Mixtures of Stick–Breaking Processes

Maria Kalli & Stephen G. Walker [1]

Abstract

We consider mixtures of stick–breaking processes as a generalization of the mixture of Dirichlet process model. We provide a sampling algorithm which covers all such models provide specific reasons for using particular choices of prior. Numerical illustrations involving real data sets are presented.

Keywords: Dirichlet process; Stick–breaking process; Markov chain Monte Carlo; Mixture model.

1. INTRODUCTION.

The well known and widely used mixture of Dirichlet process (MDP) model was first introduced by Lo (1984). Since the advent of Markov chain Monte Carlo methods within the mainstream statistics literature (Smith and Roberts, 1993), and the specific application to the MDP model (Escobar, 1988; Escobar, 1994; Escobar and West, 1995), the model has become one of the most popular in Bayesian nonparametrics.

Variations of the original algorithm of Escobar have been numerous; for example, MacEachern (1994); Müller and MacEachern (1998); and Neal (2000). All of these algorithms rely on integrating out the random distribution function from the model, removing the infinite dimensional problem. Recent ideas have left the infinite dimensional distribution in the model and found ways of sampling a sufficient but finite number of variables at each iteration of a Markov chain with the correct stationary distribution. See Papaspiliopoulos and Roberts (2008), Walker (2007), and Kalli, Griffin and Walker (2008).

In this paper we consider mixtures of stick–breaking processes and establish reasons for selecting a particular type of process. This shift away from the Dirichlet process has been impossible in the past due to the lack of sampling algorithms which

[1]Maria Kalli is Lecturer, Center of Health Services Studies, University of Kent, Canterbury, U. K. (email: M.Kalli@kent.ac.uk), and Stephen G. Walker is Professor of Statistics, Institute of Mathematics, Statistics & Actuarial Science, University of Kent, Canterbury, U. K. (email: S.G.Walker@kent.ac.uk).

can deal with general stick–breaking processes. However, the algorithm described is able to handle the more general set–up.

The lay–out of the paper is as follows. In Section 2 we describe some preliminaries to help with later sections. In Section 3 we provide details of our ideas for choosing particular stick–breaking processes and Section 4 describes the sampling algorithm for mixture of stick–breaking process models. Section 5 contains numerical illustrations and, finally, in Section 6, we provide a brief discussion.

2. PRELIMINARIES.

We write $P \sim D(\vartheta, P_0)$ to denote that P is a Dirichlet process (Ferguson, 1973) with parameters $\vartheta > 0$, the scale parameter, and P_0, a distribution on the real line. It is well known that P has a stick–breaking representation (Sethuraman, 1994) given by

$$P = \sum_{j=1}^{\infty} w_j \, \delta_{\theta_j},$$

where the $\{\theta_j\}$ are independent and identically distributed from P_0 and

$$w_1 = v_1, \;\; w_j = v_j \prod_{l<j} (1 - v_l)$$

with the $\{v_j\}$ being independent and identically distributed from beta$(1, \vartheta)$.

The MDP model, with kernel $K(y; \theta)$, is given by

$$f_P(y) = \int K(y; \theta) \, \mathrm{d}P(\theta)$$

with $P \sim D(\vartheta, P_0)$. It is possible to remove P from this model via simple integration and so the stick–breaking representation of P is not used in this case. However, the stick–breaking representation is essential to estimation via the non–marginal models of Papaspiliopoulos and Roberts (2008) and Walker (2007). The idea is that we can write

$$f_{v,\theta}(y) = \sum_{j=1}^{\infty} w_j \, K(y; \theta_j)$$

and the key is to find exactly which (finite number of) variables need to be sampled to produce a valid Markov chain with correct stationary distribution.

As a prior for a distribution function, the Dirichlet process has a number of attractive properties which make it highly suitable for the mixture model. The key

property is the connection with the Pólya–urn scheme (Blackwell and MacQueen, 1973). However, none of these "nice" mathematical properties are to do with the model and the suitability for using it as a prior for P. The weights of the mixture; that is, the $\{w_j\}$ are quite simplistic, and has a distribution which depends on the single parameter ϑ. This can easily be generalized to $v_j \sim \text{beta}(a_j, b_j)$.

The reason why little progress has been made beyond the Dirichlet process is due to the absence of a sampling algorithm. It is not possible to integrate out the random distribution function and hence it is necessary to use one of the non–marginal algorithms. The algorithm described in Walker (2007) is actually no more complicated for general stick–breaking processes than it is for the Dirichlet process.

The details are given in Walker (2007) but we briefly describe the basis for the algorithm. The joint density

$$f_{v,\theta}(y,u) = \sum_{j=1}^{\infty} \mathbf{1}(u < w_j)\, K(y;\theta_j)$$

is the starting point. Given u, the number of mixtures is finite, the indices being $A_u = \{k : w_k > u\}$. One has

$$f_{v,\theta}(y|u) = N_u^{-1} \sum_{j \in A_u} K(y;\theta_j),$$

and the size of A_u is $N_u = \sum_{j=1}^{\infty} \mathbf{1}(w_j > u)$.

One can then introduce a further latent variable which indicates which of these finite number of mixtures provides the observation to give the joint density

$$f_{v,\theta}(y,u,d) = \mathbf{1}(u < w_d)\, K(y;\theta_d).$$

Hence, a complete likelihood function for (v,θ) is available as a simple product of terms and, crucially, the choice of d is finite. Without the u the choice would be infinite and so would lead to difficulties in the implementation of a Markov chain Monte Carlo algorithm.

Now that it is possible to work with general stick–breaking processes it is incumbent to understand how to choose the distribution of the $\{v_j\}$. We want to use $v_j \sim \text{beta}(a_j, b_j)$ independently and so it is a matter of selecting the $\{a_j, b_j\}$. For P to be a proper random distribution function it is sufficient that

$$\sum_{j=1}^{\infty} \log(1 + a_j/b_j) = +\infty;$$

see Ishwaran and James (2001).

3. PRIOR SETTING.

It is a difficult task to know exactly how to determine values for the $\{a_j, b_j\}$. Our idea is to understand the weights via a Bayesian *parametric* model; say $\tilde{w}_j(\phi)$ and with a probability on ϕ, say $\pi(\phi)$. We then compute

$$\xi_j = \mathrm{E}(\tilde{w}_j(\phi)) = \int \tilde{w}_j(\phi)\, \pi(\mathrm{d}\phi).$$

We will now ensure that our weights for the mixture model are such that $\mathrm{E}(w_j) = \mathrm{E}(\tilde{w}_j)$ hence providing a motivation for the selection of the $\{a_j, b_j\}$. If we write $\tau_j = a_j/(a_j + b_j)$ then we require the $\{\tau_j\}$ to satisfy $\tau_1 = \xi_1$ and for $j > 1$

$$\tau_j \prod_{l<j} (1 - \tau_l) = \xi_j.$$

The result is that we set, for $j > 1$,

$$\tau_j = \left(1 - \sum_{l<j} \xi_l\right)^{-1} \xi_j.$$

It is clear that $\tau_j < 1$ since $\xi_j < 1 - \sum_{l<j} \xi_l$ which follows since $\sum_l \xi_l = 1$. It is easy to check that in this way we satisfy the condition

$$\sum_{j=1}^{\infty} \log(1 + a_j/b_j) = +\infty.$$

For $a_j/b_j = \tau_j/(1 - \tau_j)$ and $\log(1 + \tau_j/(1 - \tau_j)) = -\log(1 - \tau_j)$ so

$$\sum_{j=1}^{M} -\log(1 - \tau_j) = -\log \prod_{j=1}^{M} (1 - \tau_j) = -\log\left(1 - \sum_{j=1}^{M} \xi_j\right) \to +\infty.$$

Re–parameterizing $a_j = c_j \tau_j$ and $b_j = c_j(1 - \tau_j)$, we need to be able to specify the $\{c_j\}$. We look at $\mathrm{E}(w_j^2)$ and this is given, after some routine algebra, by $(\mathrm{E}w_j^2) = \xi_j\, H_j$, where

$$H_j = \frac{1 + c_j \tau_j}{1 + c_j} \prod_{l<j} \left(1 - \frac{c_l \tau_l}{1 + c_l}\right).$$

Hence, writing $c_l/(1 + c_l) = q_l$, we have

$$\mathrm{Var}(w_j) = \xi_j \left[(1 - q_j) + q_j \tau_j\right] \prod_{l<j} (1 - q_l \tau_l) - \xi_j^2.$$

This gives us the means to control the variance of the $\{w_j\}$ and hence determine how close in probability the weights are to the $\{\xi_j\}$. For a particular choice of variances, we can select the $\{q_j\}$ to satisfy

$$q_j = (1-\tau_j)^{-1}\left[1 - \frac{\mathrm{Var}(w_j) + \xi_j^2}{\xi_j \prod_{l<j}(1-q_l\tau_l)}\right].$$

It is not possible to take arbitrary parametric $V_j = \mathrm{Var}(w_j(\phi))$ in order to establish the $\{q_j\}$, since there are some constraints. Clearly, since $0 < q_j < 1$ we need

$$0 < 1 - \frac{V_j + \xi_j^2}{\xi_j \prod_{l<j}(1-q_l\tau_l)} < 1 - \tau_j$$

and hence we need

$$\tau_j C_j < V_j < C_j,$$

where $C_j = \xi_j \prod_{l<j}(1-q_l\tau_l) - \xi_j^2$.

One particular idea, which we shall rely on for the numerical illustrations, is to take large variances in an attempt to be non–informative. This amounts to choosing q_j to be small, but not zero, and hence we take $c_j = \epsilon$, for some small ϵ, for all j. This follows since $\mathrm{Var}(w_j) < \xi_j(1-\xi_j)$ and we obtain this limit as $q_j \downarrow 0$.

Example 1. The example we will consider first is a geometric–beta model where $\tilde{w}_j(\phi) = \phi(1-\phi)^{j-1}$, so for $\phi \sim \mathrm{beta}(a,b)$ we have

$$\xi_j = \mathrm{E}(\tilde{w}_j) = \frac{\Gamma(a+b)\,\Gamma(a+1)\,\Gamma(b+j-1)}{\Gamma(a)\,\Gamma(b)\,\Gamma(a+b+j)}.$$

In the special case that $a = b = 1$ then $\xi_j = 1/[j(j+1)]$. This is an interesting example, since we match with the Dirichlet process expected weights, which are

$$\mathrm{E}(w_j) = \rho(1-\rho)^{j-1},$$

with $\rho = 1/(1+\vartheta)$, when we take $\phi = \rho$ a.s.

Example 2. Here we consider a Poisson–gamma model so that, for $j \geq 1$,

$$\tilde{w}_j(\phi) = \frac{\phi^{j-1}}{(j-1)!}\, e^{-\phi}$$

and $\phi \sim \text{Ga}(a, b)$. Hence,

$$\xi_j = \text{E}(\tilde{w}_j) = \frac{b^a}{\Gamma(a)\,(j-1)!} \int \phi^{j-1} e^{-\phi} \phi^{a-1}\, e^{-b\phi}\, \text{d}\phi = \frac{b^a\,\Gamma(a+j-1)}{(j-1)!\,\Gamma(a)\,(b+1)^{a+j-1}}.$$

In the special case that $a = b = 1$ we have $\xi_j = 2^{-j}$ and then it is easy to see that $\tau_j = 1/2$ for all j. Hence, taking $c_j = \epsilon$ for all j, this results in $a_j = b_j = \epsilon/2$, which fits neither the Dirichlet process nor the two–parameter Pitman–Yor process (Pitman and Yor, 1996), which has $a_j = 1 - \sigma$ and $b_j = \vartheta + j\sigma$ for some $\sigma < 1$.

Example 3. Another possibility is to match the expected weights with those that are commonly employed in a normal mixture model: that is given a random integer $M \geq 1$, the mixture model has M components with weights $\{\tilde{w}_{1M}, \ldots, \tilde{w}_{MM}\}$. Hence,

$$\xi_j = \sum_{M \geq j} \text{E}(\tilde{w}_{jM})\, P(M).$$

For example, if $\text{E}(\tilde{w}_{jM}) = 1/M$ and $P(M) = e^{-\psi}\psi^{M-1}/(M-1)!$ then

$$\xi_j = \psi^{-1}\, e^{-\psi} \sum_{M \geq j} \psi^M / M! = \psi^{-1} S(j; \psi),$$

where $S(j; \psi)$ is the probability a Poisson random variable with parameter ψ is greater than or equal to j.

4. SIMULATION ALGORITHM

We briefly describe the simulation algorithm, but only provide the sampling procedure without derivation since this has appeared elsewhere (Kalli, Griffin and Walker 2008). We sample one of the full conditionals in a different and more efficient manner than that in Walker (2007). We sample $\pi(v, u| \cdots)$ as a block and this involves sampling $\pi(v| \cdots \text{ exclude } u)$ and then $\pi(u|v, \cdots)$, where $\pi(v| \cdots \text{ exclude } u)$ is obtained by integrating out u from $\pi(v, u| \cdots)$. The distribution $\pi(v| \cdots \text{ exclude } u)$ will be the standard full conditional for a stick-breaking process (see Ishwaran and James (2001)). Standard MCMC theory on blocking suggests that this should lead to a more efficient sampler.

Recall that we have the model

$$f(y) = \sum_{j=1}^{\infty} w_j\, K(y; \theta_j),$$

where the θ_j are independent and identically distributed from p_0, the $\{w_j\}$ have a stick–breaking process based on the Dirichlet process, described in Section 2.

The variables that need to be sampled at each sweep of a Gibbs sampler are

$$\{(\theta_j, v_j), j = 1, 2, \ldots; (d_i, u_i), i = 1, \ldots, n\}.$$

1. $\pi(\theta_j | \cdots) \propto p_0(\theta_j) \prod_{d_i=j} K(y_i; \theta_j)$.

2. $\pi(v_j | \cdots \text{ exclude } u) \propto \text{beta}(v_j; a_j, b_j)$, where

$$a_j = 1 + \sum_{i=1}^{n} \mathbf{1}(d_i = j)$$

and

$$b_j = \vartheta + \sum_{i=1}^{n} \mathbf{1}(d_i > j).$$

3. $\pi(u_i | \cdots) \propto \mathbf{1}(0 < u_i < w_{d_i})$.

4. $\text{P}(d_i = k | \cdots) \propto \mathbf{1}(k : w_k > u_i)\, K(y_i; \theta_k)$.

Obviously, we can not sample all of the (θ_j, v_j). But it is not required to in order to proceed with the chain. We only need to sample up to the integer N for which we have found all the appropriate w_k in order to do step 4 exactly. Since the weights sum to 1 if we find N_i such that $\sum_{k=1}^{N_i} w_k > 1 - u_i$ then it is not possible for any of the w_k, for $k > N_i$, to be greater than u_i.

There are some important points to make here. First, it is a trivial extension to consider more general stick–breaking processes for which $v_j \sim \text{beta}(\alpha_j, \beta_j)$ independently. Then, in this case, we would have

$$a_j \to a_j + \sum_{i=1}^{n} \mathbf{1}(d_i = j)$$

and

$$b_j \to b_j + \sum_{i=1}^{n} \mathbf{1}(d_i > j).$$

This easy extension to more general priors is not a feature of alternative sampling algorithms. Secondly, the algorithm is remarkably simple to implement; all full conditionals are standard.

5. NUMERICAL ILLUSTRATIONS.

First, in Section 5.1, we will look at the effect different choices of (a, b, ϵ), for the Poisson–gamma and geometric–beta models, have on the mean $E(w_j)$ and variance $\text{Var}(w_j)$ of the stick-breaking weights.

Then, in Section 5.2, we consider the effect of (a, b, ϵ) on the density estimate of $f(y)$. In these illustration, plots of the predictive density $f(y)$ are provided. The data sets we chose for this study are:

(1) the galaxy data set, which consists of the velocities of 82 distant galaxies diverging from our own galaxy. This is the most commonly used data set in density estimation studies see Escobar and West (1995) and Green and Richardson (2001).

(2) the S&P 500 index daily returns data set (as described in Section 5.2)

We chose the galaxy data set as it has a small sample size yet is multi–modal. The S&P 500 data set, on the other hand, is a large data set and is uni–modal; thus it would be interesting to see how (a, b, ϵ) effect density estimation in these cases. We will produce plots of the predictive density for both data sets and for the S&P 500 data set we provide additional tables of the median, skewness and kurtosis of the predictive density which we compare with the values of the empirical distribution.

For the analysis of both data sets we use the normal kernel $K(y|\theta)$ with components $\theta = (\mu, \lambda)$, and $P_0(\mu, \lambda) = \text{N}(\mu|\nu, \xi^2) \times \text{G}(\lambda|\gamma, \beta)$; where $\text{G}(\gamma, \beta)$ denotes the gamma distribution. Hence, for the parameters of our MDP mixture we take $\nu = \text{median}(y)$, $\xi = R$, $\gamma = 2$, and $\beta = 1$ and $R =$ range of the data, which is similar to the Escobar and West (1995) choice. We use the geometric-beta model setting exclusively in this sub–section.

5.1 Prior means and variance of weights

Figure 1 displays the effect of changes in b under the geometric-beta prior setting, while (a, ϵ) are kept constant. We begin with $b = 2$ and double it each time to 4, 8, and 16. In the case when $a = 2$ and $\epsilon = 2$, we observe that the smaller the value of b, the higher the starting values of $E(w_j)$ and $\text{Var}(w_j)$. Also, at smaller values of b the decay in both mean and variance of the stick-breaking weights is much sharper

than at higher values. This means that we have more significant weights the larger the value of b is. Once we increase a to 8, the only change we observe is the increase in the starting value of $E(w_j)$; the change in $\text{Var}(w_j$ is marginal, and the decay is exactly the same as in the case of $a = 2$ and $\epsilon = 2$.

Figure 2 displays the effect of changes in b, under the Poisson-gamma prior setting, while (a, ϵ) are kept constant. Concentrating on $E(w_j)$, we see that the effect of an increase in b has the opposite effect to that of the geometric-beta prior setting; that is the starting value of $E(w_j)$ increases as b increases. The decay is also much sharper the bigger b gets, which leads to the conclusion that in the Poisson-gamma case we get more significant weights in terms of $E(w_j)$ when b is kept small. The same is valid when we increase a to 8. What we would like to comment on in this case is the change in behavior of the $\{E(w_j)\}$. They do not always decay as was the case with $a = 2$; As b gets bigger, the first few values of the $\{E(w_j)\}$ sequence increase and then decrease. Moving to the $\{\text{Var}(w_j)\}$ sequence, the effect of an increase in b is the same regardless the choice of a. The effect is also the same as with the geometric-beta prior setting, the sequence decays much sharply the bigger b is, resulting in only the first few $\text{Var}(w_j)$ to be significant. Finally the increase of a to 8 has the same effect on $\text{Var}(w_j)$ as that of $E(w_j)$.

Figure 3 studies the effect of ϵ on $\{\text{Var}(w_j)\}$ under both prior settings. As ϵ gets larger, we see that $\text{Var}(w_j)$ gets smaller. The point to make is that under the Poisson-gamma model the decay in the variance sequence is less smooth than that of the geometric-beta model. In both cases we have more significant variance values the smaller ϵ is.

Parameter ϵ has no effect on $E(w_j)$, only on $\text{Var}(w_j)$. To study the effect in relation to changes in (a, b) we compared the first four values of the aforementioned moments when $\epsilon = 4$ and $\epsilon = 16$. The results are shown in Tables 1and 2. Form these tables we can see that as a increases and b decreases the decay in all three moments slows down. The $\text{Var}(w_j)$ increases the closer the values of these parameters are; for instance, a quick look at table 1 shows that when $(a = 2, b = 8)$, $\text{Var}(w_1) = 0.032$ and when $(a = 4, b = 6)$, $\text{Var}(w_1) = 0.048$. Keeping b constant and increasing a slows the decay; whereas keeping a constant and increasing b speeds the decay. The effect of ϵ then is rather obvious; the smaller it is the greater $\text{Var}(w_j)$.

5.2 Density estimation

Figure 4 studies the effect of changes in b, under the geometric-beta prior setting, on the density estimate of $f(y)$. The plots of the predictive density are produced for the galaxy and S&P 500 data sets. The effect of changes in b is more evident on the galaxy data set. As (a, ϵ) are kept constant while b increases, the number of modes of $\widehat{f}(y)$ decreases. We started out with 6 modes at $b = 4$ and dropped to 3 modes at $b = 20$. Clearly the value of b effects the clustering structure, as it impacts on the variability of the stick-breaking weights. The effect of b is not that evident when we look at the estimated density of the S&P 500 data set. However, in this case we are not looking at the number of modes, we are interested in the tail behavior and skewness of of $\widehat{f}(y)$. Daily stock index returns are characterized by heavy/fat tails and slight negative skewness and we would like to see if our choice of prior will result in capturing these characteristics. Figure 8 shows the effect of b by looking at the tail of $\widehat{f}(y)$ on the log scale. Again, we see that the smaller b is the more the clusters around the tail and the heavier the tail is; a look at the kurtosis estimates of Table 3 confirms this. The skewness estimate that seems closer to that of the data set is that obtained when $b = 12$. Again, we see that as b increases, the skewness estimate increases.

Figure 5 studies the effect of changes in a under the geometric-beta prior setting, on $\widehat{f}(y)$. The plots of the predictive density are produced for the galaxy and the S&P 500 data sets. From the galaxy data we can see that a has less of an effect on the number of modes of $\widehat{f}(y)$ than b. It does effect the clustering structure but not as much as b does. What is more, it has the opposite effect; that is as a increases the number of modes increases. The effect is less abrupt as we go from 4 modes when $a = 4$ to 5 modes when $a = 20$. The S&P 500 plot of 5 do not actually say a lot; however the tail- effect of a can been seen in Figure 7 where we plot the estimated density on the log scale. Clearly the clustering at the tail increases as a increases, but the tails are not as heavy as those of Figure 8, where we study the effect of increases in b. Table 4 confirms this tail observation, the kurtosis estimates decrease as a increases. Regarding the skewness estimates, those tend to oscillate from negative to positive.

Figure 6 studies the effect of changes in ϵ under the geometric-beta prior setting.

The plots of the predictive density are produced for the galaxy and the S&P 500 data sets. The effect ϵ has on $\widehat{f}(y)$ is similar to that of a, however it is more obvious. Looking at the plots of the galaxy data, we see that the number of modes increases as ϵ increases, from 3 modes to 5, but up to a point. As ϵ jumps from 10 to 20 these modes appear to decrease again from 5 to 4. For the S&P 500 data it would be best to look at Figure 9, showing the clustering effect at the tail, on log scale. We appear to have more clusters at the tail as ϵ increases; an effect similar to that of the galaxy data set. The tails are more heavy than those of Figure 7, but they appear less heavy than those of Figure 8 when in fact they are not as the kurtosis estimates in Table 5 increase as ϵ increases. This is probably due to the different normals that are formed within our mixture. The skewness increases with ϵ, and exhibits the same behavior as the modes; that is it starts to decrease at some point. Although some of these figures are not close to the statistics obtained form the data set, they hint to same conclusion; The choice of prior matters and does have an effect on the end result, therefore it is preferable to incorporate this knowledge in the prior from the start.

6. DISCUSSION.

The class of mixtures of stick–breaking processes is rich and extends beyond the well known Dirichlet process mixture model. Fortunately, we can now provide an algorithm which covers all these models and which are all one and the same. In this respect the Dirichlet case is no longer special. It does have special mathematical properties, but none of these are persuasive from a modeling perspective. Now there is the real possibility of choosing a model based on qualitative prior information and this we suggest is best done by matching the means of the random weights with those which arise from some simple, and hence well understood, parametric model. A bound for the variances has been given and hence it is possible to allow arbitrary uncertainty within the constraints. One idea is to allow for maximum variance possible, which is in keeping with the philosophy of a nonparametric model. We have shown using Poisson–gamma and geometric–beta models that a rich variety of mean weights can be obtained.

The prior has been seen to have an important effect on the results (e.g. density estimation) and so it is incumbent on Bayesian nonparametric modelers to incorporate knowledge in the right way, without recourse to restrictions (i.e. the Dirichlet model)

or guessing appropriate specifications in general stick–breaking processes.

References

Blackwell, D. and MacQueen, J.B. (1973). Ferguson distributions via Pólya–urn schemes. *Annals of Statistics* **1**, 353–355.

Escobar, M.D. 1988. Estimating the means of several normal populations by nonparametric estimation of the distribution of the means. Unpublished Ph.D. dissertation, Department of Statistics, Yale University.

Escobar, M.D. (1994). Estimating normal means with a Dirichlet process prior. *Journal of the American Statistical Association* **89**, 268–277.

Escobar, M.D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* **90**, 577–588.

Ferguson, T.S. (1973). A Bayesian analysis of some nonparametric problems. *Annals of Statistics* **1**, 209–230.

Ishwaran, H. and James, L.F. (2001). Gibbs sampling methods for stick–breaking priors. *Journal of the American Statistical Association* **96**, 161–173.

Kalli, M., Griffin, J.E., and Walker, S.G. (2008). Slice Sampling the Mixture of Dirichlet Process Model: A Comparative Study and New Ideas. *Technical Report.* University of Kent.

Lo, A.Y. (1984). On a class of Bayesian nonparametric estimates I. Density estimates. *Annals of Statistics* **12**, 351–357.

MacEachern, S.N. (1994). Estimating normal means with a conjugate style Dirichlet process prior. *Communications in Statistics: Simulation and Computation* **23**, 727–741.

MacEachern, S.N. and Müller, P. (1998). Estimating mixtures of Dirichlet process models. *Journal of Computational and Graphical Statistics* **7**, 223–238.

MacEachern, S.N. (2000). Dependent Dirichlet processes. *Technical Report.* Ohio State University.

Neal, R. (2000). Markov chain sampling methods for Dirichlet process mixture models. *Journal of Computational and Graphical Statistics* **9**, 249–265.

Papaspiliopoulos, O. and Roberts, G.O. (2008). Retrospective Markov chain Monte Carlo methods for Dirichlet process hierarchical models. To appear in *Biometrika*.

Pitman, J. and Yor, M. (1996). Random discrete distributions derived from self–similar random sets. *Electronic Journal of Probability* **1**, 1–28.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica* **4**, 639–650.

Smith, A.F.M. and Roberts, G.O. 1993. Bayesian computations via the Gibbs sampler and related Markov chain Monte Carlo methods. Journal of the Royal Statistical Society, Series B 55, 3–23.

Walker, S.G. (2007). Sampling the Dirichlet mixture model with slices. *Communications in Statistics: Simulation and Computation* **36**, 45–54.

Table 1: Effect of parameter ϵ on $E(w_j^2)$ and $Var(w_j)$

	$\epsilon = 4$		
	$(a = 2, b = 8)$	$(a = 3, b = 7)$	$(a = 4, b = 6)$
$\mathbf{E(w_j^2)}$			
$j = 1$	0.6720	0.5320	0.4080
$j = 2$	0.0409	0.0596	0.0722
$j = 3$	0.0040	0.0092	0.0159
$j = 4$	0.0005	0.0018	0.0042
$\mathbf{V(w_j)}$			
$j = 1$	0.0320	0.0420	0.0480
$j = 2$	0.0197	0.0231	0.0246
$j = 3$	0.0027	0.0052	0.0077
$j = 4$	0.0004	0.0012	0.0024

Table 2: Effect of parameter ϵ on $E(w_j^2)$ and $Var(w_j)$

	$\epsilon = 16$		
	$(a = 2, b = 8)$	$(a = 3, b = 7)$	$(a = 4, b = 6)$
$\mathbf{E(w_j^2)}$			
$j = 1$	0.6494	0.5024	0.3741
$j = 2$	0.0267	0.0429	0.0544
$j = 3$	0.0019	0.0053	0.0102
$j = 4$	0.0002	0.0009	0.0023
$\mathbf{V(w_j)}$			
$j = 1$	0.0094	0.0124	0.0141
$j = 2$	0.0055	0.0064	0.0070
$j = 3$	0.0006	0.0013	0.0019
$j = 4$	0.0001	0.0003	0.0006

Table 3: S&P 500 geometric-beta model - changes in b

	Actual	$(2, 4, 10)$	$(2, 8, 10)$	$(2, 12, 10)$	(2,20,10)
median	0.0440	0.0444	0.0419	0.0379	0.0307
st.dev	1.1303	1.5606	1.3177	1.3769	1.1801
skewness	-4.1290	7.4382	-5.0690	-4.8293	-3.6073
kurtosis	90.3710	333.9853	247.0066	186.3995	118.8424

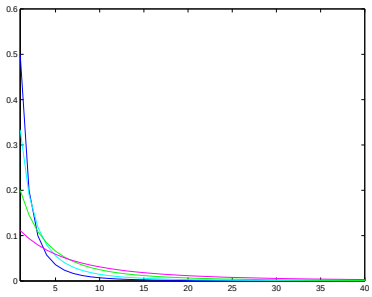
Table 4: S&P 500 geometric-beta model - changes in a

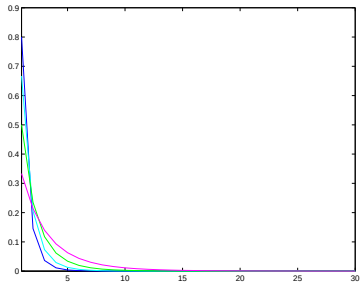
	Actual	$(4, 8, 4)$	$(8, 8, 4)$	$(12, 8, 4)$	(20,8,4)
median	0.0440	0.0502	0.0439	0.0371	0.0424
st.dev	1.1303	1.6502	1.8182	1.4871	1.7829
skewness	-4.1290	-16.8005	0.1554	-4.6809	0.6542
kurtosis	90.3710	801.0859	825.1846	306.4145	264.5234

Table 5: S&P 500 geometric-beta model - changes in ϵ

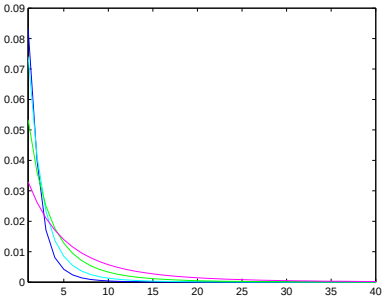
	Actual	$(4, 8, 0.1)$	$(4, 8, 1)$	$(4, 810)$	(4,8,20)
median	0.0440	0.0392	0.0548	0.0525	0.0507
st.dev	1.1303	1.1306	1.1791	1.6506	1.3879
skewness	-4.1290	-3.9127	5.8531	9.7297	2.1671
kurtosis	90.3710	84.116	359.2197	782.5149	440.7116

Mean - $E(w_j)$





Variance - $Var(w_j)$



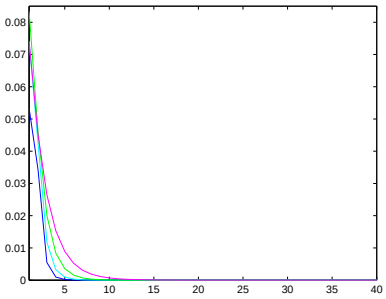
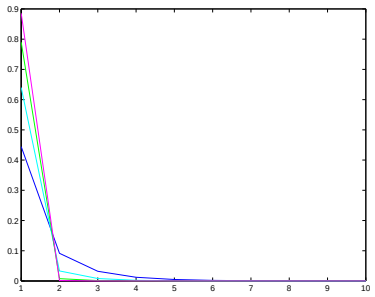
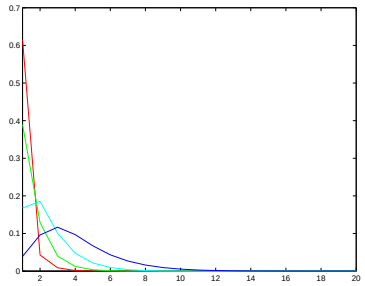


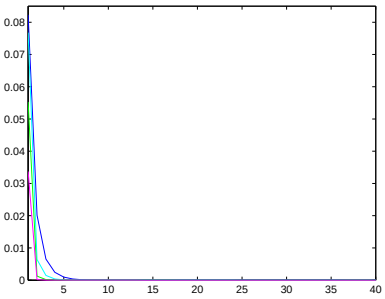
Figure 1: Effects of changes in b on $E(w_j)$ and $Var(w_j)$ under the geometric-beta prior setting. Graphs on left are for $a = 2$, $\epsilon = 2$ and graphs on right are for $a = 8$, $\epsilon = 2$. For $b = 2$ (blue), for $b = 4$ (cyan), for $b = 8$ (green) and for $b = 16$ (magenta)

Mean - $E(w_j)$





Variance - $Var(w_j)$



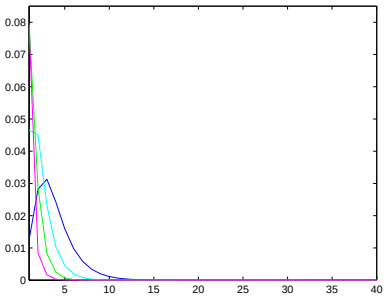
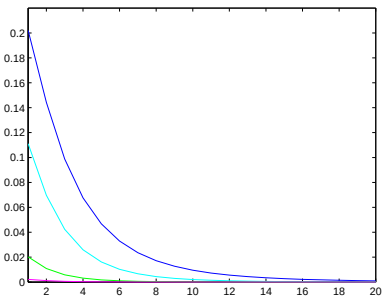


Figure 2: Effects of changes in b on $E(w_j)$ and $Var(w_j)$ under the poisson-gamma prior setting. Graphs on left are for $a = 2\,\epsilon = 2$ and graphs on right are for $a = 8$, $\epsilon = 2$. For $b = 2$ (blue), for $b = 4$ (cyan), for $b = 8$ (green) and for $b = 16$ (magenta)

Variance - $Var(w_j)$



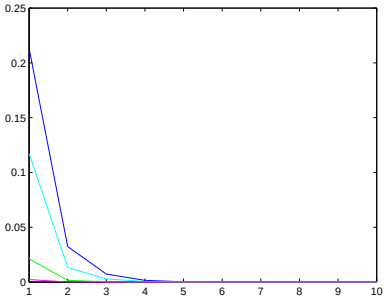
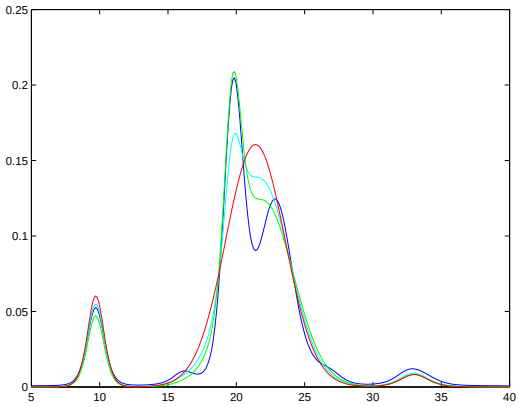


Figure 3: Effects of changes in ϵ on $Var(w_j)$ under the geometric-beta prior setting (left) and poisson-gamma prior setting (right). For both settings $a = 4\,\text{and}\,b = 8$. For $\epsilon = 0.1$ (blue), for $\epsilon = 1$ (cyan), for $\epsilon = 10$ (green) and for $\epsilon = 100$ (magenta)

Galaxy data set



S& P 500 data set

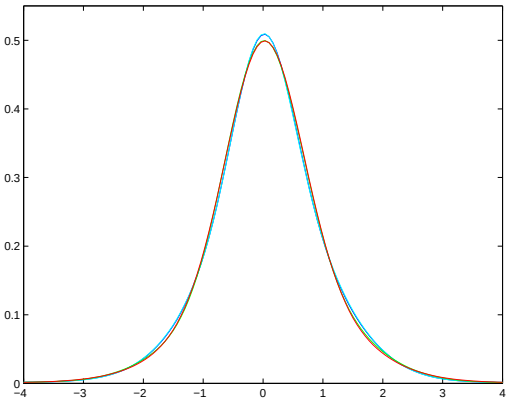
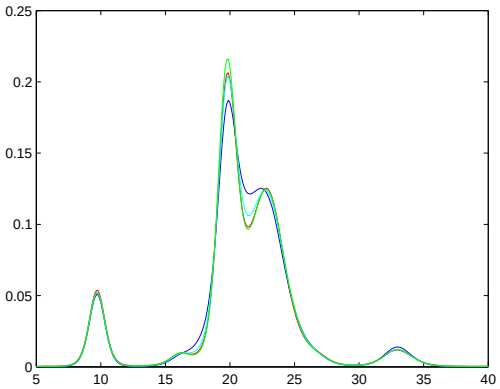


Figure 4: Effects of changes in b on $f(y)$ under the geometric-beta prior setting, using the G_0 prior of Escobar and West (1995). The values of a, and, ϵ are kept constant at 2, and, 10 respectively. For $b = 4$ (blue), for $b = 8$ (cyan), for $b = 12$ (green) and for $b = 20$ (red)

Galaxy data set



S& P 500 data set

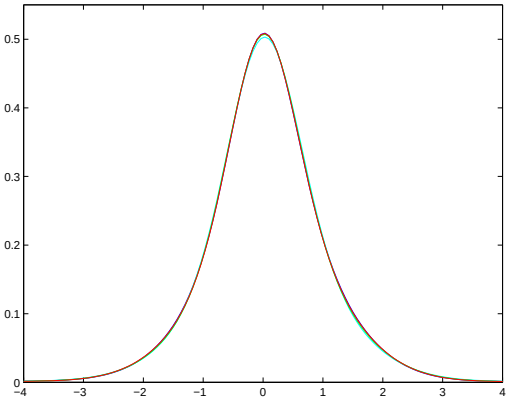
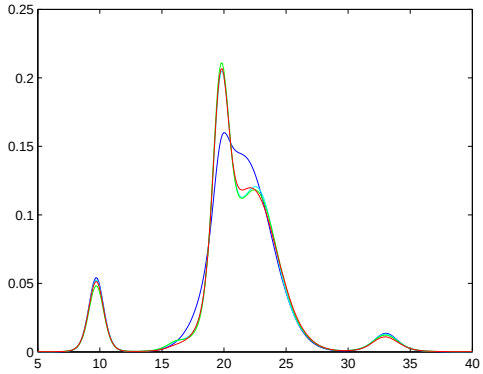


Figure 5: Effects of changes in a on $f(y)$ under the geometric-beta prior setting, using the G_0 prior of Escobar and West (1995). The values of b, and, ϵ are kept constant at 8, and, 4 respectively. For $a = 4$ (blue), for $a = 8$ (cyan), for $a = 12$ (green) and for $a = 20$ (red)

Galaxy data set



S& P 500 data set

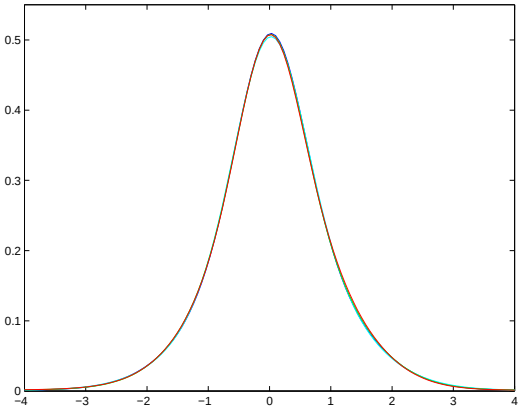


Figure 6: Effects of changes in c on $f(y)$ under the geometric-beta prior setting, using the G_0 prior of Escobar and West (1995). The values of a, and, b are kept constant at 4, and, 8 respectively. For $\epsilon = 0.1$ (blue), for $\epsilon = 1$ (cyan), for $\epsilon = 10$ (green) and for $\epsilon = 20$ (red)

The plots of thee effect of changes in (a, b, ϵ), under the geometric beta prior setting, on the tail of the S&P 500 estimated density, $\widehat{f}(y)$ are on a log scale. (Figures 7, 8 and 9)

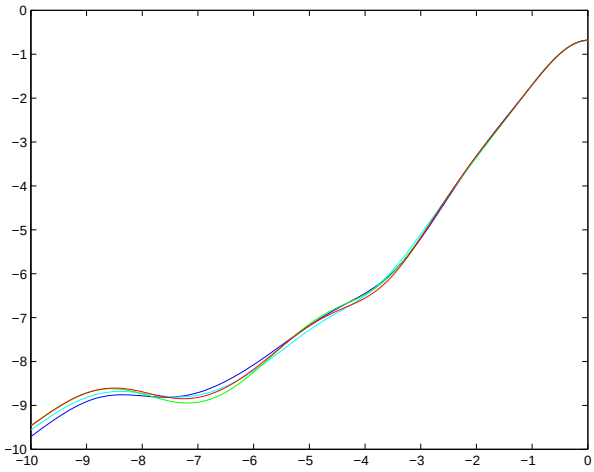


Figure 7: S&P 500: tails as a changes - geometric–beta model

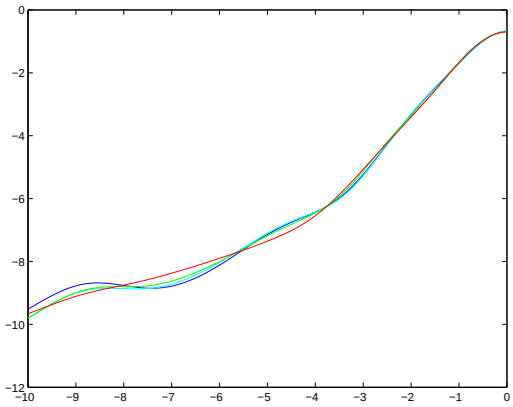


Figure 8: S&P 500: tails as b changes - geometric–beta model

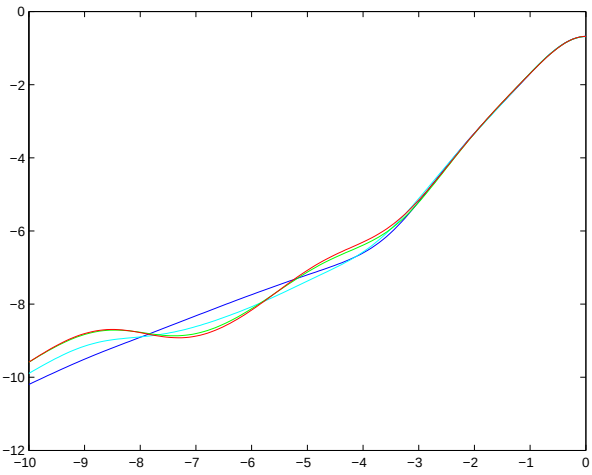


Figure 9: S&P 500: tails as ϵ changes - geometric–beta model