Implementing the Collapsed Gibbs Sampler for Latent Dirichlet Allocation

Steve Bronder
sab2287@columbia.edu

Antonio Moretti
am4134@columbia.edu

May 13, 2016

Abstract

Latent Dirichlet Allocation is a two stage hierarchical clustering process that is typically fit through MCMC or variational inference. We discuss how the collapsed Gibbs sampler is derived and implemented, including the method of integrating out multinomial parameters as an example of Rao-Blackwellization. This allows for all samples to be drawn from simple conditional distributions while also supporting the update of each topic distribution after each word is assigned to a particular topic. We illustrate how to perform posterior inference on a collection of documents from JSTOR. We review a few open problems and compare several heuristics to check rates of convergence.

1 Introduction

Topic models are a family of unsupervised machine learning methods for summarizing and organizing large collections of documents. Topic models aim to mimic the writing process in which an author draws upon a set of topics based on the focus of the narrative. Words represent various topics and provide clues about the underlying themes that comprise the document. Introduced by Blei in 2003, Latent Dirichlet Allocation [**?**] (LDA) has become a widely popular topic model for text processing due to its simplicity. The same model is derived in 2000 by Pritchard for isolating microbial species using genotype data [**?**]. At its core LDA is a two level Bayesian heirarchical model for clustering with hidden factors. Formally, a topic is defined as a hidden probability distribution over a fixed vocabulary. Documents are modeled as mixtures of topic distributions, which in turn are modeled as distributions over words. This two stage hierarchical process supports a soft rather than a hard assignment of each document to multiple topics where words are the only variables observed.

Outline We give a brief description of the LDA model in Section 2. In Section 3 we discuss the challenge of posterior inference and motivate the MCMC approach. Section 4 describes the properties of the Gibbs sampler including the collapsed sampler and Rao Blackwellization. Some mathematical derivations are provided for completeness in Section 7.

2 LDA Model

We introduce some notation in order to state the model. Following Carpenter [**?**], let K be the number of topics and let M be the number of documents. Let N_m be the number of words in the mth document and let J be the distinct number of words in the corpus. Let $W_{m,n}$ be the observed document term matrix where rows represent documents, columns represent words and the components are the counts of the nth word in the mth document. Let $Z_{m,n} \in 1:K$ be the topic to which the word of $W_{m,n}$ is assigned. Let $\theta_m \in [0,1]^K$ be the topic distribution for the mth document and $\phi_k \in [0,1]^J$ be the word distribution for the kth topic. The hyperparameter $\alpha \in \mathbb{R}^K$ is a vector of prior counts for topics in documents. Similarly $\beta \in \mathbb{R}^J$ is a vector of prior counts for words in a given topic. LDA is defined by the following generative model.

1. For each topic k draw a word distribution:
   (a) $\phi_k \sim Dirichlet(\beta)$
2. For each document m:
   (a) $\theta_m \sim Dirichlet(\alpha)$
   (b) For each of the n words in m:
      i. $z_{m,n} \sim Mult(\theta_m)$
      ii. $w_{m,n} \sim Mult(\phi_{z_{m,n}})$

The joint distribution of the graphical model factorizes as products of the conditional distributions:

$$p(\theta|\alpha)p(z|\theta)p(w|z,\phi)p(\phi|\beta) \tag{1}$$

Define $c_{k,m,j}$ to be the count for the number of times word j is assigned to topic k in document m:

$$c_{k,m,j} = \sum_{n=1}^{Nm} I\Big(z_{m,n} = k \;\&\; w_{m,n} = j\Big)$$

Note that we can marginalize the distribution of counts over any of the three variables k, m and j. For example,

$$c_{k,m,*} = \sum_{j=1}^{J} c_{k,m,j}$$

3 Posterior Inference for LDA

In order to compute the posterior distribution and fit an LDA model we must find the conditional distribution of hidden variables given the observations:

$$P(\theta, z, \phi|w, \alpha, \beta) \propto \prod_{k=1}^{K} p(\phi_k|\beta) \prod_{m=1}^{M} p(\theta_m|\alpha) \prod_{n=1}^{N} p(z_{m,n}|\theta_m)p(w_{m,n}|z_{m,n}, \phi_k) \tag{2}$$

Recall the first two terms are Dirichlet. The third term represents a draw of the topic assignment θ_d for each of the N_m words in the document. The last term represents the likelihood of drawing a word given the joint distribution of observations and hidden variables.

It is easy to see that evaluating the above presents computational challenges. Consider for simplicity that we are dealing with one document where the topics ϕ_k are fixed. The per document posterior is then proportional to the following.

$$p(\theta|\alpha)\prod_{n=1}^{N} p(z_n|\theta)p(w_n|z_n,\phi_k) \tag{3}$$

In order to ensure that the above is normalized it is necessary to compute the evidence.

$$\int_\theta p(\theta|\alpha)\prod_{n=1}^{N}\sum_{z=1}^{K} p(z_n|\theta)p(w_n|z_n,\phi_k)d\theta \tag{4}$$

We can move the summation outside the integral to obtain the form of the evidence below:

$$\sum_{z=1}^{K}\int_\theta p(\theta|\alpha)\prod_{n=1}^{N} p(z_n|\theta)p(w_n|z_n,\phi_k)d\theta \tag{5}$$

The above is a hypergeometric function shown by Dickey to be intractable [**?**]. The simplified posterior is also equivalent to the sum of N^k tractable Dirichlet integrals. Even without considering topics or multiple documents, we are forced to compute an exponential number of Dirichlet integrals. This is often solved through approximate inference such as MCMC or mean field variational methods.

4 The Gibbs Sampler

The MCMC approach builds an irreducible, aperiodic and positive recurrent chain whose stationary distribution is the posterior distribution of interest. The Gibbs sampler then uses the space of all configurations of hidden variables to represent the space of the Markov chain. We can then run the chain by iteratively sampling from the distribution of each hidden variable conditioned on the observations and the current state of each of the other hidden variables. After a suitable burn-in period, we can collect samples to obtain a Monte Carlo estimate of the posterior.

A naive sampler would condition on all of the hidden variables, however it is possible to integrate out the multinomial parameters for a faster mixing chain. We describe this process in detail in the Appendix. In addition, by using conditional expectations we can improve the variance of the Monte Carlo estimate.

$$Var(\mathbb{E}[\delta(X)|Y]) \leq Var(\delta(X)) \tag{6}$$

We review some basic properties of the Gibbs sampler and Rao-Blackwellization below. We wish to show that the support of the joint distribution f is the Cartesian product of marginal distributions.

Definition 4.1. The positivity condition for a joint distribution $f(X_1,\cdots,X_p)$ with marginal densities $f_{X_i}(x_i)$ is satisfied if the following holds: $f_{X_i}(x_i) > 0$ for all $(X_1,\cdots,X_p)$ implies that $f(X_1,\cdots,X_p) > 0$.

Theorem 4.1 (Hammersley-Clifford)**.** *Suppose that* $(X_1, \cdots, X_d)$ *satisfies the positivity condition with joint pdf* $f(X_1, \cdots, X_p)$*. For all* $(Y_1, \cdots Y_p) \in supp(f)$

$$f(X_1, \cdots, X_p) \propto \prod_{i=1}^{P} \frac{f_{X_i}|X_{-i}(X_i|X_1, \cdots X_{i-1}, Y_{i+1}, \cdots Y_p)}{f_{X_i}|X_{-i}(Y_i|X_1, \cdots X_{i-1}, Y_{i+1}, \cdots Y_p)} \tag{7}$$

Proof.

$$f(X_1, \cdots, X_{p-1}, X_p) = f_{X_p}(X_p|X_1, \cdots X_{p-1}) f(X_1, \cdots X_{p-1}) \tag{8}$$

Similarly we have that

$$f(X_1, \cdots, X_{p-1}, Y_p) = f_{X_p}(Y_p|X_1, \cdots X_{p-1}) f(X_1, \cdots X_{p-1}) \tag{9}$$

By equation (5) we can write the following.

$$f(X_1, \cdots, X_p) = f(X_1, \cdots, X_{p-1}) f_{X_p}(X_p|X_1, \cdots X_{p-1}) \tag{10}$$

Rewrite the joint distribtion on the RHS using equation (6).

$$= f(X_1, \cdots, X_{p-1}, Y_p) \frac{f_{X_p}(X_p|X_1, \cdots X_{p-1})}{f_{X_p}(Y_p|X_1, \cdots X_{p-1})} \tag{11}$$

Repeating this process of rewriting the joint in terms of conditionals:

$$= \cdots \tag{12}$$

$$= f(X_1, \cdots, X_p) \frac{f_{X_p}(X_p|Y_1, \cdots Y_p)}{f_{X_p}(Y_p|Y_1, \cdots Y_p)} \cdots \frac{f_{X_p}(X_p|X_1, \cdots X_{p-1})}{f_{X_p}(Y_p|X_1, \cdots X_{p-1})} \tag{13}$$

□

In order to make use of Rao-Blackwellization, we require the joint distribution to be uniquely specified by conditional distributions. Recall that this is a property of the LDA graphical model by construction.

Lemma 4.2. *The transition probability matrix of the Gibbs sampler is*

$$\begin{aligned} K(x^{(t-1)}, x^{(t)}) = f_{X_1|X_{-1}}(x_1|x_1^{t-1}, \cdots x_p^{t-1}) \times f_{X_2|X_{-2}}(x_2^t|x_1^t, x_3^{t-1} \cdots x_p^{t-1}) \\ \times \cdots \times f_{X_p|X_{-p}}(x_p^t|x_1^t, x_{p_1}^t, \cdots x_p^{t-1}) \end{aligned} \tag{14}$$

Proof.

$$P(X^t \in \mathcal{X}|X^{t-1} = x^{t-1}) = \int_{\mathcal{X}} f_{X_t|X_{t-1}}(x^t|x^{t-1}) dx^t \tag{15}$$

$$\begin{aligned} = \int_{\mathcal{X}} f_{X_1|X_{-1}}(x_1|x_1^{t-1}, \cdots x_p^{t-1}) \times f_{X_2|X_{-2}}(x_2^t|x_1^t, x_3^{t-1} \cdots x_p^{t-1}) \\ \times \cdots \times f_{X_p|X_{-p}}(x_p^t|x_1^t, x_{p_1}^t, \cdots x_p^{t-1}) dx^t \end{aligned} \tag{16}$$

□

There are two rules that are required to create an MCMC sampler, one to propose how to move from x to y, and another rule to accept that move as the new position. In the context of the Gibbs sampler, suppose the stationary distribution $X = (X_1, \ldots, X_d)$ is

$$\pi(X_1, \ldots, X_d) = g(X_1, X_2, \ldots, X_d)/\kappa \tag{17}$$

where κ is a normalizing constant. Given the current value $\mathbf{x} = (x_1, x_2, \ldots, x_d)$, for the next move we follow the steps below.

1. Assign $J \sim U(1, 2, \ldots, n)$.
2. Sample X_j with the proposal distribution $\pi(\tilde{x}_J | x_1, \ldots, x_{J-1}, x_{J+1}, \ldots, x_d)$ where $\tilde{x}_J$ is our proposed move from the first rule.

Unlike some MCMC samplers, the Gibbs sampler accepts this next move with probability one. The conditional distribution used for Gibbs sampling is given below and derived in the Appendix:

$$p(z_{a,b} | z_{-(a,b)}, y, \alpha, \beta) = \frac{(\alpha_{z_{a,b}} + c + {}^{-(a,b)}_{z_{a,b},a,*}) \times (\beta_{y_{a,b}} + c^{-(a,b)}_{z_{a,b},a,*}) \big/ (J \times \beta_j + c^{-(a,b)}_{z_{a,b},*,*})}{\sum_{k=1}^{K} (\alpha_{z_{a,b}} + c + {}^{-(a,b)}_{z_{a,b},a,*}) \times (\beta_{y_{a,b}} + c^{-(a,b)}_{z_{a,b},a,*}) \big/ (J \times \beta_j + c^{-(a,b)}_{z_{a,b},*,*})} \tag{18}$$

5 Implementation and Experimental Results

need to discuss our implementation and results...

describe functions for sampling from multinomial and dirichlet distributions, and the actual gibbs sampler function that builds a chain by incrementing/decrementing each of the hidden variables conditioned on all of the observed and remaining hidden variables

6 Open Questions

outline a few open questions / areas of active research wrt Gibbs sampler

Rates of convergence between $K^n(x, y) \to \pi$ usually measured by total variation distance

$$\|K_x^n - \pi\|_{TV} = \frac{1}{2} \sum_y |K^n(x, y) - \pi(y)| = \max_{A \subseteq \mathcal{X}} |K^n(x, A) - \pi(A)| \tag{19}$$

... can we derive eigenvalue bounds on rates of convergence by using a reversible chain Gibbs sampler (described by Liu 1995)?

1. Sample an index j uniformly form the distribution on $\{1, \cdots, p\}$
2. Sample $X_j^{(t)} \sim f_{X_j | X_{-j}}(\cdot | X_1^{(t-1)}, \cdots, X_{j-1}^{(t-1)}, X_{j+1}^{(t-1)}, \cdot X_p^{(t-1)})$ setting $X_i^{(t)} = X_i^{(t-1)}$ $\forall i \neq j$.

Algorithm 1 Topic Model Procedure

```
                                                          ▷ Define Variables
integer :: m,n, top_size, Z, gg, kk, i, j, ll, nn
integer, dimension(n,m) :: matrix
integer, dimension(ntopics,n) :: NZW
integer, dimension(ntopics,m) :: NZM
integer, dimension(ntopics) :: NZ
integer, dimension(m) :: NM
integer, intent(in) :: ntopics, max_iter
integer, dimension(top_size) :: topics
integer, dimension(ntopics) :: genZ
real, dimension(ntopics) :: p_z
real,intent(in) :: alpha, beta
procedure COLLAPSED GIBBS SAMPLER

    for i in 1:iterations do
        kk = 1

        for j in 1:len(Documents) do

            for ll in 1:len(words) do
                if matrix(ll, j) == 0 then Next

                for nn in 1:matrix(ll,j) do
                    Z = topics(kk)
                                    ▷ Decrement DocumentTopic and WordTopic matrices
                    NZM(Z,j) = NZM(Z,j) - 1
                    NM(j) = NM(j) - 1
                    NZW(Z,ll) = NZW(Z,ll) - 1
                    NZ(Z) = NZ(Z) - 1

                    vocab = count which(matrix(:,j) != 0)

                    p_z = (NZM(:,j) + alpha) × (NZW(ll,:) + beta) / (NZ + vsize × beta)

                    p_z = p_z / sum(p_z)
                                    ▷ Generate Realization from Multinomial Distribution
                    call GENMUL(1,p_z,ntopics, genZ)
                    Z = which(genZ == 1)
                    topics(kk) = Z
                    kk = kk + 1
                                    ▷ Increment DocumentTopic and WordTopic matrices
                    NZM(Z,j) = NZM(Z,j) + 1
                    NM(j) = NM(j) + 1
                    NZW(Z,ll) = NZW(Z,ll) + 1
                    NZ(Z) = NZ(Z) + 1
```

7 Appendix

7.1 Integrating out Multinomial Parameters

$$\int\int p(\theta|\alpha)p(z|\theta)p(w|z,\phi)p(\phi|\beta)d\theta d\phi \tag{20}$$

The joint distribution factorizes based on dependencies specified by the graphical model.

$$= \int p(\theta|\alpha)p(z|\theta)d\theta \int p(w|z,\phi)p(\phi|\beta)d\phi \tag{21}$$

Take the product over documents and topics and replace the conditional probabilities with the Multinomial and Dirichlet distributions.

$$= \prod_{m=1}^{M} \int \frac{\Gamma(\sum \alpha_k)}{\prod \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta^{\alpha_k - 1} \prod_{n=1}^{Nm} \theta^{c_{k,m,*}} d\theta \prod_{k=1}^{K} \int \frac{\Gamma(\sum \beta_j)}{\prod \Gamma(\beta_j)} \prod_{j=1}^{J} \phi^{\beta_j - 1} \prod_{m=1}^{M} \prod_{n=1}^{Nm} \phi^{c_{k,*,j}} d\phi \tag{22}$$

Rewrite the above exploiting the congucacy between prior and posterior.

$$= \prod_{m=1}^{M} \int \frac{\Gamma(\sum \alpha_k)}{\prod \Gamma(\alpha_k)} \prod_{k=1}^{K} \theta^{\alpha_k + c_{k,m,*} - 1} d\theta \prod_{k=1}^{K} \int \frac{\Gamma(\sum \beta_j)}{\prod \Gamma(\beta_j)} \prod_{j=1}^{J} \phi^{\beta_j + c_{k,*,j} - 1} d\phi \tag{23}$$

Note that both integrals are unnormalized Dirichlets. Multiply by the normalization constant and its reciprocal to cancel both integrals as they evaluate to 1 by definition.

$$= \prod_{m=1}^{M} \frac{\Gamma(\sum \alpha_k)}{\prod \Gamma(\alpha_k)} \frac{\prod \Gamma(\alpha_k + c_{k,m,*})}{\Gamma(\sum \alpha_k + c_{k,m,*})} \prod_{k=1}^{K} \frac{\Gamma(\sum \beta_j)}{\prod \Gamma(\beta_j)} \frac{\prod \Gamma(\beta_j + c_{k,*,j})}{\Gamma(\sum \beta_j + c_{k,*,j})} \tag{24}$$

Drop terms that do not depend on the distribution over counts $c_{k,m,j}$

$$\propto \prod_{m=1}^{M} \frac{\prod \Gamma(\alpha_k + c_{k,m,*})}{\Gamma(\sum \alpha_k + c_{k,m,*})} \prod_{k=1}^{K} \frac{\prod \Gamma(\beta_j + c_{k,*,j})}{\Gamma(\sum \beta_j + c_{k,*,j})} \tag{25}$$

Split the product over documents and topics including/excluding the (a,b)th component.

$$= \prod_{m \neq a} \frac{\prod \Gamma(\alpha_k + c_{k,m,*})}{\Gamma(\sum \alpha_k + c_{k,m,*})} \frac{\prod \Gamma(\alpha_k + c_{k,a,*})}{\Gamma(\sum \alpha_k + c_{k,a,*})} \times \prod_{k=1}^{k} \frac{\prod_{j \neq y_{a,b}} \Gamma(\beta_j + c_{k,*,j}) \Gamma(\beta_{y_{ab}} + c_{k,*,y_{a,b}})}{\Gamma(\sum \beta_j + c_{k,*,j})} \tag{26}$$

Drop terms that do not depend on current position (a,b)

$$\propto \frac{\prod_{k=1}^{K} \Gamma(\alpha_k + c_{k,a,*})}{\Gamma(\sum_{k=1}^{K} \alpha_k + c_{k,a,*})} \prod_{k=1}^{K} \frac{\Gamma(\beta_{y_{ab}} + c_{k,*,y_{ab}})}{\Gamma(\sum \beta_j + c_{k,*,j})} \tag{27}$$

Split the product over topics including/excluding the $c^{a,b}$th component

$$\propto \frac{\prod_{k\neq z_{ab}} \Gamma(\alpha_k + c^{-(a,b)}_{k,a,*}) \times \Gamma(\alpha_{z_{a,b}} + c^{-(a,b)}_{z_{a,b},a,*})}{\Gamma(1 + \sum \alpha_k + c^{-(a,b)}_{k,a,*})} \tag{28}$$
$$\times \prod_{k\neq z_{a,b}}^{K} \frac{\Gamma(\beta_{y_{a,b}} + c^{-(a,b)}_{k,*,y_{a,b}})}{\Gamma(\sum \beta_j + c_{k,*,j})} \frac{\Gamma(\beta_{y_{a,b}} + c^{-(a,b)}_{z_{a,b},*,y_{a,b}} + 1)}{\Gamma(1 + \sum \beta_j + c^{-(a,b)}_{z_{a,b},*,j})}$$

Use the fact that that $\Gamma(x+1) = x\Gamma(x)$

$$= \frac{\prod_{k\neq z_{a,b}} \Gamma(\alpha_k + c^{-(a,b)}_{k,a,*})\Gamma(\alpha_{z_{a,b}} + c^{-(a,b)}_{z_{a,b},a,*})(\alpha_{z_{a,b}} + c^{-(a,b)}_{z_{a,b},a,*})}{\Gamma(1 + \sum \alpha_k + c^{-(a,b)}_{k,a,*})} \tag{29}$$
$$\times \prod_{k\neq z_{a,b}} \frac{\Gamma(\beta_{y_{a,b}} + c^{-(a,b)}_{k,*,y_{a,b}})}{\Gamma(\sum \beta_j + c_{k,*,j})} \frac{\Gamma(\beta_{y_{a,b}} + c^{-(a,b)}_{z_{a,b},*,y_{a,b}})}{\Gamma(\sum \beta_j + c^{-(a,b)}_{z_{a,b},*,j})} \frac{(\beta_{y_{a,b}} + c^{-(a,b)}_{z_{a,b},*,y_{a,b}})}{\sum(\beta_j + c^{-(a,b)}_{z_{a,b},*,j})}$$

Combine the Γ terms back into the product:

$$= \frac{\prod_{k=1}^{K} \Gamma(\alpha_k + c^{-(a,b)}_{k,a,*})(\alpha_{z_{a,b}} + c^{-(a,b)}_{z_{a,b},*,y_{a,b}})}{\Gamma(1 + \sum \alpha_k + c^{-(a,b)}_{k,a,*})} \tag{30}$$
$$\times \prod_{k=1}^{K} \frac{\Gamma(\beta_{y_{a,b}} + c^{-(a,b)}_{k,*,y_{a,b}})}{\Gamma(\sum \beta_j + c^{-(a,b)}_{z_{a,b},*,j})} \frac{(\beta_{y_{a,b}} + c^{-(a,b)}_{z_{a,b},*,y_{a,b}})}{\sum(\beta_j + c^{-(a,b)}_{z_{a,b},*,j})}$$

Drop the topic denominator and general product terms which are constant:

$$\propto \frac{(\alpha_{z_{a,b}} + c+^{-(a,b)}_{z_{a,b},a,*}) \times (\beta_{y_{a,b}} + c^{-(a,b)}_{z_{a,b},a,*})}{\sum(\beta_j + c^{-(a,b)}_{z_{a,b},*,j})} = \frac{(\alpha_{z_{a,b}} + c+^{-(a,b)}_{z_{a,b},a,*}) \times (\beta_{y_{a,b}} + c^{-(a,b)}_{z_{a,b},a,*})}{\sum(\beta_j) + c^{-(a,b)}_{z_{a,b},*,*}} \tag{31}$$

The conditional distribution used for Gibbs sampling is given by normalizing the result above:

$$p(z_{a,b}|z_{-(a,b)}, y, \alpha, \beta) = \frac{(\alpha_{z_{a,b}} + c+^{-(a,b)}_{z_{a,b},a,*}) \times (\beta_{y_{a,b}} + c^{-(a,b)}_{z_{a,b},a,*}) \big/ (J \times \beta_j + c^{-(a,b)}_{z_{a,b},*,*})}{\sum_{k=1}^{K} (\alpha_{z_{a,b}} + c+^{-(a,b)}_{z_{a,b},a,*}) \times (\beta_{y_{a,b}} + c^{-(a,b)}_{z_{a,b},a,*}) \big/ (J \times \beta_j + c^{-(a,b)}_{z_{a,b},*,*})} \tag{32}$$

7.2 The Gamma, Beta and Dirichlet Function Redux

Euler would introduce the useful function below.

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt = 2 \int_0^\infty u^{2x-1} e^{-u^2} du$$

Substituting $t = u^2$ and evaluating $x = 1/2$ gives the Euler-Poisson integral.

$$\Gamma(1/2) = 2 \int_0^\infty e^{-u^2} du = \sqrt{\pi}$$

Using integration by parts it is easy to see $\Gamma(x+1) = x\Gamma(x)$ and thus $\Gamma(x+1) = x!$ when x is an integer.

$$\Gamma(n) = \int_0^\infty e^{-x}x^{n-1}dx$$

$$\Gamma(n+1) = \int_0^\infty e^{-x}x^{n}dx$$

Let $u = x^n, du = nx^{n-1}$, $dv = e^{-x}$ and $v = -e^{-x}$

$$= -x^n e^{-x}\Big|_0^\infty - \int_0^\infty -e^{-x}nx^{n-1}dx$$

$$= n\int_0^\infty e^{-x}x^{n-1}dx = n\Gamma(n)$$

Euler found that by using the Gamma function we can define the Beta function:

$$\frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)} = B(x,y) = \int_0^1 t^{a-1}(1-t)^{b-1}dt$$

The beta is constructed using Gamma functions of two variables. The Gamma function can be used to derive the Beta. Perform the substitution $t = sin^2(\theta)$:

$$B(x,y) = \int_0^1 t^{x-1}(1-t)^{y-1}dt$$

$$= \int_0^{\pi/2} sin(t)^{2x-1}cos(t)^{2y-1}dt = B(y,x)$$

$$\Gamma(x)\Gamma(y) = 4\int_0^\infty u^{2x-1}e^{-u^2}du \int_0^\infty v^{2y-1}e^{-v^2}dv$$

$$= 4\int_0^\infty\int_0^\infty e^{-(u^2+v^2)}u^{2x-1}v^{2y-1}dudv$$

Let $u = r\ cos(\theta)$ and $v = r\ sin(\theta)$:

$$
\begin{aligned}
&= 4 \int_0^{\infty} \int_0^{\pi/2} e^{-r^2} r^{2(x+y)-1} cos^{2x-1}(\theta) sin^{2y-1}(\theta) dr d\theta \\
&= 2 \int_0^{\infty} r^{2(x+y)-1} e^{-r^2} dr \times 2 \int_0^{\pi/2} cos^{2x-1}(\theta) sin^{2y-1}(\theta) d\theta \\
\Gamma(x)\Gamma(y) &= \Gamma(x+y) B(y,x) \\
\frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)} &= B(y,x)
\end{aligned}
$$

□

Recall that the Dirichlet is constructed from Gamma functions of d variables:

$$p(x|\alpha) = \frac{\Gamma(\sum_i \alpha_i)}{\prod_i \Gamma(\alpha_i)} \prod_{i=1}^{d} x_i^{\alpha_i - 1}$$

The Dirichlet distribution is the conjugate prior to the Multinomial. Note the similar functional form.

$$
\begin{aligned}
&posterior \propto likelihood \times prior \\
&= \frac{\Gamma(n+1)}{\prod \Gamma(x_i+1)} \prod_{i=1}^{K} p_i^{x_i} \frac{\Gamma(\sum \alpha_i)}{\prod \Gamma(\alpha_i)} \prod x_i^{\alpha_i - 1} \\
&\propto \prod_{y_i} \prod_{j=1}^{K} p_j^{y_i^{(j)}} \prod_{j=1}^{K} p_j^{\alpha_j - 1} \\
&= \prod_{j=1}^{K} p_j^{\alpha_j - 1 + \sum y_i^{(j)}} \\
&= Dir(a_j + \sum y_i)
\end{aligned}
$$

The posterior is again a Dirichlet random variable whose new parameters are increased by the sum of the multinomial parameters.