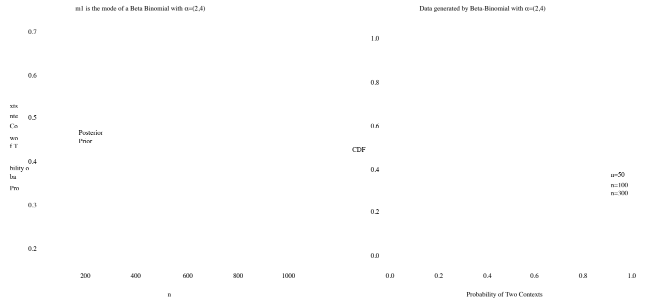


Fig 3. Plots of the posterior probability of two contexts for the EP prior. The true data generating process in each plot is 1 context with $\alpha_1 = (2, 4)$. The left plot shows the

posterior probability of two contexts as a function of n. The right plot shows the CDF of the probability of two contexts; the vertical line shows the prior probability of two contexts.

3.2.1. Behavior of the HUP Model. The joint distribution of w and z is given by

$$(25)\qquad p_{HUP}(z,w|n,T,\alpha) = p(w|z,\alpha,n,T)\pi_{HUP}(z|n,T)$$

$$= \left(\prod_{j=1}^{T}\left(\prod_{k=1}^{V}\frac{\Gamma(n_{j,k}(z,w)+\alpha_{j,k})}{\Gamma(\alpha_{j,k})}\right)\frac{\Gamma(\sum_{k=1}^{V}\alpha_{j,k})}{\Gamma(n_j(z)+\sum_{k=1}^{V}\alpha_{j,k})}\right)$$

$$\times\frac{1}{T}\frac{1}{b(n,p)}\frac{1}{N(n_1,\dots,n_T)}\frac{(T-p)!}{T!}$$

An asymptotic approximation for the joint distribution of z and w is given in Proposition 6.

Proposition 6 (Asymptotic Form of the HUP Joint). Suppose that z uses p contextx, that $n_j(z)$ grows unboundedly for $j = j_1,\dots,j_t$, and that $V_{j_i} = \{k : n_{j_i,k}(z,w)$ grows unboundedly$\}$. Then the asymptotic approximation for the HUP joint is proportional to

$$(26)\qquad p_{HUP}(z,w|n,T,\alpha) \propto \frac{1}{n!n^{p-1}e^{n}}$$

$$\times\left(\prod_{i=1}^{t}\left(\prod_{k\in V_{j_i}} n_{j_i,k}(z,w)^{n_{j_i,k}(z,w)+\alpha_{j_i,k}-\frac{1}{2}}\right)n_{j_i}(z)^{1-\sum_{k=1}^{V}\alpha_{j_i,k}}\right)$$

The asymptotic approximation in Proposition 6 is used to derive an integral approximation for the joint density of w and a given set of contexts

$j_1,\dots,j_p$ under the HUP. This asymptotic formula utilizes two simplifications. The first is the Stirling's approximations. The second is that the redundancy numbers $R(n_{j_1}(w),\dots,n_{j_t}(w))$ can be ignored. Both follow from the fact that the integral approximation is finite. The Stirling's approximations are only invalid near the boundary of the simplex and such z have probability 0 asymptotically. Similarly, the redundancy factor is greater than one only on lower dimensional subspaces of the simplex and integrating over these subspaces provides mass 0.

The set $R_p = \{z : n_j(z) > 0$ for exactly p of the j$\}$ decomposes as $R_p = \cup R_p(j_1,\dots,j_p)$ where the j_i are unique elements of $1,\dots,T$. The set $R_p(j_1,\dots,j_p)$ decomposes as a union over partitions with sets of sizes $n_1,\dots,n_p > 0$ with $n_1 + \cdots + n_p = n$, call such a set $B_{j_1,\dots,j_p}(n_1,\dots,n_p)$, which is the collection of all allocation vectors using contexts $j_1,\dots,j_p$ allocating n_j features to one of the contexts for each j_i (that is, $n_{j_i}(z) \in \{n_1,\dots,n_p\}$). Using these definitions, the HUP can be written as

$$\pi_{HUP}(R_p|n,T) = \frac{1}{T}$$

$$\pi_{HUP}(R_p(j_1,\dots,j_p)|R_p,n,T) = \frac{(T-p)!p!}{T!}$$

$$\pi_{HUP}(B_{j_1,\dots,j_p}(n_1,\dots,n_p)|R_p(j_1,\dots,j_p,|R_p,n,T) = \frac{1}{b(n,p)}$$

$$\pi_{HUP}(z|B_{j_1,\dots,j_p}(n_1,\dots,n_p),R_p(j_1,\dots,j_p,|R_p,n,T) = \frac{1}{p!N(n_1,\dots,n_p)}$$

Theorem 3 (Asymptotic Integral of HUP Joint). The joint probability of w and $R_p(j_1,\dots,j_p)$ is asymptotically approximated by the integral

(27)

$$p_{HUP}(R_p(j_1,\dots,j_p),w|n,T) \approx \frac{(T-p)!}{T!Ta(p)} \times \left(\prod_{i=1}^{p}\left(\frac{\Gamma(\sum_{k=1}^{V}\alpha_{j_i,k})}{\prod_{k=1}^{V}\Gamma(\alpha_{j_i,k})}\right)\right)\left(\left(\frac{2\pi}{n}\right)^{\frac{V-1}{2}}\prod_{k=1}^{V}\left(\frac{m_k}{n}\right)^{m_k+\sum_{i=1}^{p}\alpha_{j_i,k}-\frac{1}{2}}\right)$$
$$\times \int\left(\prod_{k=1}^{V}\prod_{i=1}^{p}(\varrho_{j_i,k})^{\alpha_{j_i,k}-1}\right)\left(\prod_{i=1}^{p}\left(\sum_{k=1}^{V}\varrho_{j_i,k}\frac{m_k}{n}\right)^{1-\sum_{k=1}^{V}\alpha_{j_i,k}}\right)\prod_{k=1}^{V}\prod_{i=1}^{p-1}d\varrho_{j_i,k}$$

Details on the proof of Theorem 3 are provided in the Supplementary Materials.

Remark. Because $\gamma_k = \frac{m_k}{n} \rightarrow \gamma^*_k > 0$ with probability 1, the posterior probability of any set of contexts is non-zero asymptotically. This value is approximated (up to proportion) by

$$p_{HUP}(R_p(j_1,\dots,j_p),w|n,T) \propto p!(p-1)!(T-p)! \times \left(\prod_{i=1}^{p} \frac{\Gamma(\sum_k \alpha_{j_i,k})}{\prod_k \Gamma(\alpha_{j_i,k})}\right)\left(\prod_{k=1}^{V} (\gamma_k)^{\sum_{i=1}^{p} \alpha_{j_i,k}}\right) \times \int \left(\prod_{k=1}^{V}\prod_{i=1}^{p} (\varrho_{j_i,k})^{\alpha_{j_i,k}-1}\right)\left(\prod_{i=1}^{p}\left(\sum_{k=1}^{V} \varrho_{j_i,k}\gamma_k\right)^{1-\sum_{k=1}^{V}\alpha_{j_i,k}}\right)\prod_{k=1}^{V}\prod_{i=1}^{p-1} d\varrho_{j_i,k}$$

Although this shows that the HUP does not provide consistent inference for a single sample, the posterior for the HUP does not degenerate to T contexts as do the posteriors for the LDA and EP priors.

3.2.2. Behavior of the NUHP Model. This prior is asymptotically equivalent (with probability 1) to the HUP whenever $a_t = 1$ for all $t = 1,\dots,T$. The prior decomposes the T-simplex into a set of faces, placing a hierarchical uniform prior on the set of faces and a (restricted) Dirichlet prior on the interior of each face. Due to this decomposition, one has an asymptotic integral approximation for the joint of w and a set of contexts $j_1,\dots,j_p$.

Theorem 4 (Asymptotic Integral of NUHP Joint). The joint probability of w and $R_p(j_1,\dots,j_p)$ is asymptotically approximated by the integral

$$p_{NUHP}(R_p(j_1,\dots,j_p),w|n,T) \approx \frac{p!(T-p)!}{T!T}\,\Gamma\left(\sum_{i=1}^{p} a_{j_i}\right)\left(\prod_{i=1}^{p}\Gamma(a_{j_i})\right)^{-1} \times \left(\prod_{i=1}^{p}\left(\frac{\Gamma(\sum_{k=1}^{V}\alpha_{j_i,k})}{\prod_{k=1}^{V}\Gamma(\alpha_{j_i,k})}\right)\right)\left(\frac{2\pi}{n}\right)^{\frac{V-1}{2}}\left(\prod_{k=1}^{V}\left(\frac{m_k}{n}\right)^{m_k+\sum_{i=1}^{p}\alpha_{j_i,k}-\frac{1}{2}}\right) \times \int \left(\prod_{k=1}^{V}\prod_{i=1}^{p}(\varrho_{j_i,k})^{\alpha_{j_i,k}-1}\right)\left(\prod_{i=1}^{p}\left(\sum_{k=1}^{V}\varrho_{j_i,k}\frac{m_k}{n}\right)^{a_{j_i}-\sum_{k=1}^{V}\alpha_{j_i,k}}\right)\prod_{k=1}^{V}\prod_{i=1}^{p-1} d\varrho_{j_i,k} \tag{28}$$

3.3. Small Sample Posterior Comparison. In addition to considering the asymptotic behavior of the models, it is also instructive to understand their small sample behaviors. The posterior probability of $p = T$ for $T = 2,3$ is shown in Figures 4 and 5, respectively. In these figures, it is assumed that there are V = 2 possible features and the context specific Dirichlet parameters are taken to be $\alpha_j \in \{(1,1),(1,2)\}$ for T = 2 and $\alpha_j \in \{(1,1),(1,2),(2,1)\}$ for T = 3. The context mixture parameters are given by a = 1 and $\lambda = 1$. As can be seen in the figures, while the posterior probabilities for the HUP and NUHP provide reasonable finite values for prob(p = T|w) those for the LDA and EP are both increasing to 1 as n increases. As expected, the EP posterior increases to 1 slower than the LDA posterior.

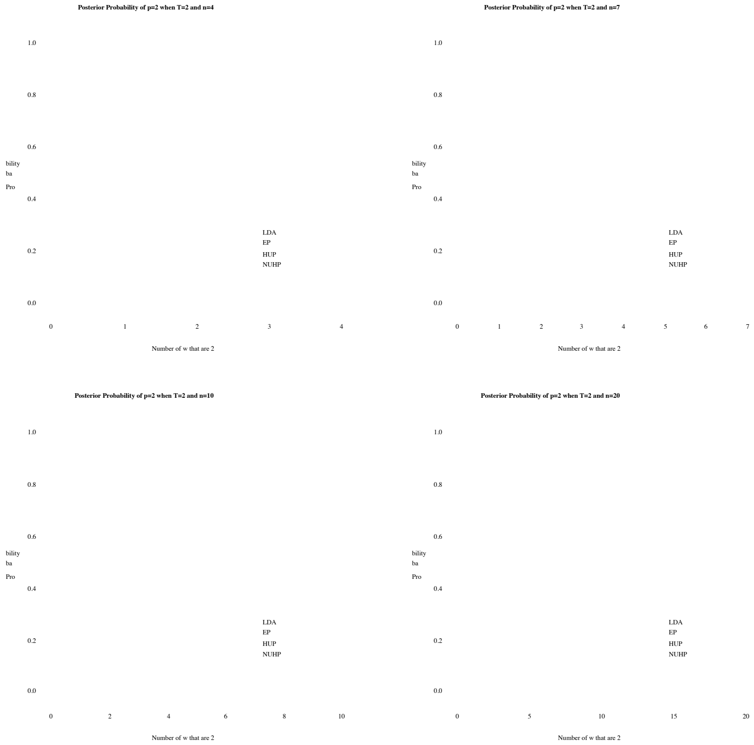


Fig 4. Posterior Behavior of the Models in a Single Sample for T = 2

4. Analysis of a Collection of Samples. In a collection of samples, there are two distinct ways to define common contexts. First, when $\alpha_t = \alpha$ for all t, the only way to define common contexts is by the β_t. Second, if the model is expanded to allow for differing α_t, then these α_t can distinguish

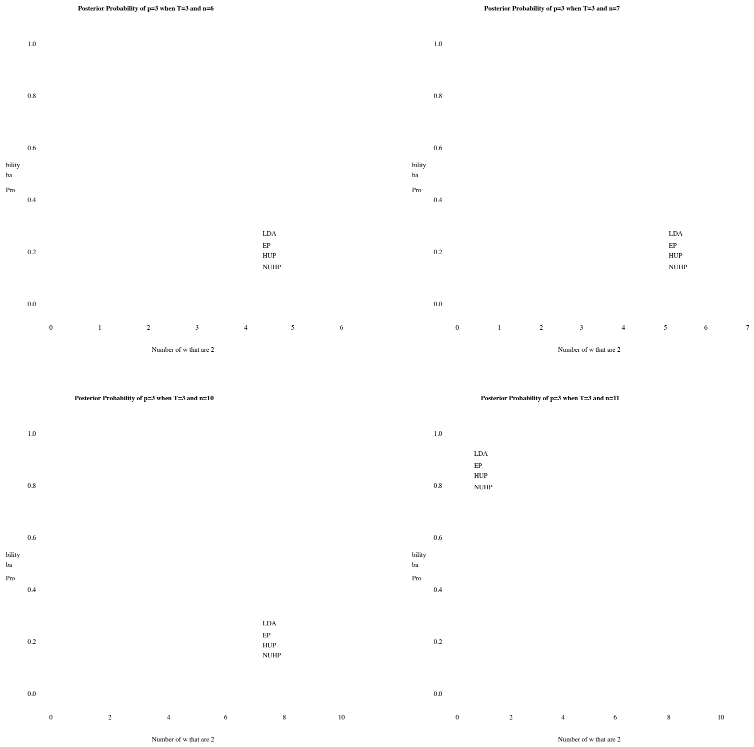


Fig 5. Posterior Behavior of the Models in a Single Sample for T = 3

the contexts, and the feature distribution for a given context can differ between samples. The important consideration here is that although different samples have different β_t, these β_t are assumed to be drawn from a common distribution. This analysis focuses on the latter construction of the model. This accounts for the "burstiness" of features across samples and is called the DCMLDA model (Doyle and Elkan, 2009).

There are two situations of interest in a collection of samples. First, when a single set of contexts generates the collection. Second, when the samples are partitioned into clusters and each cluster is generated by a unique set of contexts. In order to consider these problems, one has to develop an appropriate prior for the generating contexts for a collection of samples. The samples are then generated after conditioning on this generating set of contexts. Thus, the data generating process is necessarily hierarchical in nature.

The hierarchical nature of the question makes the HUP and NUHP models ideal for considering analysis in a collection of samples. Throughout this section assume that the S samples are generated by context sets d = $(d_1,...,ds)$ and have lengths n = $(n_1,...,ns)$. The asymptotic results assume that both n_s and S increase to ∞, although the assumption on n_s can be relaxed. Because each sample provides only a point estimate of γ_v for $v \in [V]$, the requirement of $S \to \infty$ is necessary to obtain posterior consistency.

Setting the prior is not an easy task for either the classification or clustering problem. At first glance, prior independence seems reasonable, defining the appropriate priors as $\pi(d|n,T) \propto \prod_{s=1}^{S} \pi(d_s|n_s,T)$. However, this definition will clearly lead to inconsistency for the LDA and EP models and could potentially provide bad behavior for the HUP and NUHP models. Moreover, this definition of the prior does not reflect the assumptions being made about the data generating process.

As an alternative, begin with the assumptions made by the classification and clustering problems and define the prior in a hierarchical manner. The following hierarchical construction provided the motivation for the NUHP model; it essentially transforms the LDA model into the NUHP in a collection of samples. Thus, the remainder of this analysis is focused on the HUP and NUHP models.

Denote the observed features of sample s by w_s. The sampling distribution is given by

$$p(w_1,...,ws|n_1,...,ns,d_1,...,ds,T) = \prod_{s=1}^{S} p(w_s|n_s,d_s,T)$$

and the posterior distribution is given by

$$
\begin{aligned}
&p(d_1,\dots,d_S|n_1,\dots,n_S,w_1,\dots,w_S,T)\\
&= \left(\prod_{s=1}^{S} p(w_s|n_s,d_s,T)\right)\frac{p(d_1,\dots,d_S|n_1,\dots,n_S,T)}{p(w_1,\dots,w_S|n_1,\dots,n_S,T)}\\
&= \left(\prod_{s=1}^{S} \frac{p(w_s,d_s|n_s,T)}{p(d_s|n_s,T)}\right)\frac{p(d_1,\dots,d_S|n_1,\dots,n_S,T)}{p(w_1,\dots,w_S|n_1,\dots,n_S,T)}\\
&= \left(\prod_{s=1}^{S} \frac{p(d_s|w_s,n_s,T)p(w_s|n_s,T)}{p(d_s|n_s,T)}\right)\frac{p(d_1,\dots,d_S|n_1,\dots,n_S,T)}{p(w_1,\dots,w_S|n_1,\dots,n_S,T)}\\
&= \left(\prod_{s=1}^{S} p(d_s|w_s,n_s,T)\right)\times\frac{p(d_1,\dots,d_S|n_1,\dots,n_S,T)}{\prod_{s=1}^{S} p(d_s|n_s,T)}\\
&\quad\times\frac{\prod_{s=1}^{S} p(w_s|n_s,T)}{p(w_1,\dots,w_S|n_1,\dots,n_S,T)}
\end{aligned}
$$

The term $\frac{\prod_{s=1}^{S} p(w_s|n_s,T)}{p(w_1,\dots,w_S|n_1,\dots,n_S,T)}$ does not depend on $d_1,\dots,d_S$ and as such does not effect the ratio of posterior probabilities. The posterior comprises of the product of the posterior in each sample times a correction factor $\frac{p(d_1,\dots,d_S|n_1,\dots,n_S,T)}{\prod_{s=1}^{S} p(d_s|n_s,T)}$ which accounts for the difference between the prior for a collection of samples and the product of priors in individual samples. For a single sample the HUP and NUHP models have

$$
p(d_s|n_s,T) = \frac{(T-p)!p!}{T!T}
$$

where d_s comprises of p contexts.

4.1. A Single Set of Generating Contexts. Assume that $d = d_1 = d_2 = \cdots = d_S$ for all samples $1,\dots,S$. In order to complete the prior specification, a probability must be assigned to $p(d_1,\dots,d_S|n_1,\dots,n_S,T) = p(d|T)$. One can define this prior in a uniform manner:

$$
p(j_1,\dots,j_p|n_1,\dots,n_S, S, T) = \frac{1}{M}
$$

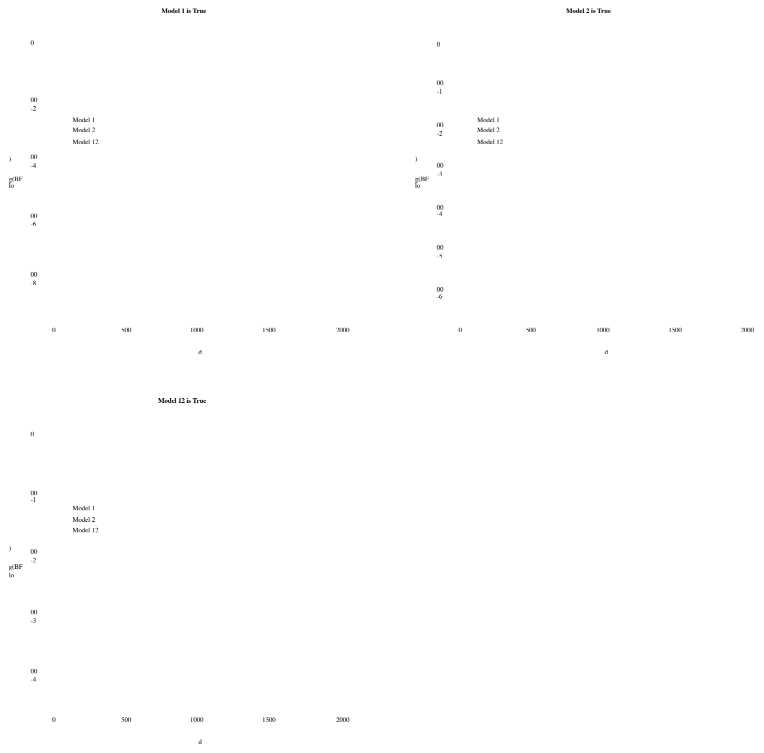
where $j_1,\dots,j_p$ are a particular choice of p out of the T contexts, of which there are $M = 2^T - 1$ choices. This choice is inconsequential when T is fixed and is made as a convenience. Thus, the correction factor in the posterior is given by

$$\frac{p(d_1 = \cdots = d_S = (j_1,\ldots,j_p)|n_1,\ldots,n_S, S, T)}{\prod_{s=1}^{S} p(d_s = (j_1,\ldots,j_p)|n_s,T)} = \frac{1}{M}\left(\frac{(T-p)!p!}{T!T}\right)^{-S}$$

whenever the collection is generated by a particular set of p contexts.

The asymptotic consistency of the HUP (and NUHP for a = 1) in a collection of samples generated by a single context is demonstrated for T = 2 and T = 3 in Figures 6 and 7, respectively.

Fig 6. Consistency for T = 2. The x-axis is S = d samples.

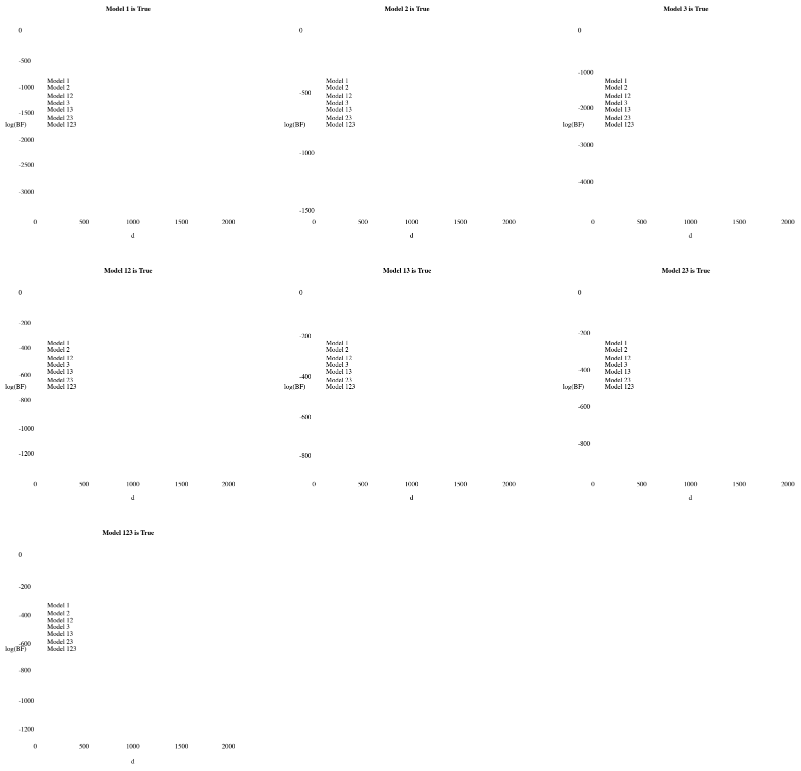


The asymptotic consistency of the HUP and NUHP priors in a collection of samples generated by a single context can be proven under the assumption that $\sum_k \alpha_{j,k} = a_j$ for all j. This assumption is made throughout the rest of this subsection. With this restriction, the joint of a single document is given by

$$p(d_s, w_s | n_s, T) \propto \frac{p!(T-p)!\Gamma(\sum_i a_{j_i})}{\prod_{k=1}^{V} \Gamma(\sum_i \alpha_{j_i,k})} \left(\prod_{k=1}^{V} (m_{s,k})_{n_s} \right)^{\sum_{i=1}^{p} \alpha_{j_i,k}}$$

where $\Gamma(\sum_i a_{j_i})$ is replaced with $(p-1)!$ for the HUP.

Fig 7. Consistency for T = 3



To further simplify the analysis, assume that n_s is large for each s, so that $\frac{m_{s,k}}{n_s}$ can be replaced by $\gamma_{s,k}$. In order to prove consistency, one has to determine the sampling distribution of $\gamma_{s,k}$ under an assumed model. Make the following assumptions about the data generating process.

1. Whenever p contexts generate the collection, $\frac{n_{s,j_i}}{n_s} = \nu_{s,j_i}$ is sampled from a Dirichlet($a_{j_1},\dots,a_{j_p}$) distribution on the p simplex. Thus, the assumed data generating model is one of the models under consideration.
2. For a given j_i, assume that the vector comprising of $\frac{n_{s,j_i,k}}{n_{s,j_i}} = p_{s,j_i,k}$ is sampled according to the Dirichlet(α_{j_i}) prior.

4.1.1. Consistency for the HUP Model. In this section, it is assumed that the data generating process has $a_{j_i} = 1$ for some collection of topics. The

posterior probability that a given collection is generated by the contexts $j_1,\dots,j_p$ is approximated by

$$p(j_1,\dots,j_p|w_1,\dots,w_S, S, T) \propto \prod_{s=1}^{S}\left[\frac{(p-1)!}{\prod_{k=1}^{V}\Gamma(\sum_i \alpha_{j_i,k})}\left(\prod_{k=1}^{V}(\gamma_{s,k})^{\sum_{i=1}^{p}\alpha_{j_i,k}}\right)\right]$$

The vectors γ_s are iid random variables. In order to compare two sets of contexts, one can consider the difference in their log posterior probabilities, which is given by

$$\begin{aligned}\log&\left(\frac{p(j_1,\dots,j_p|w_1,\dots,w_S, S, T)}{p(t_1,\dots,t_p|w_1,\dots,w_S, S, T)}\right)\\ &= S\left[\log\left(\frac{(p-1)!}{(p-1)!}\right) + \sum_{k=1}^{V}\log\left(\frac{\Gamma(\sum_1^p \alpha_{t_i,k})}{\Gamma(\sum_1^p \alpha_{j_i,k})}\right)\right]\\ &\quad + \sum_{s=1}^{S}\sum_{k=1}^{V}\left[\sum_1^p \alpha_{j_i,k} - \sum_1^p \alpha_{t_i,p}\right]\log(\gamma_{s,k})\end{aligned}$$

Thus, the log of the ratio of posterior probabilities can be expressed as a constant times S

$$S\left[\log\left(\frac{(p-1)!}{}\right) + \sum^{V}\log\left(\frac{\Gamma(\sum_1^p \alpha_{t_i,k})}{}\right)\right]$$

$$(p-1)! \quad {}_{k=1} \quad \Gamma(\Sigma \; {}_1 \alpha_{j_i,k})$$

plus a sum of iid random variables

$$\sum_{s=1}^{S} X_s = \sum_{s=1}^{S} \sum_{k=1}^{V} \left[\sum_{1}^{p} \alpha_{j_i,k} - \sum_{1}^{p} \alpha_{t_i,p} \right] \log(\gamma_{s,k})$$

where $X_s = \sum_{k=1}^{V} \left[\sum_1^p \alpha_{j_i,k} - \sum_1^p \alpha_{t_i,p} \right] \log(\gamma_{s,k})$. The random variables γ_s are generated in the following manner. When $j_1,\ldots,j_p$ generate the collection, then

$$\gamma_{s,k} = \sum_{i=1}^{p} p_{s,j_i,k} \nu_{s,j_i}$$

where

$$p_{s,j_i} = (p_{s,j_i,1},\ldots,p_{s,j_i,V})$$

is distributed Dirichlet(α_{j_i}) and

$$\nu_s = (\nu_{s,j_1},\ldots,\nu_{s,j_p})$$

is distributed uniformly on the p simplex. In order to prove consistency, one has to determine the properties of

$$\log\left(\sum_{i=1}^{p} p_{s,j_i,k} \nu_{s,j_i}\right)$$

Since the X_s are iid with finite variance, it suffices to show that

$$\left[\log\left(\frac{(p-1)!}{(p-1)!}\right) + \sum_{k=1}^{V} \log\left(\frac{\Gamma(\sum_1^p \alpha_{t_i,k})}{\Gamma(\sum_1^p \alpha_{j_i,k})}\right)\right] + E[X_s]$$

is positive whenever $(j_1,\ldots,j_p)$ generates the collection and negative whenever $(t_1,\ldots,t_p)$ generates the collection. The following technical lemma provides the result.

Lemma 2. Suppose that $X \sim \text{Beta}(a, A - a)$, $Y \sim \text{Beta}(b, B - b)$, and $V \sim \text{Beta}(A, B)$. Define $Z = V X + (1 - V)Y$. Then $Z \sim \text{Beta}(a + b, A - a + B - b)$.

Proof. The characteristic function of Z is given by

$$\phi z(t) = E\left[\exp(\imath Zt)\right]$$

$$= E\left[\phi x(V\,t)\phi((1-V)t)\right]$$

$$= E\left[\sum_{n,m=1}^{\infty} \frac{(a)_{(n)}(V\,\imath t)_{n}}{(A)_{(n)}n!} \frac{(b)_{(m)}((1-V)\imath t)_{m}}{(B)_{(m)}m!}\right]$$

$$= \frac{\Gamma(A+B)}{\Gamma(A)\Gamma(B)} \sum_{n,m=1}^{\infty} \frac{(a)_{(n)}}{(A)_{(n)}n!} \frac{(b)_{(m)}}{(B)_{(m)}m!} \frac{\Gamma(A+n)\Gamma(B+m)}{\Gamma(A+B+n+m)} (\imath t)_{n+m}$$

$$= \sum_{n,m=1}^{\infty} \frac{(a)_{(n)}(b)_{(m)}}{(A+B)_{(n+m)}n!m!} (\imath t)_{n+m}$$

where $(a)_{(n)} = a(a+1)\cdots(a+n-1)$ Applying Lemma 1 and summing over $n + m = k$ one gets

$$\phi z(t) = \sum_{k} \sum_{n,m=1}^{\infty} \frac{(a+b)_{(n+m)}}{(A+B)_{(n+m)}(n+m)!} (\imath t)_{n+m}$$

and thus $Z \sim \text{Beta}(a + b, A - a + B - b)$.

Corollary 2. Supose that $\sum_{k} \alpha_{j,k} = 1$ for all j. When $j_1,\ldots,j_p$ generate the sample, then $\gamma_s \sim \text{Dirichlet}(\sum_{i=1}^{p} \alpha_{j_i})$ if $\nu_s \sim \text{Dirichlet}(1_p)$.

Equipped with the distribution of γ_s for any choice of contexts, consistency of the HUP in a collection of samples follows.

Theorem 5. Supose that $\sum_{k} \alpha_{j,k} = 1$ for all j. Further suppose that whenever contexts $j_1,\ldots,j_p$ generate the sample then $\frac{n_{s,j_i}}{n_s} = \nu_{s,j_i}$ is drawn uniformly from the p-simplex. Then the HUP provides consistent inference for the set of contexts that generates a collection of samples.

Proof. Suppose that $j_1,\ldots,j_p$ generate the sample and let $t_1,\ldots,t_p$ be any other set of contexts. From Corollary 2, γ_s are iid Dirichlet($\sum_{i} \alpha_j$). The quantity

$$\left[\left((p-1)! \right)^{V} \left(\Gamma(\sum^{p} \right) \right]$$

$$E_s = \log \quad (p-1)! \quad + \sum_{k=1} \log \quad \Gamma(\textstyle\sum_1^p \alpha_{t_i,k}) \ \alpha_{j_i,k}) \quad + E[X_s]$$

is the Kullback-Leibler divergence between X and Y where $X \sim \text{Dirichlet}(\sum_i \alpha_{j_i})$ and $Y \sim \text{Dirichlet}(\sum_i \alpha_{t_i})$. Thus E_s is positive and the theorem follows.

4.1.2. Consistency for the NUHP Model. The proof of consistency for the NUHP model follows arguments which are similar to those for the HUP model.

Theorem 6. Suppose that $a_{j_i} = \sum_{k=1}^{V} \alpha_{j_i,k}$ for all i = 1,...,T. Further suppose that whenever contexts $j_1,...,j_p$ generate the sample then $\frac{n_{s,j_i}}{n_s} =$ ν_{s,j_i} is drawn from a Dirichlet($a_{j_1},...,a_{j_p}$) distribution. Then the NUHP provides consistent inference for the set of contexts that generates a collection of samples.

4.2. Clustering in a Collection of Samples. In order to complete the specification for a mixture of contexts, one has to both cluster samples together and assign a set of contexts to each subset of samples in a cluster. Once again, the prior is constructed in a hierarchical fashion. There are at most $M = 2^T - 1$ possible context allocations for subsets of samples in a collection because there are only M unique choices for the set of contexts that can generate any particular sample. Assume that the there are m clusters $g_1,...,g_m$ in a particular partition of [S] = {1,...,S} with sizes $\eta_1,...,\eta_m$ and context sets $c_1,...,c_m$. The HUP for this clustering problem is given by

$$p(g|n_1,...,n_S,T) = \frac{1}{Mb(S,m)N(\eta_1,...,\eta_m)}$$

All that remains is to define a prior for the sets of contexts c that generate the data in the given clusters. Since each cluster must be assigned a unique set of contexts, there are M(M − 1)···(M − m + 1) ways to assign context sets for a partition with m clusters. Using the uniform prior provides

$$p(c|g,n_1,...,n_S,T) = \frac{(M-m)!}{M!}$$

The posterior probability of the partition and context allocation c,g, where the samples $d_{g_i,1},\dots,d_{g_i,\eta_i}$ are in subset g_i, is given by

$$p(c,g|w_1,\dots,w_s,n_1,\dots,n_s,S,T) \propto \left(\prod_{i=1}^{m}\prod_{s\in g_i} p(d_s|w_s,n_s,T)\right) \times \left(\prod_{i=1}^{m}[(T-p_i)!p_i!]\right)^{-\eta_i} \frac{(M-m)!}{b(S,m)N(\eta_1,\dots,\eta_m)}$$

where p_i is the number of contexts used in c_i.

The choices of (c,g) can be represented by the space $M_{S,M}$ of S × M matrices which are zero except for a single element in each row that is one. The number of clusters, m, is given by the number of columns of such a matrix that contain at least one non-zero element. The cluster sizes are given by the column sums of the matrix. A random walk on the space of clusters of samples in the collection can be performed by doing a random walk on the space $M_{S,M}$.

5. Conclusion. The LDA model is inconsistent for determining the contexts that are used in a sample for the latent allocation problem. The posterior distribution under this prior and the Ewens-Pitman prior both degenerate to the maximum number of allowed contexts. In contrast, the HUP and NUHP produce posterior distributions for a single sample that are non-degenerate. When considering a collection of samples, the HUP and NUHP produce consistent inference for the set of contexts that generates the collection in a range of cases as the number of samples increases to infinity, making them ideal priors to use for clustering of samples into groups which are generated by different context sets. This consistency in a collection is strongly tied to the per-sample perplexity of the collection. The hierarchical priors choose the set of contexts that asymptotically minimize per-sample perplexity.

However, there are open questions to be addressed even for the simple “bag of words” model considered in this paper. The first concerns the behavior of the HUP and NUHP models in a collection of samples when the

feature Dirichlet parameters do not sum to the context Dirichlet parameters ($a_{j_i} = \sum_{k=1}^{V} \alpha_{j_i,k}$). The distribution of $\gamma_{s,k}$ under this assumption and consistency are open questions. Second, what is the posterior behavior when a is mispecified? How do the different priors behave for misspecified models and is there a way to specify a prior which is consistent for the contexts generating the data without needing to specify a? Further, the consistency properties of the posterior behavior of extensions of this hierarchical modeling strategy to other latent allocation models, such as time varying content analysis and n-gram models need to be addressed.

Supplemental Materials.

Lemma 3.

Lemma 3. Let b(n, p) be the number of Young's diagrams with n boxes and p rows. For fixed p

$$\lim_{n\to\infty} \frac{b(n, p)}{n_{p-1}} = a(p)$$

where $a(p)_{-1} = p!(p-1)!$.

Proof. Because b(n, p) is weakly increasing in n, it suffices to prove the result for n = kp. Proceed by induction

$$\begin{aligned}
\frac{b(n, p)}{n_{p-1}} &= \frac{b(kp, p)}{(kp)_{p-1}} \\
&= \frac{b(kp-1, p-1) + b((k-1)p, p)}{(kp)_{p-1}} \\
&= \frac{b(kp-1, p-1) + b((k-1)p-1, p-1) + b((k-2)p, p)}{(kp)_{p-1}} \\
&= \frac{1}{(kp)_{p-1}} \sum_{l=1}^{k} b(lp-1, p-1) \\
&= \frac{1}{(kp)_{p-1}} \sum_{l=1}^{k} (lp-1)_{p-2} \frac{b(lp-1, p-1)}{(lp-1)_{p-2}} \\
&\approx \frac{1}{(kp)_{p-1}} \sum_{l=1}^{k} (lp-1)_{p-2} a(p-1) \\
&\approx \frac{a(p-1)}{pk_{p-1}} \sum_{l=1}^{k} (l)_{p-2}
\end{aligned}$$

The sum $\sum_{l=1}^{k}(l)_{p-2}$ is a polynomial of degree p − 1 in k with leading coefficient $\frac{1}{p-1}$. Therefore,

$$\frac{b(n, p)}{n^{p-1}} \approx \frac{a(p-1)}{p(p-1)}$$

Finally, because a(1) = 1 it follows that $a(p)^{-1} = p!(p-1)!$.

Proof of Proposition 6.

Proof. This follows from simple algebra and the Stirling approximation to the Γ function. In the following, C(z,w) is a constant that does not vary with n but can change from line to line.

$$
\begin{aligned}
p_{HUP}(z,w|n,T,\alpha) &= p_{HUP}(w|z,\alpha,n,T)\pi_{HUP}(z|n,T)\\
&= \left(\prod_{j=1}^{T}\left(\prod_{k=1}^{V}\frac{\Gamma(n_{j,k}(z,w)+\alpha_{j,k})}{\Gamma(\alpha_{j,k})}\right)\frac{\Gamma(\sum_{k=1}^{V}\alpha_{j,k})}{\Gamma(n_j(z)+\sum_{k=1}^{V}\alpha_{j,k})}\right)\\
&\times \frac{1}{T}\frac{1}{b(n,p)}\frac{1}{N(n_1,\ldots,n_T)}\frac{(T-p)!}{T!}\\
&= \left(\prod_{i=1}^{p}\left(\prod_{k=1}^{V}\frac{\Gamma(n_{j_i,k}(z,w)+\alpha_{j_i,k})}{\Gamma(\alpha_{j_i,k})}\right)\frac{\Gamma(\sum_{k=1}^{V}\alpha_{j_i,k})}{\Gamma(n_{j_i}(z)+\sum_{k=1}^{V}\alpha_{j_i,k})}\right)\\
&\times \frac{1}{T}\frac{1}{b(n,p)}R(n_{j_1},\ldots,n_{j_p})\binom{n}{n_{j_1},\ldots,n_{j_p}}^{-1}\frac{(T-p)!}{T!}\\
&= \left(\prod_{i=1}^{t}\left(\prod_{k\in V_{ji}}\frac{\Gamma(n_{j_i,k}(z,w)+\alpha_{j_i,k})}{\Gamma(\alpha_{j_i,k})}\right)\left(\prod_{k/\in V_{ji}}\frac{\Gamma(n_{j_i,k}(z,w)+\alpha_{j_i,k})}{\Gamma(\alpha_{j_i,k})}\right)\right.\\
&\times \left.\frac{\Gamma(\sum_{k=1}^{V}\alpha_{j_i,k})}{\Gamma(n_{j_i}(z)+\sum_{k=1}^{V}\alpha_{j_i,k})}\right)\\
&\times \left(\prod_{i=t+1}^{p}\left(\prod_{k=1}^{V}\frac{\Gamma(n_{j_i,k}(z,w)+\alpha_{j_i,k})}{\Gamma(\alpha_{j_i,k})}\right)\frac{\Gamma(\sum_{k=1}^{V}\alpha_{j_i,k})}{\Gamma(n_{j_i}(z)+\sum_{k=1}^{V}\alpha_{j_i,k})}\right)\\
&\times \frac{1}{T}\frac{1}{b(n,p)}R(n_{j_1},\ldots,n_{j_p})\binom{n}{n_{j_1},\ldots,n_{j_p}}^{-1}\frac{(T-p)!}{T!}\\
&= \frac{C(z,w)}{n!b(n,p)}\left(\prod_{i=1}^{t}\left(\prod_{k\in V_{ji}}\Gamma(n_{j_i,k}(z,w)+\alpha_{j_i,k})\right)\frac{n_{j_i}(z)!}{\Gamma(n_{j_i}(z)+\sum_{k=1}^{V}\alpha_{j_i,k})}\right)\\
&\to \frac{C(z,w)}{n!n^{p-1}}\left(\prod_{i=1}^{t}\left(\prod_{k\in V_{ji}}\sqrt{\frac{2\pi}{n_{j_i,k}(z,w)+\alpha_{j_i,k}}}\left(\frac{n_{j_i,k}(z,w)+\alpha_{j_i,k}}{e}\right)^{n_{ji,k}(z,w)+\alpha_{ji,k}}\right)\right.\\
&\times \left.\frac{\sqrt{2\pi n_{j_i}(z)}\left(\frac{n_{ji}(z)}{e}\right)^{n_{ji}(z)}}{\sqrt{\frac{2\pi}{n_{ji}(z)+\sum_{k=1}^{V}\alpha_{ji,k}}}\left(\frac{n_{ji}(z)+\sum_{k=1}^{V}\alpha_{ji,k}}{e}\right)^{n_{ji}(z)+\sum_{k=1}^{V}\alpha_{ji,k}}}\right)\\
&= \frac{C(z,w)}{n!n^{p-1}e^{n}}\left(\prod_{i=1}^{t}\left(\prod_{k\in V_{ji}}n_{j_i,k}(z,w)^{n_{ji,k}(z,w)+\alpha_{ji,k}-\frac{1}{2}}\right)n_{j_i}(z)^{1-\sum_{k=1}^{V}\alpha_{ji,k}}\right)
\end{aligned}
$$

Proof of Theorem 3.

Proof. For convenience, define

$$A(\{n_{j,k}\}) = \{z : n_{j,k}(z,w) = n_{j,k}\}$$

The Stirling approximation for $n_{j_i,k}(z,w)$ provides

$$\begin{aligned}
p(A(\{n_{j_i,k}\}),w|n,T,\alpha) = {} & \frac{(T-p)!}{T!Tb(n,p)} \left(\prod_{i=1}^{p}\left(\frac{\Gamma(\sum_{k=1}^{V}\alpha_{j_i,k})}{\prod_{k=1}^{V}\Gamma(\alpha_{j_i,k})}\right)\right) \\
& \times \left(\prod_{i=1}^{p}\left(\frac{\prod_{k=1}^{V}\Gamma(n_{j_i,k}+\alpha_{j_i,k})}{\Gamma(n_{j_i}+\sum_{k=1}^{V}\alpha_{j_i,k})}\right)\right) \\
& \times \frac{\prod_{i=1}^{p}\Gamma(n_{j_i}+1)}{\Gamma(n+1)} R(n_{j_1},\dots,n_{j_p}) \\
& \times \prod_{k=1}^{V}\frac{\Gamma(m_k+1)}{\prod_{i=1}^{p}\Gamma(n_{j_i,k}+1)} \\
\approx {} & \frac{(T-p)!}{T!Tb(n,p)} \left(\prod_{i=1}^{p}\left(\frac{\Gamma(\sum_{k=1}^{V}\alpha_{j_i,k})}{\prod_{k=1}^{V}\Gamma(\alpha_{j_i,k})}\right)\right)(2\pi)^{\frac{V-1}{2}} R(n_{j_1},\dots,n_{j_p}) \\
& \times \left(\prod_{i=1}^{p}\left(\frac{\prod_{k=1}^{V} n_{j_i,k}^{n_{j_i,k}+\alpha_{j_i,k}-\frac{1}{2}}}{n_{j_i}^{n_{j_i}+\sum_{k=1}^{V}\alpha_{j_i,k}-\frac{1}{2}}}\right)\right) \frac{\prod_{i=1}^{p} n_{j_i}^{n_{j_i}+\frac{1}{2}}}{n^{n+\frac{1}{2}}} \prod_{k=1}^{V}\frac{m_k^{m_k+\frac{1}{2}}}{\prod_{i=1}^{p} n_{j_i,k}^{n_{j_i,k}+\frac{1}{2}}} \\
= {} & \frac{(T-p)!}{T!Tb(n,p)} \left(\prod_{i=1}^{p}\left(\frac{\Gamma(\sum_{k=1}^{V}\alpha_{j_i,k})}{\prod_{k=1}^{V}\Gamma(\alpha_{j_i,k})}\right)\right)(2\pi)^{\frac{V-1}{2}} R(n_{j_1},\dots,n_{j_p}) \\
& \times \left(\prod_{k=1}^{V}\prod_{i=1}^{p} n_{j_i,k}^{\alpha_{j_i,k}-1}\right)\left(\prod_{i=1}^{p} n_{j_i}^{1-\sum_{k=1}^{V}\alpha_{j_i,k}}\right)\left(\prod_{k=1}^{V} m_k^{m_k+\frac{1}{2}}\right) n^{-n-\frac{1}{2}} \\
= {} & \frac{(T-p)!}{T!Tb(n,p)} \left(\prod_{i=1}^{p}\left(\frac{\Gamma(\sum_{k=1}^{V}\alpha_{j_i,k})}{\prod_{k=1}^{V}\Gamma(\alpha_{j_i,k})}\right)\right)(2\pi)^{\frac{V-1}{2}} R(n_{j_1},\dots,n_{j_p}) \\
& \times \left(\prod_{k=1}^{V}\prod_{i=1}^{p} n_{j_i,k}^{\alpha_{j_i,k}-1}\right)\left(\prod_{i=1}^{p}\left(\sum_{k=1}^{V} n_{j_i,k}\right)^{1-\sum_{k=1}^{V}\alpha_{j_i,k}}\right)\left(\prod_{k=1}^{V} m_k^{m_k+\frac{1}{2}}\right) n^{-n-\frac{1}{2}}
\end{aligned}$$

In order to compute the probability for $R_p(j_1,\dots,j_p)$, one must sum over the $A(\{n_{j_i,k}\})$. This sum can be approximated an integral after making the change of variables $\varrho_{j_i,k} = \frac{n_{j_i,k}}{m_k}$. Since $\sum_{i=1}^{p} \varrho_{j_i,k} = 1$, this is an integral over the Cartesian product of V simplices. The Jacobian of this change of variables is $\prod_k m_k^{p-1}$. Setting the redundancy numbers to 1 provides an approximation of

$$
\begin{aligned}
&p(R_p(j_1,\dots,j_p), w|n, T, \alpha)\\
&\approx \frac{(T-p)!}{T!Tb(n,p)}\left(\prod_{i=1}^{p}\left(\frac{\Gamma(\sum_{k=1}^{V}\alpha_{j_i,k})}{\prod_{k=1}^{V}\Gamma(\alpha_{j_i,k})}\right)\right)(2\pi)^{\frac{V-1}{2}}n^{-n-\frac{1}{2}}\left(\prod_{k=1}^{V} m_k^{m_k+\frac{1}{2}}\right)\\
&\quad\times\int\left(\prod_{k=1}^{V}\prod_{i=1}^{p}(\varrho_{j_i,k}m_k)^{\alpha_{j_i,k}-1}\right)\left(\prod_{i=1}^{p}\left(\sum_{k=1}^{V}\varrho_{j_i,k}m_k\right)^{1-\sum_{k=1}^{V}\alpha_{j_i,k}}\right)\prod_k m_k^{p-1}\prod_{k=1}^{V}\prod_{i=1}^{p-1}d\varrho_{j_i,k}\\
&= \frac{(T-p)!}{T!Tb(n,p)}\left(\prod_{i=1}^{p}\left(\frac{\Gamma(\sum_{k=1}^{V}\alpha_{j_i,k})}{\prod_{k=1}^{V}\Gamma(\alpha_{j_i,k})}\right)\right)(2\pi)^{\frac{V-1}{2}}n^{-n-\frac{1}{2}}\left(\prod_{k=1}^{V} m_k^{m_k+\sum_{i=1}^{p}\alpha_{j_i,k}-\frac{1}{2}}\right)\\
&\quad\times\int\left(\prod_{k=1}^{V}\prod_{i=1}^{p}(\varrho_{j_i,k})^{\alpha_{j_i,k}-1}\right)\left(\prod_{i=1}^{p}\left(\sum_{k=1}^{V}\varrho_{j_i,k}m_k\right)^{1-\sum_{k=1}^{V}\alpha_{j_i,k}}\right)\prod_{k=1}^{V}\prod_{i=1}^{p-1}d\varrho_{j_i,k}\\
&= \frac{(T-p)!}{T!T}\frac{n^{p-1}}{b(n,p)}\left(\prod_{i=1}^{p}\left(\frac{\Gamma(\sum_{k=1}^{V}\alpha_{j_i,k})}{\prod_{k=1}^{V}\Gamma(\alpha_{j_i,k})}\right)\right)\left(\frac{2\pi}{n}\right)^{\frac{V-1}{2}}\left(\prod_{k=1}^{V}\left(\frac{m_k}{n}\right)^{m_k+\sum_{i=1}^{p}\alpha_{j_i,k}-\frac{1}{2}}\right)\\
&\quad\times\int\left(\prod_{k=1}^{V}\prod_{i=1}^{p}(\varrho_{j_i,k})^{\alpha_{j_i,k}-1}\right)\left(\prod_{i=1}^{p}\left(\sum_{k=1}^{V}\varrho_{j_i,k}\frac{m_k}{n}\right)^{1-\sum_{k=1}^{V}\alpha_{j_i,k}}\right)\prod_{k=1}^{V}\prod_{i=1}^{p-1}d\varrho_{j_i,k}\\
&\approx \frac{(T-p)!}{T!Ta(p)}\left(\prod_{i=1}^{p}\left(\frac{\Gamma(\sum_{k=1}^{V}\alpha_{j_i,k})}{\prod_{k=1}^{V}\Gamma(\alpha_{j_i,k})}\right)\right)\left(\frac{2\pi}{n}\right)^{\frac{V-1}{2}}\left(\prod_{k=1}^{V}\left(\frac{m_k}{n}\right)^{m_k+\sum_{i=1}^{p}\alpha_{j_i,k}-\frac{1}{2}}\right)\\
&\quad\times\int\left(\prod_{k=1}^{V}\prod_{i=1}^{p}(\varrho_{j_i,k})^{\alpha_{j_i,k}-1}\right)\left(\prod_{i=1}^{p}\left(\sum_{k=1}^{V}\varrho_{j_i,k}\frac{m_k}{n}\right)^{1-\sum_{k=1}^{V}\alpha_{j_i,k}}\right)\prod_{k=1}^{V}\prod_{i=1}^{p-1}d\varrho_{j_i,k}
\end{aligned}
$$

The term

$$\frac{1}{T}\left(\frac{2\pi}{n}\right)^{\frac{V-1}{2}}\prod_{k=1}^{V}\left(\frac{m_k}{n}\right)^{m_k+\frac{1}{2}}$$

is common to the probabilities for all of the choices for $R_p(j_1,\dots,j_p)$. Lemma 4 shows that the integral is bounded.

Lemma 4. Suppose that $x_1,\dots,x_m,y_1,\dots,y_n,\chi_1,\dots,\chi_m,\zeta_1,\dots,\zeta_n$ are all positive with $\sum x_j + \sum y_i$ less than 1. Then

$$\prod_j x_j^{\chi_j} \le \left(\sum x_j + \sum y_i\right)^{\sum \chi_j - \sum \zeta_i} \le \prod_i y_i^{-\zeta_i}$$

Proof.

$$\left(\sum x_j\right)^{\sum \chi_j} \le \left(\sum x_j + \sum y_i\right)^{\sum \chi_j} \le \left(\sum x_j + \sum y_i\right)^{\sum \chi_j - \sum \zeta_i}$$

$$\le \left(\sum x_j + \sum y_i\right)^{-\sum \zeta_i} \le \left(\sum y_i\right)^{-\sum \zeta_i}$$

The lemma will follow so long as

$$\prod_i a_i^{b_i} \le \left(\sum_i a_i\right)^{\sum_i b_i}$$

for positive a_i, b_i, which follows directly from induction.

Theorem 7. The integral

$$I(m_k, n; \{\alpha_{ji,k}\}) = \int \left(\prod_{k=1}^{V} \prod_{i=1}^{p} (\varrho_{ji,k})^{\alpha_{ji,k} - 1}\right) \left(\prod_{i=1}^{p} \left(\sum_{k=1}^{V} \varrho_{ji,k} \frac{m_k}{n}\right)^{1 - \sum_{k=1}^{V} \alpha_{ji,k}}\right) \prod_{k=1}^{V} \prod_{i=1}^{p-1} d\varrho_{ji,k}$$

is bounded above and below.

Proof. Let $\gamma_{ji,k}$ be positive constants with $\sum_{k=1}^{V} \gamma_{ji,k} = 1$, then

$$\left(\sum_{k=1}^{V} \varrho_{ji,k} \frac{m_k}{n}\right)^{1 - \sum_{k=1}^{V} \alpha_{ji,k}} = \left(\sum_{k=1}^{V} \varrho_{ji,k} \frac{m_k}{n}\right)^{\sum_{k=1}^{V} (\gamma_{ji,k} - \alpha_{ji,k})}$$

Let $L_i = \{k : \gamma_{ji,k} > \alpha_{ji,k}\}$ and $U_i = \{k : \gamma_{ji,k} < \alpha_{ji,k}\}$, then

$$\prod_{k \in L_i} \left(\varrho_{ji,k} \frac{m_k}{n}\right)^{\gamma_{ji,k} - \alpha_{ji,k}} \le \left(\sum_{k=1}^{V} \varrho_{ji,k} \frac{m_k}{n}\right)^{1 - \sum_{k=1}^{V} \alpha_{ji,k}} \le \prod_{k \in U_i} \left(\varrho_{ji,k} \frac{m_k}{n}\right)^{\gamma_{ji,k} - \alpha_{ji,k}}$$

Define

$$l_{j_i,k} = \max\{\gamma_{j_i,k}, \alpha_{j_i,k}\} \qquad \text{and} \qquad u_{j_i,k} = \min\{\gamma_{j_i,k}, \alpha_{j_i,k}\}$$

then

$$\left(\prod_{i=1}^{p} \prod_{k\in L_i} \left(\frac{m_k}{n}\right)^{\gamma_{j_i,k}-\alpha_{j_i,k}}\right) \prod_{k=1}^{V} D(l_{j_1,k},\dots,l_{j_p,k}) \le I\,(m_k,n;\{\alpha_{j_i,k}\})$$

and

$$I\,(m_k,n;\{\alpha_{j_i,k}\}) \le \left(\prod_{i=1}^{p} \prod_{k\in U_i} \left(\frac{m_k}{n}\right)^{\gamma_{j_i,k}-\alpha_{j_i,k}}\right) \prod_{k=1}^{V} D(u_{j_1,k},\dots,u_{j_p,k})$$

where

$$D(\xi_1,\dots,\xi_p) = \frac{\prod_{i=1}^{p} \Gamma(\xi_i)}{\Gamma(\sum_{i=1}^{p} \xi_i)}$$

All that remains now is to justify the use of the Stirling approximation and setting the redundancy numbers to unity. The latter is simple to justify. Because the redundancy number is greater than one only if at least two of the cluster sizes are the same, this induces linear equality constraints on the $\varrho_{j_i,k}$. This reduces the power of m_k in the Jacobian of the transformation from $n_{j_i,k}$ to $\varrho_{j_i,k}$. This reduces the power of n for the integral over this subspace. Since the integral is finite over this subspace, the extra power of n reduces its contribution to the joint probability to 0.

Justifying the use of the Stirling approximation is a slightly more delicate matter. In order to obtain a certain accuracy using the approximation, one has to restrict to $n_{j_i,k}$ large enough. Fix N and consider all $n_{j_i,k} \ge N$ sufficient for the integral formula to be valid. Divide the space of z into two set S and U where S represents the set of z with all $n_{j_i,k}(z,w) \ge N$ and U represents the set of z with at least one $n_{j_i,k}(z,w) < N$. Because the integral over set of $U_\varrho = \{\varrho_{j_i,k} : \varrho_{j_i,k} < \frac{N}{m_k}\}$ is 0 asymptotically the sum over S converges to the integral in (27). Because the ratio between the sum over U and the integral over U_ϱ is bounded, the fact that the integral over

U_0 converges to 0 implies that the sum over U converges to 0. Therefore, the integral formula is valid as an asymptotic approximation to the joint probability.

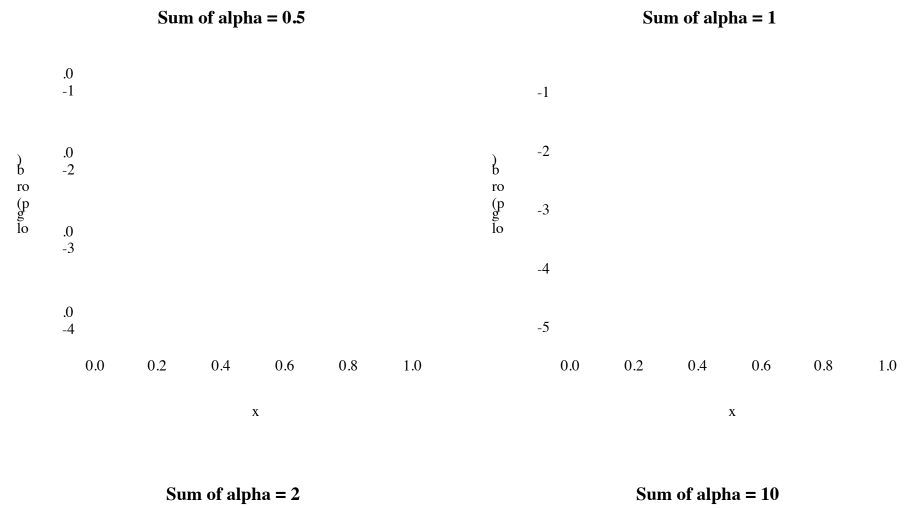
Effect of α_j on the HUP Posterior for a Single Sample. In order to understand the effect of $\sum_k \alpha_{j,k}$ on the asymptotic joint probability, consider the situation where there are two features and two or three possible contexts. Define the α matrices to be

$$\alpha = \begin{pmatrix} \frac{1}{3} & \frac{2}{3} \\ \frac{3}{4} & \frac{1}{4} \end{pmatrix} \qquad \alpha = \begin{pmatrix} \frac{1}{3} & \frac{2}{3} \\ \frac{3}{4} & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{2} \end{pmatrix}$$

for two and three contexts, respectively. Figures 8 and 9 show the effects of scaling the matrices by 0.5,1,2,10 for two and three contexts, respectively.

Page 36

Increasing the scaling factor causes the regions where different context mixtures are preferred to change in accordance with the Dirichlet sampling distribution of β. When $\alpha_{j,k} < 1$ then the region where the mixture is preferred to a single model grows since the Dirichlet distribution for any single context forces mass towards the endpoints of the interval. Vertical lines have been inserted at $\frac{1}{3}$ and $\frac{3}{4}$ for two contexts and at $\frac{1}{3}$, $\frac{1}{2}$, and $\frac{3}{4}$ for three contexts.



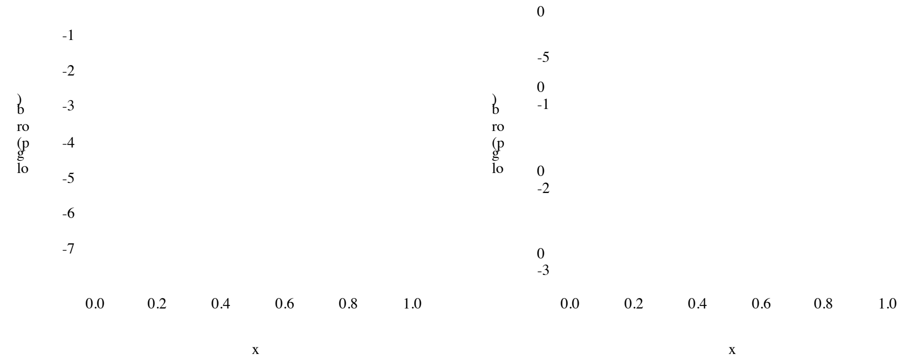
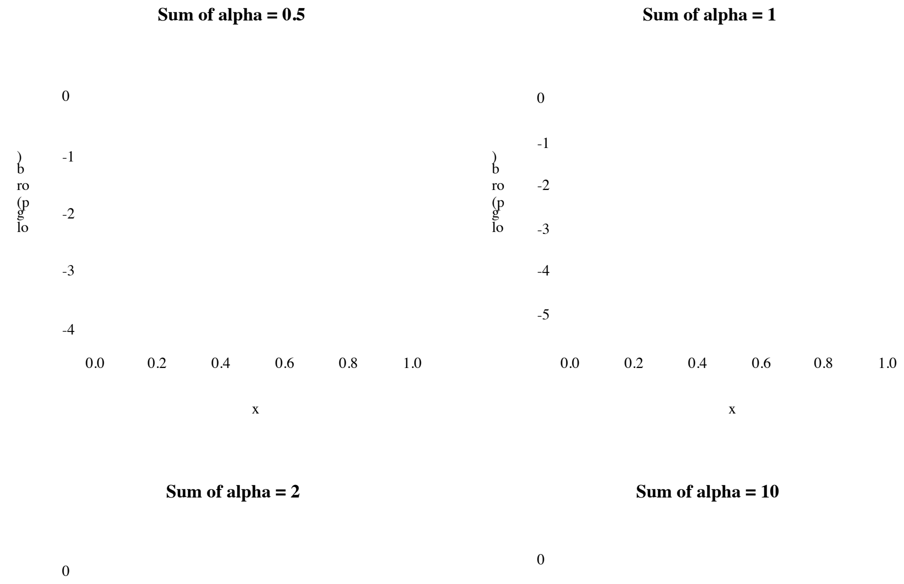


Fig 8. Effect of $\sum_k \alpha_{j,k}$ for T = 2


Sum of alpha = 0.5

Sum of alpha = 1

Sum of alpha = 2

Sum of alpha = 10

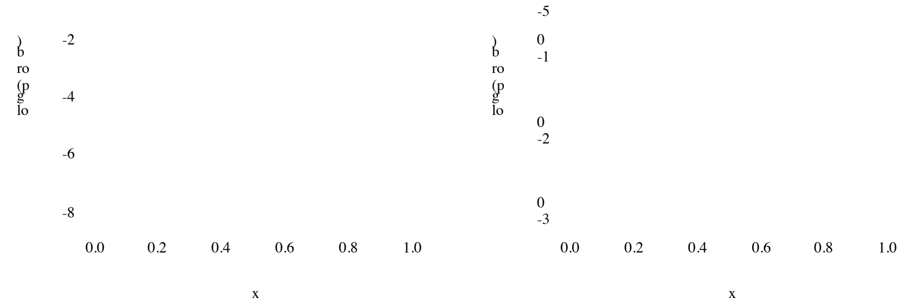


Fig 9. Effect of $\sum_k \alpha_{j,k}$ for T = 3

The integral

$$\int \left(\prod_{k=1}^{V} \prod_{i=1}^{p} (\varrho_{ji,k})^{\alpha_{ji,k}-1} \right) \left(\prod_{i=1}^{p} \left(\sum_{k=1}^{V} \varrho_{ji,k}\gamma_k \right)^{1-\sum_{k=1}^{V} \alpha_{ji,k}} \right) \prod_{k=1}^{V} \prod_{i=1}^{p-1} d\varrho_{ji,k}$$

can be achieved through direct simulation using the Cartesian product of V simplexes. For a given context, if $\sum_k \alpha_{ji,k} \leq 1$, then $\left(\sum_{k=1}^{V} \varrho_{ji,k}\gamma_k\right)^{1-\sum_{k=1}^{V} \alpha_{ji,k}} \leq$

1. However, if $\sum_k \alpha_{ji,k} > 1$, then $\left(\sum_{k=1}^{V} \varrho_{ji,k}\gamma_k\right)^{1-\sum_{k=1}^{V} \alpha_{ji,k}}$ is unbounded. Because of this, sampling from the simplexes using $\alpha_{ji,k}$ converges slowly. Using Lemma 4, one can replace $\alpha_{ji,k}$ by $\frac{\alpha_{ji,k}}{\sum_k \alpha_{ji,k}}$ whenever $\sum_k \alpha_{ji,k} > 1$ and use importance sampling, producing far more stable estimates of the integral.

References.

Barry, D. and Hartigan, J. (1992). Product partition models for change point problems. The Annals of Statistics, pages 260–279.

Blei, D., Ng, A., and Jordan, M. (2003). Latent dirichlet allocation. The Journal of Machine Learning Research, 3:993–1022.

Canini, K., Shi, L., and Griffiths, T. (2009). Online inference of topics with latent dirichlet allocation. In Proceedings of the International Conference on Artificial Intelligence and Statistics, pages 65–72.

Casella, G., Moreno, E., and Girón, F. (2011). Cluster analysis, model selection, and prior distributions on models. University of Florida Technical Report.

Doyle, G. and Elkan, C. (2009). Accounting for burstiness in topic models. In Proceedings of the 26th Annual International Conference on Machine Learning, pages 281–288. ACM.

Endres, F., Plagemann, C., Stachniss, C., and Burgard, W. (2009). Unsupervised discovery of object classes from range data using latent dirichlet allocation. In Robotics: Science and Systems Conference.

Hartigan, J. (1990). Partition models. Communications in Statistics-Theory and Methods, 19(8):2745–2756.

Li, L. and Fei-Fei, L. (2007). What, where and who? classifying events by scene and object recognition. In Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on, pages 1–8. IEEE.

Li, L., Socher, R., and Fei-Fei, L. (2009). Towards total scene understanding: Classification, annotation and segmentation in an automatic framework. In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pages 2036–2043. IEEE.

Mariote, L., Medeiros, C., and da Torres, R. (2007). Parallelized variational em for latent dirichlet allocation: An experimental evaluation of speed and scalability. In Data Mining Workshops, 2007. ICDM Workshops 2007. Seventh IEEE International Conference on, pages 349–354. IEEE.

Newman, D., Asuncion, A., Smyth, P., and Welling, M. (2007). Distributed inference for latent dirichlet allocation. Advances in Neural Information Processing Systems, 20(1081-1088):17–24.

Pitman, J. (1996). Some developments of the blackwell-macqueen urn scheme. Lecture Notes-Monograph Series, pages 245–267.

Porteous, I., Newman, D., Ihler, A., Asuncion, A., Smyth, P., and Welling, M. (2008). Fast collapsed gibbs sampling for latent dirichlet allocation. In KDD'08: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 569–577.

Teh, Y., Newman, D., and Welling, M. (2007). A collapsed variational bayesian inference algorithm for latent dirichlet allocation. Advances in neural information processing systems, 19:1353.

Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006). Hierarchical dirichlet processes. Journal of the American Statistical Association, 101(476):1566–1581.

Wang, X. and Grimson, E. (2007). Spatial latent dirichlet allocation. Advances in Neural Information Processing Systems, 20:1577–1584.

Xing, D. and Girolami, M. (2007). Employing latent dirichlet allocation for fraud detection in telecommunications. Pattern Recognition Letters, 28(13):1727–1734.

Zhang, T., Lu, H., and Li, S. (2009). Learning semantic scene models by object classification and trajectory clustering. In Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pages 1940–1947. IEEE.

Address of the First and Second authors,
E-mail: ajwomack@stat.ufl.edu; emoreno@ugr.es