Technical Report No. 0104, Department of Statistics, University of Toronto

Defining Priors for Distributions Using Dirichlet Diffusion Trees

Radford M. Neal
Department of Statistics and Department of Computer Science
University of Toronto, Toronto, Ontario, Canada
`http://www.cs.utoronto.ca/~radford/`
`radford@stat.utoronto.ca`

12 March 2001

Abstract. I introduce a family of prior distributions over univariate or multivariate distributions, based on the use of a "Dirichlet diffusion tree" to generate exchangeable data sets. These priors can be viewed as generalizations of Dirichlet processes and of Dirichlet process mixtures. They are potentially of general use for modeling unknown distributions, either of observed data or of latent values. Unlike simple mixture models, Dirichlet diffusion tree priors can capture the hierarchical structure that is present in many distributions. Depending on the "divergence function" employed, a Dirichlet diffusion tree prior can produce discrete or continuous distributions. Empirical evidence is presented that some divergence functions produce distributions that are absolutely continuous, while others produce distributions that are continuous but not absolutely continuous. Although Dirichlet diffusion trees are defined in terms of a continuous-time stochastic process, inference for finite data sets can be expressed in terms of finite-dimensional quantities, which should allow computations to be performed by reasonably efficient Markov chain Monte Carlo methods.

1 Introduction

A Bayesian model for an unknown distribution is determined by a prior distribution over possible distributions. For such a model to be useful in practice, the prior must be an adequate approximation to our actual prior beliefs about the unknown distribution, and it must be possible to compute the predictive distribution for new data with reasonable efficiency. The desire for flexibility in the prior in order to capture complex prior beliefs (eg, that the distribution may, or may not, exhibit a hierarchical structure) must be balanced against the computational difficulties that very general priors may pose.

One of the simplest priors for an unknown distribution is the Dirichlet process (Ferguson 1973), which is very easy to handle computationally, but which produces distributions that are discrete with probability one, and hence is unsuitable for typical density modeling problems. This problem can be remedied by convolving the distribution with some continuous kernel, thereby spreading out the peaks to produce a continuous distribution, or more generally, by using the Dirichlet process to define a mixture distribution with a countably infinite number of components, each of some fairly simple parametric form (Antoniak 1973; Ferguson 1983). Computations for such Dirichlet process mixture models are often feasible using Markov chain sampling methods (Escobar and West 1995; Bush and MacEachern 1996; MacEachern and Müller 1998; Neal 2000).

Dirichlet process mixture models are not always ideal, however, because they use a prior distribution in which the parameters of one mixture component are independent of the parameters of other components. For many problems, we would expect instead that the components will be hierarchically organized. For instance, consider data on traits of organisms belonging to various species, which we might expect to model using one mixture component for each species present. Our prior belief in such a situation is likely to be that the parameters of the mixture components representing different species are not independent, but instead exhibit dependencies due to the (possibly unknown) pattern of evolutionary relationships among these species. Furthermore, we might believe that a species may not actually be adequately represented by a single mixture component of some simple form (eg, Gaussian), but that instead the organisms in each species may come in sub-groups, corresponding, for example, to different environmental conditions. We would expect that distributions of traits will differ slightly from one sub-group to another, but that these sub-group differences will generally be less than the differences between species. The sub-groups may themselves be divided into sub-sub-groups, and so forth, perhaps without limit, though the number and precision of observations may limit the degree to which the finer details of this hierarchy can be discerned.

We see therefore that for many problems an infinite number of simple mixture components are needed to model the complexities of the actual density. Dirichlet process mixture models do have an infinite number of components, but since the prior for the parameters of these components does not capture the hierarchical structure, inference based on a Dirichlet process mixture model will be inefficient — it will take more data than it should to force the model to create appropriate components, since the model does not "know" that these components are likely to be similar to other components. Furthermore, even when the data is sufficient to force the posterior distribution to be concentrated near the correct density, the fact that this density has low prior probability can slow the convergence of simple Markov chain samplers, sometimes drastically. (Jain and Neal (2000) discuss this problem, and introduce a more complex Markov chain sampler that alleviates it.)

In this paper, I introduce a new prior distribution over distributions that can be seen as a hierarchical generalization of Dirichlet process mixture models. The latent structure underlying the data for this model is what I call a "Dirichlet diffusion tree", whose leaves are the data points, and whose non-leaf nodes represent groupings of data points in a hierarchy. This latent tree structure generalizes the one-level grouping of data points into clusters that underlies a Dirichlet process mixture model. For some problems, the latent diffusion tree structure may be of intrinsic interest. For other problems, this structure is merely a device for obtaining a more suitable prior distribution over distributions.

A variety of hierarchical clustering methods are widely used (see, for instance, Jain 1988), but these are typically not based on any probabilistic model of the data. Williams (2000) discusses hierarchical Bayesian mixture models that are similar in objective to the models described here, but which have a finite (though unknown) number of components. Branching diffusion processes that produce trees somewhat similar to those studied here have been investigated as evolutionary models, for example by Edwards (1970), but these processes differ in crucial ways from the Dirichlet diffusion trees discussed here.

Polya trees (Ferguson 1974; Maudling, Sudderth, and Williams 1992; Lavine 1992, 1994) are another generalization of the Dirichlet process that produces distributions that have hierarchical structure, and which can be continuous. However, Polya tree priors have an unfortunate dependence on an arbitrary set of division points, at which discontinuities in the density functions occur. Walker, *et al* (1999) review these and other related priors over distributions, including the "reinforced random walks" of Coppersmith and Diaconis, which resemble the diffusion trees discussed here, but which again differ in crucial respects. My hope is that Dirichlet diffusion trees will provide priors that are more suitable for many problems than those produced by these other methods, but which will still be computationally tractable, though computations will not be as easy as for simple Dirichlet process or Polya tree models.

I will first show, in Section 2, how a Dirichlet diffusion tree can be used to define a prior distribution over distributions for real data vectors. In Section 3, I will illustrate the properties of the Dirichlet diffusion tree priors obtained using certain "divergence functions". The absolute continuity of distributions obtained using these divergence functions will be investigated empirically in Section 4. In Section 5, I will discuss how Dirichlet diffusion tree priors relate to priors based on Dirichlet processes, and to Polya tree priors. I conclude in Section 6 by discussing how Dirichlet diffusion tree models can be extended to handle categorical data and other more complex situations, and by outlining how computations for Dirichlet diffusion tree models should be possible using Markov chain Monte Carlo methods.

The programs (written in R) that were used to produce the results in this paper are available from my web page.

2 The Dirichlet diffusion tree prior

I will define the Dirichlet diffusion tree prior over distributions by giving a procedure for randomly generating a data set of n points, each a vector of p real numbers, in which the data points are drawn independently from a common distribution drawn from the prior. The procedure generates these random data sets one point at time, with each point being drawn from its conditional distribution given the previously generated points, in a manner analogous to the "Polya urn" procedure for defining a Dirichlet process prior (Blackwell and MacQueen 1973). Such a procedure will be consistent with the assumption that the data points are independently drawn from some unknown distribution as long as the distribution over data sets produced by the procedure is exchangeable, since such an exchangeable prior on data sets is equivalent to a prior over distributions by de Finetti's Representation Theorem (Bernardo and Smith 1994, Section 4.3). To obtain a picture of a distribution drawn the the prior, one can simply generate a very large data set, and then look at its histogram or scatterplot.

2.1 Generation of data points using diffusion trees

A data set drawn from the Dirichlet diffusion tree prior is produced by following paths to each data point that are generated by a diffusion process. Paths to different data points are linked in a tree structure, according to when each path diverges from previous paths. Generation of these paths will be described below by a sequential procedure, in which paths are generated for each data point in turn, but it will be shown in Section 2.3 that the ordering of the data points is in fact immaterial.

The first data point is generated by a simple Gaussian diffusion process (ie, Brownian motion). This process begins at some origin, which I will here fix at zero, but which in general should be chosen based on our prior belief about the likely location of the data points. From this origin, the diffusion process operates for a predetermined length of time, which without loss of generality can be fixed at one. If at time t, the process has reached the point $X_1(t)$, the point reached an infinitesimal time, dt, later will be $X_1(t+dt) = X_1(t) + N_1(t)$, where $N_1(t)$ is a Gaussian random variable with mean zero and covariance $\sigma^2 I dt$, were σ^2 is a parameter of the diffusion process. The $N_1(t)$ at different times are independent. The end point of the path, $X_1(1)$, is the sum of the infinitesimal increments $N_1(t)$, and is easily seen to have a Gaussian distribution with mean zero and covariance $\sigma^2 I$. This end point is the first point in the data set.

The second point in the data set is also generated by following a path from the origin. This path, $X_2(t)$, initially follows the path leading to the first point, $X_1(t)$. However, the two paths will diverge at some time, T_d, after which the path to the second point is independent of the remainder of the first path. In other words, the infinitesimal increments for the second path, $N_2(t)$, are equal to the corresponding increments for the first path, $N_1(t)$, for $t < T_d$, but thereafter, $N_2(t)$ and $N_1(t)$ are independent. The distribution of the divergence time, T_d, can be expressed in terms of a "divergence function", $a(t)$, which is analogous to the "hazard function" well known in survival analysis. At each time t before divergence occurs, the probability that the path to the second data point will diverge from the path to the first data point within the next infinitesimal time period of duration dt is given by $a(t)dt$.

The path to the third data point initially follows the paths to the first two data points. At times, t, before the first two paths diverged, the probability that the third path diverges in an infinitesimal interval of duration dt is $a(t)dt/2$ — note the division by two, which comes from this path having been traversed on the way to two earlier data points. If the third path does not diverge before the time when the first two paths diverged, the third path will begin to follow one or the other of these earlier paths, with the choice being made at random with equal probabilities, since equal numbers (namely one) of previous paths went each way. Once the third path is following the path of just one of the previous data points, the probability of it diverging in an interval of duration dt is $a(t)dt$. Whenever the third path diverges, its subsequent motion is independent of the other paths.

In general, the ith point in the data set is obtained by following a path from the origin that initially coincides with the path to the previous $i-1$ data points. If the new path has not diverged at a time when paths to past data points diverged, the new path chooses between these past paths with probabilities proportional to the numbers of past paths that went each way. If at time t, the new path is following a path traversed by m previous paths, the probability that it will diverge from this path within an infinitesimal interval of duration dt is $a(t)dt/m$. Once divergence occurs, the new path moves independently of previous paths.

Figure 1 illustrates the diffusion tree process for a data set of $n = 4$ data points, each a single real number (ie, $p = 1$).

The choice of divergence function, $a(t)$, determines the nature of the distributions generated by this process. If $\int_0^1 a(t)\, dt$ is infinite, divergence before $t = 1$ is guaranteed, and hence the prior will be concentrated on continuous distributions, for which the probability of two independently chosen data points being equal is zero. Dirichlet processes and Dirichlet process mixtures can be obtained using divergence functions with infinite peaks, which result in non-zero divergence probabilities at particular exact times. The properties of distributions produced using various choices for $a(t)$ are explored in Sections 3, 4, and 5.

2.2 Probability of generating a given tree of data points

Although the Dirichlet diffusion tree prior is described above in terms of a continuous diffusion process, the probability of obtaining a given data set along with its underlying tree structure can easily be expressed directly, as long as the details of the paths taken are suppressed. What remains is the structure of the tree (ie, how data points are hierarchically grouped), the times at which paths diverged, and the locations of these divergence points and of the final data points. Figure 2 shows this less-detailed representation of the example shown in Figure 1.

The probability of obtaining a given tree and data set can be written as a product of two factors, one pertaining to the tree, the other to the data. The tree factor is the probability of obtaining the given tree structure along with the given divergence times. The data factor is the probability of obtaining the given locations for the divergence points and final data points when the tree structure and divergence times are as given.

The tree factor can be found without reference to the locations of the points, since the divergence function depends only on t. Since the procedure described in Section 2.1 generates data points in sequence, this factor is most easily obtained in the same way, though we will see in Section 2.3 that the order does not actually matter. We will need the probability

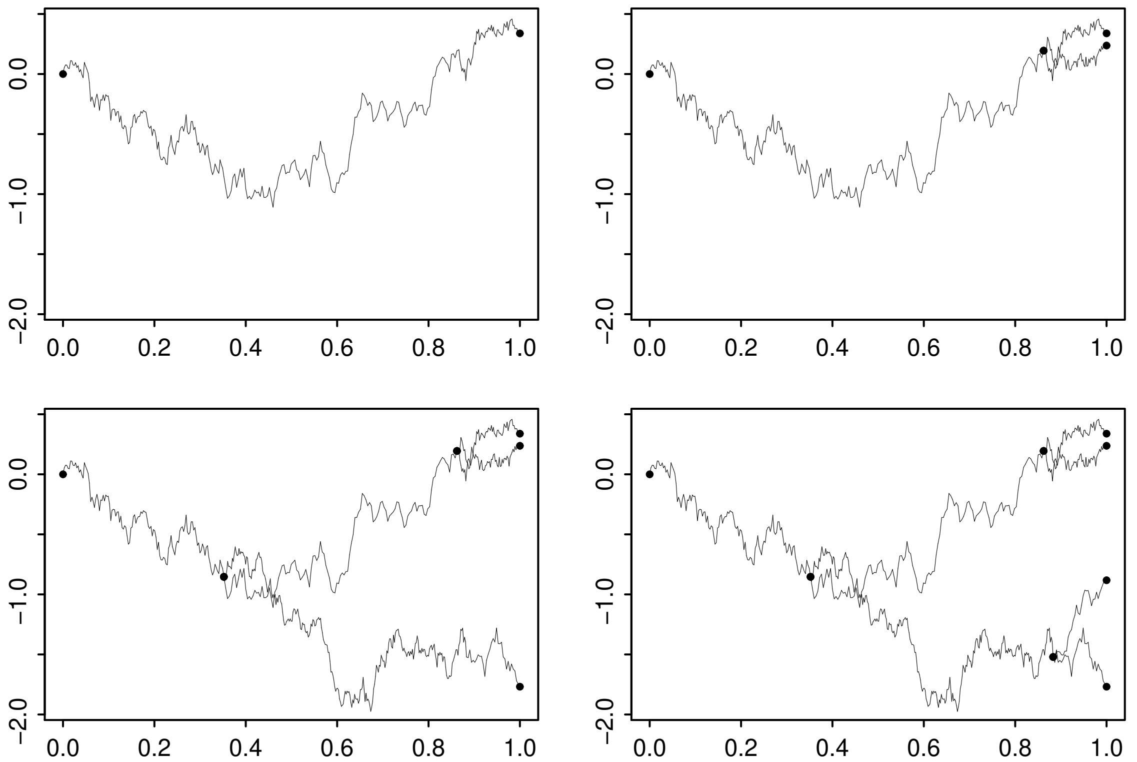


Figure 1: Generation of a data set of four real numbers from the Dirichlet diffusion tree prior with $\sigma = 1$ and $a(t) = 1/(1-t)$. Diffusion time is on the horizontal axis, data values on the vertical axis. The upper-left plot shows the path to the first data point, whose value is 0.34. The upper-right plot shows this path together with the path to the second data point (0.24), which diverges from the first path at $t = 0.86$. In the lower-left plot, the path to the third data point diverges from the first two paths at $t = 0.35$, and reaches a final value of -1.77. In the lower-right plot, the path to the fourth data point (-0.87) follows the path to the third data point until $t = 0.88$.

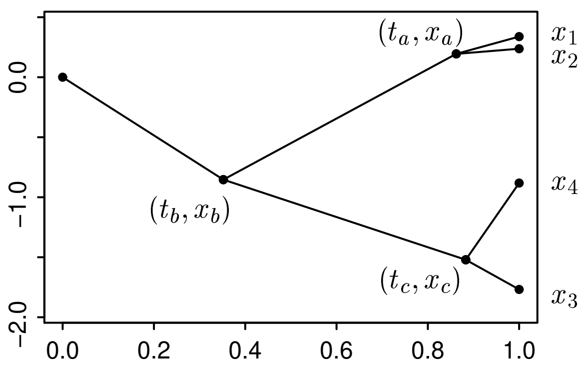


Figure 2: The data x_1, x_2, x_3, and x_4 generated above and its underlying tree, with details of the paths suppressed except at the divergence points a, b, and c.

that a new path following a path previously traversed m times will not diverge between time s and time t. This can be derived as follows, by dividing the time from s to t into k intervals of duration $(t-s)/k$, and letting k go to infinity:

$$
\begin{aligned}
P(\text{no divergence}) &= \lim_{k\to\infty} \prod_{i=0}^{k-1} \Big(1 - a(s+i(t-s)/k)(t-s)/k/m\Big) && (1)\\
&= \exp\Big(\lim_{k\to\infty} \sum_{i=0}^{k-1} \log\Big(1 - a(s+i(t-s)/k)(t-s)/k/m\Big)\Big) && (2)\\
&= \exp\Big(-\int_s^t a(u)/m\, du\Big) \;=\; \exp((A(s)-A(t))/m) && (3)
\end{aligned}
$$

where $A(t) = \int_0^t a(u)\,du$ (the cumulative divergence function). Using this, the probability (density) of obtaining the tree structure and divergence times shown for the example in Figure 2 is as follows:

$$
\begin{aligned}
&\exp(-A(t_a))\ a(t_a) \ \times\ \exp(-A(t_b)/2)\ (a(t_b)/2)\\
&\qquad \times\ \exp(-A(t_b)/3)\ (1/3)\ \exp(A(t_b)-A(t_c))\ a(t_c)
\end{aligned}
\qquad (4)
$$

The first factor above is the probability that the second path will not diverge before time t_a, times the probability density that it will diverge at time t_a. The second factor is the probability that the third path will not diverge from the first two paths before time t_b, times the probability density that it will diverge at time t_b. The third factor is the probability that the fourth path will not diverge from the previous three paths before time t_b, times the probability that at this divergence point, the fourth path will follow the path it followed, which was taken by one of the three previous paths, times the probability that it will not diverge from this path between times t_b and t_c, times the probability density that it will diverge at time t_c.

Given the tree structure and divergence times, the data factor is easily found as a product of Gaussian densities, by noting that the distribution for the location of a point that has diffused for time d starting from location x is Gaussian with mean x and covariance $\sigma^2 Id$. Letting $\phi(x,\mu,\Sigma)$ be the Gaussian probability density for a vector x with mean μ and covariance Σ, we can write the data factor for the example in Figure 2 as follows:

$$
\begin{aligned}
&\phi(x_b,0,\sigma^2 I t_b) \times \phi(x_a,x_b,\sigma^2 I(t_a-t_b)) \times \phi(x_1,x_a,\sigma^2 I(1-t_a)) \times \phi(x_2,x_a,\sigma^2 I(1-t_a))\\
&\qquad \times \phi(x_c,x_b,\sigma^2 I(t_c-t_b)) \times \phi(x_3,x_c,\sigma^2 I(1-t_c)) \times \phi(x_4,x_c,\sigma^2 I(1-t_c))
\end{aligned}
\qquad (5)
$$

With the model formulated as here, the p components of the data vectors are independent once one conditions on the tree structure — ie, all dependencies are modeled by means of the latent structure. However, it would be easy to introduce correlations between components in expression (5) if this were desired.

2.3 Proof of exchangeability

To show that the procedure for generating a data set using a Dirichlet diffusion tree defines a valid prior over densities, we must show that the probability density for a data set does not change when the order of the data points is permuted — ie, that the prior is "exchangeable". I will show this by proving the stronger property that the probability density for producing a data set along with its underlying tree structure and the times and locations of the divergence points is the same for any ordering of data points. By summing over all possible tree structures and integrating over times and locations of divergences it then follows that the prior for data sets is exchangeable.

The probability density for a data set along with an underlying tree can be written as a product of factors, each pertaining to one segment of the tree. The tree in Figure 2 has seven segments, connecting pairs of points as follows:

$$
\begin{array}{lll}
 & & (t_a, x_a) - (1, x_1) \\
 & (t_b, x_b) - (t_a, x_a) & \\
 & & (t_a, x_a) - (1, x_2) \\
(0,0) - (t_b, x_b) & & \\
 & & (t_c, x_c) - (1, x_4) \\
 & (t_b, x_b) - (t_c, x_c) & \\
 & & (t_c, x_c) - (1, x_3)
\end{array}
$$

For each segment, $(t_u, x_u) - (t_v, x_v)$, there is a factor in the probability density that corresponds to the probability density for the diffusion process starting at x_u to move to x_v in time $t_v - t_u$, which is $\phi(x_v, x_u, \sigma^2 I(t_v - t_u))$. The product of these factors is the overall data factor in the density, which for the example of Figure 2 is given by (5). Since the set of segments making up the tree does not depend on the order of the data points, neither will this data factor in the probability density.

A segment of the tree traversed by more than one path will also be associated with a factor in the overall probability density pertaining to the lack of divergence of these paths before the end of the segment. If such a segment ends before $t = 1$ (as will always be the case for $a(t)$ that produce continuous distributions), there will also be a factor for the probability density for one path to diverge at the end of this segment, along with factors for the probabilities of later paths taking the branches they did. The product of such factors for all segments is the overall factor relating to the tree structure. For the example of Figure 2, the tree factor given by (4) can be rewritten as a product of factors associated with the three segments traversed by more than one path, $(0,0) - (t_b, x_b)$, $(t_b, x_b) - (t_a, x_a)$, and $(t_b, x_b) - (t_c, x_c)$, as follows:

$$
\begin{aligned}
&\exp(-A(t_b))\ \exp(-A(t_b)/2)\ (a(t_b)/2)\ \exp(-A(t_b)/3)\ (1/3) \\
&\quad \times\ \exp(A(t_b) - A(t_a))\ a(t_a)\ \times\ \exp(A(t_b) - A(t_c))\ a(t_c)
\end{aligned}
\tag{6}
$$

This is simply a rearrangement of the factors in (4), with $\exp(-A(t_a))$ rewritten as $\exp(-A(t_b)) \exp(A(t_b) - A(t_a))$. The first factor above, associated with the segment $(0,0) - (t_b, x_b)$, can be interpreted as the the probability that the path to the second data point does not diverge from the first path before t_b, times the probability that the path to the third data point does not diverge before t_b, times the probability density that the third path does diverge at t_b, times the probability that the fourth path does not diverge before t_b, times the probability that the fourth path follows the third path rather than the first or second path at this divergence point. We need to show that this factor does not actually depend on the ordering of the data points, even though it is expressed in an order-dependent way above.

Consider a segment, $(t_u, x_u) - (t_v, x_v)$, that was traversed by $m > 1$ paths. The probability that the $m-1$ paths after the first do not diverge before t_v is $\prod_{i=1}^{m-1} \exp((A(t_u) - A(t_v))/i)$, which does not depend on the order of the data points. If $t_v = 1$, this is the whole factor for this segment. Otherwise, suppose that the first $i-1$ paths do not diverge at t_v, but that path i does diverge at t_v. (Note that i will be at least two, and that some path must diverge at t_v, as otherwise the segment would not end at that time.) The probability density for this divergence is $a(t_v)/(i-1)$. Subsequent paths take one or the other branch at this divergence point, with probabilities proportional to the number that have gone each way previously. Suppose that n_1 paths in total go the way of the first path, and n_2 go the way of path i. (Note that $n_1 \geq i-1$, and $n_1 + n_2 = m$.) The probability that path j (with $j > i$) goes the way of the first path will be $c_1/(j-1)$, where c_1 is the number of paths that went that way before, which will vary from $i-1$ to $n_1 - 1$. The probability that path j goes the other way will be $c_2/(j-1)$, where c_2 is the number of paths that went the other way before, which will vary from 1 to $n_2 - 1$. The product of these branching probabilities for all $j > i$ will be

$$\prod_{c_1=i-1}^{n_1-1} c_1 \cdot \prod_{c_2=1}^{n_2-1} c_2 \Big/ \prod_{j=i+1}^{m} (j-1) \;=\; \frac{(n_1-1)!}{(i-2)!} \cdot (n_2-1)! \cdot \frac{(i-1)!}{(m-1)!} \tag{7}$$

$$= \; (i-1) \cdot \frac{(n_1-1)!\,(n_2-1)!}{(m-1)!} \tag{8}$$

When this is multiplied by the probability density of $a(t_v)/(i-1)$ for path i diverging at time t_v, the two factors of $i-1$ cancel, leaving a result that does not depend on i, and hence is independent of the order of the data points. The above argument has to be modified if $a(t)$ has an infinite peak, allowing more than one path to diverge at exactly the same time, but the result is the same.

2.4 Another way of generating data sets with Dirichlet diffusion trees

The continuous time picture of a Dirichlet diffusion tree prior that is illustrated in Figure 1 is not convenient for actual generation of data sets. Instead, one can work in terms of the times and locations of the divergence points and the final data points only, as in Figure 2, and use the inverse cumulative divergence function $A^{-1}(e)$, rather than $a(t)$. This also avoids problems when $a(t)$ has infinite peaks.

A data set can easily be generated once the tree structure and the times at which divergences occur have been generated — the locations of the divergence points can be generated (in order of increasing time), followed by the final data points, by simply sampling from the appropriate Gaussian distributions, as illustrated by expression (5).

The tree structure and divergence times can be generated one path at a time, by following the current tree of paths from its root. The path to data point i will initially follow the same path as the previous $i-1$ data points. If we ignore for the moment the possibility of the new path not diverging from this initial path until after the previous paths diverged, the cumulative distribution function of the time of divergence will be $C(t) = 1 - \exp(-A(t)/(i-1))$. We can generate a divergence time, t_d, for the new path by generating a random variate, U, uniformly over $(0, 1)$, and then letting $t_d = C^{-1}(U) = A^{-1}(-(i-1)\log(1-U))$. Since $-\log(1-U)$ is exponentially distributed, this is equivalent to generating an exponential random variate, E, and letting $t_d = A^{-1}((i-1)E)$. If the divergence time generated in this way is past the point where previous paths diverged, one of the previous paths is chosen at random, with probabilities proportional to the numbers that went each way previously. We then generate a time for divergence from the new segment in similar fashion. In general, when the new path reaches a segment that extends from time t_u to time t_v and that has been traversed m times previously, we generate a new exponential random variate, E, and compute the divergence time of the new path as $t_d = A^{-1}(A(t_u) + mE)$. If t_d is greater than t_v, the new path does not diverge within this segment, in which case we continue with a segment randomly chosen from among those that branch at time t_v.

Note that this procedure requires computation only of A^{-1}, not of A or a. (The apparent need above to compute $A(t_u)$ can be bypassed, since t_u will have previously been found by applying A^{-1}.) We can therefore specify a Dirichlet diffusion tree prior by defining $A^{-1}(e)$, which will be an ordinary function even when $a(t)$ has infinite peaks.

3 Examples of Dirichlet diffusion tree priors

The properties of a Dirichlet diffusion tree prior vary according to the choice of divergence function, $a(t)$. In this section, priors defined by various choices for $a(t)$ will be investigated by looking at data sets that were generated from these priors using the method of Section 2.4.

3.1 Priors with $a(t) = c/(1-t)$

Divergence functions of the form $a(t) = c/(1-t)$ have integrals that diverge only logarithmically as $t \to 1$: $A(t) = \int_0^t a(u)\,du = -c\log(1-t)$ and $A^{-1}(e) = 1 - \exp(-e/c)$. Distributions drawn from such a prior will be continuous — ie, the probability that two data points drawn such a distribution will be identical is zero. These distributions will, however, show some degree of tight clustering, varying with the choice of c, due to the possibility that divergence will not occur until quite close to $t = 1$.

Figure 3 shows histograms of two one-dimensional data sets of 4000 points drawn from this prior with $c = 1$. Similarly, Figure 4 shows histograms of data sets drawn from this prior with $c = 1/3$. The smaller value for c leads to paths diverging later, and hence to more pronounced clustering of data points at a small scale. In Section 4, empirical evidence will be presented that the distributions generated with $c = 1$ are absolutely continuous, but those generated with $c = 1/3$ are not, and hence do not have density functions.

Figure 5 shows the generation of a two-dimensional data set of 1000 points, drawn from

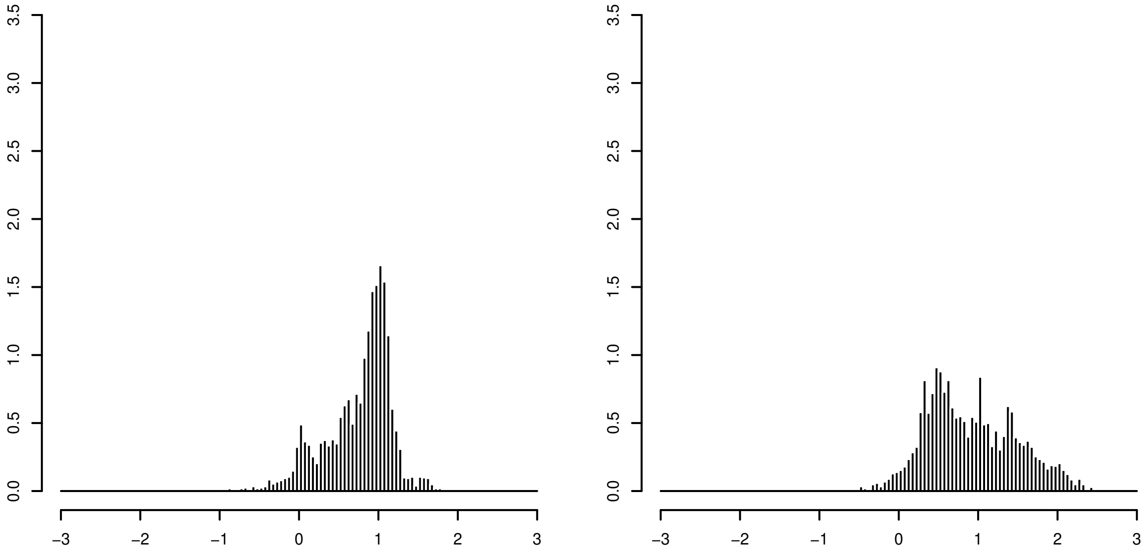


Figure 3: Probability density histograms of two data sets of 4000 points drawn from the Dirichlet diffusion tree prior with $\sigma = 1$ and $a(t) = 1/(1-t)$.

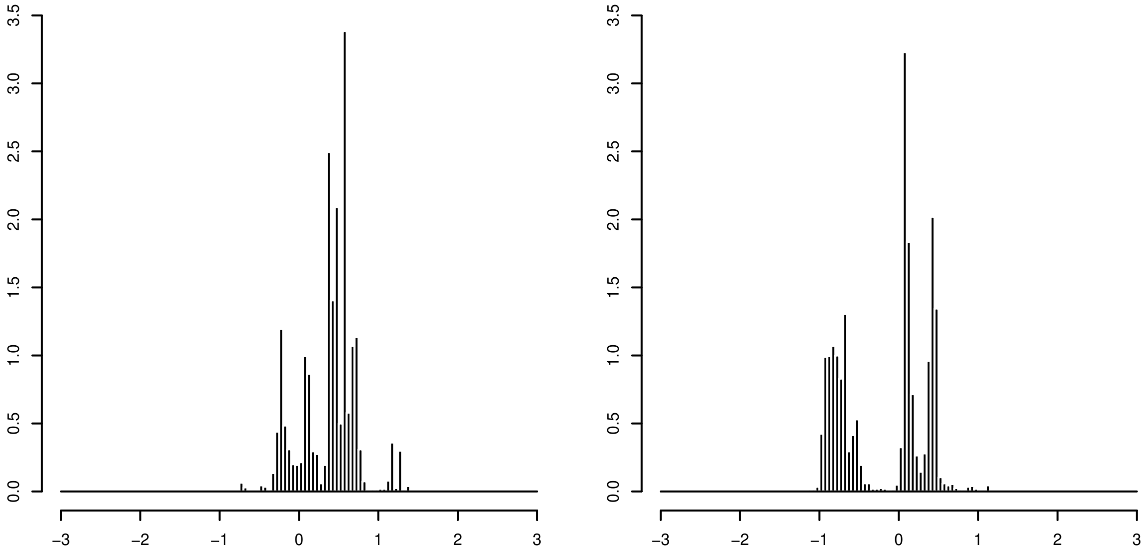


Figure 4: Probability density histograms of two data sets of 4000 points drawn from the Dirichlet diffusion tree prior with $\sigma = 1$ and $a(t) = (1/3)/(1-t)$.

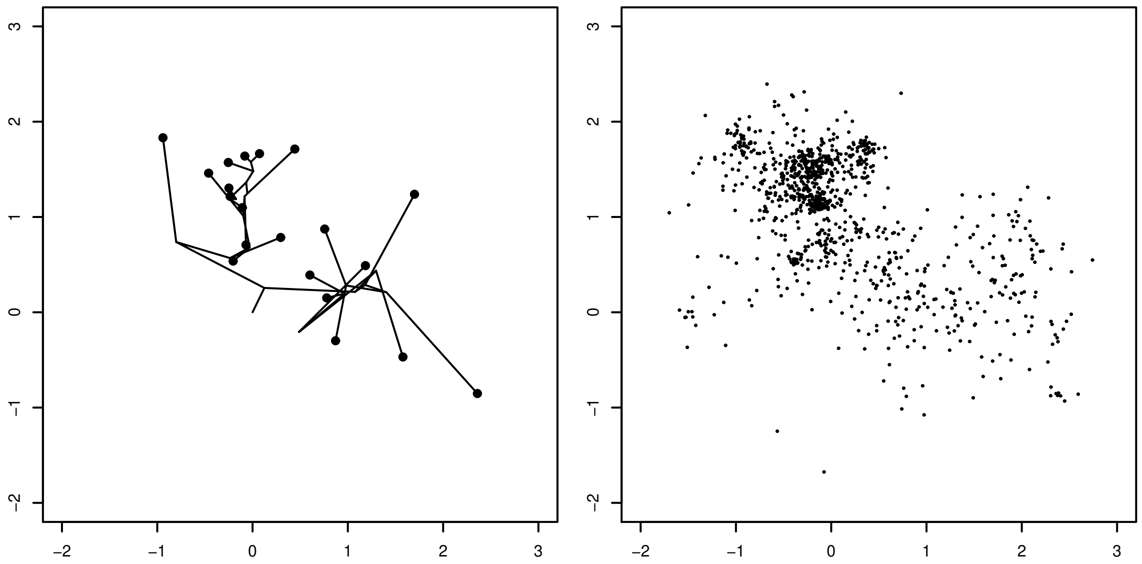


Figure 5: Generation of a two-dimensional data set from the Dirichlet diffusion tree prior with $\sigma = 1$ and $a(t) = 1/(1-t)$. The plot on the left shows the first twenty data points generated along with the underlying tree structure. The plot on the right shows 1000 data points obtained by continuing the procedure beyond the twenty points shown on the left.

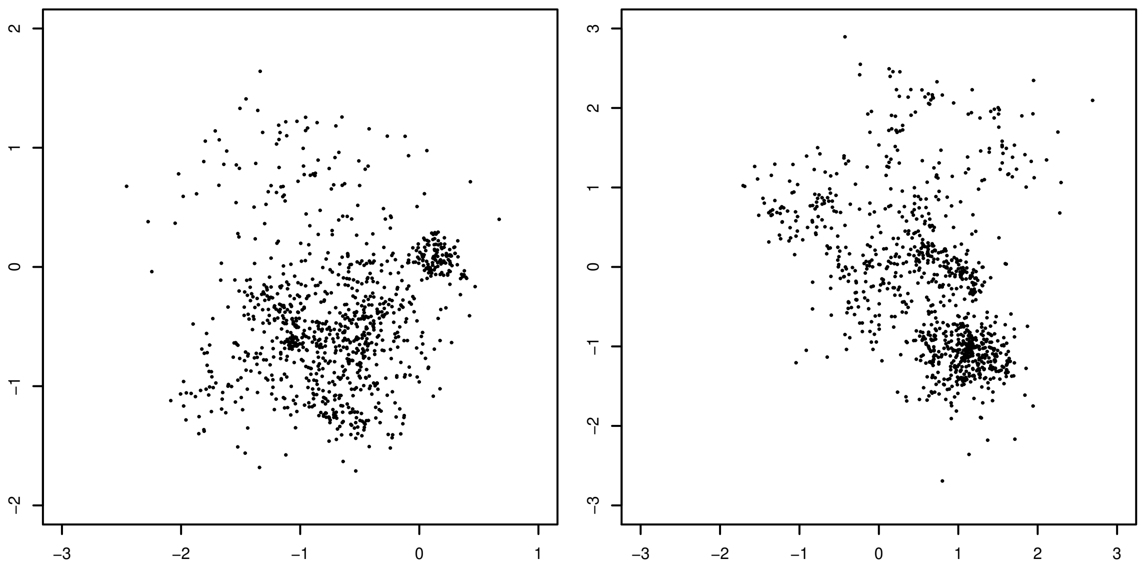


Figure 6: Two more data sets of 1000 points that were drawn independently from the same prior as for Figure 5. Note that the scales differ for the three data sets.

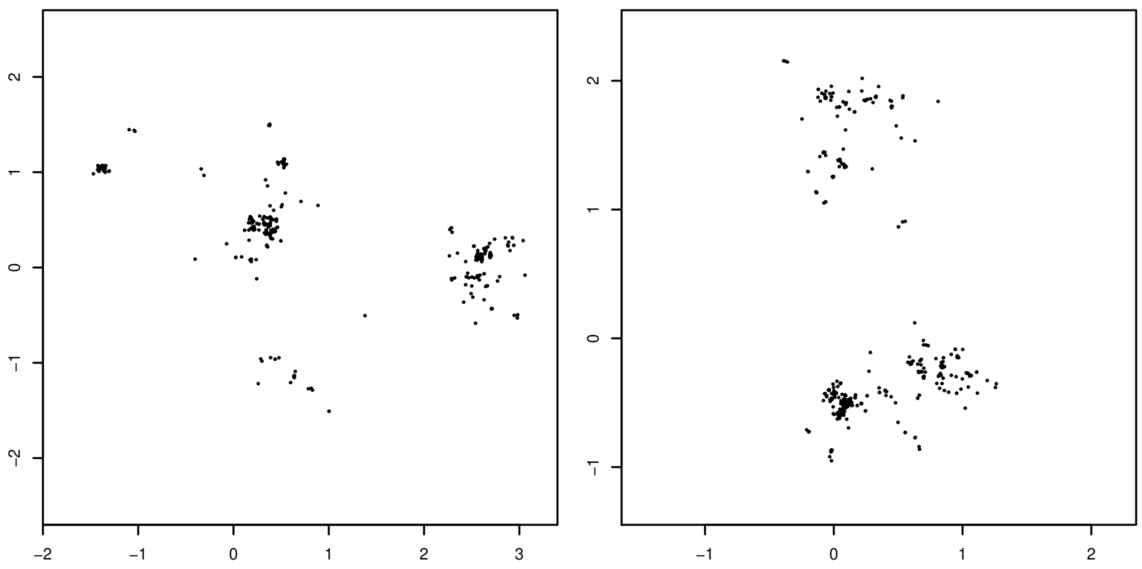


Figure 7: Two data sets of 1000 points that were drawn independently from the Dirichlet diffusion tree prior with $\sigma = 1$ and $a(t) = (1/4)/(1-t)$. Note that the scales differ for the two data sets.

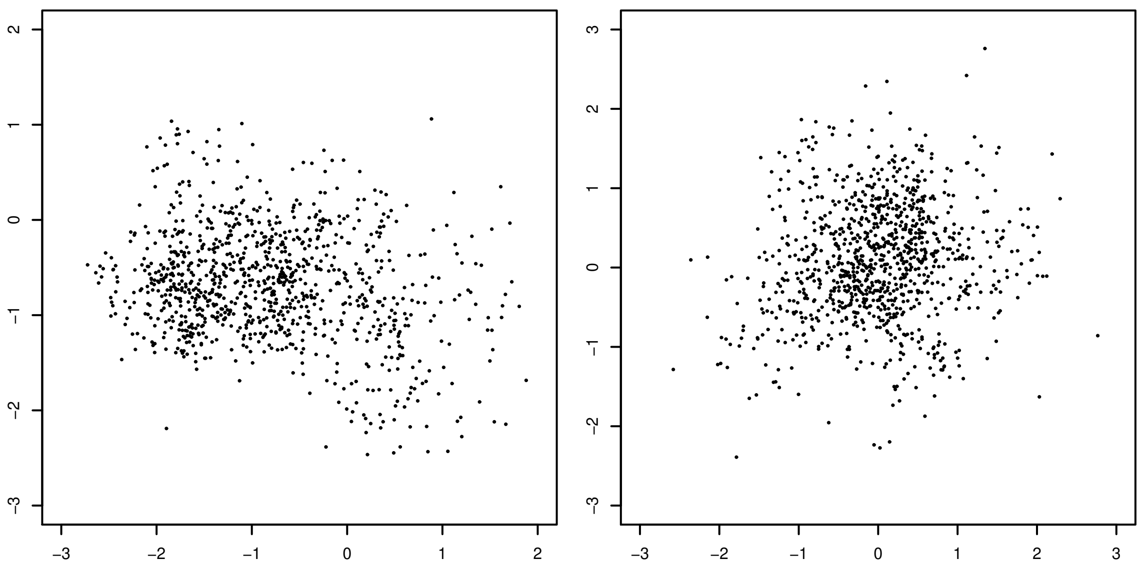


Figure 8: Two data sets of 1000 points that were drawn independently from the Dirichlet diffusion tree prior with $\sigma = 1$ and $a(t) = (3/2)/(1-t)$. Note that the scales differ for the two data sets.

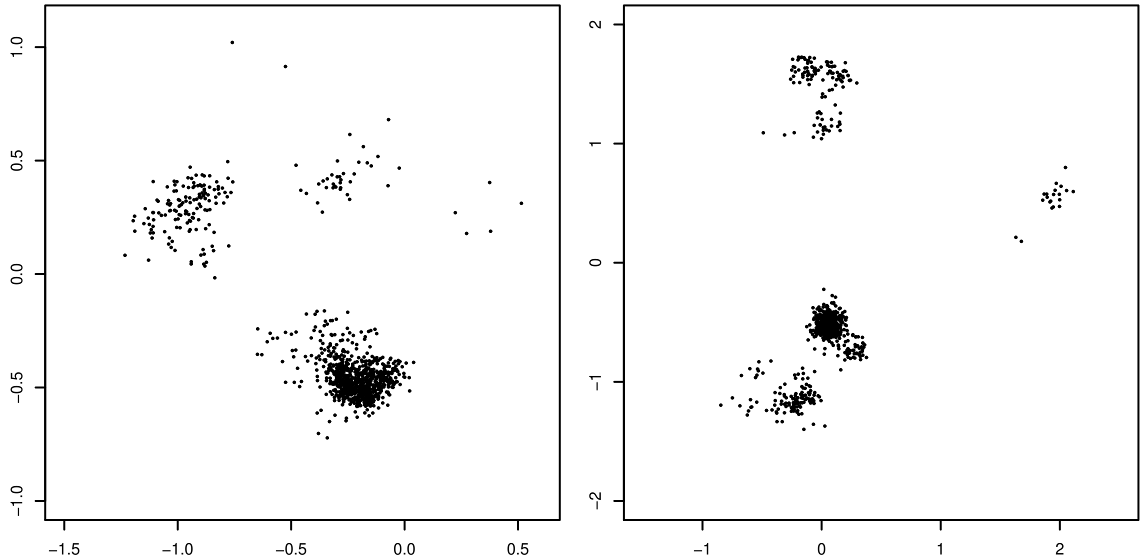


Figure 9: Two data sets of 1000 points that were drawn independently from the Dirichlet diffusion tree prior with $\sigma = 1$ and $a(t) = (1/2) + (1/200)/(1-t)^2$. Note that the scales differ for the two data sets.

this prior with $c = 1$. The plot on the left shows the underlying tree generated along with the first 20 data points; the plot on the right shows all 1000 data points. Figure 6 shows two more data sets from the same prior. These data sets show the hierarchical structure that the Dirichlet diffusion tree prior can produce — there are not just multiple modes in this data, but further structure within each mode. This hierarchy is more obvious in Figure 7, which shows data sets drawn from the prior with $c = 1/4$, producing tighter clusters. Figure 8 shows that when c is instead set to the larger value of 3/2, the distributions generated are smoother, and do not exhibit any obvious hierarchical structure. However, the underlying tree still has an effect, as it is responsible for the non-Gaussian nature of these distributions.

3.2 Priors with $a(t) = b + c/(1-t)^2$

Priors with different characteristics can be obtained using a divergence function of the form $a(t) = b + c/(1-t)^2$, for which $A(t) = bt - c + c/(1-t)$ and

$$A^{-1}(e) \;=\; \begin{cases} \dfrac{b+c+e-\sqrt{(b+c+e)^2-4be}}{2b} & \text{if } b \neq 0 \\ 1 - c/(e+c) & \text{if } b = 0 \end{cases} \qquad (9)$$

Figure 9 shows two data sets generated from such a prior with $b = 1/2$ and $c = 1/200$. These choices produce fairly well-separated clusters, exhibiting a clear hierarchical structure, but with the points within each cluster being more smoothly distributed than in Figure 7.

4 Testing absolute continuity of distributions produced from Dirichlet diffusion tree priors

The distributions produced by a Dirichlet diffusion tree prior will be continuous (with probability one) when the divergence function, $a(t)$, is such that $\int_0^1 a(t)\,dt$ is infinite, since this implies that two data points drawn from a distribution drawn from the prior will have probability zero of being identical. However, it does not follow that distributions drawn from the prior will be absolutely continuous, which is what is required for them to have density functions (see, for instance, Billingsley 1995, Sections 31 and 32). In this section, I will investigate empirically whether or not the examples of Dirichlet diffusion tree priors discussed in the previous section produce absolutely continuous distributions.

Absolute continuity will be tested by looking at distances to nearest neighbors in a sample from the distribution. Suppose we have a sample of n data vectors, each consisting of p real numbers, that were drawn independently from a distribution drawn from some prior. For each data vector, x, in this sample, we compute the Euclidean distances from it to all the other data vectors in the sample, and from this find the ratio, r, of the distance to the nearest other data vector to the distance to the second-nearest other data vector. If the distribution from which these vectors were drawn has a continuous density function, and if the sample size, n, is sufficiently large, the density in the region near x where its two nearest neighbors are located will be approximately constant. It follows that the conditional distribution of the nearest data vector given the position of the second-nearest data vector will be approximately uniform over the sphere whose centre is at x and whose radius is the distance to the second-nearest neighbor. This implies in turn that the distribution of r^p will be uniform over $(0,1)$.

The empirical cumulative distribution function of r^p for a large sample therefore provides evidence of whether or not the distribution from which the sample came is absolutely continuous. This evidence will fall short of proof, partly, of course, because there will be no guarantee that the sample was large enough to reveal the asymptotic behaviour. However, even apart from the inevitable uncertainty of empirical tests, the argument above does not exclude the possibility that the distribution of r^p might somehow be uniform even if the distribution is not absolutely continuous. Also, the argument shows only that the distribution of r^p should be uniform if the distribution is absolutely continuous and its density function is continuous. A finite number of discontinuities in the density function would not matter (given a large enough sample), but the situation when the density function has an infinite number of discontinuities is unclear.

Empirical cumulative distributions of r^p for one-dimensional data sets generated using $a(t) = c/(1-t)$ are shown in Figure 10, for various values of c. The plot for each value of c shows empirical distributions from two samples of size 4000. For $c = 1$ and $c = 1/3$, these are the samples shown in Figures 3 and 4. Clearly, the distribution of r^p when $c = 1/5$ is far from uniform, providing evidence (subject to the caveats above) that distributions produced using this Dirichlet diffusion tree prior are not absolutely continuous. The plots

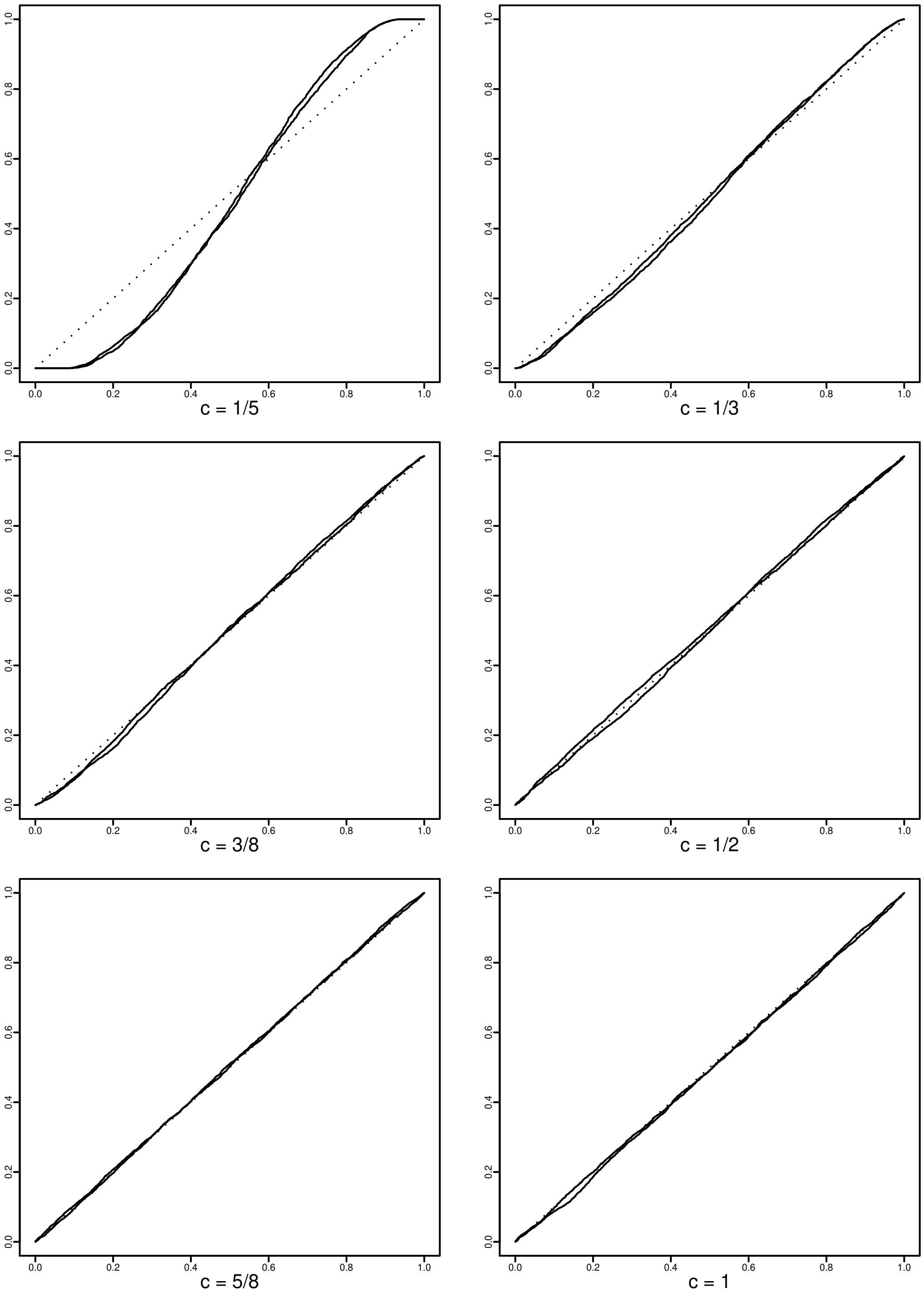


Figure 10: Empirical cumulative distributions of the ratio of distance to nearest neighbor to distance to second-nearest neighbor. Each plot shows cumulative distributions for two samples of size 4000 drawn from two univariate distributions that were drawn from priors with $a(t) = c/(1-t)$, with the indicated values for c. The dotted diagonal lines show the cumulative distribution for the uniform distribution.

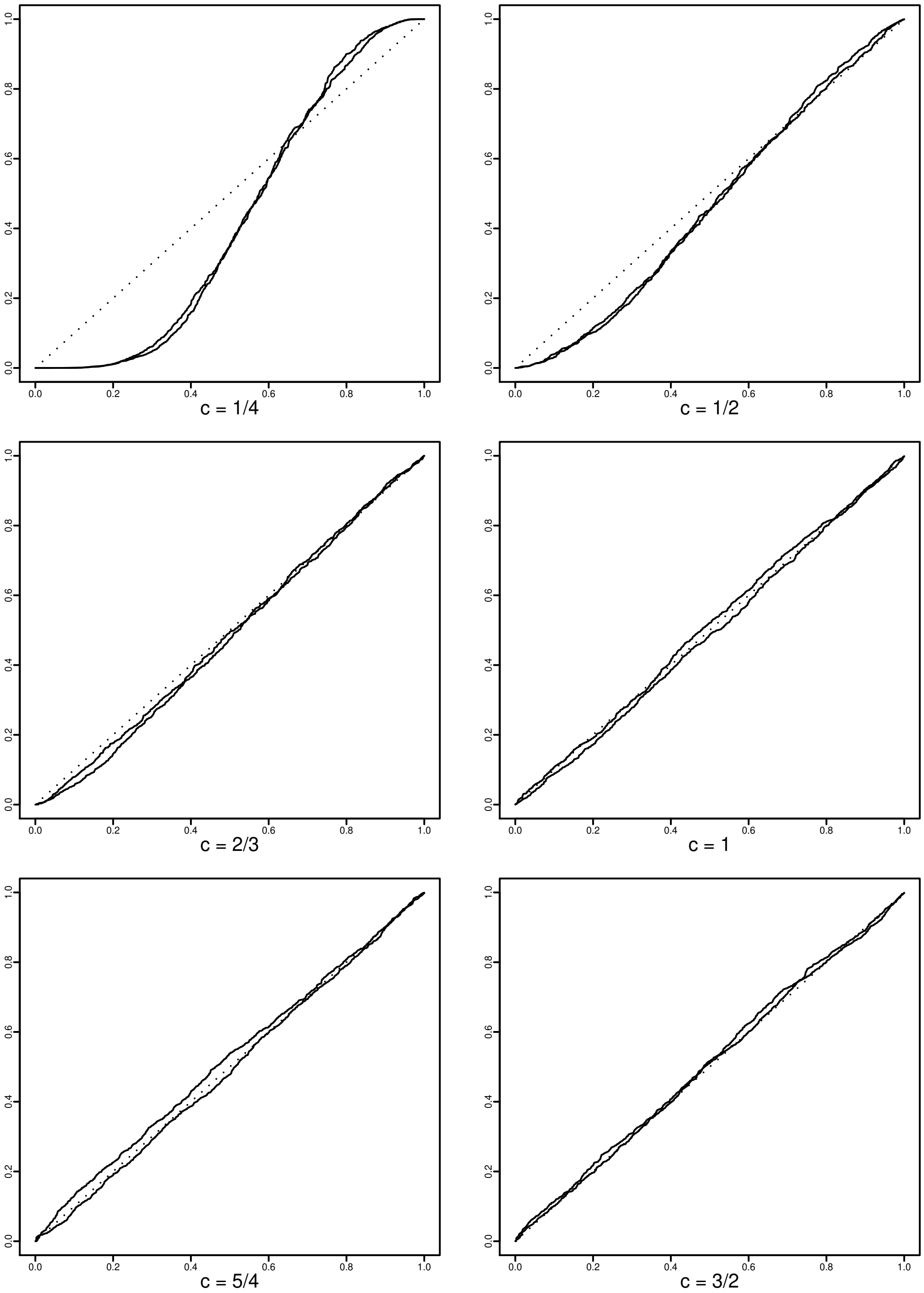


Figure 11: Empirical cumulative distributions of r^p for samples from bivariate (ie, $p = 2$) distributions. Each plot shows cumulative distributions for two samples of size 1000 drawn from two distributions that were drawn from priors of the form $a(t) = c/(1-t)$, with the indicated values for c. The dotted diagonal lines show the cumulative distribution for the uniform distribution.

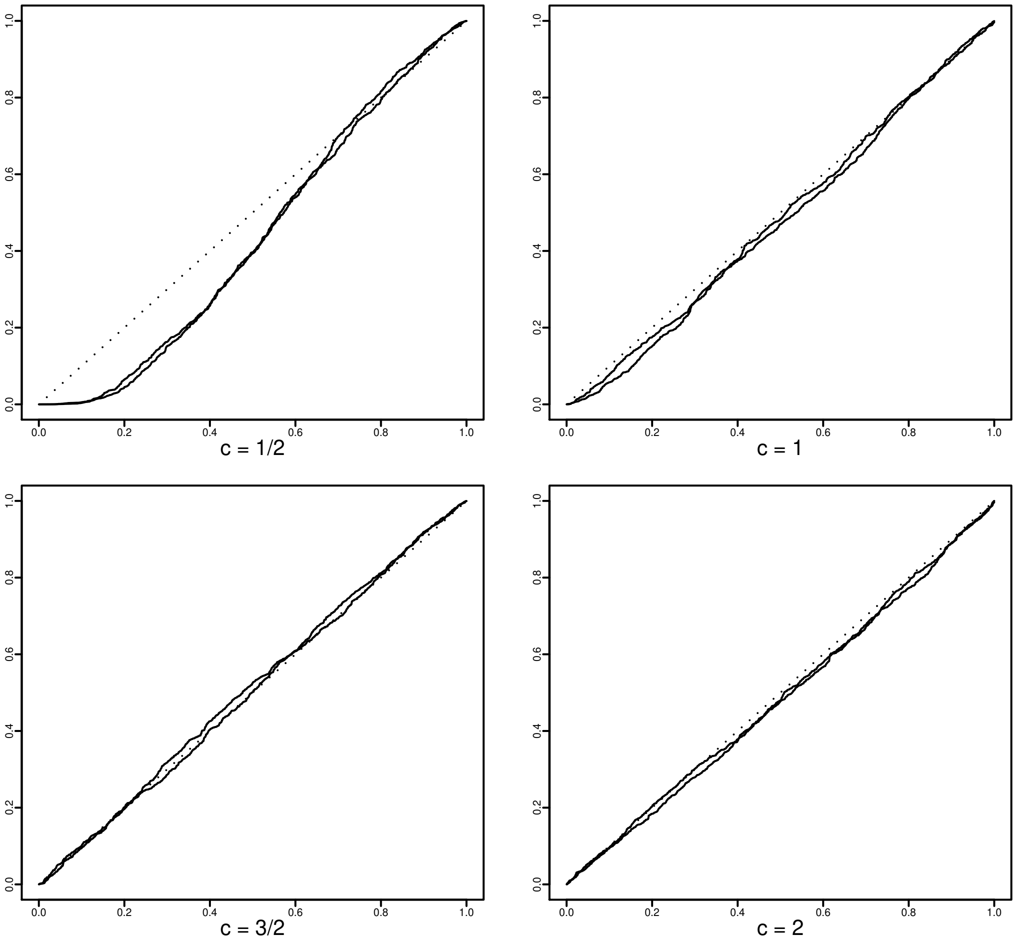
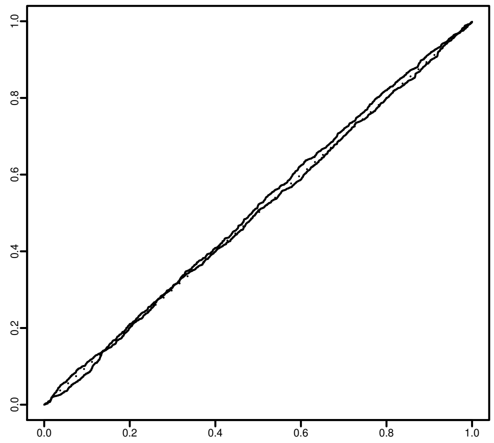


Figure 12: Empirical cumulative distributions of r^p for samples from three-dimensional (ie, $p = 3$) distributions. Each plot shows cumulative distributions for two samples of size 1000 drawn from two distributions that were drawn from priors of the form $a(t) = c/(1-t)$, with the indicated values for c. The dotted diagonal lines show the cumulative distribution for the uniform distribution.

Figure 13: Empirical cumulative distributions of r^p for two bivariate samples of size 1000 drawn from two distributions that were drawn from the prior with $a(t) = (1/2) + (1/200)/(1-t)^2$. The dotted diagonal line shows the cumulative distribution for the uniform distribution.



also indicate (somewhat less clearly) that the distributions produced with $c = 1/3$ and $c = 3/8$ are not absolutely continuous. For $c = 1/2$, $c = 5/8$, and $c = 1$, no clear departure from a uniform distribution for r^p is seen, consistent with distributions produced from these priors being absolutely continuous.

Figure 11 shows empirical cumulative distributions of r^p for two-dimensional data sets, again generated using $a(t) = c/(1-t)$. (The samples are those shown in Figures 6, 7, and 8 for the corresponding values of c.) When $c = 1/4$ and $c = 1/2$, the distributions of r^p are clearly non-uniform; they also appear to be non-uniform when $c = 2/3$, though somewhat less clearly. This is evidence that the priors with these values for c produce distributions that are not absolutely continuous. For $c = 1$, $c = 5/4$, and $c = 3/2$, there is no evidence that the distributions of r^p are non-uniform, consistent with these priors producing distributions that are absolutely continuous.

For three-dimensional distributions produced with $a(t) = c/(1-t)$, Figure 12 shows that the distribution of r^p is non-uniform for $c = 1/2$ and likely for $c = 1$, but there is no evidence of non-uniformity for $c = 3/2$ and $c = 2$.

From these results, I conjecture that Dirichlet diffusion tree priors for p-dimensional distributions with $a(t) = c/(1-t)$ produce distributions that are not absolutely continuous when $c < p/2$, but which are absolutely continuous when $c > p/2$. That $c = p/2$ is an important point can be seen from the distribution of the difference between two data vectors drawn from a common distribution drawn from the prior. Due to exchangeability, we can consider these data vectors to be the first two produced by the Dirichlet diffusion tree. Conditional on the first two paths diverging at time t, the distribution of the difference between the first two data vectors will be Gaussian with mean zero and covariance $2\sigma^2(1-t)I$. The density for the divergence time is $a(t)\exp(-A(t)) = c(1-t)^{c-1}$. Integrating over t gives the unconditional density for the difference, v:

$$f(v) \;=\; \int_0^1 c(1-t)^{c-1}\;(4\pi\sigma^2(1-t))^{-p/2}\,\exp(-|v|^2/4\sigma^2(1-t))\,dt \tag{10}$$

There is a singularity in the density at $v = 0$ if $c \le p/2$, as the integral will then diverge. Since the density above integrates over possible distributions, the presence of this singularity does not directly say anything about a single distribution drawn from the prior, but in combination with the empirical evidence, it suggests that $c = p/2$ is the critical point with regard to absolute continuity.

Note that since the Dirichlet diffusion tree procedure generates the p variables independently once the latent tree structure has been generated, the p-dimensional distributions produced by priors with $c > 1/2$ will, if this conjecture is true, have absolutely continuous marginal distributions, even when the joint distribution is not absolutely continuous.

I conjecture that, in contrast, all Dirichlet diffusion tree priors using $a(t) = b + c/(1-t)^2$ with $c > 0$ produce absolutely continuous distributions. This is supported by the uniformity of r^p see in Figure 13, which is based on the two samples from such a prior shown in Figure 9.

5 Relationships to other priors for distributions

In this Section, I will show how Dirichlet diffusion trees include as special cases certain priors based on Dirichlet processes, and discuss how they compare with priors based on Polya trees, which are another generalization of Dirichlet processes.

5.1 Relationship to Dirichlet processes

A Dirichlet process prior with a Gaussian base probability measure can be viewed as a degenerate Dirichlet diffusion tree prior. The equivalent of a Dirichlet process with concentration parameter α and base probability measure that is Gaussian with mean zero and covariance $\sigma^2 I$ is obtained using a Dirichlet diffusion tree with variance σ^2 and a divergence function that is zero except for an infinite peak of mass $\log(1+\alpha)$ at $t = 0$. This divergence function can be specified implicitly by the following inverse cumulative divergence function:

$$A^{-1}(e) \;=\; \begin{cases} 0 & \text{if } e < \log(1+\alpha) \\ 1 & \text{if } e \geq \log(1+\alpha) \end{cases} \tag{11}$$

As discussed in Section 2.4, A^{-1} is all that is needed to generate data sets from the prior, and hence is also sufficient to define the prior theoretically. Since $A^{-1}(e)$ becomes one at a finite value for e, the distributions produced are discrete, a well-known property of the Dirichlet process.

The restriction to Gaussian base probability measures in this construction is not severe, since the equivalent of a Dirichlet process with any other base probability measure can easily be obtained by simply transforming the data values.

5.2 Relationship to Dirichlet process mixtures

Dirichlet process mixture models of one simple form are special cases of Dirichlet diffusion tree models, in which the divergence function, $a(t)$, has infinite peaks at $t = 0$ and at one later time, and is zero elsewhere.

This class of simple Dirichlet process mixtures has component distributions that are Gaussian with unknown mean, μ, and known covariance, $\sigma_x^2 I$. The prior distribution for the mean of a component is Gaussian with mean zero and covariance $\sigma_\mu^2 I$. The Dirichlet process concentration parameter will be denoted by α.

The Dirichlet diffusion tree equivalent of such a model is obtained by letting $\sigma^2 = \sigma_\mu^2 + \sigma_x^2$ and defining $a(t)$ implicitly by means of A^{-1} as follows:

$$A^{-1}(e) \;=\; \begin{cases} 0 & \text{if } e < \log(1+\alpha) \\ \sigma_\mu^2/\sigma^2 & \text{if } e \geq \log(1+\alpha) \end{cases} \tag{12}$$

Figure 14 shows a data set generated from this prior, along with the tree structure underlying the first twenty data points. This tree structure has only two layers, unlike the

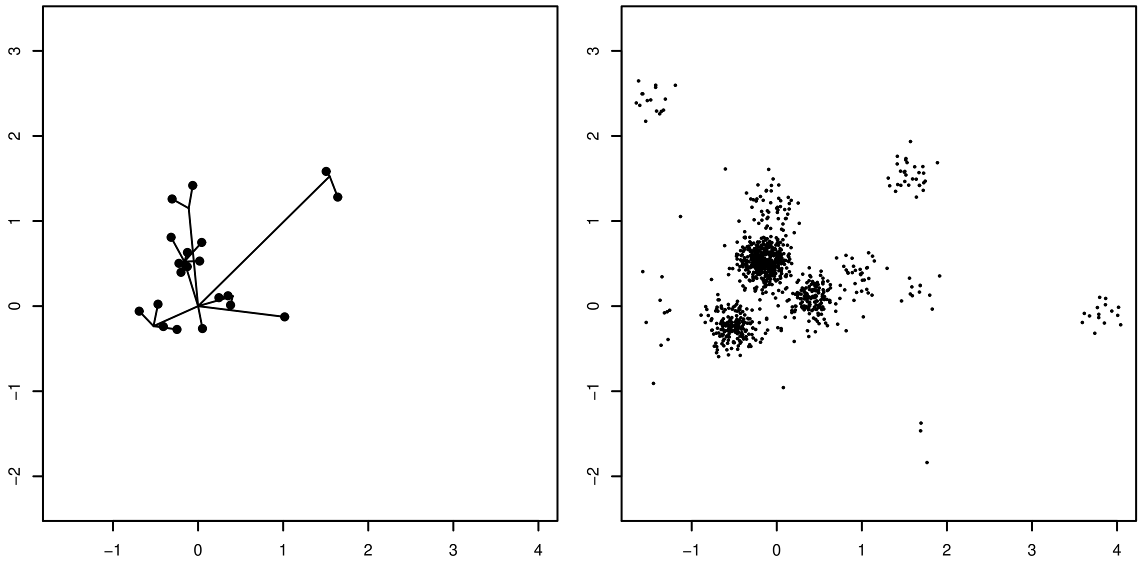


Figure 14: Generation of a two-dimensional data set from the Dirichlet diffusion tree prior that is equivalent to the simple Dirichlet process mixture with $\sigma_\mu = 0.99$, $\sigma_x = 0.14$, and $\alpha = 1$. The plot on the left shows the first twenty data points generated along with the underlying tree structure. The plot on the right shows 1000 data points obtained by continuing the procedure beyond the twenty points shown on the left.

more general Dirichlet diffusion tree model illustrated in Figure 5. The effect of this is that components of the mixture do not have any internal structure, or looked at another way, that components have no tendency to cluster together, apart from their tendency to cluster near the prior mode.

In more general Dirichlet process mixture models, the variance of the component distributions might also vary. A somewhat similar effect arises in general Dirichlet diffusion tree models as a result of varying divergence times. However, to obtain an exact equivalent of these Dirichlet process mixture models, the Dirichlet diffusion tree model would need to be generalized in some way, for instance by allowing the variance of the diffusion to vary along a path, as is briefly discussed below in Section 6.

5.3 Relationship to Polya tree priors

Polya trees (Ferguson 1974; Maudling, Sudderth, and Williams 1992; Lavine 1992, 1994) resemble Dirichlet diffusion trees in the way that data points are generated by following paths down a tree, and in the way that paths to new data points have a tendency to follow previous paths, but with some probability of divergence. A fairly minor difference is that the depth of a node in a Polya tree is given by an integer from one up, rather than by a real valued time in $[0, 1]$, as for Dirichlet diffusion trees. A more fundamental difference is that the structure of a Polya tree is predetermined. For example, a Polya tree for distributions

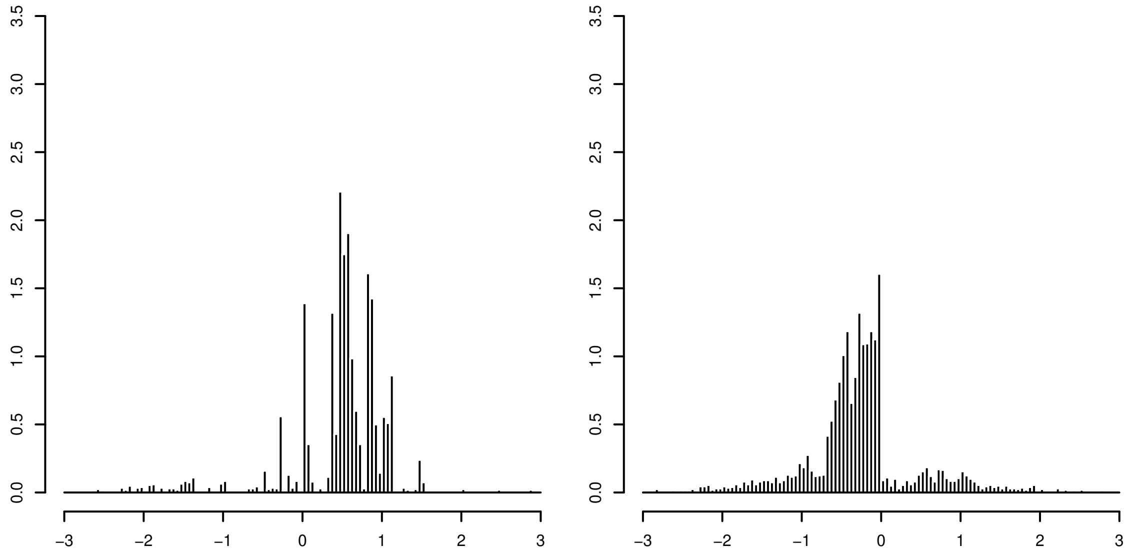
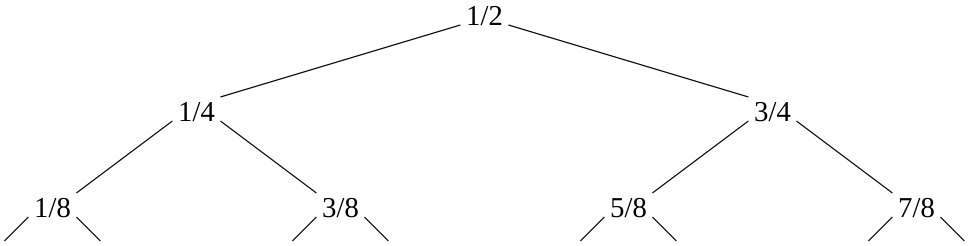


Figure 15: Probability density histograms of data sets of size 4000 generated from Polya tree priors. The data set on the left was generated from a prior with $\alpha(k) = 1$; the data set on the right was generated with $\alpha(k) = k^2$.

over $(0, 1)$ might look as shown below (similar to Figure 1 of Ferguson 1974), where data values are shown at each node:



The tree extends downwards indefinitely, covering the interval $(0, 1)$ more and more finely.

A distribution over distributions is defined using this tree by giving a function, $\alpha(k)$, that controls the probabilities that a path down the tree to a data point will follow the left or the right branch at level k in the tree. One formulation is that at each node, the probability of going left is drawn from the beta$(\alpha(k), \alpha(k))$ distribution, independently for each of the infinite number of nodes. Equivalently, a set of data points drawn from a distribution drawn from the prior can be generated by following paths from the top of the tree, with the probability of the ith path going left from the node it reached at level k being $(\ell + \alpha(k))/(i - 1 + 2\alpha(k))$, where ℓ is the number of previous paths that went left at that node.

Figure 15 shows two data sets generated using Polya trees with $\alpha(k) = 1$ and $\alpha(k) = k^2$. The points from $(0, 1)$ were transformed by applying the inverse of the Gaussian cumulative distribution function, in order to produce distributions comparable to those produced using Dirichlet diffusion trees.

The distribution on the left of Figure 15, produced by the Polya tree prior with $\alpha(k) = 1$, resembles the distributions shown in Figure 4, produced by the Dirichlet diffusion tree prior with $a(t) = (1/3)/(1-t)$, showing a similar hierarchical structure. With probability one, a distribution drawn from the Polya tree prior with $\alpha(k) = 1$ will be continuous but not absolutely continuous, since Polya tree priors with $\alpha(k)$ constant are in the class of priors over distribution functions for which this was shown by Dubins and Freedman (1967).

The distribution on the right of Figure 15, produced with $\alpha(k) = k^2$, somewhat resembles the distributions shown in Figure 3, which were drawn from a Dirichlet diffusion tree with $a(t) = 1/(1-t)$. A striking difference is apparent, however, in that the density shown on the right of Figure 15 has a large discontinuity at zero, and smaller discontinuities are present throughout the density. As discussed by Ferguson (1974), these discontinuities are a consequence of the way the interval is partitioned by the Polya tree. Although when $\alpha(k) = k^2$ the distributions produced are absolutely continuous with probability one, the densities for these distributions have discontinuities at all division points, which are countably infinite in number.

We therefore see that Polya tree priors resemble Dirichlet diffusion tree priors in some important respects, but the way the data space is partitioned by their fixed tree structure has undesirable consequences. Dirichlet diffusion trees avoid this problem. On the other hand, the fixed tree structure makes prediction for new data points much easier for Polya trees that for Dirichlet diffusion trees, for which prediction requires integrating over possible tree structures underlying the data.

6 Discussion

In this paper I have shown that Dirichlet diffusion trees can be used to define a variety of priors for distributions, whose properties vary with the choice of divergence function. Further work is needed to clarify the properties of these priors, and in particular to confirm or refute the conjectures I have made concerning absolute continuity of the distributions produced.

In practice, the characteristics of the unknown distribution will seldom be known exactly, and it will hence be appropriate to give higher-level prior distributions to the parameters of the divergence function — such as c for the priors of Section 3.1, or b and c for the priors of Section 3.2 — allowing the characteristics of the distribution to be inferred from the data. Alternatively, one might try to use some nonparametric model for the divergence function.

If the data were observed with some amount of noise, or the data values were rounded, it would be appropriate to regard the Dirichlet diffusion tree as defining the prior for the distribution of the unrounded, noise-free values, not for the data actually observed. Somewhat similarly, when the data is categorical, unobserved latent values can be introduced that determine the probabilities of the observed data (eg, via a logistic or probit model). The distribution of these vectors of latent values could then be given a Dirichlet diffusion tree prior, indirectly defining a prior for the joint distribution of the categorical values.

One could also envision more elaborate transformations from the data whose distribution has a Dirichlet diffusion tree prior to the observed data. For instance, univariate data, z, could be modeled as $z = x\exp(y)$, with the distribution of (x, y) having a bivariate Dirichlet diffusion tree prior. The spread of a cluster of z values will then depend on the magnitude of the latent value y for that cluster. More fundamentally, the Dirichlet diffusion tree could include latent values that affect the diffusion or divergence process itself — for instance, one could let $a(t) = \exp(y)/(1-t)$, with y being one of the values that is changing with t via the diffusion process, or use a time-varying diffusion variance, $\sigma(t) = \exp(y)$, with y again being an additional variable changing by diffusion. However, extensions of this sort undermine the possibility of using a finite representation of the relevant aspects of the tree underlying a finite data set (as in Figure 2); as a result, computations would probably be possible only by using some approximation based on a discretization of time.

The finite representation of a Dirichlet diffusion tree shown in Figure 2 should allow computations to be performed in a reasonably efficient manner using Markov chain Monte Carlo methods (see, for instance, Gilks, *et al* 1996). A Markov chain would need to be defined that converges to the posterior distribution over tree structures and over times and locations of the non-leaf nodes in the tree. It should be possible to do Metropolis updates for the times at which paths diverge, and these updates could be defined so as to change the tree structure when they result in the time order of nodes changing. The positions of non-leaf nodes could be updated by Gibbs sampling. Alternatively, these positions could be efficiently integrated over, as done by Williams (2000), eliminating them from the state of the Markov chain. If the Dirichlet diffusion tree does not define the prior for the distribution of the observed data, but rather for underlying latent values, these latent values would also be updated as part of the Markov chain, as would any unknown parameters of the divergence function.

Acknowledgements

This research was supported by the Natural Sciences and Engineering Research Council of Canada and the Institute for Robotics and Intelligent Systems.

References

Antoniak, C. E. (1974) "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems", *Annals of Statistics*, vol. 2, pp. 1152-1174.

Bernardo, J. M. and Smith, A. F. M. (1994) *Bayesian Theory*, Chichester: John Wiley.

Billingsley, P. (1995) *Probability and Measure*, third edition, New York: John Wiley.

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

Bush, C. A. and MacEachern, S. N. (1996) "A semiparametric Bayesian model for randomised block designs", *Biometrika*, vol. 83, pp. 275-285.

Dubins, L. E. and Freedman, D. A. (1967) "Random distribution functions", in L. M. LeCam and J. Neyman (editors) *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics*, Volume II, Part 1, pp. 183-214, University of California Press.

Edwards, A. W. F. (1970) "Estimation of the branch points of a branching diffusion process" (with discussion), *Journal of the Royal Statistical Society B*, vol. 32, pp. 155-174.

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

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

Ferguson, T. S. (1974) "Prior distributions on spaces of probability measures", *Annals of Statistics*, vol. 2, pp. 615-629.

Ferguson, T. S. (1983) "Bayesian density estimation by mixtures of normal distributions", in H. Rizvi and J. Rustagi (editors) *Recent Advances in Statistics*, pp. 287-303, New York: Academic Press.

Gilks, W. R., Richardson, S., and Spiegelhalter, D. J. (1996) *Markov Chain Monte Carlo in Practice*, London: Chapman and Hall.

Jain, A. K. (1988) *Algorithms for Clustering Data*, Englewood Cliffs, New Jersey: Prentice Hall.

Jain, S. and Neal, R. M. (2000) "A split-merge Markov chain Monte Carlo procedure for the Dirichlet process mixture model", Technical Report No. 2003, Dept. of Statistics, University of Toronto, 32 pages.

Lavine, M. (1992) "Some aspects of Polya tree distributions for statistical modeling", *Annals of Statistics*, vol. 20, pp. 1222-1235.

Lavine, M. (1994) "More aspects of Polya tree distributions for statistical modeling", *Annals of Statistics*, vol. 22, pp. 1161-1176.

MacEachern, S. N. and Müller, P. (1998) "Estimating mixture of Dirichlet process models", *Journal of Computational and Graphical Statistics*, vol. 7, pp. 223-238.

Mauldin, R. D., Sudderth, W. D., and Williams, S. C. (1992) "Polya trees and random distributions", *Annals of Statistics*, vol. 20, pp. 1203-1221.

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

Walker, S. G., Damien, P., Purushottam, W. L., and Smith, A. F. M. (1999) "Bayesian nonparametric inference for random distributions and related functions" (with discussion), *Journal of the Royal Statistical Society B*, vol. 61, pp. 485-527.

Williams, C. K. I. W. (2000) "A MCMC approach to hierarchical mixture modelling", in S. A. Solla, T. K. Leen, and K-R. Muller (editors), *Advances in Neural Information Processing Systems 12*, MIT Press.