Density Modeling and Clustering Using Dirichlet Diffusion Trees

Radford M. Neal

Bayesian Statistics 7, 2003, pp. 619-629.

Presenter: Ivo D. Shterev

Outline

- Motivation.
- Data points generation.
- Probability of generating Dirichlet diffusion trees (DDT).
- Exchangeability.
- Data points generation from the DDT structure.
- Examples.
- Testing for absolute continuity.
- Simple relationships to other processes.
- Discussion.

Motivation

- Dirichlet Process (DP) produces distributions which are discrete with probability one, hence unsuitable for density modeling.
- Convolution of the distribution with continuous kernel can produce densities, i.e. using DP mixtures with countably infinite number of components.
- Parameters of one DP mixture component are independent of parameters of other components, because the parameter priors are independent.
- DP mixtures are inefficient when data exhibits hierarchical structure.
- Polya Trees (PT) is a generalization of DP that can produce hierarchical distributions, but these distributions have discontinuities.
- An alternative is to use DDT.

Data Points Generation

- Suppose we want to generate n data points $\boldsymbol{X} \in \mathbb{R}^p$ from DDT.
- $\boldsymbol{X}_1$ is generated from a Gaussian diffusion process.

$$\boldsymbol{X}_1(t+dt) \;\; = \;\; \boldsymbol{X}_1(t) + \boldsymbol{N}_1(dt) \;, 0 \leq t \leq T$$

- where $\boldsymbol{N}_1(dt) \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I} dt)$, σ^2 is a parameter of the diffusion process, dt is infinitely small, and T is the period.
- It can be seen that $\boldsymbol{X}_1(t+dt) \sim \mathcal{N}(\boldsymbol{0}, \sigma^2 \boldsymbol{I} dt)$.
- $\boldsymbol{X}_2(t)$ is generated by following a path from the origin, initially following the path of $\boldsymbol{X}_1(t)$, but diverging at some time.
- We need to introduce "divergence function" $a(t)$.
- $\boldsymbol{X}_2(t)$ will diverge during $t+dt$ with probability $P_d = a(t)dt$.
- After divergence, the remaining paths are independent.

Data Points Generation

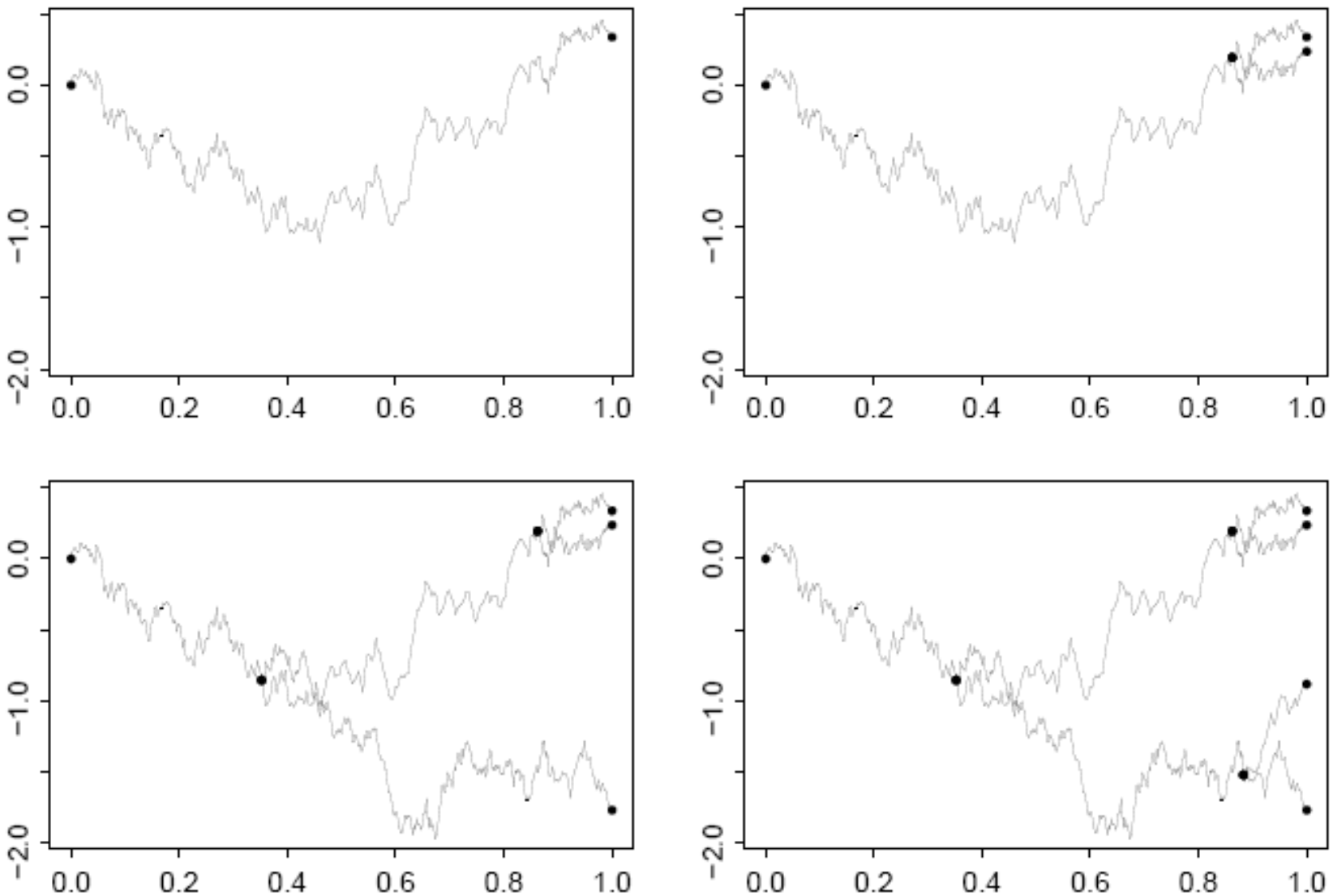
- $\boldsymbol{X}_3(t)$ initially follows $\boldsymbol{X}_1(t)$ and $\boldsymbol{X}_2(t)$.
- $\boldsymbol{X}_3(t)$ can diverge before $\boldsymbol{X}_1(t)$ and $\boldsymbol{X}_2(t)$ have diverged, with $P_d = \frac{a(t)dt}{2}$, or follow $\boldsymbol{X}_1(t)$ or $\boldsymbol{X}_2(t)$ with probability $P_f = 0.5$.
- If $\boldsymbol{X}_3(t)$ follows $\boldsymbol{X}_1(t)$ or $\boldsymbol{X}_2(t)$, we have $P_d = a(t)dt$.
- Generally

$$\begin{aligned} P_d &= \frac{a(t)dt}{m_i} \\ i_f &= \arg\max_i m_i \text{ , with probability } P_f \propto m_i \end{aligned}$$

- where m_i is the number of previous points following the ith path.
- Once divergence occurs, the new path moves independently from previous paths.

Data Points Generation

- Examples for $n = 4$, $p = 1$, $T = 1$.



Probability of Generating DDT

- Probability of no divergence between times t and s ($s < t$) over a path previously traversed m times.

$$
\begin{aligned}
P_{nd}(s,t) &= \lim_{k\to\infty} \prod_{i=0}^{k-1} \left(1 - a(s + i\frac{t-s}{k})\frac{t-s}{km}\right) \\
&= \exp\left(\sum_{i=0}^{\infty} \log\left(1 - a(s + i\frac{t-s}{k})\frac{t-s}{km}\right)\right) \\
&= \exp\left(-\int_s^t \frac{a(u)}{m} du\right) \\
&= \exp\left(-\frac{A(t) - A(s)}{m}\right)
\end{aligned}
$$

- where $A(t) = \int_0^t a(u)du$ is the "cumulative divergence function".

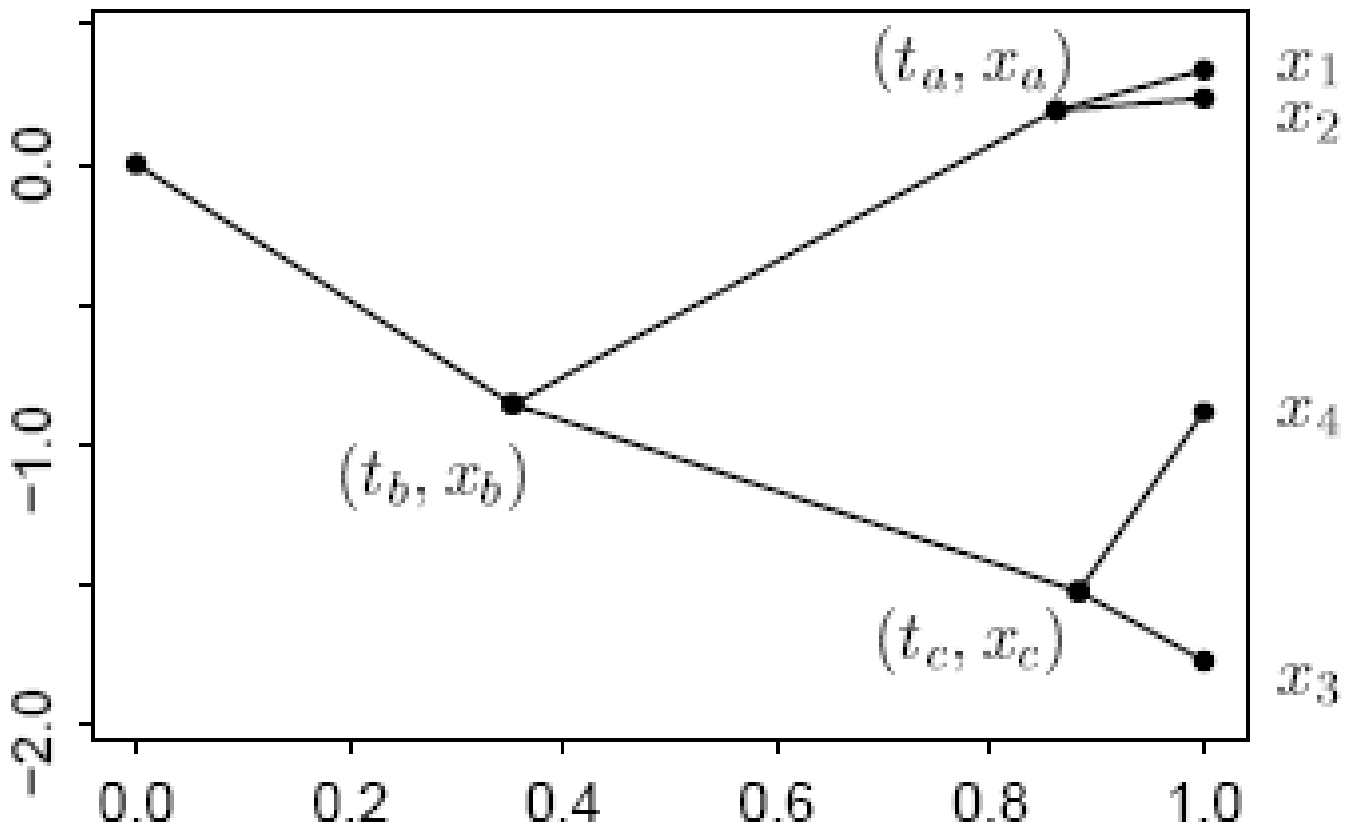
Probability of Generating DDT

- $a(t)$ plays a central role.
- Sufficient condition for divergence, i.e. $Pr\big(\boldsymbol{X}_i(t) = \boldsymbol{X}_j(t)\big) = 0$ for any $i \neq j$ is

$$\int_0^T a(t) \quad = \quad \infty$$

Probability of Generating DDT

- We don’t need all details of the paths - we need only the tree structure and divergence times.



Probability of Generating DDT

- Probability (density) of obtaining the tree structure and divergence times (also called "tree factor").

$$
\begin{aligned}
P_t &= \exp\big(-A(t_a)\big)a(t_a) \text{ , second path} \\
&\times \exp\big(-\frac{A(t_b)}{2}\big)\frac{a(t_b)}{2} \text{ , third path} \\
&\times \exp\big(-\frac{A(t_b)}{3}\big)\frac{1}{3}\exp\big(A(t_b)-A(t_c)\big)a(t_c) \text{ , fourth path}
\end{aligned}
$$

- "Data factor" is given as

$$
\begin{aligned}
DF &= \mathcal{N}(\boldsymbol{x}_b, \mathbf{0}, \sigma^2\boldsymbol{I}t_b)\mathcal{N}\big(\boldsymbol{x}_a, x_b, \boldsymbol{I}(t_a-t_b)\big)\mathcal{N}\big(\boldsymbol{x}_1, \boldsymbol{x}_a, \sigma^2\boldsymbol{I}(1-t_a)\big) \\
&\times \mathcal{N}\big(\boldsymbol{x}_2, \boldsymbol{x}_a, \sigma^2\boldsymbol{I}(1-t_a)\big)\mathcal{N}\big(\boldsymbol{x}_c, \boldsymbol{x}_b, \sigma^2\boldsymbol{I}(t_c-t_b)\big) \\
&\times \mathcal{N}\big(\boldsymbol{x}_3, \boldsymbol{x}_c, \sigma^2\boldsymbol{I}(1-t_c)\big)\mathcal{N}\big(\boldsymbol{x}_4, \boldsymbol{x}_c, \sigma^2\boldsymbol{I}(1-t_c)\big)
\end{aligned}
$$

- where $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{x}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$, with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$.

Exchangeability

- We can rewrite the "tree factor" as

$$\begin{aligned} P_t &= \exp\big(-A(t_b)\big)\exp\big(-\frac{A(t_b)}{2}\big)\frac{a(t_b)}{2}\exp\big(-\frac{A(t_b)}{3}\big)\frac{1}{3} \\ &\times \exp\big(A(t_b)-A(t_a)\big)a(t_a)\exp\big(A(t_b)-A(t_c)\big)a(t_c) \end{aligned}$$

- The first term is associated with the segment $(0,\mathbf{0})-(t_b,\boldsymbol{x}_b)$. If the term remains unchanged after any reordering of the data points (paths), then we have exchangeability.

Exchangeability

- Consider the segment $(t_u, \boldsymbol{x}_u) - (t_v, \boldsymbol{x}_v)$, that was traversed by $m > 1$ paths.
- the probability that the $m-1$ paths after the first will not diverge within the segment is

$$P_{nd}^{m-1} = \prod_{i=1}^{m-1} \exp\Big(-\frac{A(t_v) - A(t_u)}{i}\Big)$$

- if $t_v = T$, this is the whole factor for this segment, otherwise, there must be some divergence at t_v.
- suppose the $i-1$ ($i > 2$) paths do not diverge at t_v, but the ith path diverges at t_v, therefore $P_d^i = \frac{a(t_v)}{i-1}$.
- suppose n_1 is the number of points following the $i-1$ paths and n_2 is the number of points following the ith path ($n_1 \geq i-1$, $n_1 + n_2 = m$).

Exchangeability

- the probability that path j ($j > i$) follows the $i-1$ paths is $P_f = \frac{c_1}{j-1}$, where ($i - 1 \leq c_1 \leq n_1 - 1$).
- the probability that path j ($j > i$) follows the ith path is $P_f = \frac{c_2}{j-1}$, where ($1 \leq c_2 \leq n_2 - 1$).
- their product for all c_1, all c_2, and all $j > i$ gives

$$
\begin{aligned}
P_d^{\text{all } j} &= \frac{\prod_{c_1=i-1}^{n_1-1} c_1 \prod_{c_2=1}^{n_2-1} c_2}{\prod_{j=i+1}^{m} (j-1)} \\
&= \frac{(n_1-1)!}{(i-2)!}(n_2-1)!\frac{(i-1)!}{(m-1)!} \\
&= (i-1)\frac{(n_1-1)!(n_2-1)!}{(m-1)!}
\end{aligned}
$$

Exchangeability

- taking the product of all probabilities associated with the segment under consideration, we obtain the "tree factor" associated with the segment, i.e.

$$
\begin{aligned}
P_t &= P_d^i P_d^{\text{all } j} P_{nd}^{m-1} \\
&= \frac{a(t_v)}{i-1}(i-1)\frac{(n_1-1)!(n_2-1)!}{(m-1)!}\prod_{i=1}^{m-1}\exp\Big(-\frac{A(t_v)-A(t_u)}{i}\Big) \\
&= \frac{a(t_v)(n_1-1)!(n_2-1)!}{(m-1)!}\prod_{i=1}^{m-1}\exp\Big(-\frac{A(t_v)-A(t_u)}{i}\Big),
\end{aligned}
$$

- which is independent of i, i.e. exchangeability.
- the above analysis does not incorporate the case when more than one path diverges at the same time, i.e. when $a(t)$ has an infinite peak, but the proof can be modified to incorporate this.

Data Points Generation from the DDT Structure

- Probability (now called "cumulative distribution function") of path i diverging at time t is

$$C(t) \quad = \quad 1 - \exp\big(- \frac{A(t)}{i-1} \big)$$

- where the assumption is that path i initially follows the previous $i-1$ paths.
- Divergence time can be generated as

$$t_d \quad = \quad C^{-1}(U) = A^{-1}\big(- (i-1)\log(T-U) \big)$$

- where $U \sim \mathcal{U}(0,T)$, and $T = 1$ in the examples.
- it is more convenient to work with $A^{-1}(\cdot)$, since it avoids the infinite peaks of $a(t)$.

Examples

- $a(t) = \frac{c}{1-t}$, where c is a constant.
- cumulative divergence function

$$\begin{aligned} A(t) &= \int_0^t a(u)du = -c\log(1-t) \\ A^{-1}(e) &= 1 - \exp(-\frac{e}{c}) \end{aligned}$$

- distributions drawn from such a prior will be continuous, since $\int_0^1 a(t) = \infty$.

Examples

- one-dimensional points

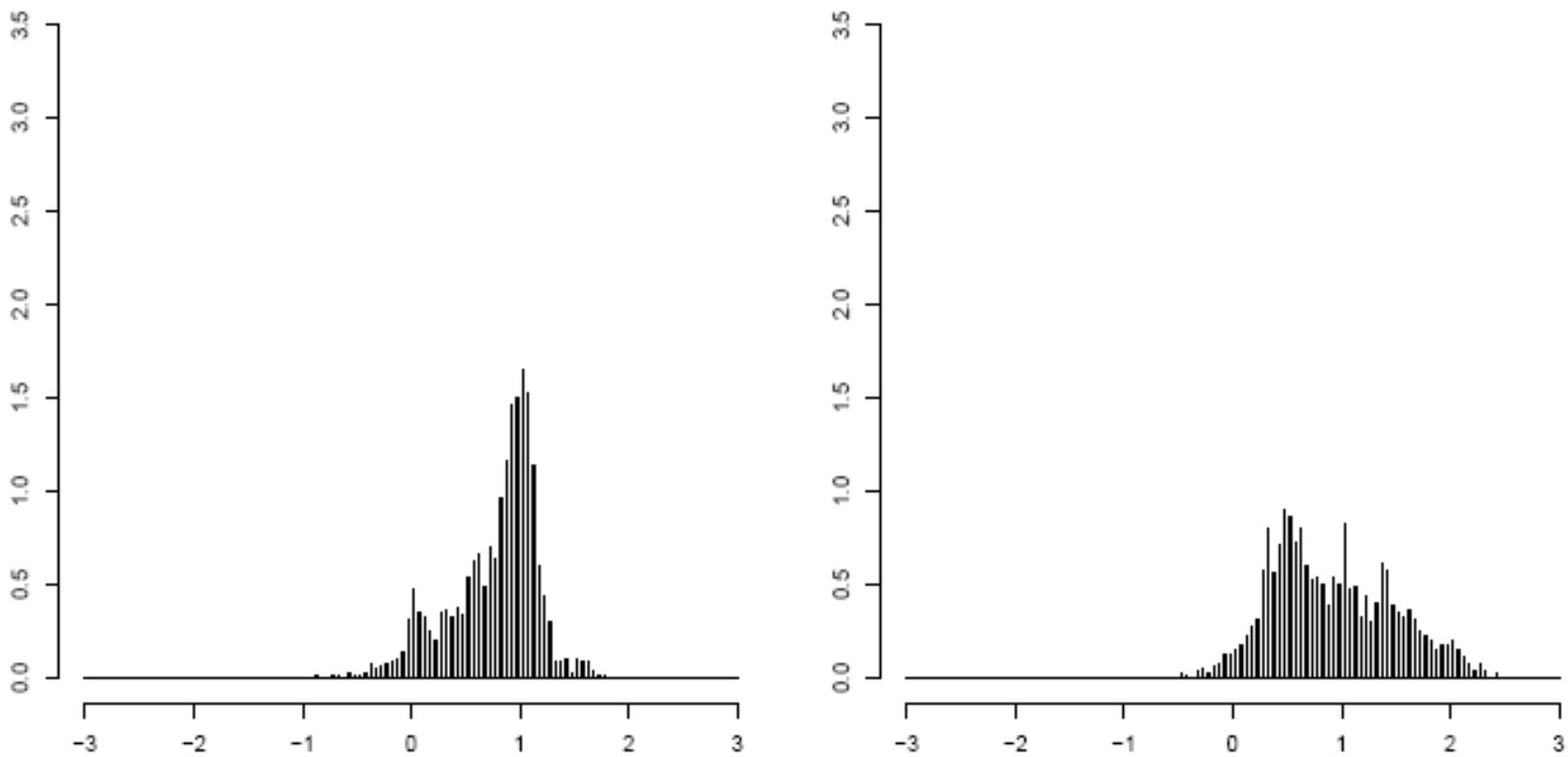


Figure 3: Probability density histograms of two data sets of 4000 points drawn from the Dirichlet diffusion tree prior with $\sigma = 1$ and $a(t) = 1/(1-t)$.

Examples

- one-dimensional points

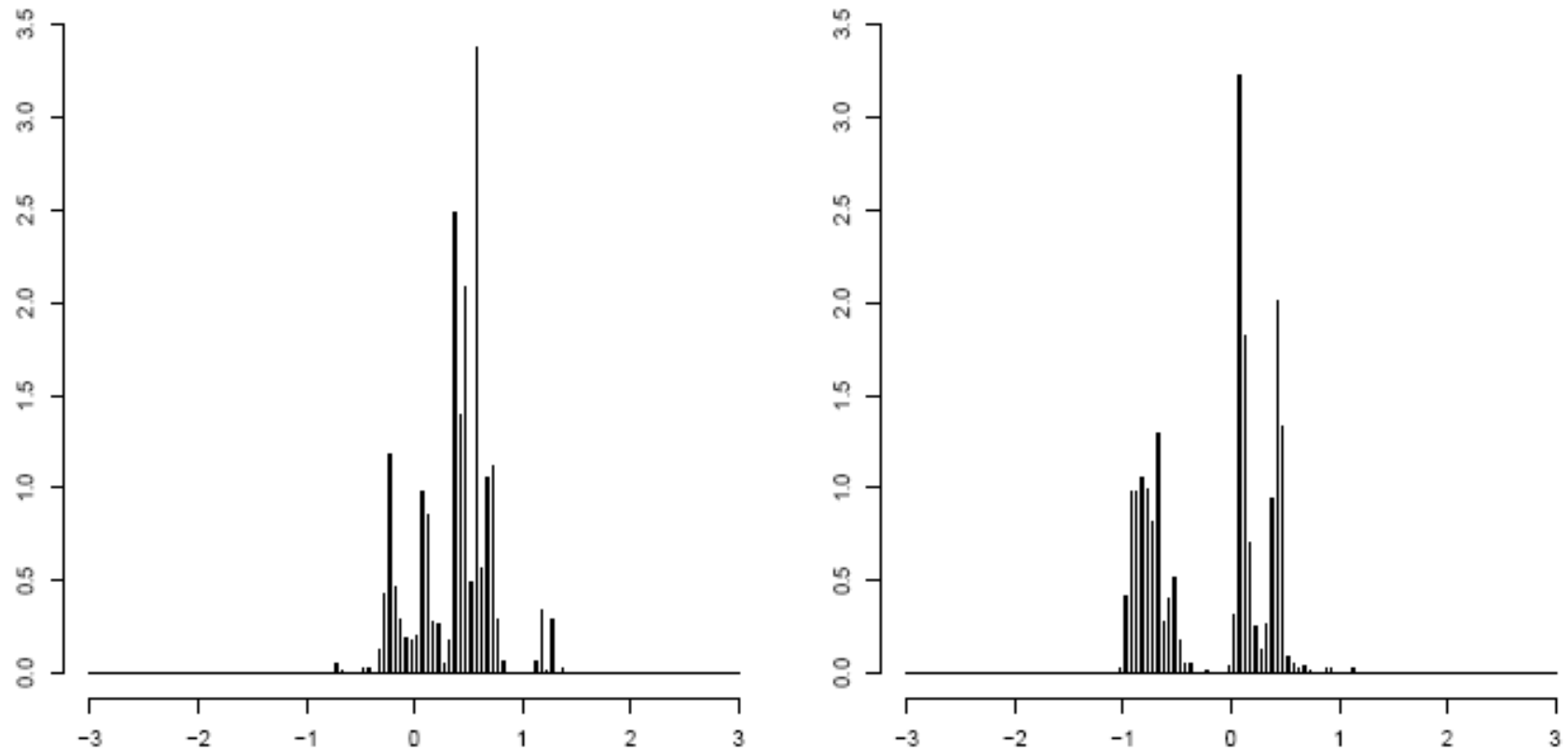


Figure 4: Probability density histograms of two data sets of 4000 points drawn from the Dirichlet diffusion tree prior with $\sigma = 1$ and $a(t) = (1/3)/(1-t)$.

Examples

- two-dimensional points

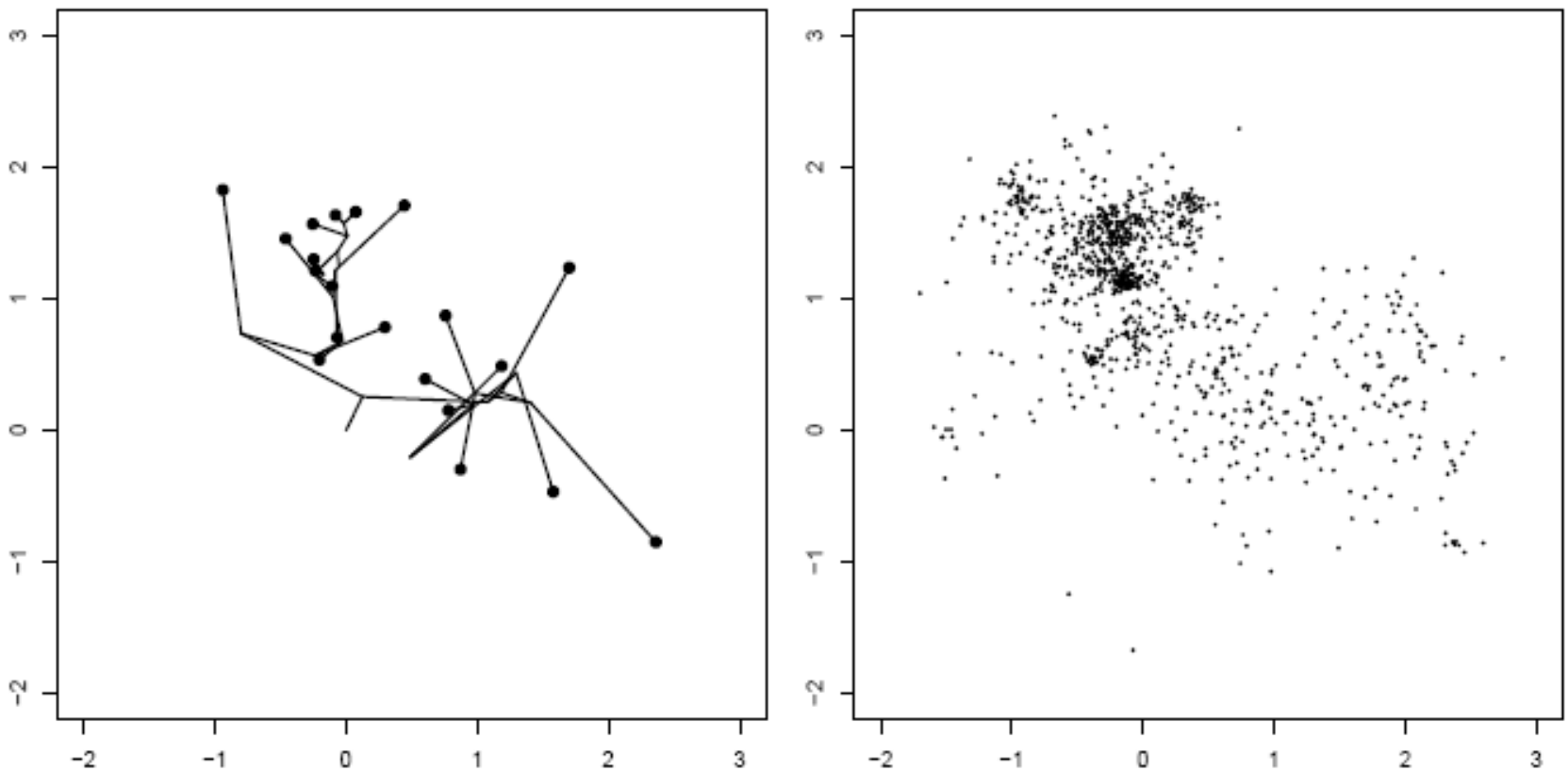


Figure 5: Generation of a two-dimensional data set from the Dirichlet diffusion tree prior with $\sigma = 1$ and $a(t) = 1/(1-t)$. The plot on the left shows the first twenty data points generated along with the underlying tree structure. The plot on the right shows 1000 data points obtained by continuing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

Examples

- two-dimensional points

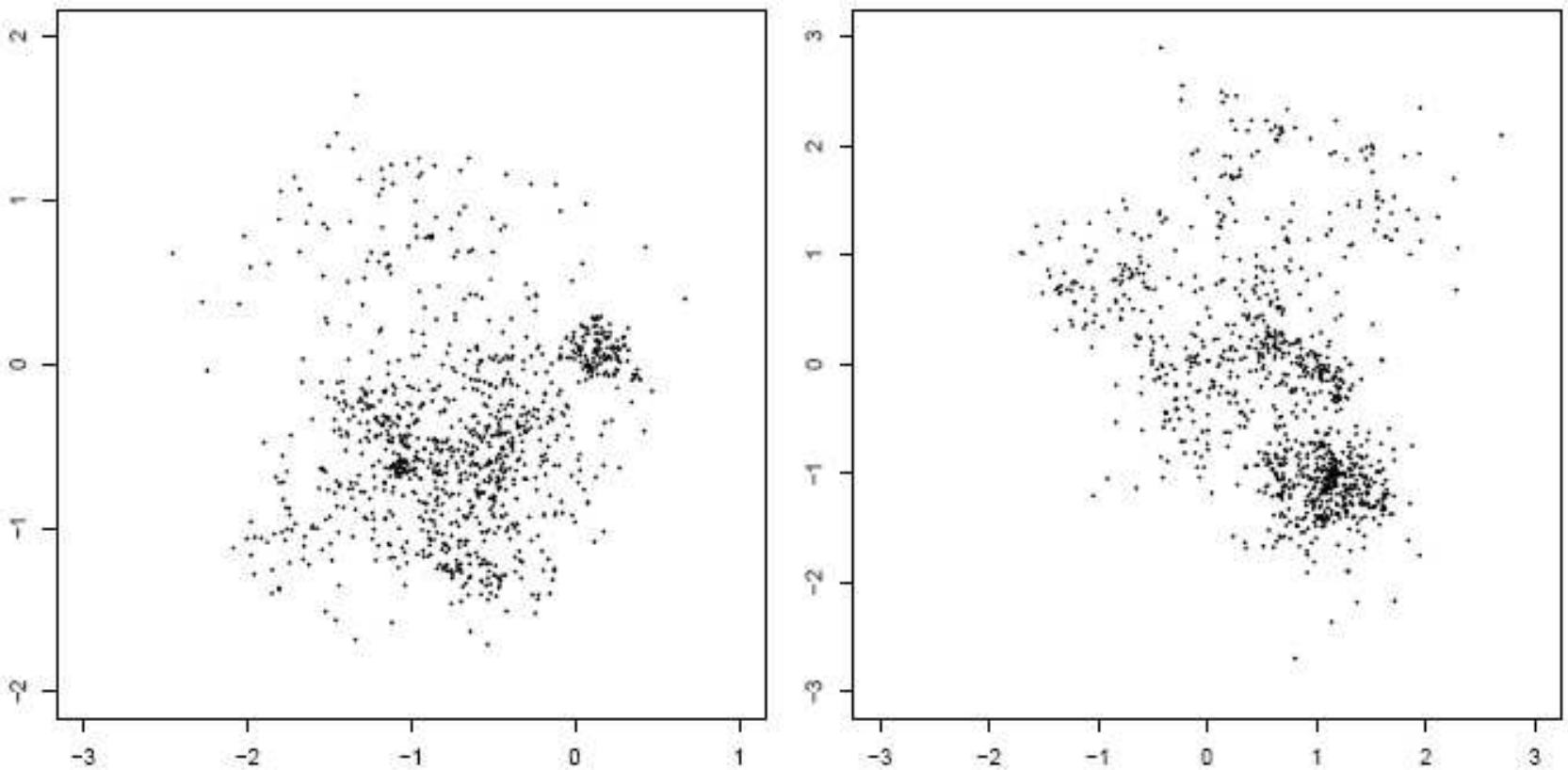


Figure 6: Two more data sets of 1000 points that were drawn independently from the same prior as for Figure 5. Note that the scales differ for the three data sets.

Examples

- two-dimensional points

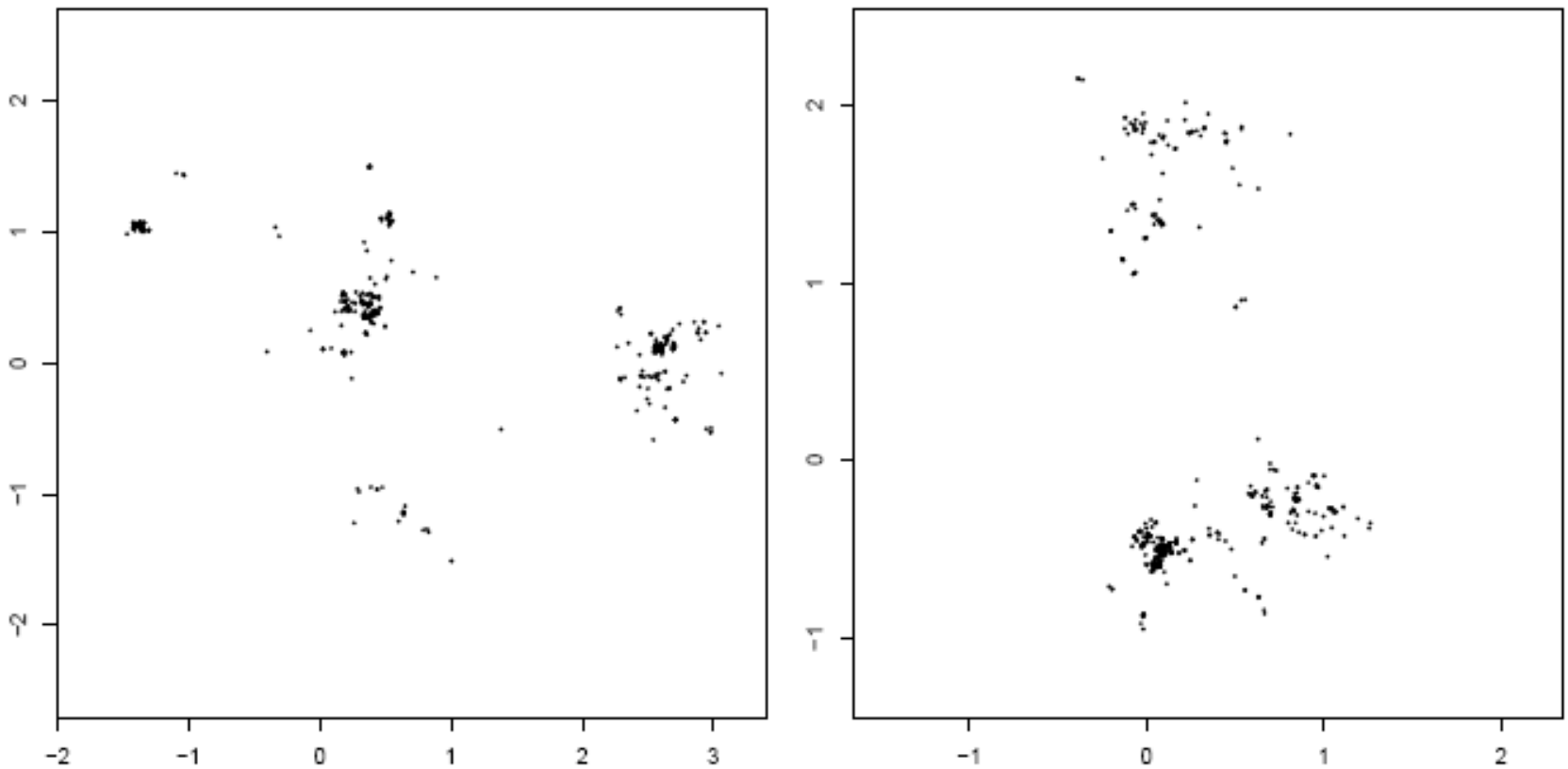


Figure 7: Two data sets of 1000 points that were drawn independently from the Dirichlet diffusion tree prior with $\sigma = 1$ and $a(t) = (1/4)/(1-t)$. Note that the scales differ for the two data sets.

Examples

- two-dimensional points

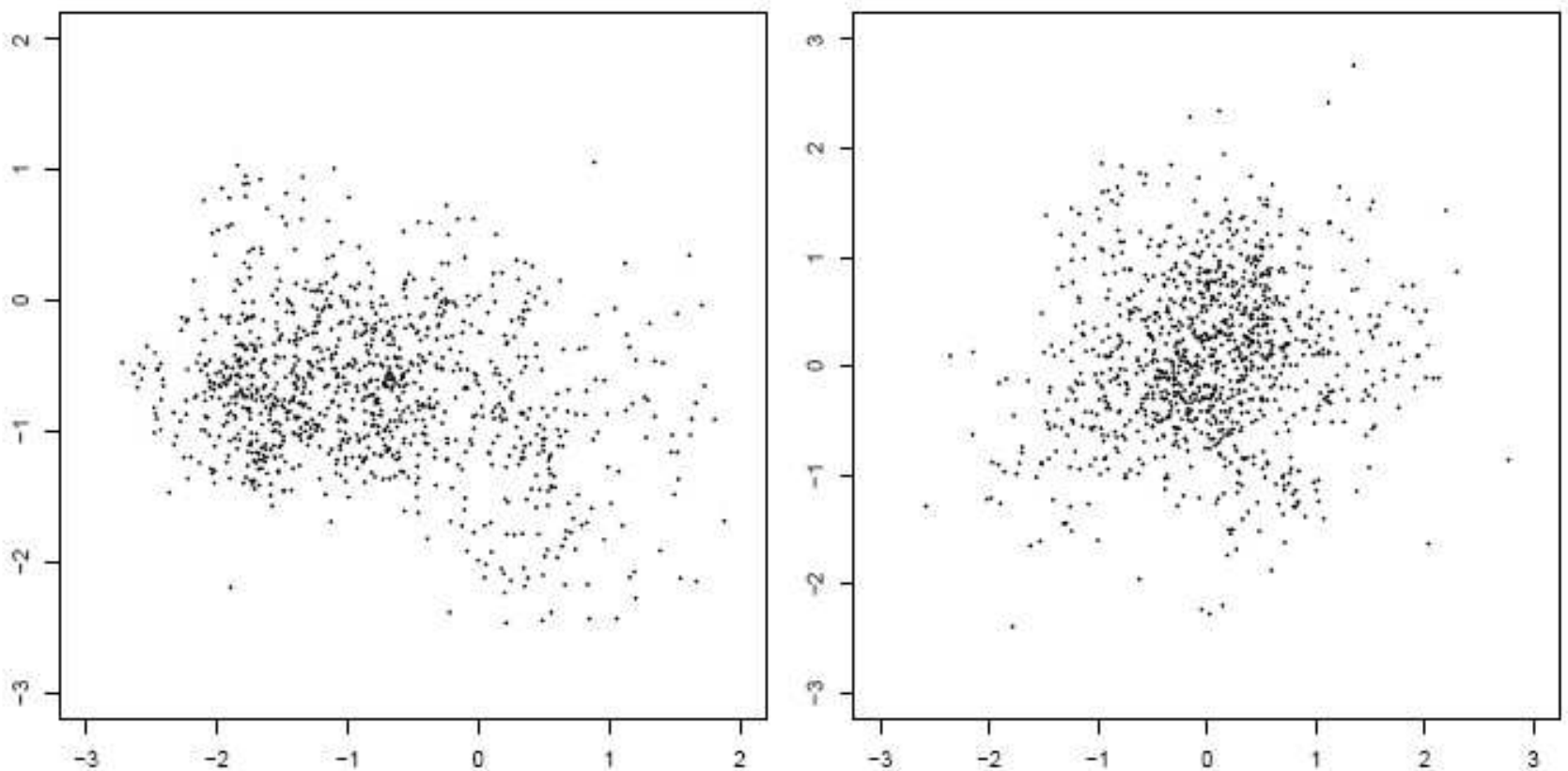


Figure 8: Two data sets of 1000 points that were drawn independently from the Dirichlet diffusion tree prior with $\sigma = 1$ and $a(t) = (3/2)/(1-t)$. Note that the scales differ for the two data sets.

Examples

- $a(t) = b + \frac{c}{(1-t)^2}$, where b and c are constants.
- cumulative divergence function

$$A(t) \;=\; bt - c + \frac{c}{1-t}$$

$$A^{-1}(t) = \begin{cases} \frac{b+c+e-\sqrt{(b+c+e)^2-4be}}{2b} & \text{if } b \neq 0 \\ 1 - \frac{c}{e+c} & \text{if } b = 0 \end{cases}$$

- good choices are $b = \frac{1}{2}$ and $c = \frac{1}{200}$, giving well-separated clusters with points smoothly distributed within each cluster.

Examples

- two-dimensional points

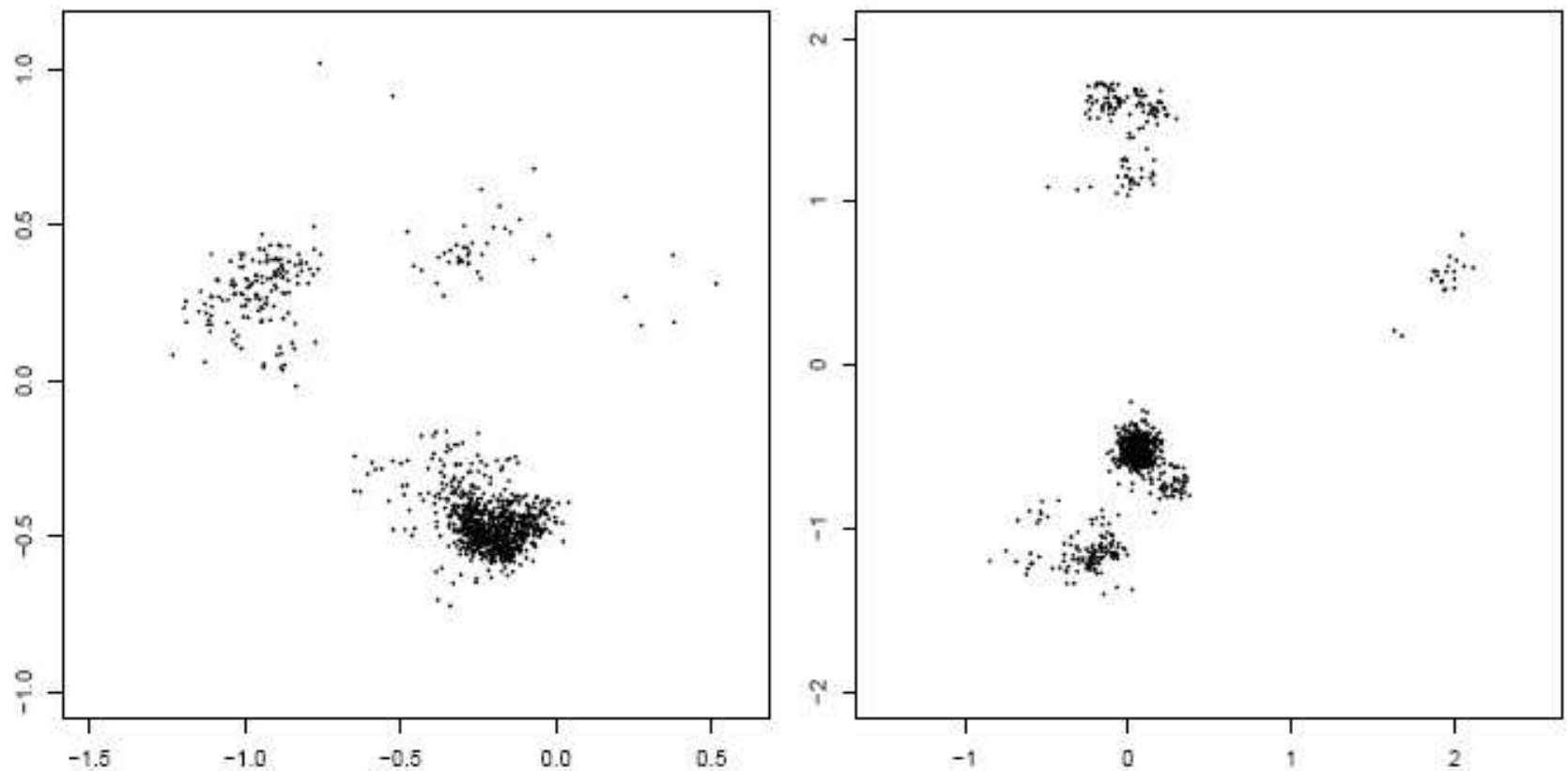


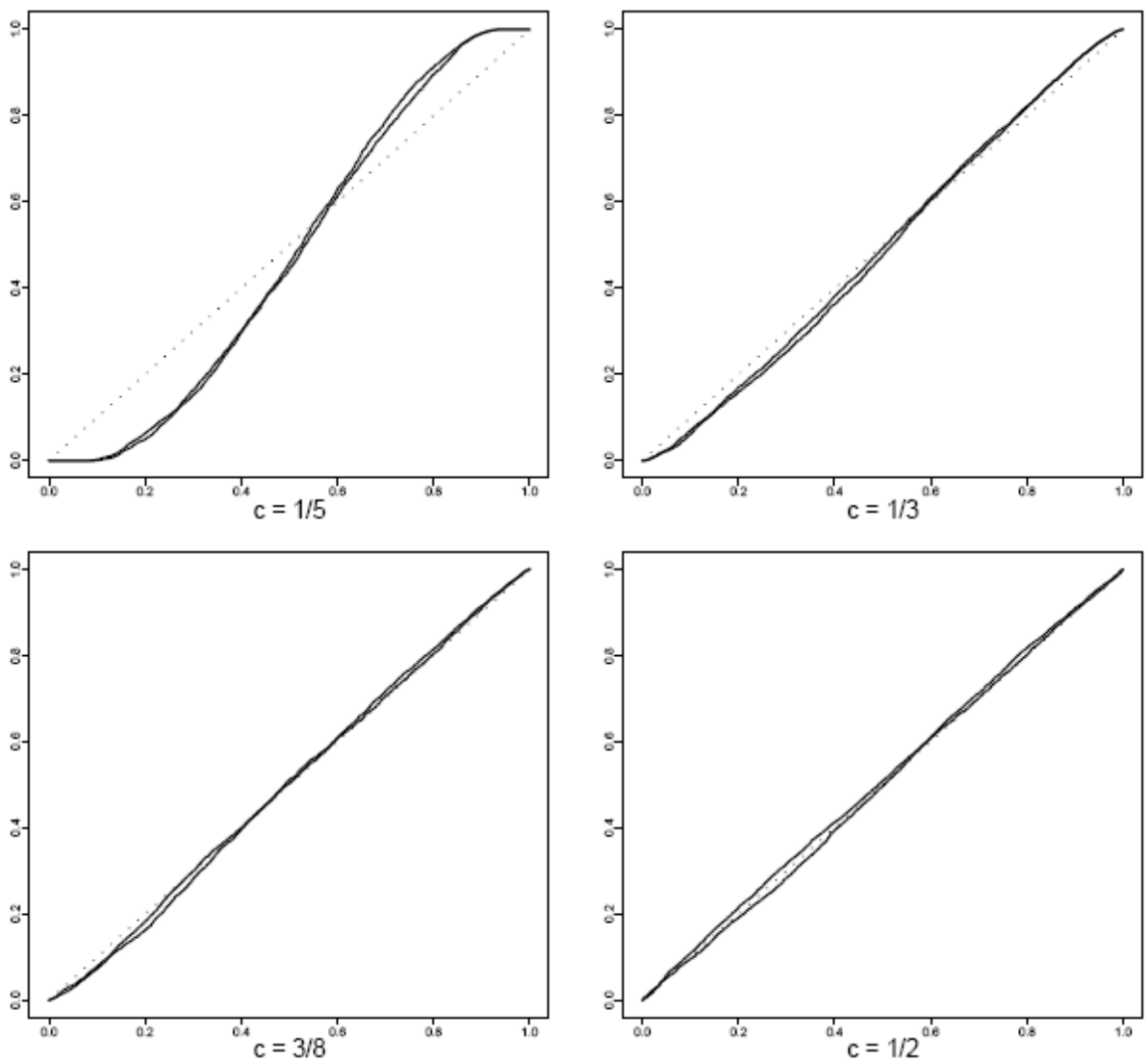
Figure 9: Two data sets of 1000 points that were drawn independently from the Dirichlet diffusion tree prior with $\sigma = 1$ and $a(t) = (1/2) + (1/200)/(1-t)^2$. Note that the scales differ for the two data sets.

Testing for Absolute Continuity

- Distributions produced from a DDT prior will be continuous (with probability one) if $\int_0^T a(t)dt = \infty$. However, this does not imply absolute continuity.
- Absolute continuity is required for distributions drawn from a DDT prior to have densities.
- Absolute continuity is tested by looking at distances to nearest neighbors in a sample from the distribution.
  - for each x, compute Euclidean distances to the two nearest neighbors.
  - compute their ratio $r < 1$.
  - to have absolute continuity, there should be $r^p \sim \mathcal{U}(0, 1)$.
- This is just an empirical test, not a rigorous proof.

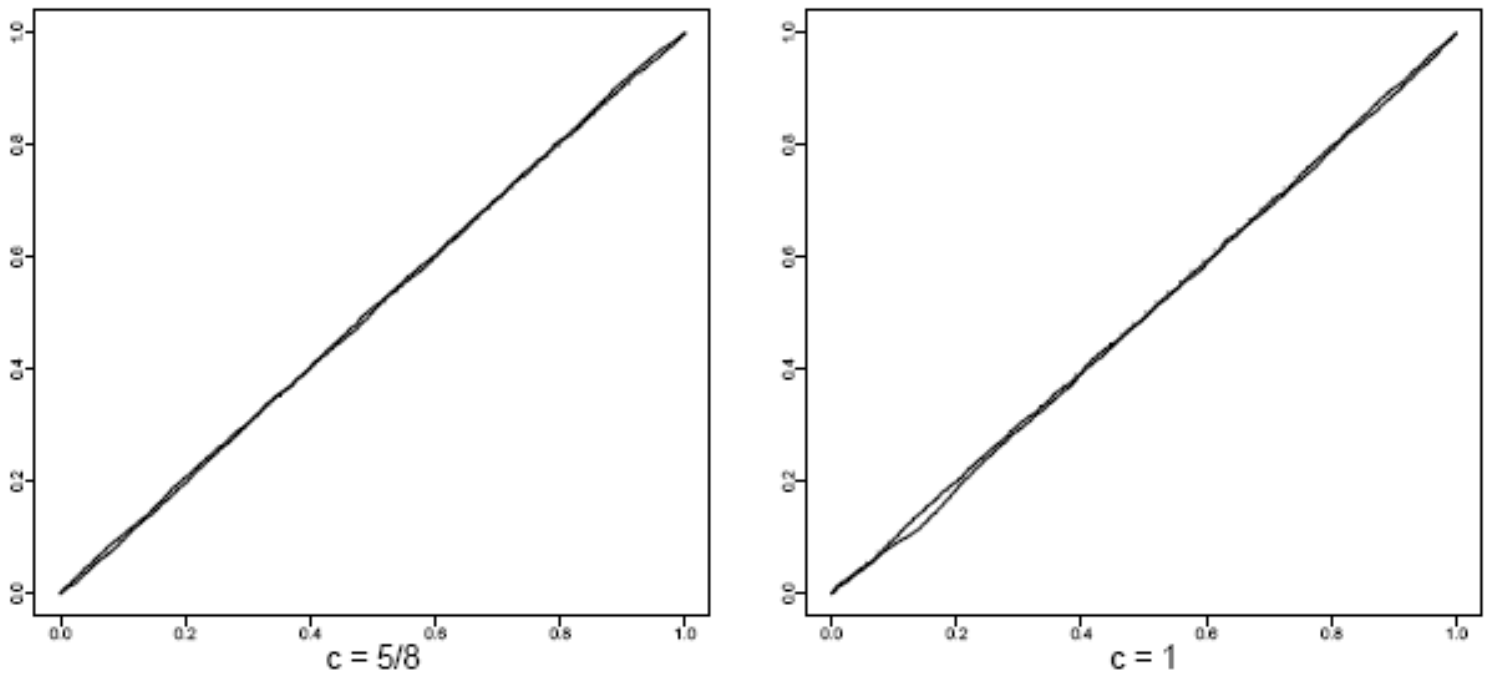
Testing for Absolute Continuity

- cdf of r^p, $p = 1$, $a(t) = c/(1-t)$, sample size 4000.



Testing for Absolute Continuity

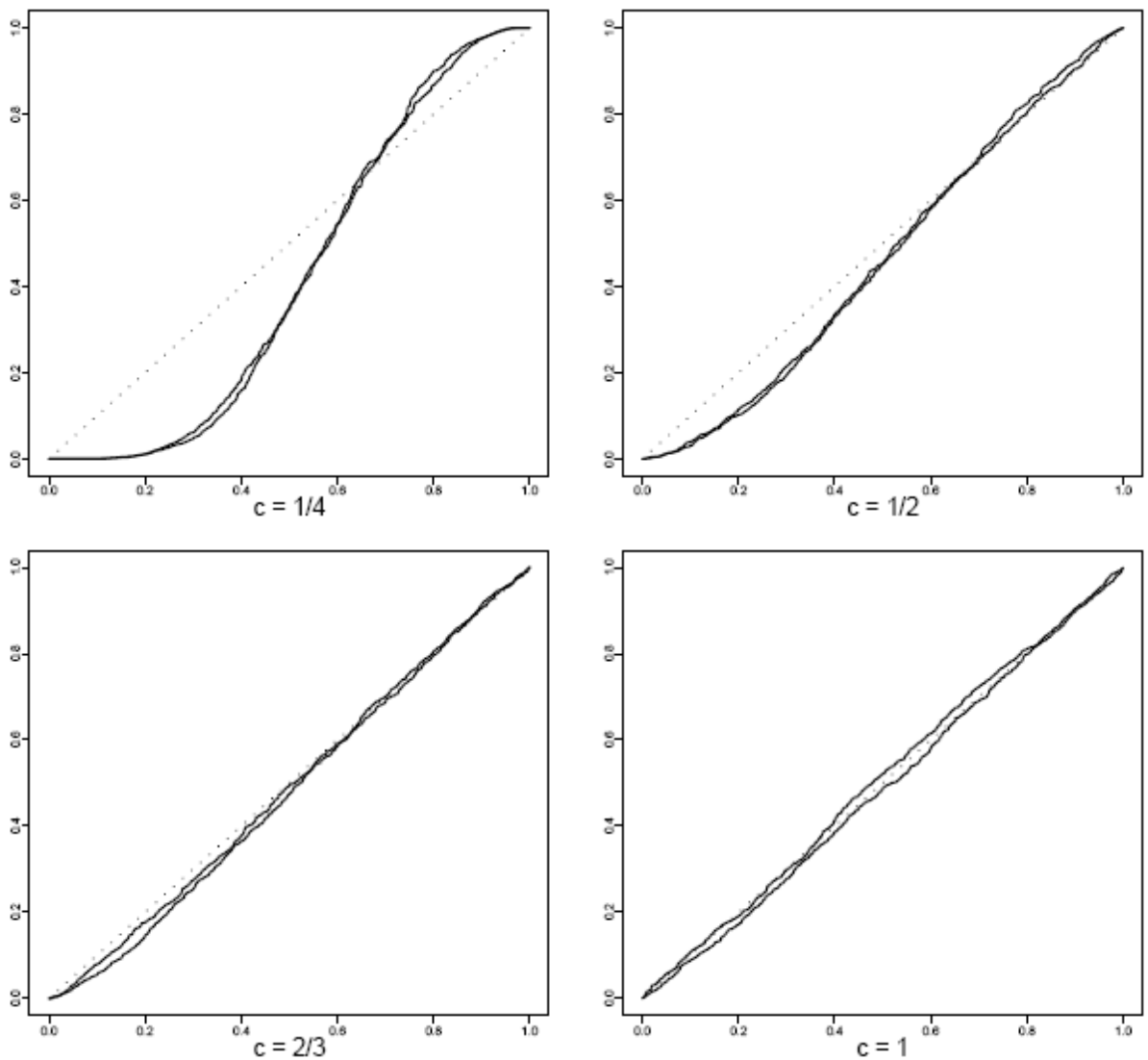
- cdf of r^p, $p = 1$, $a(t) = c/(1-t)$, sample size 4000.



- for $c = \frac{1}{2}$, $c = \frac{5}{8}$, and $c = 1$, there is absolute continuity.

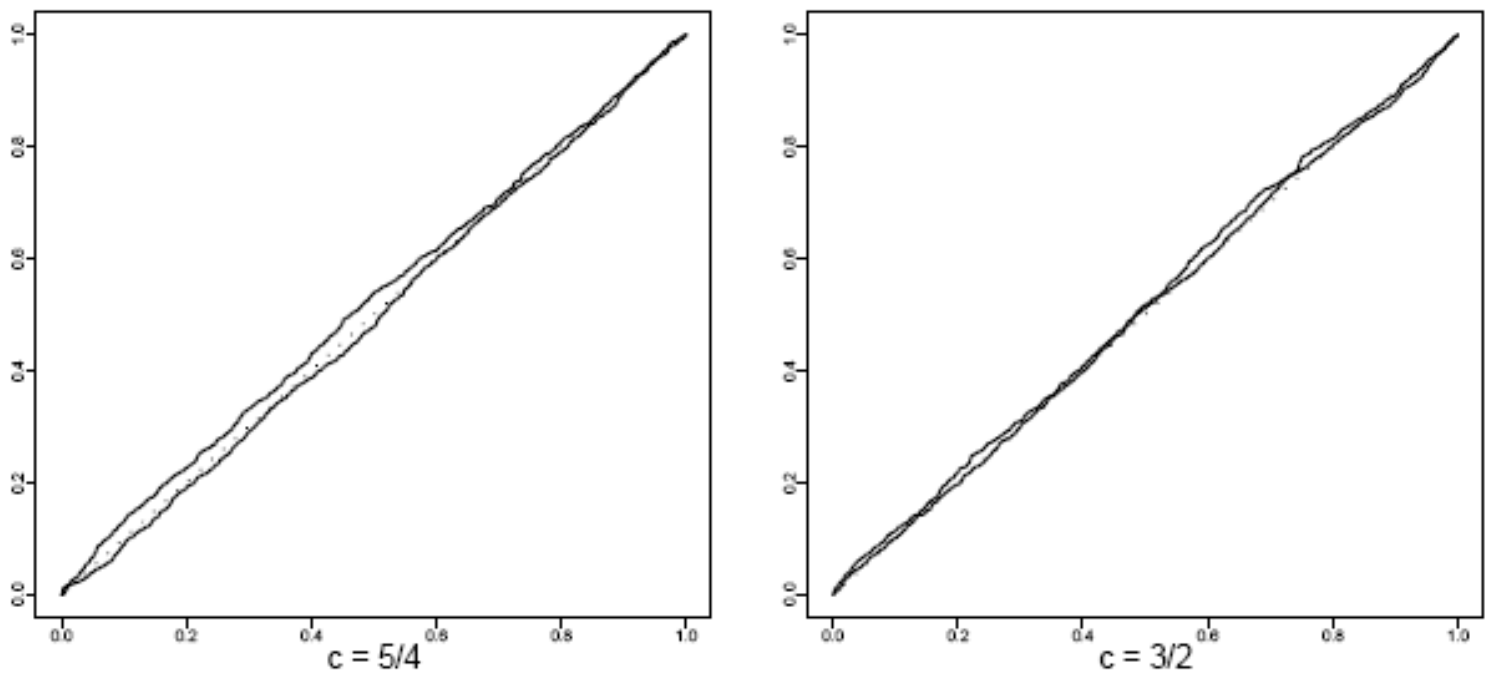
Testing for Absolute Continuity

- cdf of r^p, $p = 2$, $a(t) = c/(1-t)$, sample size 1000.



Testing for Absolute Continuity

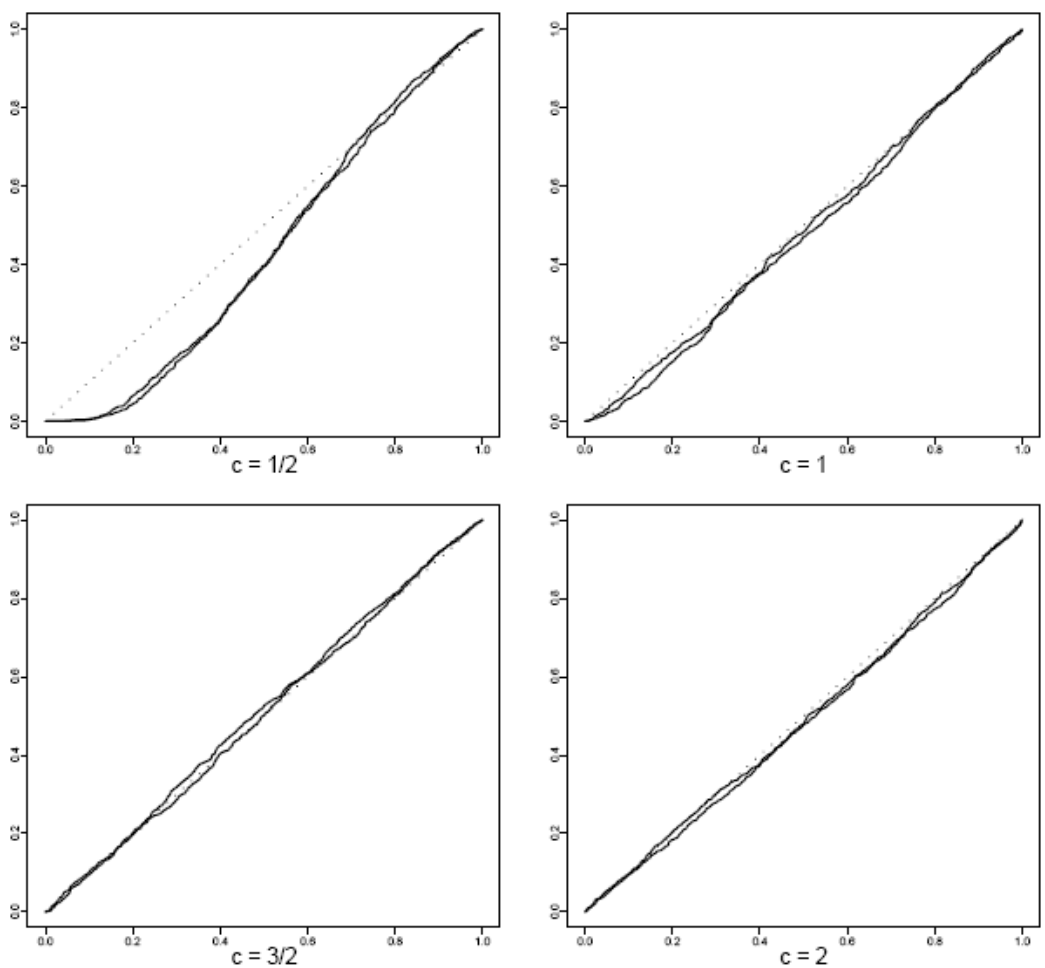
- cdf of r^p, $p = 2$, $a(t) = c/(1-t)$, sample size 1000.



- for $c = \frac{3}{2}$, $c = \frac{5}{4}$, and $c = 1$, there is absolute continuity.

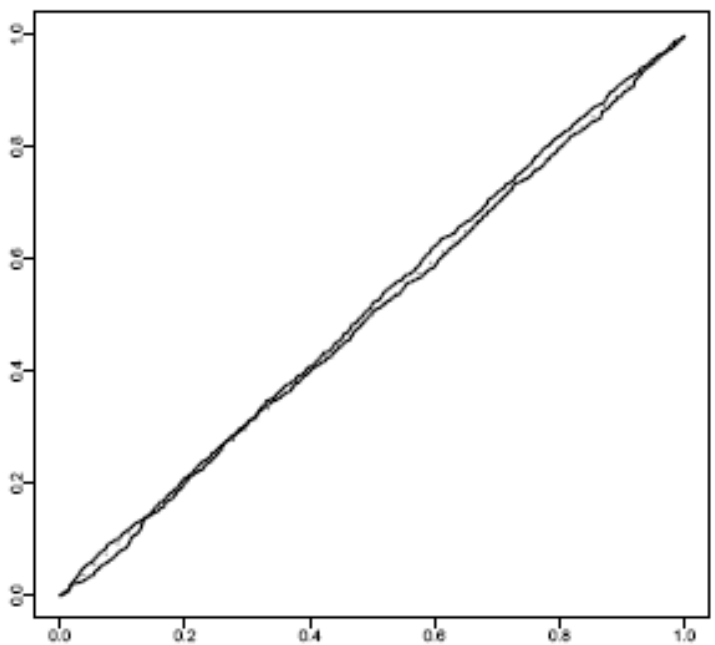
Testing for Absolute Continuity

- cdf of r^p, $p = 3$, $a(t) = c/(1-t)$, sample size 1000, for $c = 3/2$ and $c = 2$, there is absolute continuity.



Testing for Absolute Continuity

- cdf of r^p, $p = 2$, $a(t) = \frac{1}{2} + \frac{1}{200(1-t)^2}$, sample size 1000, there is absolute continuity.



- Conjecture: for $a(t) = \frac{c}{1-t}$, there is absolute continuity iff $c > \frac{p}{2}$.

Testing for Absolute Continuity

- Consider the distribution of the difference $\boldsymbol{\nu}$ between two paths, conditioned on their divergence time t. We have $f(\boldsymbol{\nu}|t) \sim \mathcal{N}\big(\mathbf{0}, 2\sigma^2(1-t)\boldsymbol{I}\big)$.
- density of divergence time is $a(t)\exp\big(-A(t)\big) = c(1-t)^{c-1}$.
- therefore

$$
\begin{aligned}
f(\boldsymbol{\nu}) &= \int_0^1 c(1-t)^{c-1} \frac{\exp\big(-\frac{|\boldsymbol{\nu}|^2}{4\sigma^2(1-t)}\big)}{\big(4\pi\sigma^2(1-t)\big)^{\frac{p}{2}}} dt \\
&= \int_0^1 c(1-t)^{c-1-\frac{p}{2}} \frac{\exp\big(-\frac{|\boldsymbol{\nu}|^2}{4\sigma^2(1-t)}\big)}{(4\pi\sigma^2)^{\frac{p}{2}}} dt \\
&= \frac{c}{(4\pi\sigma^2)^{\frac{p}{2}}} \int_1^\infty u^{\frac{p}{2}-1-c} \exp\big(-\frac{|\boldsymbol{\nu}|^2}{4\sigma^2} u\big) du
\end{aligned}
$$

Testing for Absolute Continuity

- continuing, we have

$$
\begin{aligned}
f(\boldsymbol{\nu}) &= \frac{c}{(4\pi\sigma^2)^{\frac{p}{2}}} \int_1^\infty u^{\frac{p}{2}-1-c} \exp\Big(-\frac{|\boldsymbol{\nu}|^2}{4\sigma^2}u\Big)du \\
&= \begin{cases}
\frac{c}{(4\pi\sigma^2)^{\frac{p}{2}}} \Big(\frac{|\boldsymbol{\nu}|^2}{4\sigma^2}\Big)^{c-\frac{p}{2}} \Gamma\Big(\frac{p}{2}-c, \frac{|\boldsymbol{\nu}|^2}{4\sigma^2}\Big) & \text{if } \frac{p}{2} > c, \boldsymbol{\nu} \neq \mathbf{0} \\
\infty & \text{if } \frac{p}{2} > c, \boldsymbol{\nu} = 0 \\
\frac{c}{(4\pi\sigma^2)^{\frac{p}{2}}(c-\frac{p}{2})} & \text{if } \frac{p}{2} < c, \boldsymbol{\nu} = 0
\end{cases}
\end{aligned}
$$

- so at $\boldsymbol{\nu} = \mathbf{0}$ (for $\frac{p}{2} > c$), $f(\boldsymbol{\nu})$ is undefined.
- therefore $c = \frac{p}{2}$ seems to be a critical point with respect to continuity.
- Conjecture: In contrast, DDT priors using $a(t) = b + \frac{c}{(1-t)^2}$ ($c > 0$) always produce absolutely continuous distributions.

Simple Relationships to other Processes

- DDT with variance σ^2, $a(t) = 0$ except for an infinite peak of mass $\log(1+\alpha)$ at $t = 0$, is equivalent to $DP\big(\alpha, \mathcal{N}(\mathbf{0}, \sigma^2 \boldsymbol{I})\big)$.
- to generate a simple DP mixture with components $\sim \mathcal{N}(\boldsymbol{\mu}, \sigma_x^2 \boldsymbol{I})$ and prior $\sim \mathcal{N}(\mathbf{0}, \sigma_\mu^2 \boldsymbol{I})$, generate a DDT with

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

- where $\sigma^2 = \sigma_\mu^2 + \sigma_x^2$.

Simple Relationships to other Processes

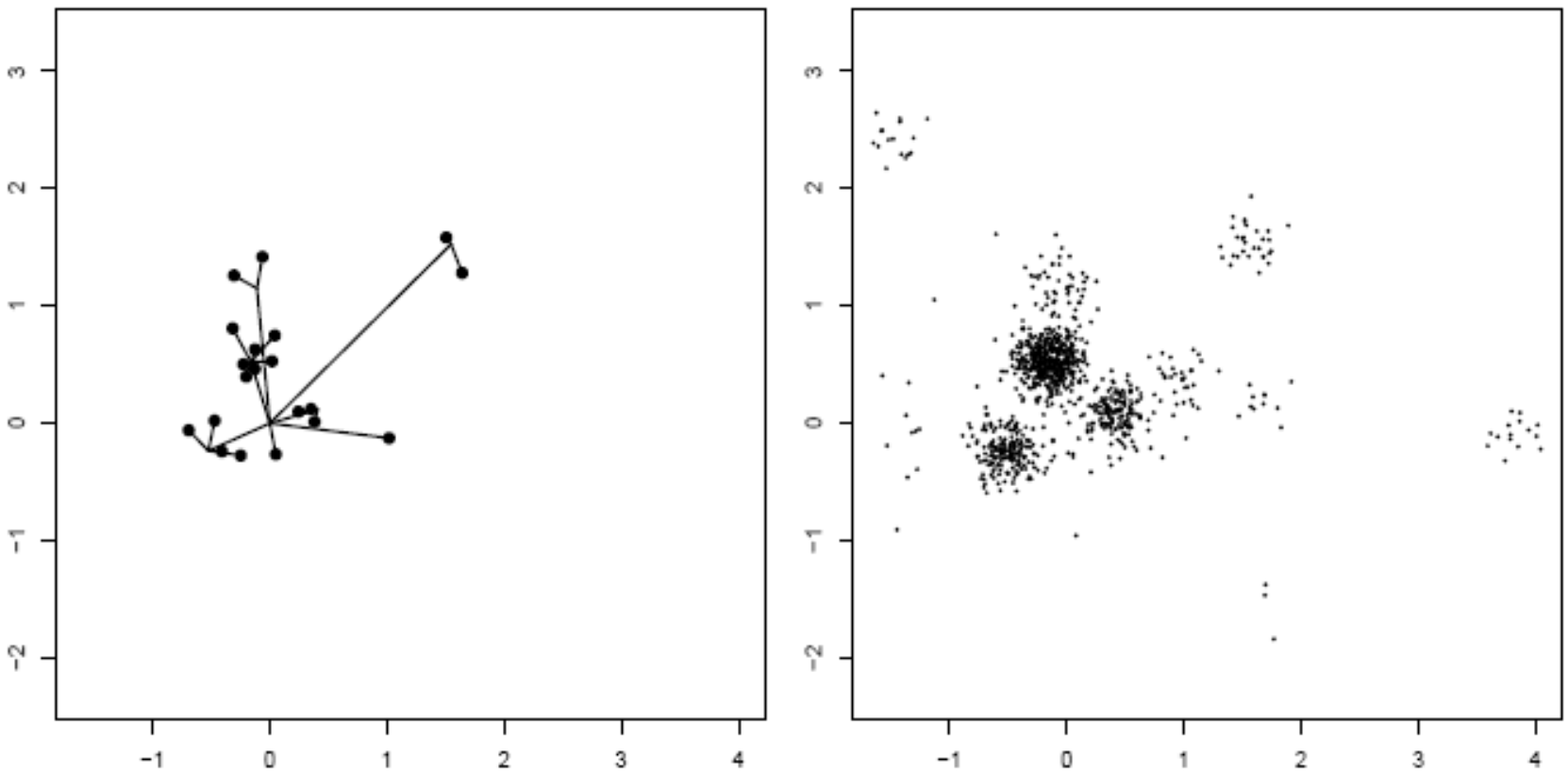


Figure 14: Generation of a two-dimensional data set from the Dirichlet diffusion tree prior that is equivalent to the simple Dirichlet process mixture with $\sigma_\mu = 0.99$, $\sigma_x = 0.14$, and $\alpha = 1$. The plot on the left shows the first twenty data points generated along with the underlying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

Discussion

- DDT can be used to produce variety of distributions. Their properties vary with the choice of divergence function.
- The parameters of the divergence function can also be given prior distributions, for greater flexibility.
- Alternatively, one can substitute the divergence function with a stochastic process.
- DDT can model vectors of latent variables, if the observed data is perturbed by noise.
- Univariate data (even parameters of the DDT itself) can be modeled by multivariate DDTs, like $z = x\exp(y)$, where $(x, y) \sim DDT$.
- σ^2 can be made time-varying, or with a DDT prior.
- Gaussian and Wishart priors can be used for the mean and covariance respectively of $\boldsymbol{N}(t)$.