Statistical Inference of Discretely Observed Compound Poisson Processes and Related Jump Processes

Suraj Shah

March 24, 2019

Abstract

Contents

1 Introduction

Definition 1.1 (Counting Process)**.** A *counting process* is a stochastic process $\{N(t) : t \geq 0\}$ with values that are non-negative, integer and non-decreasing i.e. $\forall s, t \geq 0 : s \leq t$:

1. $N(t) \geq 0$,
2. $N(t) \in \mathbb{N}$,
3. $N(s) \leq N(t)$.

Definition 1.2 (Poisson Process)**.** A *Poisson process with intensity* λ is a counting process $\{N(t) : t \geq 0\}$ with the following properties:

1. $N(0) = 0$,
2. It has independent increments i.e. $\forall n \in \mathbb{N} : 0 \leq t_1 \leq t_2 \leq \cdots \leq t_n$, $N(t_n) - N(t_{n-1}), N(t_{n-1}) - N(t_{n-2}), \ldots, N(t_1)$ are independent,
3. The number of occurrences in any interval of length t is a Poisson random variable with parameter λt i.e. $\forall s, t : s \leq t$, $N(t) - N(s) \sim$ Poisson$(\lambda(t - s))$.

Lemma 1.1. *A Poisson process with intensity* λ *has exponentially distributed inter-arrival times with rate* λ*.*

Definition 1.3 (Compound Poisson Process)**.** Let $N(t) : t \geq 0$ be a d-dimensional Poisson process with intensity λ.

Let $Y_1, Y_2, \ldots$ be a sequence of i.i.d random variables taking values in $\mathbb{R}^d$ with common distribution F.

Also assume that the Y_i's are independent of the Poisson process $\{N(t) : t \geq 0\}$.

Then, a *Compound Poisson process (CPP)* is a stochastic process $\{X(t) : t \geq 0\}$ such that

$$X(t) = \sum_{i=1}^{N(t)} Y_i$$

where, by convention, we take $X(t) = 0$ if $N(t) = 0$.

Suppose we take discrete observations of a CPP i.e. we consider $X(\Delta), X(2\Delta), \ldots$ where $X(t) : t \geq 0$ is a CPP. We want to estimate F. Note that the jump size $X(n\Delta) - X((n-1)\Delta)$ is equivalent in distribution to a Poission random

sum of intensity Δ:

$$\begin{aligned} X(n\Delta) - X((n-1)\Delta) &= \sum_{i=1}^{N(n\Delta)} Y_i - \sum_{i=1}^{N((n-1)\Delta)} Y_i \\ &= \sum_{i=1}^{N(n\Delta)-N((n-1)\Delta)} Y_i \\ &=^d \sum_{i=1}^{N} Y_i \end{aligned}$$

where $N \sim \text{Poisson}(\Delta)$

2 Spectral Approach

Now we have formulated the problem, we visit some methods for estimating the unknown density f. Since adding a Poisson number of Y's is referred to as compounding, much of the literature refers to the problem of recovering density f of Y's from observations of X as decompounding.

The approach of decompounding was famously proposed by Buchmann and Grübel to estimate the density f for discrete and continuous cases of the distribution F of the Y's.

Van Es built on this idea for fixed sampling rate $\Delta = 1$ using the Lévy - Khintchine formula. We explain the idea behind this method and show its strength through various examples.

2.1 Van Es

2.1.1 Construction of Density Estimator via suitable inversion of characteristic functions

We first note the following property:

Proposition 2.1. *For Poisson random sum X, the characteristic function of X, denoted by ϕ_X, is given by $\phi_X(t) = \mathbb{E}e^{itX} = e^{-\lambda+\lambda\phi_f(t)}$*

Proof.

$$
\begin{aligned}
\phi_X(t) &= \mathbb{E}e^{itX} \\
&= \mathbb{E}\left[\exp\left(it\sum_{i=1}^{N(\lambda)} Y_i\right)\right] \\
&= \mathbb{E}\left[\prod_{i=1}^{N(\lambda)} \exp(itY_i)\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\prod_{i=1}^{N(\lambda)} \exp(itY_i)\middle| N(\lambda)\right]\right] \\
&= \mathbb{E}\left[\prod_{i=1}^{N(\lambda)} \mathbb{E}\left[\exp(itY_1) \mid N(\lambda)\right]\right] && \text{(by i.i.d assumption of the } Y_i\text{'s)} \\
&= \mathbb{E}\left[\prod_{i=1}^{N(\lambda)} \phi_f(t)\right] && (Y_1 \text{ and } N(\lambda) \text{ are independent)} \\
&= \mathbb{E}\left[\exp(N(\lambda)\ln\phi_f(t))\right] \\
&= \exp(\lambda(e^{\ln\phi_f(t)} - 1)) && \text{(MGF of a Poisson random variable)} \\
&= e^{-\lambda+\lambda\phi_f(t)}
\end{aligned}
$$

□

We can rewrite $\phi_X(t)$ as:

$$
\begin{aligned}
\phi_X(t) &= e^{-\lambda}(e^{\lambda\phi_f(t)} - 1 + 1) \\
&= e^{-\lambda} + e^{-\lambda}(e^{\lambda\phi_f(t)} - 1) \\
&= e^{-\lambda} + e^{-\lambda}\frac{e^{\lambda} - 1}{e^{\lambda} - 1}(e^{\lambda\phi_f(t)} - 1) \\
&= e^{-\lambda} + \frac{1 - e^{-\lambda}}{e^{\lambda} - 1}(e^{\lambda\phi_f(t)} - 1) \qquad (1)
\end{aligned}
$$

Since a zero jump size provides no additional information on the density f, we want to gain information about X conditional on the event that there is at least one jump. Seeing that $X \mid N(\lambda) > 0$ has a density is somewhat intuitive, but we provide a proof of this.

Lemma 2.1. *The random variable $X \mid N(\lambda) > 0$ has a density.*

Proof. By the Radon-Nikodym Theorem, a random variable X has a density if and only if $\mathbb{P}(X \in A) = 0$ for every Borel set A with Lebesgue measure zero.

Suppose that $\text{Leb}(A) = 0$. Then

$$\begin{aligned}\mathbb{P}(X \in A | N(\lambda) > 0) &= \frac{1}{\mathbb{P}(N(\lambda) > 0)} \sum_{n=1}^{\infty} \mathbb{P}\left(Y_1 + \cdots + Y_n \in A, N(\lambda) = n\right) \\ &= \frac{1}{\mathbb{P}(N(\lambda) > 0)} \sum_{n=1}^{\infty} \mathbb{P}\left(Y_1 + \cdots + Y_n \in A\right) \mathbb{P}(N(\lambda) = n)\end{aligned}$$

Note that for each $n, Y_1 + \cdots + Y_n$ has a density so $\mathbb{P}\left(Y_1 + \cdots + Y_n \in A\right) = 0$. Thus the result follows.

□

Let g be the density of $X \mid N(\lambda) > 0$.
Let $\phi_g(t) = \mathbb{E}\left[e^{itX} \mid N(\lambda) > 0\right] = \frac{\mathbb{E}\left[e^{itX} \mathbb{1}(N(\lambda)>0)\right]}{\mathbb{P}(N(\lambda)>0)}$.
Then

$$\begin{aligned}\phi_X(t) &= \mathbb{E}\left[e^{itX} \mathbb{1}(N(\lambda) = 0)\right] + \mathbb{E}\left[e^{itX} \mathbb{1}(N(\lambda) > 0)\right] \\ &= \mathbb{P}(N(\lambda) = 0) + \mathbb{P}(N(\lambda) > 0)\phi_g(t) \\ &= e^{-\lambda} + (1 - e^{-\lambda})\phi_g(t)\end{aligned}$$

Therefore, using (1), we get that

$$\phi_g(t) = \frac{1}{e^{\lambda} - 1}(e^{\lambda \phi_f(t)} - 1) \tag{2}$$

Thus, we can see from this that if we were to obtain an estimator for $\phi_g(t)$, then by suitable inversion of the formula in (2), we would obtain an estimator for $\phi_f(t)$.

In order to rewrite (2) in terms of $\phi_f(t)$, we must be able to invert the complex exponential function since $\phi_f(t)$ takes complex values. However, such function is not invertible since it is not bijective: in particular it is not injective as $e^{w+2\pi i} = e^w \ \forall w \in \mathbb{C}$.

Therefore, we use the following lemmas concerning the distinguished logarithm:

Lemma 2.2. *If $h_1 : \mathbb{R} \to \mathbb{C}$ and $h_2 : \mathbb{R} \to \mathbb{C}$ are continuous functions such that $h_1(0) = h_2(0) = 0$ and $e^{h_1} = e^{h_2}$, then $h_1 = h_2$.*

Proof. See Appendix. □

Lemma 2.3. *If $\phi : \mathbb{R} \to \mathbb{C}$ is a continuous function such that $\phi(0) = 1$ and $\phi_g(t) \neq 0 \ \forall t \in \mathbb{R}$ then there exists a unique continuous function $h : \mathbb{R} \to \mathbb{C}$ with $h(0) = 0$ and $\phi(t) = e^{h(t)}$ for $t \in \mathbb{R}$.*

Proof. See Appendix. □

Therefore, for such a function ϕ as described in the Lemma, we say that the unique function h is the distinguished logarithm and we denote $h(t) = \text{Log}(\phi(t))$. Note also that for ϕ and ψ satisfying the assumptions of the Lemma, we have $\text{Log}(\phi(t)\psi(t)) = \text{Log}(\phi(t)) + \text{Log}(\psi(t))$ as expected.

Therefore, noting that $\phi(t) = e^{\lambda(\phi_f(t)-1)}$ is a continuous function satisfying $\phi(0) = 1$ and $\phi(t) \neq 0 \ \forall t \in R$, we get that

$$\begin{aligned}\lambda(\phi_f(t) - 1) &= \text{Log}\left(e^{\lambda(\phi_f(t)-1)}\right) && \text{(Lemma 2.2)}\\ &= \text{Log}\left(e^{-\lambda}\left[(e^{\lambda} - 1)\phi_g(t) + 1\right]\right)\\ &= -\lambda + \text{Log}\left((e^{\lambda} - 1)\phi_g(t) + 1\right)\end{aligned}$$

Therefore,

$$\phi_f(t) = \frac{1}{\lambda}\text{Log}\left((e^{\lambda} - 1)\phi_g(t) + 1\right) \tag{3}$$

By Fourier inversion, for integrable ϕ_f we have

$$f(x) = \frac{1}{2\pi\lambda}\int_{-\infty}^{\infty} e^{-itx}\text{Log}\left((e^{\lambda} - 1)\phi_g(t) + 1\right)dt \tag{4}$$

This suggests that if we can estimate $\phi_g(t)$, then we have an estimate of f.

2.1.2 Kernel density estimators

We provide the intuition behind choosing our estimator for g on observations of non-zero jump size as a kernel density estimator.

Let X be a random variable with probability density p with respect to the Lebesgue measure on $\mathbb{R}$. The corresponding distribution function is $F(x) = \int_{-\infty}^{x} p(t)dt$.

Consider n i.i.d observations $X_1, \ldots, X_n$ with same distribution as X. The empirical distribution function is given by

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x)$$

By the Strong Law of Large Numbers, since for fixed x, $I(X_i \leq x)$ are i.i.d , we have that

$$F_n(x) \to \mathbb{E}\left[I(X_1 \leq x)\right] = \mathbb{P}(X \leq x) = F(x)$$

almost surely as $n \to \infty$.

Therefore, $F_n(x)$ is a consistent estimator of $F(x)$ for every $x \in \mathbb{R}$. Also note that $p(x) = \frac{\mathrm{d}}{\mathrm{d}x}F(x)$, so for sufficiently small $h > 0$ we can write an approximation

$$p(x) \approx \frac{F(x+h) - F(x-h)}{2h}$$

Thus, intuitively we can replace F by our empirical distribution function F_n to give us an estimator $\hat{p}_n(x)$ of $p(x)$

$$\begin{aligned}\hat{p}_n(x) &= \frac{F_n(x+h) - F_n(x-h)}{2h} \\ &= \frac{1}{2nh}\sum_{i=1}^{n} I(x-h < X_i \leq x+h) \\ &= \frac{1}{nh}\sum_{i=1}^{n} K_0\left(\frac{x-X_i}{h}\right)\end{aligned}$$

where $K_0(u) = \frac{1}{2}I(-1 < u \leq 1)$.

A simple generalisation is to replace K_0 by some arbitrary (but well-chosen) integrable function $K : \mathbb{R} \to \mathbb{R}$ such that $\int K(u)du = 1$ and $K(u) = K(-u)$ for every $u \in \mathbb{R}$. Such a function K is called a *kernel* and the parameter h is called a *bandwidth* of the estimator

$$\hat{p}_n(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right) \tag{5}$$

We call this estimator a *kernel density estimator*.

Thus, for some kernel w with characteristic function ϕ_w and observations $Z_1, \ldots, Z_n$, we estimate density g by the kernel density estimator

$$g_{nh}(x) = \frac{1}{nh}\sum_{i=1}^{n} w\left(\frac{x-Z_i}{h}\right)$$

Letting $\phi_{\text{emp}}(t) = \frac{1}{n}\sum_{j=1}^{n} e^{itZ_j}$ be the empirical characteristic function, we get that

$$\begin{aligned}\phi_{g_{nh}}(t) &= \int_{-\infty}^{\infty} e^{itx} g_{nh}(x)dx \\ &= \int_{-\infty}^{\infty} e^{itx}\frac{1}{nh}\sum_{j=1}^{n} w\left(\frac{x-Z_j}{h}\right)dx \\ &= \frac{1}{n}\sum_{j=1}^{n} e^{itZ_j}\int_{-\infty}^{\infty} e^{ithy}w(y)dy \qquad \left(\text{by the substitution } y = \frac{x-Z_j}{h}\right) \\ &= \phi_{\text{emp}}(t)\phi_w(ht)\end{aligned}$$

In view of (4) It is tempting to introduce an estimator $\hat{f}_{nh}$ of f

$$\hat{f}_{nh}(x) = \frac{1}{2\pi\lambda}\int_{-\infty}^{\infty} e^{-itx}\text{Log}\Big((e^{\lambda}-1)\phi_{\text{emp}}(t)\phi_w(ht)+1\Big)dt \tag{6}$$

but this brings two main issues:

1. In light of Lemma 2.3, we may have some Borel set A with non-zero Lebesgue measure such that $(e^\lambda - 1)\phi_{\text{emp}}(t)\phi_w(ht) + 1$ is zero for $t \in A$. The distinguished logarithm is undefined under such sets and thus our estimator of f is undefined in this case.

2. There is no guarantee that the integral is finite. For example,

$$\phi_{g_{nh}}(t) = \frac{\exp(e^{it}) - 1}{e^\lambda - 1}$$

would give $\hat{f}_{nh}(1)$ to be infinity.

In order to prove asymptotic properties, we must adjust our estimators by bounding $\hat{f}_{nh}$ for each n using a suitable sequence $(M_n)_{n\geq 1}$. However, for our discussion, we note such limitations and provide simulations for examples where these two cases do not occur.

2.1.3 Simulation Results

We note that for $\lambda < \log 2$, the distinguished logarithm in (6) reduces to the principal branch of the logarithm. This is the logarithm whose imaginary part lies in the interval $(-\pi, \pi]$. We also note, as written above, that bounding $\hat{f}_{nh}$ by a suitable sequence is not needed in practice. Therefore, we can use (6) to compute our estimator with the principal branch of the logarithm, provided $\lambda < \log 2$.

We use the following kernel w given by

$$w(t) = \frac{48t(t^2 - 1)\cos t - 144(2t^2 - 5)\sin t}{\pi t^7}$$

This kernel has a fairly complicated form but its characteristic function $\phi_w(t)$ has a much simpler expression given by

$$\phi_w(t) = (1 - t^2)^3 \mathbb{1}\{|t| < 1\}$$

We can rewrite (6) as $\hat{f}_{nh}(x) = \hat{f}^{(1)}_{nh}(x) + \hat{f}^{(2)}_{nh}(x)$ where

$$\hat{f}^{(1)}_{nh}(x) = \frac{1}{2\pi\lambda}\int_0^\infty e^{-itx}\text{Log}\Big((e^\lambda - 1)\phi_{\text{emp}}(t)\phi_w(ht) + 1\Big)dt \tag{7}$$

$$\begin{aligned}\hat{f}^{(2)}_{nh}(x) &= \frac{1}{2\pi\lambda}\int_{-\infty}^0 e^{-itx}\text{Log}\Big((e^\lambda - 1)\phi_{\text{emp}}(t)\phi_w(ht) + 1\Big)dt \\ &= \frac{1}{2\pi\lambda}\int_0^\infty e^{itx}\text{Log}\Big((e^\lambda - 1)\phi_{\text{emp}}(-t)\phi_w(ht) + 1\Big)dt \end{aligned} \tag{8}$$

since ϕ_w is symmetric. We use a bandwidth of 0.14. Such a bandwidth is arbitrary and we may use better methods to compute a bandwidth estimator that yields better results. This can be done via cross-validation.

We approximate (7) and (8) by the Trapezoid Rule:

Trapezoid Rule. *Let $\{t_j\}_{j=0}^{N-1}$ be a set of N equally spaced values partitioning $[a,b]$, with spacing $\Delta t_k = \Delta t = \frac{b-a}{N}$. Then, for integrable function f we get the following approximation*

$$\int_a^b f(x)dx \approx \Delta t \left(\frac{f(t_0) + f(t_{N-1})}{2} + \sum_{j=1}^{N-2} f(t_j) \right) \tag{9}$$

We can approximate (7) (and similarly (8)) by computing the integrand from 0 to some sufficiently large M.

We may also allow for (9) to be written as a 'nice' sum in order to compute the Fast Fourier Transform. Thus, we write

$$\int_a^b f(t)dt \approx \Delta t \left(\sum_{j=0}^{N-1} f(t_j) \right) \tag{10}$$

Applying this to (7) and (8) we get for $t_j = j\eta$ for some spacing parameter η

$$\hat{f}_{nh}^{(1)}(x) \approx \frac{\eta}{2\pi\lambda} \sum_{k=0}^{N-1} e^{-it_j x} \text{Log}\Big((e^{\lambda} - 1)\phi_{\text{emp}}(t_j)\phi_w(ht_j) + 1\Big), \tag{11}$$

$$\hat{f}_{nh}^{(2)}(x) \approx \frac{\eta}{2\pi\lambda} \sum_{k=0}^{N-1} e^{it_j x} \text{Log}\Big((e^{\lambda} - 1)\phi_{\text{emp}}(-t_j)\phi_w(ht_j) + 1\Big), \tag{12}$$

We apply the Fast Fourier Transform to evaluate our functions $\hat{f}_{nh}^{(1)}$ and $\hat{f}_{nh}^{(2)}$ at points $\{x_k\}_{k=0}^{N-1}$.

Fast Fourier Transform. *Let $\{x_k\}_{k=0}^{N-1}$ be a sequence of complex numbers. The Fast Fourier Transform computes the sequence $\{Y_j\}_{j=0}^{N-1}$ where*

$$Y_j = \sum_{k=0}^{N-1} x_k e^{-ij\frac{2\pi k}{N}} \tag{13}$$

The inverse transform is given by

$$Y_j = \frac{1}{N} \sum_{k=0}^{N-1} x_k e^{ij\frac{2\pi k}{N}} \tag{14}$$

Thus, we employ a regular spacing with parameter δ so that our values $\{x_k\}_{k=0}^{N-1}$ evenly spaced and given by

$$x_k = \frac{-N\delta}{2} + \delta k$$

Thus we have

$$\hat{f}_{nh}^{(1)}(x_k) \approx \frac{1}{2\pi\lambda} \sum_{k=0}^{N-1} e^{-ijk\eta\delta} e^{it_j \frac{N\delta}{2}} \psi^{(1)}(t_j)\eta, \tag{15}$$

$$\hat{f}_{nh}^{(2)}(x_k) \approx \frac{1}{2\pi\lambda} \sum_{k=0}^{N-1} e^{ijk\eta\delta} e^{-it_j \frac{N\delta}{2}} \psi^{(2)}(t_j)\eta, \tag{16}$$

Therefore, we take $\eta\delta = \frac{2\pi}{N}$ and we apply FFT on the sequence $\left\{e^{it_j \frac{N\delta}{2}} \psi^{(1)}(t_j)\eta\right\}_{j=0}^{N-1}$ to get values for $\hat{f}_{nh}^{(1)}$ and we apply IFFT on the sequence $\left\{e^{-it_j \frac{N\delta}{2}} \psi^{(2)}(t_j)\eta\right\}_{j=0}^{N-1}$ to get values for $\hat{f}_{nh}^{(2)}$.

We take N to be a power of 2 for computational speed up in calculating the Discrete Fourier Transforms and we choose η relatively small so that δ can be relatively larger and so points are relatively separate from one another.

The results for $N = 16384$ and $\eta = 0.01$ based on 1000 observations can be shown in Figure 1.

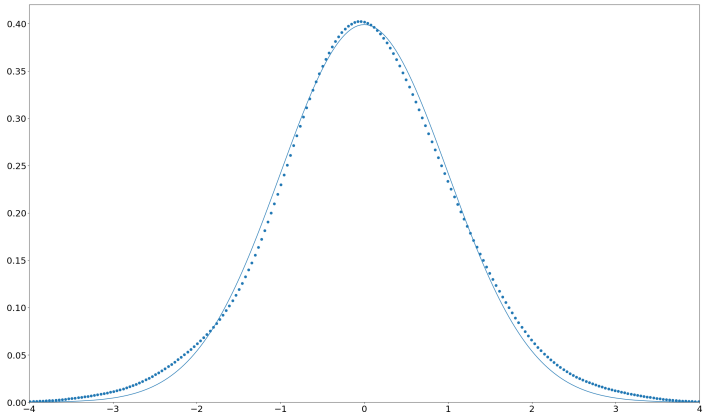

Figure 1: Density Estimator of Standard Normal

The second example we consider is the case of f being a mixture of two normal densities

$$f(\cdot) = \rho_1 \psi(\cdot; \mu_1, \sigma_1^2) + \rho_2 \psi(\cdot; \mu_2, \sigma_2^2)$$

where $\rho_1 = \frac{2}{3}, \rho_2 = \frac{1}{3}, \mu_1 = 0, \mu_2 = 3, \sigma_1^2 = 1, \sigma_2^2 = \frac{1}{9}$. We use a bandwidth of 0.1.

3 Bayesian Approach

We now consider the non-parametric Bayesian approach to estimating 'true' parameter values f_0 as well as λ_0 from discrete evenly-spaced observations of the CPP.

The advantages of using a non-parametric Bayesian approach are two-fold:

1. We obtain a distribution over all probability densities rather than just a point estimate as in the spectral approach,
2. We obtain automatic uncertainty quantification through the posterior credible sets

3.1 Bayesian Nonparametrics

In the Bayesian framework, the data X follows a distribution determined by a parameter θ, which is itself considered to be generated from a prior distribution Π. The corresponding posterior distribution is the conditional distribution of θ, given X. This framework is identical in parametric and nonparametric Bayesian statistics, the only difference being the dimension of the parameter θ.

Because, in the nonparametric case, the prior distribution is a distribution over an infinite-dimensional space, we must carefully construct the setup in order to apply Bayes Theorem.

Let $(\mathcal{F}, \mathcal{B})$ be a measurable space, where $\mathcal{F}$, in our following discussion, will be a set of functions denoting the parameter space. Then the prior distribution Π is a probability measure on $(\mathcal{F}, \mathcal{B})$.

Let $(\mathfrak{X}, \mathcal{X})$ denote our sample space. Suppose we have data X - this data is a random variable from some probability space $(\Omega, \mathcal{A}, \mathbb{P})$ so $X : \Omega \to \mathfrak{X}$. Our model/likelihood P_θ of X given θ is a Markov kernel with source $(\mathcal{F}, \mathcal{B})$ and target $(\mathfrak{X}, \mathcal{X})$.

Definition 3.1. Let $(X, \mathcal{A}), (Y, \mathcal{B})$ be measurable spaces. A Markov kernel with source $(X, \mathcal{A})$ and target $(Y, \mathcal{B})$ is a map $\kappa : X \times \mathcal{B} \to [0, 1]$ with the following properties:

1. The map $x \mapsto \kappa(x, B)$ is $(X, \mathcal{A})$-measurable for every $B \in \mathcal{B}$
2. The map $B \mapsto \kappa(x, B)$ is a probability measure on $(Y, \mathcal{B})$ for every $x \in X$.

Therefore, $A \mapsto P_\theta(A)$ is a probability measure on $(\mathfrak{X}, \mathcal{X})$.

For pair (X, θ), we have a well-defined joint distribution on the product space $(\mathcal{F} \times \mathfrak{X}, \mathcal{B} \otimes \mathcal{X})$ given by

$$P(B, A) = P(\theta \in B, X \in A) = \int_B P_\theta(A) d\Pi(\theta)$$

and we note that the integral is well-defined since the map $\theta \mapsto P_\theta(A)$ is $(\mathcal{F}, \mathcal{B})$-measurable.

From this we get the marginal distribution of X given by

$$P(X \in A) = \int P_\theta(A) d\Pi(\theta)$$

i.e. a probability measure P_X on $(\mathfrak{X}, \mathcal{X})$

$$P(\theta \in B|X) = \mathbb{E}_P \left[\mathbb{1}(\theta \in B) \mid \sigma(X)\right] \qquad \text{a.s.}$$

Note that $\sigma(X) = \mathcal{X}$ and we can write the posterior distribution $\Pi(B|X) = P(\theta \in B|X)$. Under certain conditions, there exists a regular version of the conditional distribution i.e. a Markov kernel μ from source $(\mathfrak{X}, \mathcal{X})$ and target $(\mathcal{F}, \mathcal{B})$ such that the map $B \mapsto \mu(x, B)$ is a probability measure on $(\mathcal{F}, \mathcal{B})$ and equals $\Pi(B|X)$ almost surely.

A sufficient condition for this is when $\mathcal{F}$ is a Polish space equipped with Borel σ-algebra $\mathcal{B}$.

3.2 Gugushvili

We consider the problem of estimating density f_0 from the non-parametric class $\mathcal{F}$ of location-scale mixtures of normal densities.

3.2.1 Mixtures

A powerful technique, especially for computation, is to construct a prior on probability densities is through mixtures.

For given probability density functions $\{\psi_\theta : \theta \in \Theta\}$ (such that the map $x \mapsto \psi_\theta(x) = \psi(\theta, x)$ is measurable in both arguments) and a probability distribution F on the parameter space $(\Theta, \mathcal{T})$, let

$$p_F(x) = \int f(\theta, x) dF(x)$$

Then a prior on F induces a prior on densities. The function ψ is referred to as the kernel function and the measure F is called the mixing distribution.

In our case we use a location-scale mixture of normal densities. Therefore, our kernel is of the form $x \mapsto \frac{1}{\sigma}\psi\left(\frac{x-\mu}{\sigma}\right)$ where ψ denotes the probability density function of a standard normal and $\theta = (\mu, \sigma)$.

Such a family is rich since by Fejer's theorem,

$$\int \frac{1}{\sigma}\psi\left(\frac{x-\mu}{\sigma}\right) f(\mu) d\mu \to f(\cdot) \text{ as } \sigma \to 0 \text{ in } L^1$$

Thus, a prior on densities may be induced by putting a prior on the mixing distribution F and a prior on σ restricted to positive values.

The mixing distribution F may be given a Dirichlet process prior since it leads to efficient computational algorithms as well as rich convergence properties. Thus, for our discussion, such a prior is particularly attractive.

3.2.2 Priors on Spaces of Probability Measures

As we have seen in the previous section, placing a prior on the mixing distribution F induces a prior on the space of probability densities. This construction comes with some technical complications.

We provide a rigorous discussion of its construction nad then introduce methods of constructing priors, in particular stick breaking and the Pólya urn scheme.

Let $(\mathfrak{X}, \mathcal{X})$ be a Polish space and consider priors on the collection $\mathfrak{M}$ of all probability measures on $(\mathfrak{X}, \mathcal{X})$. A prior Π on $\mathfrak{M}$ can be viewed as the distribution of a *random measure* P.

Definition 3.2. A random measure, in this case, can be thought of as a random element from some probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to our collection of probability measures $(\mathfrak{M}, \mathcal{B}(\mathfrak{M}))$ equipped with a suitable Borel σ-field. This is equivalent to being a (a.s.) locally finite Markov kernel from a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ to our sample space $(\mathfrak{X}, \mathcal{X})$.

Thus we can view P as a map $P : \Omega \times \mathcal{X} \to \mathbb{R}$ such that P_ω is a measure on $(\mathfrak{X}, \mathcal{X})$ for every $\omega \in \Omega$. Thus $P_\omega \in \mathfrak{M}$.

Now, also from the definition, for fixed $A \in \mathcal{X}$, $P(A) : \Omega \to \mathbb{R}$ is a random variable. Therefore, we get that $(P(A) : A \in m\mathcal{X})$ is a stochastic process on the underlying probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

Then Π is a measure on $(\mathfrak{M}, \mathcal{M})$ where $\mathcal{M}$ is some Borel σ-field for the weak topology etc... and it makes sense to talk about

$$\Pi(M) = \Pr(P \in M)$$

for some probability measure Pr on $(\mathfrak{M}, \mathcal{M})$. Therefore, we can speak of drawing a measure P from the prior Π and then sampling observations $X \in \mathfrak{X}$ from P.

3.2.3 Countable Sample Spaces

A probability distribution on a countable sample space (equipped with the σ-field of all its subset) can be represented as a infinite-length probability vector $s = (s_1, s_2, \dots)$, thus assigning probability weights to each element in the sample space. This can be trivially seen in, for example, a Poisson distribution where we get the probability vector:

$$\left(e^{-\lambda} \frac{\lambda^k}{k!} \right)_{k=0}^{\infty}$$

Therefore, a prior on the set $\mathfrak{M}$ of all probability measures on a countable sample space can therefore be identified with the distribution of a random

element with values in the countable-dimensional unit simplex

$$\mathbb{S}_\infty = \left\{ s = (s_1, s_2, \dots) : s_j \geq 0, j \in \mathbb{N}, \sum_{j=1}^{\infty} s_j = 1 \right\} \tag{17}$$

Characterising the σ-field on $\mathbb{S}_\infty$ by the σ-algebra generated by the coordinate maps $s \mapsto s_i, i \in \mathbb{N}$, then this shows that a map p from some probability space into $\mathbb{S}_\infty$ is a random element if and only if every coordinate variable p_i is a random variable. Hence a prior corresponds to an infinite sequence of nonnegative random variables $p_1, p_2, \dots$ that add up to 1.

We can use this fact to sample from our prior Π by obtaining a sequence $Y_1.Y_2, \dots$ of random variables such that their sum $\sum_{i=1}^{\infty} Y_i = 1$. We do this via several methods.

3.2.4 Construction through Normalization

Given nonnegative random variables $Y_1, Y_2, \dots$ such that $\sum_{j=1}^{\infty} Y_j$ is positive and converges a.s. we can define a prior on $\mathbb{S}_\infty$ by putting

$$p_k = \frac{Y_k}{\sum_{j=1}^{\infty} Y_j}, \qquad k \in \mathbb{N} \tag{18}$$

A simple sufficient condition for the a.s. convergence of the random series is that $\sum_{j=1}^{\infty} \mathbb{E}(Y_j) < \infty$. The construction follows by the following lemma:

Lemma 3.1. *If $Y_1, Y_2, \dots$ are independent, positive random variables with $\sum_{j=1}^{\infty} Y_j < \infty$ a.s. and marginal densities, i.e. the density of $(Y_i)_{i \in I}$ where $I \subset \mathbb{N}$, is positive everywhere in $(0, \infty)$, then the support of the prior defined by (18) is equal to the full space $\mathbb{S}_\infty$.*

3.2.5 Construction through Stick Breaking

We perform the following algorithm to distribute the total mass 1, conceptually visualised with a stick of length 1, randomly to each element of $\mathbb{N}$.

1. We first break the stick at the point given by the random variable V_1 where $0 \leq V_1 \leq 1$ and assign mass V_1 to p_1.
2. We think of the remaining mass $1 - V_1$ as a new stick and break it into two pieces of relative lengths V_2 and $1 - V_2$ according to the value of random variable V_2. We assign mass $(1 - V_1)V_2$ to the point p_2
3. We repeat in this way so that point p_j has mass

$$p_j = \left(\prod_{l=1}^{j-1} (1 - V_l) \right) V_j \tag{19}$$

3.2.6 Countable Dirichlet Process

We construct the countable Dirichlet distribution via stick breaking: WE choose the variables $V_1, V_2, \dots$ independently with $V_j \sim \text{Be}(\alpha_j, \sum_{l=j+1}^{\infty} \alpha_l$ and define $(p_1, p_2, \dots)$ by (19).

It follows easily that for any $k \in \mathbb{N}$,

$$\left(p_1, \dots, p_k, 1 - \sum_{j=1}^{k} p_j \right) \sim \text{Dir} \left(k+1; \alpha_1, \dots, \alpha_k, \sum_{j=k+1}^{\infty} \alpha_j \right) \tag{20}$$

and so we have a Dirichlet Process

Definition 3.3. A random measure P on $(\mathfrak{X}, \mathcal{X})$ is said to possess a Dirichlet process distribution $\text{DP}(\alpha)$ with base measure α on the measurable space $(\mathfrak{X}, \mathcal{X})$ if, for every finite measurable partition $A_1, \dots, A_k$ of $\mathfrak{X}$, we have that the joint distribution of random variables $P(A_1), \dots, P(A_k)$ satisfy

$$(P(A_1), \dots, P(A_k)) \sim \text{Dir}\left(k+1; \alpha(A_1), \dots, \alpha(A_k)\right)$$

Therefore, in our case, the random variables $(p_1, p_2, \dots)$ follow a Dirichlet process.

Suppose that we have a sample of i.i.d observations $X_1, \dots, X_n$ from p. The likelihood L, therefore, is given by

$$L(p) = \prod_{j=1}^{\infty} p_j^{N_j}$$

where $N_j = \sum_{i=1}^{n} X_i^{(j)}$ i.e the sum of the jth component of all the observations.

The vector $N = (N_1, N_2, \dots)$ is a sufficient statistic, and hence the posterior given N is the same as the posterior given the original observations.

For any $l \in \mathbb{N}$,

$$\left(N_1, N_l, n - \sum_{j=1}^{l} N_j \right) \sim \text{MN}_{l+1} \left(n; p_1, \dots, p_l, 1 - \sum_{j=1}^{\infty} p_j \right). \tag{21}$$

By conjugacy of the finite Dirichlet distribution with the multinomial likelihood, we get that the posterior density of $(p_1, \dots, p_l, 1 - \sum_{j=1}^{l} p_j)$ given $N_1, \dots, N_l$ is given by

$$\text{Dir} \left(l+1; \alpha_1 + N_1, \dots, \alpha_l + N_l, \sum_{j=l+1}^{\infty} \alpha_j + n - \sum_{j=1}^{l} N_j \right)$$

Using part (i) of Proposition G.3, we can aggregate entries in the last cell on a partition $\{1\}, \dots, \{k\}, \{k+1, k+2, \dots\}$ for $(k \leq l)$ to give that the posterior density of $\left(p_1, \dots, p_k, 1 - \sum_{j=1}^{k} p_j\right)$ given $N_1, \dots, N_l$ is given by

$$\text{Dir}\left(k+1; \alpha_1 + N_1, \dots, \alpha_k + N_k, \sum_{j=k+1}^{\infty} \alpha_j + n - \sum_{j=1}^{k} N_j\right)$$

Since this density only depends on $N_1, dotsc, N_k$ we get that this posterior density of $\left(p_1, \dots, p_k, 1 - \sum_{j=1}^{k} p_j\right)$ given $N_1, \dots, N_l$ is the same for every $l \geq k$. Thus, since we have that $\sigma(N_1, \dots, N_l)$ is a filtration in l and $\sigma(N)$ is the minimal σ-algebra generate by this filtration then by Lévy's zero-one law, the above posterior density is the same for $\left(p_1, \dots, p_k, 1 - \sum_{j=1}^{k} p_j\right)$ given N.

The posterior parameters depend on the prior parameters and the new parameter is given simply by $\alpha \mapsto \alpha + N$ (element-wise).

In analogy with thte finite dimensional Dirichlet distribution, it is natural to call the prior the *Dirichlet process* on $\mathbb{N}$ (or the *countable Dirichlet process.* We shall write $(p_1, p_2, \dots) \sim \text{DP}(\alpha)$ where $\alpha = (\alpha_1, \alpha_2, \dots)$.

From the properties of the finite-dimensional Dirichlet distribution, the posterior mean can be written as

$$\mathbb{E}\left(p_j | X_1, \dots, X_n\right) = \frac{a_j + N_j}{\sum_{l=1}^{\infty} a_l + n}$$

This seems intuitively correct, since higher N_j will provide a greater mass on p_j.

3.2.7 Dirichlet Process on Arbitrary Sample Spaces

We saw previously that for a countable Dirichlet process prior, the posterior distribution is again a countable Dirichlet process. This result extends also to Dirichlet process on an arbitrary space.

Consider observations $X_1, X_2, \dots, X_n$ sampled independently from a distribution P that was drawn from a Dirichlet prior distribution. By some abuse of language, we will say that this sample is from the Dirichlet process.

Theorem 1. *(Conjugacy) The DP($\alpha + \sum_{i=1}^{n} \delta_{X_i}$) process is a version of the posterior distribution given an i.i.d sample $X_1, \dots, X_n$ from the DP(α)-process.*

3.2.8 Construction through Pólya Urn Scheme

1. $X_1 \sim \bar{\alpha}$
2. $X_{n+1} | X_1, \dots, X_n \sim \frac{\alpha + \sum_{i=1}^{n} \delta_{X_i}}{|\alpha| + n}$

By de Finetti's theorem, there exists a random probability measure P such that $X_i|P \overset{\text{i.i.d}}{\sim} P$. We see in Section 4.2.4 that the law of P is the DP(α)-process.

3.2.9 Stick-Breaking Representation

The stick-breaking representation of a Dirichlet process expresses it as a random discrete measure with stick-breaking weights, based on the beta-distribution.

The representation gives an easy method to simulate a Dirichlet process, at least approximately. We use this to simulate our Dirichlet process

Theorem 2. *If $\theta_1, \theta_2, \ldots \overset{iid}{\sim} \bar{\alpha}$ and $V_1, V_2, \ldots \overset{i.i.d}{\sim} Be(1, M)$ are independent random variables and $W_j = V_j \prod_{l=1}^{j-1}(1 - V_l)$ then $\sum_{j=1}^{\infty} W_j \delta_{\theta_j} \sim DP(M\bar{\alpha})$.*

Note that

Theorem 3. *(Discreteness) Almost every realization from the DP(α) is a discrete measure: $DP_\alpha(P \in \mathfrak{M} : P \text{ is discrete}) = 1$.*

This is perhaps disappointing, especially if the intention is to model absolutely continuous probability measures. In particular, the Dirichlet process cannot be used as a prior for density estimation. We therefore actually use a Dirichlet Process mixture.

3.2.10 Dirichlet Process Mixtures

The discreteness of the Dirichlet process makes it useless as a prior for estimating a density. This can be remedied by convolving it with a kernel. Although it is rarely possible to characterize the resulting posterior distribution analytically, a variety of efficient and elegant algorithms allows us to approximate it numerically.

We follow the same definition in Section 3.2.1 with F equipped with Dirichlet process prior but we also allow the kernel to depend on an additional parameter $\phi \in \Phi$, giving mixtures $p_{F,\phi}(x) = \int \psi_i(x; \theta, \phi) dF(\theta)$.

For $x \mapsto \psi_i(x; \theta, \phi)$ probability density functions (relative to a given σ-finite dominating measure ν, consider

$$X_i \overset{\text{i.i.d}}{\sim} p_{i,F,\phi}(x) = \int \psi_i(x; \theta, \phi) dF(\theta)$$

We equip F and ϕ with independent priors $F \sim \text{DP}(\alpha)$ and $\phi \sim \pi$.

The resulting model can be equivalently written in terms of n latent variables $\theta_1, \ldots, \theta_n$ as

$$X_i|\theta_i, \phi, F \overset{\text{ind}}{\sim} \psi_i(\cdot; \theta_i, \phi), \qquad \theta_i|F, \phi \overset{\text{iid}}{\sim} F, \qquad F \sim \text{DP}(\alpha), \qquad \phi \sim \pi \quad (22)$$

The latent variables $\theta_1, \ldots, \theta_n$ help to make the description simpler, since $F|\theta_1, \ldots, \theta_n \sim \mathrm{DP}(\alpha + \sum_{i=1}^n \delta_{\theta_i})$

The posterior distribution of any object of interest can be described in terms of the posterior distribution of (F, ϕ) given $X_1, \ldots, X_n$.

Note that for any measurable function ψ

$$\mathbb{E}\left[\int \psi dF|\theta_1, \ldots, \theta_n, X_1, \ldots, X_n\right] = \frac{1}{|\alpha| + n}\left[\int \psi d\alpha + \sum_{j=1}^n \psi(\theta_j)\right] \quad (23)$$

We use this to simulate from the posterior distribution of $(\theta_1, \ldots, \theta_n)$ based on a weighted generalized Pólya urn scheme.

Theorem 4. *(Gibbs sampler) In the model (22) the conditional posterior distribution of θ_i is given by*

$$\theta_i|\theta_{-1}, \psi, X_1, \ldots, X_n \sim \sum_{j \neq i} q_{i,j}\delta_{\theta_j} + q_{i,0}G_{b,i} \quad (24)$$

where $(q_{i,j} : j \in \{0, 1, \ldots, n\} \setminus \{i\})$ is the probability vector satisfying

$$q_{i,j} \propto \begin{cases} \psi_i(X_i; \theta_j, \phi), & j \neq i, j \geq 1, \\ \int \psi_i(X_i; \theta, \phi) d\alpha(\theta), & j = 0, \end{cases}$$

and $G_{b,i}$ is the baseline posterior measure given by

$$dG_{b,i}(\theta|\phi, X_i) \propto \ \psi_i(X_i; \theta, \phi) d\alpha(\theta)$$

See MCMC Methods in section 5.2 to simulate from the posterior distribution in the Dirichlet process. The basic algorithm uses the Gibbs sampling in above theorem to generate $\theta_1, \ldots, \theta_n$ given $X_1, \ldots, X_n$ and then we can readily obtain the posterior F given $X_1, \ldots, X_n$.

3.3 Gugushvili Problem Formulation

We consider the problem of estimating density f_0 from the non-parametric class $\mathcal{F}$ of location mixtures of normal densities. These are not location-scale since we fix σ.

Example. Consider distribution F on $(\mathbb{R}, \mathcal{B})$ given by

$$F(\cdot) = \sum_{j=1}^n p_j \delta_{\mu_j}(\cdot)$$

where $\sum_{j=1}^n p_j = 1$ and δ_{μ_i} denotes the dirac measure centred at $\mu_i \in \mathbb{R}$.

Then we get that

$$\int \frac{1}{\sigma}\psi\left(\frac{x-\mu}{\sigma}\right) dF(\mu) = \sum_{j=1}^{n} p_j \frac{1}{\sigma}\psi\left(\frac{x-\mu_j}{\sigma}\right)$$

so we get a weighted sum of normal densities. Therefore, mixtures basically generalise this idea of returning a weighted sum of densities from a specific parametric family.

3.3.1 Prelims and Notation

Let $\mathbb{P}_f$ be the law of the jumps Y_1 of the CPP.

Let $\mathbb{Q}^{\Delta}_{\lambda,f}$ be the law of the increments Z_1 of the discretely observed points of the CPP. These Z_i have same distribution as a Poisson random sum

Let $\mathbb{Q}^{\Delta,n}_{\lambda,f}$ be the joint law of the increments $\mathcal{Z}^{\Delta}_n = (Z^{\Delta}_1, \ldots, Z^{\Delta}_n)$ of the discretely observed points of the CPP.

Let $\mathbb{R}^{\Delta}_{\lambda,f}$ be the law of $(X_t : t \in [0,\Delta])$ i.e. the law of the CPP restricted to interval $[0,\Delta]$.

By (2.1) (we need to change it to consider Δ) we know that the characteristic function of the Poisson random sum X is given by

$$\phi_X(t) = e^{-\lambda\Delta+\lambda\Delta\phi_f(t)}$$

We also saw this can be written as

$$\phi_X(t) = e^{-\lambda\Delta} + \frac{1-e^{-\lambda\Delta}}{e^{\lambda\Delta}-1}(e^{\lambda\Delta\phi_f(t)}-1)$$

Since we are only considering mixtures of normal densities as our 'true' density f we know that $\phi_f(t) \to 0$ as $t \to \infty$ (remember for $N(\mu,\sigma^2)$, $\phi_f(t) = e^{it\mu-\frac{1}{2}\sigma^2t^2}$). Therefore, we get that $\phi_X(t) \to e^{-\lambda\Delta}$ as $t \to \infty$. Thus, we see that since $\phi_X(t)$ converges to an injective function, λ is identifiable from X. (This may be irrelevant).

Proposition 3.1. *The law $\mathbb{Q}^{\Delta}_{\lambda,f}$ of X is absolutely continuous with respect to the measure $\mu = \delta_{\{0\}} + \text{Leb}$ and has Radon-Nikodym derivative*

$$\frac{\mathrm{d}\mathbb{Q}^{\Delta}_{\lambda,f}}{\mathrm{d}\mu}(x) = e^{\lambda\Delta}\mathbb{1}_{\{0\}}(x) + (1-e^{\lambda\Delta})\sum_{m=1}^{\infty} a_m(\lambda\Delta) f^{*m}(x)\mathbb{1}_{\mathbb{R}\setminus\{0\}}(x) \tag{25}$$

where

$$a_m(\lambda\Delta) = \frac{1}{e^{\lambda\Delta}-1}\frac{(\lambda\delta)^m}{m!} \tag{26}$$

and $f * m$ denotes the m-fold convolution of f with itself.

Proof. Suppose we have $A \in \mathcal{B}$ such that $\mu(A) = 0$. Then $0 \notin A$ and A has Lebesgue-measure zero. Therefore, under event A, the Poisson random sum X is non-zero and we have

$$\begin{aligned}\mathbb{Q}^{\Delta}_{\lambda,f}(A) &= \sum_{n=1}^{\infty} \mathbb{P}(Y_1 + \cdots + Y_n \in A, N = n) \\ &= \sum_{n=1}^{\infty} \mathbb{P}(Y_1 + \cdots + Y_n \in A)\mathbb{P}(N = n) \\ &= 0\end{aligned}$$

since $Y_1 + \cdots + Y_n$ has a density. Therefore, the first statement follows. Note that for $A \in \mathcal{B}$

$$\begin{aligned}\mathbb{Q}^{\Delta}_{\lambda,f}(A) &= \mathbb{P}(N = 0) \int_A d\delta_{\{0\}} + \sum_{n=1}^{\infty} \mathbb{P}(Y_1 + \cdots + Y_n \in A)\mathbb{P}(N = n) \\ &= e^{\lambda\Delta} \int_A d\delta_{\{0\}} + (1 - e^{\lambda\Delta}) \sum_{n=1}^{\infty} a_m(\lambda\Delta) \int_A f^{*m}(x)dx && \text{Using Lemma below} \\ &= \int_A \left[e^{\lambda\Delta} \mathbb{1}_{\{0\}}(x) + (1 - e^{\lambda\Delta}) \sum_{m=1}^{\infty} a_m(\lambda\Delta) f^{*m}(x) \mathbb{1}_{\mathbb{R}\setminus\{0\}}(x) \right] d\mu(x)\end{aligned}$$

□

Lemma 3.2. *Let X and Y be two independent random variables with density functions f_X and f_Y. Then the sum $Z = X + Y$ is a random variable with density function f_Z, where f_Z is the convolution of f_X and f_Y.*

3.3.2 Algorithms for drawing from the posterior

We now discuss how to draw from the distribution of f, conditional on $\mathcal{Z}^{\Delta}_n$. As before, we assume observations at equidistant times $0, \Delta, 2\Delta, \ldots, n\Delta$. We set $Z_i = X_{i\Delta} - X_{(i-1)\Delta}$ and $Z = (Z_1, \ldots, Z_n)$. We also assume for simplicity that λ is known and fixed.

As before, since a zero increment provides no additional information on the density of f, we can discard these carefully. If we could obtain a tractable form for the conditional density of a nonzero increment Z_i then we can directly apply Bayes theorem directly and use a standard MCMC algorithm to sample from the posterior.

However, the conditional density, given by

$$p(z|f) = \frac{e^{-\lambda\Delta}}{1 - e^{-\lambda\Delta}} \sum_{k=1}^{\infty} \frac{(\lambda\Delta)^k}{k!} f^{*k}(z) \tag{27}$$

is generally rather intractable due to the infinite weighted sum of convolutions. To see this, suppose simply that f is a weighed sum of 2 normal densities:

$$f(\cdot) = \rho_1\psi(\cdot;\mu_1,\sigma) + \rho_2\psi(\cdot;\mu_2,\sigma)$$

Then we get

$$f^{*k}(\cdot) = \sum_{l=0}^{k} \binom{k}{l} \rho_1^l \rho_2^{k-1} \psi(\cdot; l\mu_1 + (k-l)\mu_2, k\sigma)$$

Plugging this into (27) we immediately see that it will be difficult to compute values (even approximately) for the density.

Therefore, we introduce auxiliary variables to circumvent the intractable form of $p(z|f)$. To motivate the choice of auxiliary variables, consider the case where the CPP is observed continuously on the interval $[0,T]$. Then we know exactly when the jumps have occurred. Suppose that jumps occur at $T_1 < T_2 < \cdots < T_N$ where $T_N \leq T$. Suppose that J_i are the corresponding jump sizes. Also suppose, for convention, that $T_0 = 0$. Then the continuous time likelihood

$$\begin{aligned} L([0,T]) &= L(\{(T_i,J_i)\}_{i=1}^N) \\ &= \prod_{i=1}^{N} e^{-\lambda(T_i - T_{i-1})} \lambda f(J_i) \\ &= \lambda^N e^{-\lambda T_N} \prod_{i=1}^{N} f(J_i) \end{aligned}$$

This likelihood is tractable and so it naturally suggests that we construct a data augmentation scheme to 'fill in' missing jump points in between our observations.

(Maybe show by picture)

We refer to the set of missing values in the $i$ith segment by $x_{(i-1,i)}$. Therefore, we get that the set of all jump values in the CPP is given by

$$x^{\text{mis}} = \bigcup_{i=1} x_{(i-1,i)}$$

Once we have this, then we are back to the continuous time case and we have a tractable likelihood. We want a Markov chain with invariant distribution $p(x^{\text{mis}}, f|Z)$. Therefore, our algorithm is as follows:

1. Initialise x^{mis}.
2. Draw $f|(x^{\text{mis}}, Z)$. We do this via maximising the likelihood explained above.

3. Draw $x^{\text{mis}}|(f, Z)$.

4. Repeat Steps 2 and 3 many times

Under weak conditions, the iterate for f are draws from the required posterior distribution. We, however, need to perform imputation in Step 3 to get draws of x^{mis}. This is nontrivial but we will see that we can do with less imputation.

3.3.3 Auxiliary Variables

Suppose Y has density f. Since f is a mixture (weighted sum) of normal densities, then drawing realizations of Y can be simulated by first drawing label l from $1, \ldots, J$ with probability ρ_l and then drawing from the $N(\mu_l, \sigma^2)$ distribution.

Our auxiliary variables are the number of jumps on segment i from the label j which we denote n_{ij}. For segment i denote

$$a_i = (n_{i1}, n_{i2}, \ldots, n_{iJ})$$

We set

$$n_i = \sum_{j=1}^{J} n_{ij}, \qquad s_j = \sum_{i=1}^{n} n_{ij}, \qquad s = \sum_{j=1}^{J} s_j = \sum_{i=1}^{n} n_i$$

so n_i denotes the total number of jumps in the ith segment, s_j denotes the total number of jumps on type/label j, s denotes the total number of jumps of all types.

3.3.4 Re-parametrisation and Prior Specification

We define

$$\psi_j = \lambda \rho_j, \qquad j = 1, \ldots, J$$

Then

$$\lambda = \sum_{j=1}^{J} \psi_j, \qquad \rho_j = \frac{\psi_j}{\sum_{j=1}^{J} \psi_j}$$

The reason for this re-parametrisation is that a CPP X whose jumps are of J types can be decomposed as $Z = \sum_{j=1}^{J} Z_j$, where the Z_j are CPPs whose jumps are of type j only and where the parameter of the Poisson random variable is ψ_j. We denote $\theta = (\psi, \mu, \tau)$ with $\psi = (\psi_1, \ldots, \psi_J)$ and $\mu = (\mu_1, \ldots, \mu_J)$ and $\tau = \frac{1}{\sigma}$.

We take priors

$$\begin{aligned} \psi_1, \ldots, \psi_J &\overset{\text{i.i.d}}{\sim} \mathcal{G}(\alpha_0, \beta_0) \\ \mu_1, \ldots, \mu_J|\tau &\overset{\text{ind}}{\sim} \mathcal{N}(\xi_j, \tfrac{1}{\kappa\tau}) \\ \tau &\sim \mathcal{G}(\alpha_1, \beta_1) \end{aligned}$$

with positive hyper-parameters $(\alpha, \beta, \kappa, \xi)$ fixed.

3.3.5 Hierarchical Model

We want to find $p(a, \theta|Z)$. We do this using

$$\begin{aligned} p(\theta, a|Z) &= p(a|\theta, Z)p(\theta|Z) \\ &\propto p(a|\theta, Z)p(Z|\theta)p(\theta) \end{aligned}$$

There are two clear steps for sampling for the required distribution. Therefore, we write our observation as a hierarchical model:

$$\begin{aligned} \theta &\sim \pi(\theta) \\ n_{ij}|\psi &\overset{\text{ind}}{\sim} \mathcal{P}(\psi_j \Delta) \\ Z_i|a_i, \mu, \tau &\overset{\text{ind}}{\sim} \mathcal{N}(a_i^T \mu, \tfrac{n_i}{\tau}) \end{aligned}$$

We construct a Metropolis-Hastings algorithm to draw from $(\theta, a|Z)$. The two main steps are:

1. Update segments: For each segment $i = 1, \ldots, n$, draw a_i conditional on (θ, Z, a_{-i})
2. Update parameters: draw θ conditional on (Z, a)

3.3.6 Updating segments

We have to draw from

$$p(a_i|\theta, Z, a_{-i}) = \frac{p(\theta, Z, a)}{p(\theta, Z, a_{-i})} \tag{28}$$

We use our hierarchical model to express $p(\theta, Z, a)$ in terms of known distributions:

$$\begin{aligned} p(\theta, Z, a) &= p(\theta)p(Z, a|\theta) \\ &= p(\theta)p(Z|a, \theta)p(a|\theta) \\ &= p(\theta) \left(\prod_{i=1}^{n} \phi\left(Z_i; a_i^T \mu, \frac{n_i}{\tau}\right) \right) \left(\prod_{i=1}^{n} \prod_{j=1}^{J} e^{-\psi_i \Delta} \frac{(\psi_i \Delta)^{n_{ij}}}{n_{ij}!} \right) \end{aligned}$$

Therefore, we get that

$$p(a_i|\theta, Z, a_{-i}) \propto \phi\left(Z_i; a_i^T \mu, \frac{n_i}{\tau}\right) \prod_{j=1}^{J} e^{-\psi_i \Delta} \frac{(\psi_i \Delta)^{n_{ij}}}{n_{ij}!}$$

We do this using a Metropolis Hastings step: First we draw a proposal n_i^* for n_i from a $\mathcal{P}(\lambda\Delta)$ distribution conditioned to have non-zero outcome. In other words we sample from a distribution with mass function

$$p(k) = \frac{e^{-\lambda\Delta}}{1 - e^{-\lambda\Delta}} \frac{(\lambda\Delta)^k}{k!}, \qquad k = 1, 2, \ldots$$

Next we draw

$$a_i^* = (n_{i1}^*, \ldots, n_{iJ}^*) \sim \mathcal{MN}\left(n_i^*; \frac{\psi_1}{\lambda}, \ldots, \frac{\psi_J}{\lambda}\right)$$

thus sampling the number of jumps in segment i of each type.

Therefore, the proposal density is given by

$$\begin{aligned} q(a_i^*|\theta) &= \frac{e^{-\lambda\Delta}}{1-e^{-\lambda\Delta}} \frac{(\lambda\Delta)^{n_i^*}}{n_i^*!} \binom{n_i^*}{n_{i1}^*, \ldots, n_{iJ}^*} \prod_{j=1}^{J} \left(\frac{\psi_j}{\lambda}\right)^{n_{ij}^*} \\ &= \frac{e^{-\lambda\Delta}}{1-e^{-\lambda\Delta}} \prod_{j=1}^{J} \left(\frac{\psi_j \Delta}{n_{ij}^*}\right)^{n_{ij}^*} \end{aligned}$$

The acceptance probability for the proposal n^* and the a_i^* is

3.3.7 Updating parameters

We update the parameters directly using the following Lemma.

Lemma 3.3. *Conditional on a, we have that $\psi_1, \ldots, \psi_J$ are independent and the following holds:*

$$\begin{aligned} \psi_j|a &\overset{ind}{\sim} \mathcal{G}(\alpha_0 + s_j, \beta_0 + n\Delta) \\ \tau|z,a &\sim \mathcal{G}(\alpha_1 + (n+J)/2, \beta_1 + (R - q^T P^{-1} q)/2) \\ \mu|\tau,z,a &\sim \mathcal{N}(P^{-1}q, \tau^{-1}P^{-1}) \end{aligned}$$

where P is the symmetric $J \times J$ matrix given by

$$P = \kappa I_{J\times J} + \tilde{P}, \qquad \tilde{P}_{jk} = \sum_i n_i^{-1} n_{ij} n_{ik}$$

q is the J-dimensional vector with

$$q_j = \kappa\xi_j + \sum_i n_i^{-1} n_{ij} z_i$$

$R > 0$ is given by

$$R = \kappa \sum_{j=1}^{J} \xi_j^2 + \sum_i n_i^{-1} z_i^2$$

and $R - q^T P^{-1} q > 0$.

Note that adding $\kappa I_{J\times J}$ ensures the invertibility of P.

Proof.

$$\begin{aligned}
p(\psi|\mu,\tau,z,a) &= p(\psi|a) \\
&\propto p(a|\psi)\pi(\psi) \\
&= \prod_{j=1}^{J}\pi(\psi_j)\left(\prod_i p(n_{ij}|\psi)\right) \\
&\propto \prod_{j=1}^{J}\pi(\psi_j)\left(\prod_i e^{-\psi_j\Delta}(\psi_j\Delta)^{n_{ij}}\right) \\
&= \prod_{j=1}^{J}\pi(\psi_j)e^{-\psi_j n\Delta}(\psi_j\Delta)^{s_j} \\
&\propto \prod_{j=1}^{J}\psi_j^{\alpha_0-1}e^{-\beta_0\psi_j}e^{-\psi_j n\Delta}(\psi_j\Delta)^{s_j} \\
&= \prod_{j=1}^{J}e^{-\psi_j(\beta_0+n\Delta)}\psi_j^{s_j+\alpha_0-1}
\end{aligned}$$

giving the required $\mathcal{G}(s_j+\alpha_0, \beta_0+n\Delta)$ distribution. For (μ,τ) we get

$$\begin{aligned}
p(\mu,\tau|z,a) &\propto p(z|a,\mu,\tau)p(\mu,\tau|a) \\
&= p(z|a,\mu,\tau)\pi(\mu|\tau)\pi(\tau) \\
&\propto \left(\prod_{i=1}^{n}\phi(z_i;a_i^T\mu,n_i/\tau)\right)\left(\tau^{\alpha_1-1}e^{-\beta_1\tau}\right)\left(\tau^{J/2}\exp\left\{-\frac{\tau\kappa}{2}\sum_{j=1}^{J}(\mu_j-\xi_j)^2\right\}\right) \\
&\propto \tau^{\alpha_1-1+(n+J)/2}\exp\left(-\beta_1\tau-\frac{D(\mu)}{2}\tau\right)
\end{aligned}$$

where

$$\begin{aligned}
D(\mu) &= \kappa\sum_{j=1}^{J}(\mu_j-\xi_j)^2+\sum_{i=1}^{n}n_i^{-1}(z_i-a_i^T\mu)^2 \\
&= \mu^T P\mu - 2q^T\mu + R
\end{aligned}$$

by easy calculation. Note that, by completing the square,

$$\mu^T P\mu - 2q^T\mu = (\mu-P^{-1}q)^T P(\mu-P^{-1}q)+q^T P^{-1}q$$

It follows by this that $\mu|\tau,z,a \sim \mathcal{N}(P^{-1}q,\tau^{-1}P^{-1})$ Also,

$$\int\exp(-\frac{\tau}{2}D(\mu))d\mu \propto e^{-\frac{\tau R}{2}}\int\exp\left(-\frac{\tau}{2}(\mu-P^{-1}q)^T P(\mu-P^{-1}q)\right)d\mu \tag{29}$$

Thus, we get that

$$
\begin{aligned}
p(\tau|z,a) &= \int p(\tau,\mu|z,a)d\mu \\
&\propto \tau^{\alpha_1-1+(n+J)/2}\exp(-\beta_1\tau)e^{-\tau R/2}(2\pi)^{J/2}\sqrt{|\tau^{-1}P^{-1}}\exp\left(\frac{1}{2}\tau q^T P^{-1}q\right) \\
&\propto \tau^{\alpha_1+(n+J)/2-1}\exp\left\{-\tau\left(\beta_1+\frac{1}{2}(R-q^TP^{-1}q)\right)\right\}
\end{aligned}
$$

giving that $\tau|z,a \sim \mathcal{G}(\alpha_1+(n+J)/2, \beta_1+(R-q^TP^{-1}q)/2)$ □