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

Contents

1 Introduction

Continuous-time jump processes are a class of càdlag random processes which exhibit almost-sure discontinuities. The compound Poisson process is one of the simplest examples of a continuous-time jump process, yet its ability to have jumps of varying size makes it an attractive candidate for modelling phenomena in the world. Examples of its use include modelling insurance claims (see [14]), natural disasters (see [17]) and the movement of financial instruments (see [13]).

Statistical inference on any non-trivial continuous-time stochastic process which is observed only on a discrete set of points is inherently difficult due to the information lost, and thus invokes an inverse problem. The simplicity of the compound Poisson process allows us to explore this inverse problem in a rigorous way, obtaining theoretical guarantees for our estimators. Furthermore, we gain a better understanding of more intricate processes which are built upon the compound Poisson process.

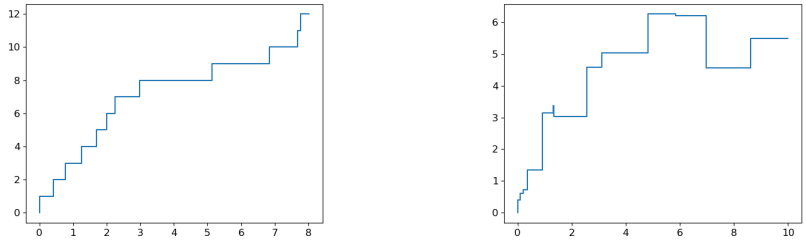
In this essay, we discuss and implement various non-parametric methods to perform inference on compound Poisson processes. In particular, we first employ a spectral approach using kernel density estimators and illustrate their performance through numerical simulations. We follow the 2007 paper of van Es et al. [16] for the first kernel density estimator and the 2014 paper of Comte et al. [4] for the second. We then visit non-parametric Bayesian inference and describe the advantages of this approach. We develop the MCMC algorithm as shown in the 2016 paper of Gugushvili et al. [11] for a practical implementation of this approach. Finally, we generalise the approach of [11] by taking a Dirichlet process mixture prior. Numerical simulations illustrate the feasibility as well as limitations of this approach.

Since this essay is computational in flavour, we focus mainly on carefully deriving the forms of the estimators and analyse their performance on various numerical simulations. We briefly present theoretical guarantees of these estimators, however, references will be given for the proofs of these results.

1.1 Compound Poisson Processes

Properties. The compound Poisson process generalises the Poisson process by allowing the jump sizes to follow a distribution rather than being of fixed unit size. We associate compound Poisson processes with the following three key properties:

1. The occurrence of the jumps follow a Poisson process,
2. The jumps sizes are independent, and follow a common distribution,
3. The inter-arrival times between the jumps and the sizes of the jumps are mutually independent random variables.



(a) Poisson Process (b) Compound Poisson Process

Figure 1.1: Simulations of Jump Processes

Definition 1.1 (Poisson Process)**.** A Poisson process with intensity λ is a non-negative, non-decreasing, integer-valued stochastic process $(N_t)_{t\geq 0}$ starting at 0 with the following properties:

1. **Independent Increments:** For every $n \in \mathbb{N}$ and $t_1, \ldots, t_n$ such that $0 \leq t_1 \leq t_2 \leq \cdots \leq t_n$, we have that

$$N_{t_n} - N_{t_{n-1}}, \ldots, N_{t_2} - N_{t_1}, N_{t_1}$$

are mutually independent,

2. **Stationary Increments:** The number of occurrences in any interval of length Δ is a Poisson random variable with parameter $\lambda\Delta$ i.e. for every $\Delta > 0$ and $t \geq 0$ we have that

$$N_{(t+1)\Delta} - N_{t\Delta} \sim \mathcal{P}(\lambda\Delta).$$

As we can see in Figure 1.1a, the jump sizes are of unit length, whilst in Figure 1.1b, the jumps to vary in both magnitude and direction.

Definition 1.2 (Compound Poisson Process)**.** Let $(N_t)_{t\geq 0}$ be a Poisson process with intensity λ. Let $Y_1, Y_2, \ldots$ be a sequence of i.i.d random variables with common distribution F. Also assume that this sequence is independent of the Poisson process.

Then, a compound Poisson process (CPP) with intensity λ and jump distribution F is a stochastic process $(X_t)_{t\geq 0}$ such that

$$X_t = \sum_{i=1}^{N_t} Y_i.$$

We will say that Y_i are the jumps of the compound Poisson process. By convention, we take $X_t = 0$ if $N_t = 0$.

1.2 The Statistical Inverse Problem

Suppose we are only able to observe the values of a CPP at times $\Delta, 2\Delta, \ldots, n\Delta$, thus giving observations

$$\{X_{i\Delta} : i = 1, \ldots, n\}.$$

Assumptions. For our purposes, we assume that the intensity λ of the CPP is known. We also assume that the jump distribution F has continuous probability density function f and assigns zero mass at the origin $\{0\}$, i.e.

$$F(\{0\}) = 0.$$

This is because, if not, then the event of a jump at some time t could potentially be indistinguishable to the event of no jump at time t.

Goal. We want to recover the density f from our observations $\{X_{i\Delta} : i = 1, \ldots, n\}$. We approach this problem using non-parametric estimation, as we want to put only few assumptions on density f. For example, we may only assume that f comes from the set of Lipschitz continuous functions. Such spaces are infinite-dimensional, and so parametric models are unsuitable for this task.

Transformation of Observations. Given our observations $\{X_{i\Delta} : i = 1, \ldots, n\}$, consider increments $\{Z_i : i = 1, \ldots, n\}$ given by

$$Z_i = X_{i\Delta} - X_{(i-1)\Delta}. \tag{1.1}$$

By the independent increments (Property 1) of the Poisson process, all Z_i are mutually independent.

Proposition 1.1. *(see, for example, [16]) For Z_i defined in* (1.1)*, we have that*

$$Z_i \overset{\mathcal{L}}{=} \mathbb{1}(N > 0) \sum_{j=1}^{N} Y_i, \tag{1.2}$$

where $N \sim \mathcal{P}(\lambda\Delta)$ and N is independent of jumps Y_i.

Proof.

$$\begin{aligned} Z_i &= X_{i\Delta} - X_{(i-1)\Delta} \\ &= (Y_{N_{(i-1)\Delta}+1} + \cdots + Y_{N_{i\Delta}}) \mathbb{1}(N_{i\Delta} > N_{(i-1)\Delta}) \\ &\overset{\mathcal{L}}{=} (Y_1 + \cdots + Y_{N_{i\Delta} - N_{(i-1)\Delta}}) \mathbb{1}(N_{i\Delta} > N_{(i-1)\Delta}) \qquad (1.3) \\ &\overset{\mathcal{L}}{=} \mathbb{1}(N > 0) \sum_{j=1}^{N} Y_i, \qquad (1.4) \end{aligned}$$

where $N \sim \mathcal{P}(\lambda\Delta)$. Line (1.3) follows by the i.i.d property of the jumps and line (1.4) follows by the stationary increments (Property 2) of a Poisson process. The independence of N and the jumps follows from the independence of the Poisson process and the jumps. □

Since all Z_i are mutually independent, it will be more ideal to deal with observations Z_i rather than $X_{i\Delta}$. Henceforth, we will refer to our observations as $\{Z_i : i = 1, \ldots, n\}$ and call such Z_i a Poisson random sum. The non-linearity of this inverse problem stems from the randomness of our Poisson random variable N in the Poisson random sum.

1.3 Properties of Poisson Random Sums

We have now converted our problem into that involving Poisson random sums. Therefore, we ought to investigate the properties of such random variables and exploit these properties for our inference. This section is important and these properties will be referred to throughout the essay.

Poisson Random Sums. Let $Y_1, Y_2, \ldots$ be a sequence of i.i.d random variables with probability density function f. As before, we will call these jumps. Let

$$Z = \mathbb{1}(N > 0) \sum_{j=1}^{N} Y_i, \qquad \text{with } N \sim \mathcal{P}(\lambda\Delta) \tag{1.5}$$

be a Poissom random sum.

Proposition 1.2. *[16] The characteristic function, ϕ_Z, of Z defined in* (1.5) *is given by*

$$\phi_Z(t) = \mathbb{E}e^{itZ} = e^{-\lambda\Delta + \lambda\Delta\phi_f(t)},$$

where ϕ_f denotes the characteristic function of a single jump.

Proof. See page 41 in Appendix. □

For the next property, we require the following Lemma.

Lemma 1.1. *Let X and Y be independent random variables with density functions f_X, f_Y and characteristic functions ϕ_X, ϕ_Y respectively. Then the sum $Z = X + Y$ is a random variable with density function f_Z and characteristic function ϕ_Z, where*

$$f_Z = f_X * f_Y, \qquad \phi_Z = \phi_X \phi_Y,$$

and $f * g$ denotes the convolution of f and g i.e.

$$(f * g)(t) = \int_{\mathbb{R}} f(\tau) g(t - \tau) d\tau.$$

Proof. See, for example, [3]. □

Proposition 1.3. *[6, Proposition 2.1] The distribution $\mathbb{P}_Z$ of Z is absolutely continuous with respect to measure $\mu = \delta_{\{0\}} + \text{Leb}$ and has Radon-Nikodym derivative*

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

where

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

and $f^{*m} = f \overbrace{* \cdots *}^{m} f$ denotes the m-fold convolution of f with itself.

Proof. See Appendix. □

In light of Proposition 1.3, we see that a zero-valued Poisson random sum Z (corresponding to $N = 0$) provides no additional information about the density f. In this case, conditional on $N > 0$, by slight modification of our proof, Z has probability density function given by

$$g(x) = \frac{e^{-\lambda\Delta}}{1 - e^{-\lambda\Delta}} \sum_{m=1}^{\infty} \frac{(\lambda\Delta)^m}{m!} f^{*m}(x). \tag{1.6}$$

2 Kernel Density Estimation

Kernel density estimation is a non-parametric method for estimating the probability density function from a finite data sample. Its ability to generate smooth curves from a discrete set of observations without assuming any parametric model makes it an ideal candidate for estimating continuous probability density functions.

Motivation. (cf. Tsybakov [1]). Let X be a random variable with probability density p with respect to the Lebesgue measure on $\mathbb{R}$. The corresponding distribution function is

$$F(x) = \int_{-\infty}^{x} p(t)dt, \qquad \text{and we have} \qquad \frac{\mathrm{d}F}{\mathrm{d}x} = p.$$

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 random variables, we have that

$$F_n(x) \to \mathbb{E}\left[I(X_1 \leq x)\right] = \mathbb{P}(X \leq x) = F(x) \text{ a.s. as } n \to \infty.$$

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

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

Thus, replacing F by our empirical distribution function F_n gives an estimator $\hat{p}_n(x)$ of $p(x)$, where

$$\begin{aligned}
\hat{p}_n(x) &= \frac{F_n(x+h) - F_n(x-h)}{2h} \\
&= \frac{1}{2nh} \sum_{i=1}^{n} I(x-h < X_i \leq x+h) \\
&= \frac{1}{nh} \sum_{i=1}^{n} k_0\left(\frac{x - X_i}{h}\right),
\end{aligned}$$

and where $k_0(u) = \frac{1}{2} I(-1 < u \leq 1)$. Note that $k_0(-u) = k_0(u)$ and $\int k_0(u)du = 1$. A simple generalisation gives us a kernel function and corresponding kernel density estimator.

Definition 2.1. (Kernel Density Estimator). Let $k : \mathbb{R} \to \mathbb{R}$ be a Lebesgue integrable function such that

$$\int_{\mathbb{R}} k(u)du = 1, \qquad \text{and} \qquad k(-u) = k(u).$$

Then we say that k is a kernel function. Let $h > 0$. Then the kernel density estimator of n i.i.d observations $X_1, \ldots, X_n$ is given by

$$\hat{p}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} k\left(\frac{x - X_i}{h}\right). \tag{2.1}$$

Examples. Simple examples include the Gaussian kernel function

$$\psi(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}.$$

Clearly, it is integrable, symmetric and integrates to 1. In the following sections, we will use the sinus kernel function

$$k(t) = \frac{\sin(\pi t)}{\pi t}, \tag{2.2}$$

and the kernel function given in [18] with expression

$$k(t) = \frac{48t(t^2-1)\cos t - 144(2t^2-5)\sin t}{\pi t^7}. \tag{2.3}$$

This expression is fairly non-trivial, but we will make use of its simpler characteristic function of the form

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

2.1 Inversion of Characteristic Functions

In this section, we follow the work of van Es et al. [16] to construct a suitable estimator for probability density function f. Returning back to our problem, suppose we have non-zero observations $\{Z_i : i = 1, \ldots, n\}$ of the compound Poisson process. For Poisson random sum

$$Z = \mathbb{1}(N > 0) \sum_{j=1}^{N} Y_i, \qquad N \sim \mathcal{P}(\lambda\Delta),$$

we split its characteristic function ϕ_Z into

$$\phi_Z(t) = \mathbb{E}\left[e^{itZ}\mathbb{1}(N=0)\right] + \mathbb{E}\left[e^{itZ}\mathbb{1}(N>0)\right]. \tag{2.4}$$

Note that

$$\phi_{Z_i}(t) = \mathbb{E}[e^{itZ}|N>0] = \frac{\mathbb{E}[e^{itZ}\mathbb{1}(N>0)}{\mathbb{P}(N>0)}.$$

Substituting this into (2.4),

$$\begin{aligned}\phi_Z(t) &= \mathbb{P}(N=0) + \mathbb{P}(N>0)\phi_{Z_1}(t) \\ &= e^{-\lambda\Delta} + (1-e^{-\lambda\Delta})\phi_{Z_1}(t).\end{aligned} \tag{2.5}$$

Using the expression for the characteristic function of a Poisson random sum in Proposition 1.2, we can rewrite (2.5) as

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

where g is the probability density function of an observation Z_i. Such density function exists by Proposition 1.3. Rearranging, we have

$$\phi_g(t) = \frac{1}{e^{\lambda\Delta}-1}(e^{\lambda\Delta\phi_f(t)} - 1). \tag{2.6}$$

This suggests that if we could suitably invert the formula in (2.6) to get an expression in terms of ϕ_f, then an estimator of ϕ_g would induce an estimator for ϕ_f. We can do this using the distinguished logarithm.

2.1.1 Distinguished Logarithm

The main issue to address in inverting relationship (2.6) is that ϕ_f takes complex values, and so we must find an inverse to the map $\exp : \mathbb{C} \to \mathbb{C}$. Such an inverse does not exist in the usual sense, since it is not injective - in particular, for every $w \in \mathbb{C}$, $e^{w+2\pi i} = e^w$. Therefore, the following results give a suitable function which has the desired properties of an inverse function to the exponential.

Lemma 2.1. *Suppose $\phi : \mathbb{R} \to \mathbb{C}$ is a continuous function such that $\phi(0) = 1$ and $\phi_g(t) \neq 0$ for every $t \in \mathbb{R}$. 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 Theorem 7.6.2 in [3]. □

For a function ϕ satisfying the assumptions of Lemma 2.1, we say that the unique function h is the distinguished logarithm and we denote

$$h(t) = \mathrm{Log}(\phi)(t).$$

The following property is an easy consequence of Lemma 2.1.

Lemma 2.2. *For ϕ_1, ϕ_2 satisfying the assumptions of Lemma 2.1, we have for $\psi(t) = \phi_1(t)\phi_2(t)$,*

$$\mathrm{Log}(\psi) = \mathrm{Log}(\phi_1) + \mathrm{Log}(\phi_2).$$

Furthermore, for $c \in \mathbb{R}$ and $\chi(t) = (\phi_1(t))^c$, we have

$$\mathrm{Log}(\chi)(t) = c\mathrm{Log}(\phi_1)(t).$$

Proof. We have that $\psi(0) = 1$ and $\psi(t) \neq 0$ for every $t \in \mathbb{R}$. Therefore, $\mathrm{Log}(\psi)$ exists by Lemma 2.1. Furthermore,

$$e^{\mathrm{Log}(\psi)(t)} = \psi(t) = \phi_1(t)\phi_2(t) = e^{\mathrm{Log}(\psi_1)(t)}e^{\mathrm{Log}(\psi_2)(t)} = e^{\mathrm{Log}(\psi_1)(t)+\mathrm{Log}(\psi_2)(t)}.$$

The first result then follows by the uniqueness of the distinguished logarithm. For the second result, note again that $\chi(0) = 1$ and $\chi(t) \neq 0$ for every $t \in \mathbb{R}$, and so $\mathrm{Log}(\chi)$ exists. Moreover,

$$e^{\mathrm{Log}(\chi)(t)} = \chi(t) = (\phi_1(t))^c = e^{c\mathrm{Log}(\psi_1)(t)}.$$

The second result similarly follows by the uniqueness of the distinguished logarithm. □

Note that $\psi(t) \triangleq e^{\phi_f(t)-1}$ is a continuous function satisfying $\psi(0) = 1$ and $\psi(t) \neq 0$ for every $t \in \mathbb{R}$. By the uniqueness of the distinguished logarithm, we see that

$$\phi_f(t) - 1 = \mathrm{Log}\left(e^{\phi_f - 1}\right)(t).$$

Now, using (2.6), we have

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

We are close to obtaining an expression for probability density function f in terms of probability density function g - we just need the ability to invert characteristic function ϕ_f. The Fourier inversion theorem allows us to do this.

2.1.2 Fourier Inversion Theorem

Let $\mathcal{F} : L^1(\mathbb{R}) \to C_0(\mathbb{R})$ denote the Fourier transform, defined by

$$\mathcal{F}[f](t) = \int_{\mathbb{R}} e^{itx} f(x)dx, \qquad t \in \mathbb{R}.$$

Note that this is equivalent to the characteristic function of a random variable with probability density function f.

Proposition 2.1. *(Fourier inversion theorem, [8]) Let $f \in L^1(\mathbb{R})$ be a continuous function. Suppose also that $\mathcal{F}[f]$ is integrable. Then*

$$f(x) = \frac{1}{2\pi}\int_{\mathbb{R}} e^{-itx}\mathcal{F}[f](t)dt. \tag{2.7}$$

We denote $\mathcal{F}^{-1}[f] : L^1(\mathbb{R}) \to C_0(\mathbb{R})$ by

$$\mathcal{F}^{-1}[f](t) = \frac{1}{2\pi}\int_{\mathbb{R}} e^{-itx} f(x)dx.$$

Then we can write the Fourier Inversion Theorem more compactly as

$$\mathcal{F}^{-1}\mathcal{F}[f] = \mathcal{F}\mathcal{F}^{-1}[f] = f.$$

Now, using the Fourier inversion theorem, for integrable ϕ_f $(= \mathcal{F}[f])$ we have

$$f(x) = \frac{1}{2\pi\lambda\Delta}\int_{-\infty}^{\infty} e^{-itx}\left(\lambda\Delta + \text{Log}\left((1-e^{-\lambda\Delta})\phi_g + e^{-\lambda\Delta}\right)(t)\right)dt. \tag{2.8}$$

Therefore, as long as ϕ_f is integrable, if we can obtain an estimator for the characteristic function ϕ_g of observations Z_i, then we obtain an estimator for probability density function f.

2.1.3 Construction of the Estimator

Using the theory of kernel density estimators at the beginning of this section, for some kernel k with characteristic function ϕ_k, bandwidth $h > 0$ and observations $Z_1, \ldots, Z_n$, we estimate density g by the kernel density estimator

$$\hat{g}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} k\left(\frac{x - Z_i}{h}\right).$$

Let ϕ_{emp} be the empirical characteristic function of the observations, given by

$$\phi_{\text{emp}}(t) = \frac{1}{n} \sum_{j=1}^{n} e^{itZ_j}.$$

Intuitively, this is the Monte Carlo approximation of the characteristic function of an observation. Then, by simple calculation, we see that

$$\phi_{\hat{g}_n}(t) = \phi_{\text{emp}}(t)\phi_k(ht), \tag{2.9}$$

where ϕ_k is the characteristic function of the kernel k:

$$\begin{aligned}
\phi_{\hat{g}_n}(t) &= \int_{-\infty}^{\infty} e^{itx} \hat{g}_n(x) dx \\
&= \int_{-\infty}^{\infty} e^{itx} \frac{1}{nh} \sum_{j=1}^{n} k\left(\frac{x - Z_j}{h}\right) dx \\
&= \frac{1}{n} \sum_{j=1}^{n} e^{itZ_j} \int_{-\infty}^{\infty} e^{ithy} k(y) dy \qquad \left(y = \frac{x - Z_j}{h}\right) \\
&= \phi_{\text{emp}}(t)\phi_k(ht).
\end{aligned}$$

Technical Issues. Taking $\phi_{\hat{g}_n}$ to be our estimator of ϕ_g, it is tempting to introduce an estimator $\hat{f}_n$ of f given by

$$\hat{f}_n(x) = \frac{1}{2\pi\lambda\Delta} \int_{-\infty}^{\infty} e^{-itx} \left(\lambda\Delta + \text{Log}\left((1 - e^{-\lambda\Delta})\phi_{\text{emp}}\phi_k(h\,\cdot) + e^{-\lambda\Delta}\right)(t)\right) dt, \tag{2.10}$$

but this brings two main issues:

1. In light of Lemma 2.1, we may have some measurable set A with non-zero Lebesgue measure such that $(1 - e^{-\lambda\Delta})\phi_{\text{emp}}(t)\phi_k(ht) + e^{-\lambda\Delta}$ 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,

$$\phi_{\hat{g}_n}(t) = \frac{\exp(e^{it}) - e^{-\lambda\Delta}}{1 - e^{-\lambda\Delta}}$$

would cause $\hat{f}_n(1)$ to be infinity.

In order to prove asymptotic properties, we must adjust our estimators by bounding $\hat{f}_n$ 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.

Properties of the Estimator. To assess theoretical performance of the estimator, we consider the mean squared error given by

$$\text{MISE}[\hat{f}_n(x)] = \mathbb{E}[(\hat{f}_n(x) - f(x))^2].$$

By algebraic manipulation, this can be decomposed into the sum of the squared bias and variance of the estimator. More formally,

$$\text{MISE}[\hat{f}_n(x)] = (\mathbb{E}[\hat{f}_n(x)] - f(x))^2 + \text{Var}(\hat{f}_n(x)).$$

Let β be some positive number and suppose that the conditions dependent on β in Section 2 of van Es et al. [16] on density f, density g, kernel k and bandwidth h are met. For brevity, we omit these conditions but they are clearly stated in the literature. Then, we get that the bias has the order bound (Proposition 2.1 [16])

$$\mathbb{E}[\hat{f}_n(x)] - f(x) = \mathcal{O}\left(h^\beta + \frac{1}{nh}\right),$$

Furthermore, if in addition we have that $nh^{1+4\beta} \to 0$, then the variance of the estimator $\hat{f}_n(x)$ has the asymptotic behaviour (Proposition 2.2 [16])

$$\text{Var}(\hat{f}_n(x)) = \frac{1}{nh}\frac{(e^{\lambda\Delta-1)^2}}{(\lambda\Delta)^2} g(x) \int k(t)^2 dt + o\left(\frac{1}{nh}\right).$$

We note the slight difference to [16] since we have not assumed that the separation distance Δ is of unit size in the construction of the estimator in this essay. By combining these two bounds, we get that the estimator is point-wise weakly consistent in the limit $h \to 0, nh \to \infty$. This shows the feasibility of this approach.

2.1.4 Numerical Simulations

We note that for $\lambda\Delta < \log 2$, the distinguished logarithm in (2.10) reduces to the principal branch of the logarithm [16]. Therefore, we can directly use logarithm from scientific computing packages, and bounding $\hat{f}_n$ by a suitable sequence is not needed. Thus, we use (2.10) directly to compute our estimator on cases where $\lambda\Delta < \log 2$. We call this the high frequency data setting since for relatively low intensity and low separation distance, we are able to observe almost all jumps of the CPP.

We rewrite (2.10) as $\hat{f}_n(x) = \hat{f}_{n,1}(x) + \hat{f}_{n,2}(x)$ where

$$\hat{f}_{n,1}(x) = \frac{1}{2\pi\lambda\Delta}\int_0^\infty e^{-itx}\log\Big((e^{\lambda\Delta}-1)\phi_{\text{emp}}(t)\phi_k(ht)+1\Big)dt, \tag{2.11}$$

$$\begin{aligned}\hat{f}_{n,2}(x) &= \frac{1}{2\pi\lambda\Delta}\int_{-\infty}^0 e^{-itx}\log\Big((e^{\lambda\Delta}-1)\phi_{\text{emp}}(t)\phi_k(ht)+1\Big)dt \\ &= \frac{1}{2\pi\lambda\Delta}\int_0^\infty e^{itx}\log\Big((e^{\lambda\Delta}-1)\phi_{\text{emp}}(-t)\phi_k(ht)+1\Big)dt. \end{aligned}\tag{2.12}$$

Line (2.12) follows since ϕ_k is symmetric. We deal with $\hat{f}_{n,1}$ - the case of $\hat{f}_{n,2}$ is very similar.

Trapezoid Rule. ([12, p. 95]) Let $\{t_j\}_{j=0}^{N-1}$ be a set of N equally spaced values partitioning $[a,b]$, with spacing $\eta = \frac{b-a}{N}$. Then, for integrable function h we get the following approximation

$$\int_a^b h(x)dx \approx \eta\sum_{j=0}^{N-1} h(t_j). \tag{2.13}$$

Approximating (2.11) using the Trapezoid rule, we get, for spacing parameter $\eta > 0$ and $t_j = j\eta$, that

$$\hat{f}_{n,1}(x) \approx \frac{\eta}{2\pi\lambda\Delta}\sum_{k=0}^{N-1} e^{-it_jx}\log\Big((e^{\lambda\Delta}-1)\phi_{\text{emp}}(t_j)\phi_k(ht_j)+1\Big). \tag{2.14}$$

We evaluate our function $\hat{f}_{nh}^{(1)}$ at points $\{x_k\}_{k=0}^{N-1}$ given by

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

for some $\delta > 0$ to be chosen later. Thus we have

$$\hat{f}_{n,1}(x_k) \approx \frac{1}{2\pi\lambda}\sum_{j=0}^{N-1} e^{-ijk\eta\delta}e^{it_j\frac{N\delta}{2}}\psi(t_j)\eta, \tag{2.15}$$

where $\psi(t) = \log\big((e^{\lambda\Delta}-1)\phi_{\text{emp}}(t)\phi_k(ht)+1\big)$. Similarly,

$$\hat{f}_{n,2}(x_k) \approx \frac{1}{2\pi\lambda}\sum_{j=0}^{N-1} e^{ijk\eta\delta}e^{-it_j\frac{N\delta}{2}}\psi(-t_j)\eta, \tag{2.16}$$

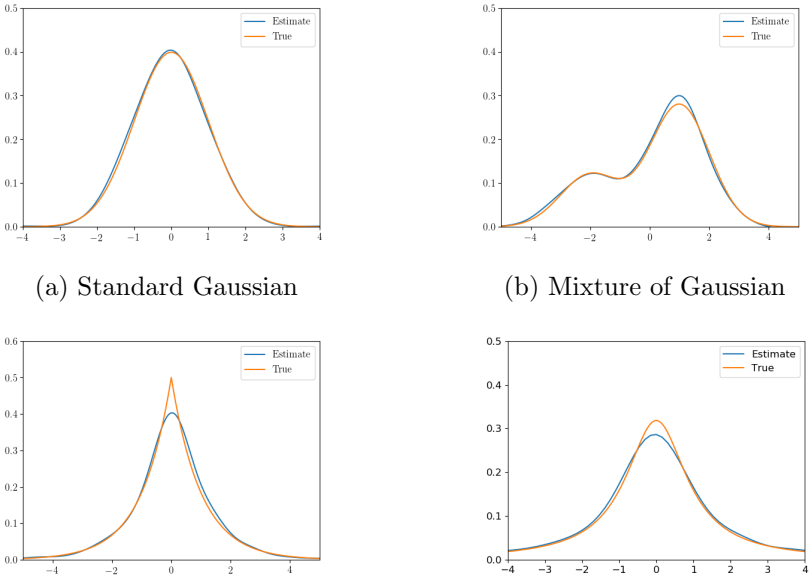
Fast Fourier Transform. Let $\{x_j\}_{j=0}^{N-1}$ be a sequence of complex numbers. The Fast Fourier Transform (FFT) computes the sequence $\{Y_k\}_{k=0}^{N-1}$ where

$$Y_k = \sum_{j=0}^{N-1} x_j e^{-ik\frac{2\pi j}{N}}. \tag{2.17}$$

The inverse transform is given by

$$Y_k = \frac{1}{N} \sum_{j=0}^{N-1} x_j e^{ik\frac{2\pi j}{N}}. \tag{2.18}$$

Taking N to be some large power of 2 and choosing η, δ such that $\eta\delta = \frac{2\pi}{N}$, we can then apply FFT to these expressions to obtain an approximation for estimator $\hat{f}_n$ of f. We choose η to be relatively small so that δ can be relatively large and thus, points at which we evaluate our density estimator are relatively separate from one another.



(a) Standard Gaussian

(b) Mixture of Gaussian

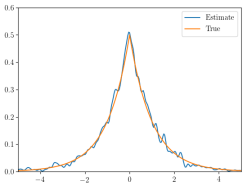
(c) Laplace

(d) Cauchy

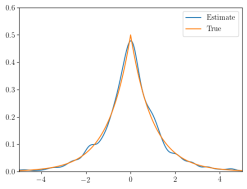
Figure 2.1: Density Estimates via Inversion of Characteristic Functions.

Examples. We simulated a CPP with intensity $\lambda = 0.5$ and obtained $n = 5000$ observations with separation distance $\Delta = 1$. We used the kernel given in (2.3). As we can see in Figure 2.1, this kernel provides satisfactory smoothing properties. For the series in (2.15) and (2.16), we took $N = 16384$, $\eta = 0.01$ and bandwidth $h = 0.14$.

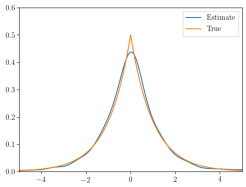
One area of concern is the inability to achieve a steep peak in the estimate of the Laplace distribution. This is mainly due to the bandwidth acting as a smoothing parameter. We can attempt to obtain a better estimate by varying the bandwidth over a range of values and choosing the one that gives the smallest error.



(a) Laplace with $h = 0.02$



(b) Laplace with $h = 0.06$



(c) Laplace with $h = 0.1$

As we can see, a smaller bandwidth captures the peak of the Laplace distribution in a quite satisfactory manner, but causes the tails of the distribution to be less smooth. A higher bandwidth results in the contrary, therefore, we must find a middle ground. We could use cross validation to fit the bandwidth which gives the best fit, as explained in [15].

2.2 Estimation of Convolution Powers

In this section we follow the work of Comte et al. [4]. Suppose we have non-zero observations $\{Z_i : i = 1, \dots, n\}$ of the compound Poisson process. By (1.6), we have that each Z_i has density g given by

$$g = \frac{e^{-\lambda\Delta}}{1 - e^{-\lambda\Delta}} \sum_{m=1}^{\infty} \frac{(\lambda\Delta)^m}{m!} f^{*m} = \frac{1}{e^{\lambda\Delta} - 1} \sum_{m=1}^{\infty} \frac{(\lambda\Delta)^m}{m!} f^{*m}. \tag{2.19}$$

We can see that density g is a weighted sum of convolution powers of density f. If we could rewrite this expression in terms of f, then an estimator of density g directly provides an estimator of f. We will show that this is possible for λ, Δ sufficiently small. Since we have convolution powers in the expression, we will see that it makes sense to use the Fourier inversion theorem to do this.

Proposition 2.2. *[6, Lemma 2.1] Provided that $\lambda\Delta < \log 2$, we have*

$$f = \sum_{m=1}^{\infty} \frac{(-1)^{m+1}}{m} \frac{(e^{\lambda\Delta} - 1)^m}{\lambda\Delta} g^{*m}.$$

Proof. We first show, using the Fourier Inversion Theorem (2.1), that the Fourier transform $\mathcal{F}$ is injective.

Suppose that we have $f \in L^1(\mathbb{R})$ such that $\mathcal{F}[f] = 0$. Then, we have that $\mathcal{F}[f] \in L^1(\mathbb{R})$ since it is the zero function. Therefore, by the Fourier Inversion Theorem, we get that

$$f(x) = \frac{1}{2\pi} \int_{\mathbb{R}} e^{-itx} \mathcal{F}[f](t) dt = 0,$$

thus showing that $f \mapsto \phi_f$ is injective on the space of Lebesgue-integrable functions, which is where our densities live.

We next use the convolution Theorem (which is equivalent to Lemma 1.1) that states that $\mathcal{F}[f * g] = \mathcal{F}[f]\mathcal{F}[g]$ for integrable functions f, g. From (2.19), the linearity of an integral and the convolution Theorem we get that

$$\begin{aligned} \mathcal{F}[g] &= \frac{1}{e^{\lambda\Delta} - 1} \sum_{m=1}^{\infty} \frac{(\lambda\Delta)^m}{m!} (\mathcal{F}[f])^m \\ &= \frac{\exp(\lambda\Delta\mathcal{F}[f]) - 1}{e^{\lambda\Delta} - 1}. \end{aligned}$$

Rearranging, we get that

$$\exp(\lambda\Delta\mathcal{F}[f]) = 1 + (e^{\lambda\Delta} - 1)\mathcal{F}[g].$$

Note that $\left\|(e^{\lambda\Delta} - 1)\mathcal{F}[g]\right\|_\infty < \left\|e^{\lambda\Delta} - 1\right\|_\infty < 1$ for $\lambda\Delta < \log 2$. Therefore, the distinguished logarithm defined in section 2.1.1 reduces to the principal branch of the logarithm. Thus, by taking logarithms

$$\mathcal{F}[f] = \frac{\log(1 + (e^{\lambda\Delta} - 1)\mathcal{F}[g])}{\lambda\Delta} = \sum_{m=1}^{\infty} \frac{(-1)^{m+1}}{m} \frac{(e^{\lambda\Delta} - 1)^m}{\lambda\Delta} \mathcal{F}[g]^m. \quad (2.20)$$

by the Taylor expansion of the logarithm, which holds since

$$\lambda\Delta < \log 2 \implies \left\|(e^{\lambda\Delta} - 1)\mathcal{F}[h]\right\|_\infty < 1.$$

Applying Fourier Inversion Theorem, since f is continuous and a probability density function so $f \in L^1(\mathbb{R})$, gives the result. □

2.2.1 Construction of the Estimator

Recall, for some kernel k with characteristic function ϕ_k, bandwidth $h > 0$ and observations $Z_1, \ldots, Z_n$, we can estimate density g by the kernel density estimator

$$\hat{g}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} k\left(\frac{x - Z_i}{h}\right).$$

Therefore, for integrable characteristic function $\phi_{\hat{g}_n}$, using the Fourier inversion theorem and (2.9) we see that the kernel density estimator can be rewritten as

$$\hat{g}_n(x) = \frac{1}{2\pi} \int_{\mathbb{R}} \phi_{\text{emp}}(t)\phi_k(ht)e^{-itx} dt.$$

Furthermore, using Lemma 1.1, we note that $\phi_{g^{*m}} = (\phi_g)^m$. Since $\hat{g}_n$ is an estimator of g, it is sensible to use $\hat{g}_n^{*m}$ as an estimator of convolution power g^*. Note that

$$\begin{aligned}\hat{g}_n^{*m}(x) &= \frac{1}{2\pi}\int_{\mathbb{R}} \phi_{\hat{g}_n^{*m}}(t)e^{-itx}dt \\ &= \frac{1}{2\pi}\int_{\mathbb{R}} (\phi_{\hat{g}_n}(t))^m\, e^{-itx}dt \\ &= \frac{1}{2\pi}\int_{\mathbb{R}} (\phi_{\text{emp}}(t)\phi_k(ht))^m\, e^{-itx}dt.\end{aligned}$$

Therefore, using $\hat{g}_n^{*m}$ as our estimator of g^{*m}, we immediately obtain, provided $\lambda\Delta < \log 2$, an estimator for f given by

$$\hat{f}_n(x) = \sum_{m=1}^{\infty} \frac{(-1)^{m+1}}{m}\frac{(e^{\lambda\Delta}-1)^m}{\lambda\Delta}\hat{g}_n^{*m}(x).$$

For small $\lambda\Delta < \log 2$, $e^{\lambda\Delta}-1$ will be close to 0. In particular, $\frac{(e^{\lambda\Delta}-1)^m}{m} \to 0$ as $m \to \infty$. Therefore, it is reasonable to truncate the series up to some sufficiently large K to give

$$\hat{f}_{n,K}(x) = \sum_{m=1}^{K} \frac{(-1)^{m+1}}{m}\frac{(e^{\lambda\Delta}-1)^m}{\lambda\Delta}\hat{g}_n^{*m}(x). \tag{2.21}$$

Properties of the Estimator. Let $f \in \mathcal{S}(\beta, L)$ where $\mathcal{S}(\beta, L)$ is the Sobolev space given by

$$S(\beta, L) = \left\{ f \in L^1(\mathbb{R}) \cap L^2(\mathbb{R}) : \int (1+x^2)^\alpha |\mathcal{F}^{-1}[f](x)|^2 dx \le L \right\}.$$

For convolution estimator $\hat{g}_n^{*m}$ of g^{*m}, we have the bound (Proposition 31 [4], [2])

$$\mathbb{E}\left(|\hat{g}_n^{*m}(t) - g^{*m}(t)|^2\right) \le C_m\left(\frac{1}{n^m} + \frac{|\mathcal{F}[g](t)|^2}{n}\right),$$

where C_m is a constant that does not depend on n or g. Furthermore, under additional conditions given explicitly in Section 4 of [4], we have that estimator $\hat{f}_{n,K}$ is minimax under L_2 norm whenever $n^{-2\beta/(2\beta+1)} > \Delta^{2K+2}$. Therefore, the estimator is optimal in the minimax sense over the Sobolev space.

2.2.2 Numerical Simulations

To compute estimators $\hat{g}_n^{*m}$, we follow the exact same procedure as in the previous section. We can express $\hat{g}_n^{*m} = \hat{g}_{n,1}^{*m} + \hat{g}_{n,2}^{*m}$ where

$$\hat{g}_{n,1}^{(m)}(x) = \frac{1}{2\pi}\int_0^\infty (\phi_{\text{emp}}(t)\phi_k(ht))^m\, e^{-itx}dt,$$

$$\begin{aligned}\hat{g}_{n,2}^{(m)}(x) &= \frac{1}{2\pi}\int_{-\infty}^0 (\phi_{\text{emp}}(t)\phi_k(ht))^m\, e^{-itx}dt \\ &= \frac{1}{2\pi}\int_0^\infty (\phi_{\text{emp}}(-t)\phi_k(ht))^m\, e^{itx}dt.\end{aligned}$$

We work with $\hat{g}_{n,1}^{(m)}$ - the case for $\hat{g}_{n,2}^{(m)}$ is very similar. We approximate $\hat{g}_{n,1}^{(m)}(x)$ using the trapezoid rule (2.13). We take a grid $t_j = j\eta$ for $j = 0, 1, \ldots, N-1$ where N is some large power of 2 and $\eta > 0$ is some constant. Then,

$$\hat{g}_{n,1}^{(m)}(x) \approx \frac{1}{2\pi}\sum_{j=0}^{N-1} (\phi_{\text{emp}}(t_j)\phi_k(ht_j))^m\, e^{-it_j x}\eta.$$

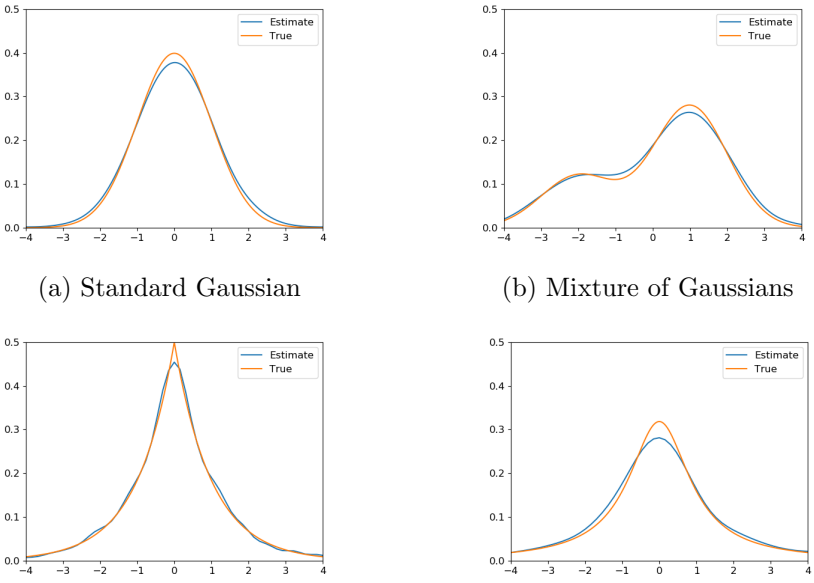
Taking $x_k = -\frac{N\delta}{2} + \delta k$ for $k = 0, 1, \ldots, N-1$ and δ is some constant to be defined later, we have

$$\hat{g}_{n,1}^{(m)}(x) \approx \frac{1}{2\pi}\sum_{j=0}^{N-1} (\phi_{\text{emp}}(t_j)\phi_k(ht_j))^m\, e^{it_j\frac{N\delta}{2}} e^{-ijk\eta\delta}\eta.$$

Similarly,

$$\hat{g}_{n,2}^{(m)}(x) \approx \frac{1}{2\pi}\sum_{j=0}^{N-1} (\phi_{\text{emp}}(-t_j)\phi_k(ht_j))^m\, e^{-it_j\frac{N\delta}{2}} e^{ijk\eta\delta}\eta.$$

We choose δ so that $\eta\delta = \frac{2\pi}{N}$ and then apply FFT to these terms to obtain an estimate for $\hat{g}_{nh}^{(m)}$. Plugging in convolution estimators into truncated sum (2.21) gives an estimator $\hat{f}_{n,K}$ of f.



(a) Standard Gaussian

(b) Mixture of Gaussians

(c) Laplace

(d) Cauchy

Figure 2.3: Density Estimates via the Estimation of Convolution Powers

Examples. We simulated a compound Poisson process with intensity $\lambda = 0.5$ and we observed equally spaced points of separation $\Delta = 1$. We obtained a sample size of 5000 non-zero increments. We used the sinus kernel function (2.2) for the kernel density estimator. We took $N = 16384$, $\eta = 0.01$, bandwidth $h = 0.14$ and a truncation $K = 10$. The simulations show a satisfactory general fit, capturing the overall shape well.

In particular, the estimator is robust to distributions with strong peaks and heavy tails. We can see this from the estimates of the Laplace and Cauchy distributions. Furthermore, the estimator nicely depicts the bimodal nature of the mixture of Gaussians.

3 Bayesian Density Estimation

The kernel density estimators provide a satisfactory attempt in recovering jump probability density function f. We will now focus on the Bayesian approach to density estimation. Advantages of using a Bayesian approach to density estimation compared to kernel density estimators include the following:

1. We obtain a distribution over the space of probability densities, rather than just a point estimate. This allows us to readily obtain uncertainty quantification through credible sets.

2. We can assert prior beliefs quantitatively through the prior distribution. This is useful in practice when we have information about the parameters in question.

3. Under suitable conditions, as shown in [11], the posterior contracts around the 'true' density at a $\sqrt{n\Delta}$-rate, where n is the number of observations and Δ is the separation size. This shows the feasibility of the Bayesian approach.

3.1 Bayes Theorem on Function Spaces

The Bayesian approach treats the unknown quantities in question as random variables. Prior beliefs about the unknown quantities are represented by the prior distribution, and the posterior distribution captures our beliefs about the unknown quantities after they have been modified in light of the observed data. If all such distributions have probability densities, then we can simply write Bayes Theorem for parameter θ and observations $X_1, \ldots, X_n$ as

$$p(\theta|X_1, \ldots, X_n) = \frac{p(X_1, \ldots, X_n|\theta)p(\theta)}{\int p(X_1, \ldots, X_n|\theta)p(\theta)d\theta}. \tag{3.1}$$

The issue that arises is that the probability density we would like to recover comes from an infinite-dimensional space. Therefore, we would need to define a distribution over such a space. In infinite-dimensional Banach spaces, there is no analogue with the Lebesgue measure so any distribution over such space does not have an equivalent probability density form. Therefore, for our purposes, we must formulate the theorem more abstractly to gain a rigorous understanding. We follow [10] and [9] to formulate the theorem on function spaces.

Background. Let $(\mathcal{X}, \mathcal{A})$ be a measurable space. This will be the space of the observations. Let $\mathcal{F}$ be the parameter space. We would like to formulate the notion the posterior distribution and prior distribution in this setting.

Prior. Suppose we have a dominated collection of probability measures $\{P_f : f \in \mathcal{F}\}$ on $(\mathcal{X}, \mathcal{A})$ with respect to some σ-finite measure μ. Let $\{p_f : f \in \mathcal{F}\}$ denote the corresponding collection of Radon-Nikodym derivatives. Suppose further that $\mathcal{F}$ is equipped with a σ-algebra $\mathcal{B}$ such that, for each $x \in \mathcal{X}$, the map $f \mapsto p_f(x)$ is measurable. Now we can define a distribution Π on $(\mathcal{F}, \mathcal{B})$, known as the prior distribution.

Joint Distribution. Define the probability space

$$(\mathcal{X} \times \mathcal{F}, \mathcal{A} \otimes \mathcal{B}, Q),$$

where

$$dQ(x, f) = p_f(x)d\mu(x)d\Pi(f).$$

Note that $x \mapsto p_f(x)$ and $f \mapsto p_f(x)$ are measurable with respect to $(X, \mathcal{A})$ and $(\mathcal{F}, \mathcal{B})$ respectively. Therefore, Q is well defined, and

$$Q(\mathcal{X}, \mathcal{F}) = \int_{\mathcal{F}} \int_{\mathcal{X}} p_f(x)d\mu(x)d\Pi(x) = \int_{\mathcal{F}} d\Pi(f) = 1.$$

Let $\pi_{\mathcal{X}}, \pi_{\mathcal{F}}$ be the coordinate projections $\mathcal{X} \times \mathcal{F} \mapsto \mathcal{X}$, $\mathcal{X} \times \mathcal{F} \mapsto \mathcal{F}$ respectively. Then, the distribution of $\pi_{\mathcal{X}}$ conditional on f has probability density

$$\begin{aligned} \frac{dQ(x, f)}{\int_{\mathcal{X}} dQ(x, f)} &= \frac{p_f(x)d\mu(x)d\Pi(f)}{d\Pi(f) \int_{\mathcal{X}} p_f(x)d\mu(x)} \\ &= p_f(x)d\mu(x), \qquad (\Pi - \text{a.s.}). \end{aligned}$$

Posterior Distribution. Similarly, the distribution of $\pi_{\mathcal{F}}$ conditional on x has probability density

$$\begin{aligned} \frac{dQ(x, f)}{\int_{\mathcal{F}} dQ(x, f)} &= \frac{p_f(x)d\mu(x)d\Pi(f)}{d\mu(x) \int_{\mathcal{F}} p_f(x)d\Pi(f)} \\ &= \frac{p_f(x)d\Pi(f)}{\int_{\mathcal{F}} p_f(x)d\Pi(f)}, \qquad \text{a.s. under the law of } \pi_{\mathcal{X}}. \end{aligned}$$

Thus we can define the distribution

$$\Pi(B|X) = \frac{\int_B p_f(x)d\Pi(f)}{\int_{\mathcal{F}} p_f(x)d\Pi(f)}, \qquad B \in \mathcal{B},$$

as the posterior distribution.

3.2 Parametric Estimation using Data Augmentation

Since explicitly describing a prior distribution on a function space is no trivial task, we first begin by simplifying our problem to the parametric case. In this section, we follow the work of Gugushvili et al. [11]. The paper uses a data augmentation scheme to make the likelihood tractable and then implements a Metropolis-Hastings-within-Gibbs algorithm for sampling from the posterior.

Mixture of Gaussians. We assume that the density f is a mixture of Gaussians:

$$f(\cdot) = \sum_{j=1}^{J} \rho_j \psi(\cdot\,; \mu_j, 1/\tau), \tag{3.2}$$

where J is known and $\psi(\cdot\ ;\mu,\sigma^2)$ denotes the normal density with mean μ and variance σ^2. For convenience, we will refer to the precision $\tau = \frac{1}{\sigma^2}$ instead of the variance. To make use of conjugate priors, we assume that the Gaussians have common precision τ. Therefore, estimating density f is equivalent to estimating parameters τ and

$$\begin{aligned}\rho &= (\rho_1,\ldots,\rho_J)^T \\ \mu &= (\mu_1,\ldots,\mu_J)^T\end{aligned}$$

As such, we now deal with parametric estimation for our parameters (ρ,μ,τ).

3.2.1 Bypassing the Intractable Likelihood

We have seen in Proposition 1.3 that for non-zero observation Z, the likelihood of Z given f is given by

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

We have seen in Section 2.2 that computing convolutions, or even estimates of convolutions, is expensive. Therefore, we would like to avoid this computation and introduce auxiliary variables to circumvent dealing with the likelihood.

Auxiliary Variables. Suppose for (possibly zero) observation Z, we knew how many terms it consists of in its Poisson sum, and how many terms arise from each of the components $1,\ldots,J$ in the mixture. In other words, for observation $Z_i = \mathbb{1}(N>0)\sum_{j=1}^{N}Y_j$ suppose we knew

$$a_i = (a_{ij} : j = 1,\ldots,J), \tag{3.3}$$

where a_{ij} denotes the number of terms in the Poisson sum occuring from component j. Then

$$Z_i \overset{\mathcal{L}}{=} \sum_{j=1}^{J}\sum_{k=1}^{a_{ij}}Y_k^{(j)}\mathbb{1}(a_{ij}>0), \qquad Y_k^{(j)} \overset{\text{ind}}{\sim} \mathcal{N}\left(\mu_j,\frac{1}{\tau_j}\right),\ j=1,\ldots,J. \tag{3.4}$$

Since a (deterministic) sum of independent Gaussian random variables is a Gaussian random variable, we get that

$$Z_i|\ a_i,\mu,\tau \sim \mathcal{N}\left(a_i^T\mu,\tau^{-1}a_i^T\mathbf{1}\right)\ \text{for } i=1,\ldots,n.$$

Our likelihood now becomes tractable, and is given by the Gaussian probability density function

$$p(Z_i|\ a_i,\mu,\tau) = \psi(Z_i;a_i^T\mu,\tau^{-1}a_i^T\mathbf{1}).$$

We will denote $a=(a_1,\ldots,a_n)$, where n the the number of observations (we allow for zero-valued observations) to be our auxiliary variable. We can now use version (3.1) of Bayes Theorem and construct a MCMC algorithm with invariant distribution $p(\mu,\rho,\tau,a|Z)$.

3.2.2 Construction of Hierarchical Model

Priors. We take the following priors for (ρ, μ, τ):

$$\begin{aligned} \rho &\sim \text{Dir}(\alpha, \ldots, \alpha) \\ \tau &\sim \mathcal{G}(\eta, \gamma) \\ \mu_j|\tau &\overset{\text{ind}}{\sim} \mathcal{N}(\xi_j, \kappa^{-1}\tau^{-1}), \end{aligned} \tag{3.5}$$

where $(\alpha, \eta, \gamma, \xi, \kappa)$ are hyperparameters.

Proposition 3.1. *For $a = (a_1, \ldots, a_n)$ defined in (3.3), we have that*

$$a_{ij}|\rho_j \overset{\text{ind}}{\sim} \mathcal{P}(\lambda\rho_j\Delta)$$

Proof. For a mixture, component j is chosen with probability ρ_j. Therefore, for m independent draws of components, the number of draws corresponding to each component has Multinomial$(m; \rho_1, \ldots, \rho_J)$ distribution. Let $n_i = \sum_{j=1}^{J} a_{ij}$. We know that $n_i \sim \mathcal{P}(\lambda\Delta)$ and that each jump term is independent. Then, by the probability mass function of a Multinomial distribution, we have

$$\begin{aligned} p(a_i|\rho) &= p(n_i|\rho)\binom{n_i}{a_{i1}, \ldots, a_{iJ}} \prod_{j=1}^{J} \rho_j^{a_{ij}} \\ &= e^{-\lambda\Delta}\frac{(\lambda\Delta)_i^n}{n_i!}\binom{n_i}{a_{i1}, \ldots, a_{iJ}} \prod_{j=1}^{J} \rho_j^{a_{ij}} \\ &= \prod_{j=1}^{J} e^{-\lambda\Delta\rho_j}\frac{(\lambda\Delta\rho_j)^{a_{ij}}}{a_{ij}!}. \end{aligned}$$

Since this is the distribution of J independent $\mathcal{P}(\lambda\Delta\rho_j)$ random variables, the result follows. □

Hierarchical Model. Let π denote the joint prior distribution on (ρ, μ, τ) specified above. Then, we can write our model as:

$$\begin{aligned} \rho &\sim \text{Dir}(\alpha, \ldots, \alpha) \\ \tau &\sim \mathcal{G}(\eta, \gamma) \\ \mu_j|\tau &\overset{\text{ind}}{\sim} \mathcal{N}(\xi_j, \kappa^{-1}\tau^{-1}) \\ a_{ij}|\ \rho &\overset{\text{ind}}{\sim} \mathcal{P}(\lambda\rho_j\Delta) \\ Z_i|a_i, \mu, \tau &\overset{\text{ind}}{\sim} \mathcal{N}(a_i^T\mu, \tau^{-1}a_i^T\mathbf{1}). \end{aligned}$$

This model gives us the following joint distribution decomposition:

$$p(\rho, \mu, \tau, a, Z) = p(\rho)p(\tau)p(\mu|\tau)p(a|\rho)p(Z|a, \mu, \tau).$$

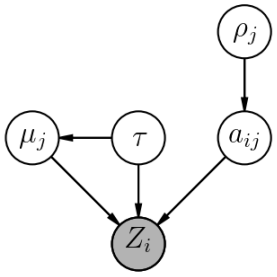


Figure 3.1: Probabilistic Graphical Model for the Hierarchical Model.

3.2.3 Contruction of MCMC Algorithm

We can use Gibbs sampling to sample from the joint distribution $p(\rho, \mu, \tau, a, Z)$ as follows:

Algorithm 1: Gibbs Sampler for Finite Mixture Hierarchical Model

Result: Samples from the posterior distribution $p(a, \mu, \tau, \rho|Z)$.
Initialise $\rho^{(0)}, \mu^{(0)}, \tau^{(0)}, a^{(0)}$;
for $t = 1, \ldots, N$ **do**
Update auxiliary variable a:

1. Sample $a^{(t)} \sim p(a|\rho^{(t-1)}, \mu^{(t-1)}, \tau^{(t-1)}, Z)$ via Metropolis-Hastings step

Update parameters ρ, τ, μ:

1. Sample $\rho^{(t)} \sim p(\rho|a^{(t)}, Z)$
2. Sample $\tau^{(t)} \sim p(\tau|a^{(t)}, Z)$
3. Sample $\mu^{(t)} \sim p(\mu|a^{(t)}, \tau^{(t)}, Z)$

end

Updating Auxiliary Variable. We would like to sample from $p(a|\rho, \mu, \tau, Z) \propto p(Z|a, \mu, \tau, \rho)p(a|\rho)$. We do this using a Metropolis-Hastings step. Note that by the Hierarchical model:

$$\begin{aligned} p(Z|a, \mu, \tau, \rho)p(a|\rho) &= \left(\prod_{i=1}^{n} p(Z_i|a_i, \mu, \tau, \rho)\right)\left(\prod_{i=1}^{n}\prod_{j=1}^{J} p(a_{ij}|\rho)\right) \\ &= \prod_{i=1}^{n}\left(\psi\left(Z_i; a_i^T\mu, a_i^T\tau^{-1}\right)\prod_{j=1}^{J} e^{-\lambda\rho_j\Delta}\frac{(\lambda\rho_j\Delta)^{a_{ij}}}{a_{ij}!}\right) \end{aligned}$$

Therefore, conditional on (ρ, μ, τ, Z), we have that each a_i is independent and so for each $i = 1, \ldots, n$, we perform a Metropolis Hastings step to sample from

$$\psi\left(Z_i; a_i^T\mu, a_i^T\tau^{-1}\right)\prod_{j=1}^{J} e^{-\lambda\rho_j\Delta}\frac{(\lambda\rho_j\Delta)^{a_{ij}}}{a_{ij}!}.$$

We construct our proposal distribution as follows:

1. We draw $n_i^\circ \sim \mathcal{P}(\lambda\Delta)$.
2. We draw $a_i^\circ = (a_{i1}^\circ, \ldots, a_{iJ}^\circ) \sim \text{Multinomial}(n_i^\circ; \rho_1, \ldots, \rho_J)$.

Then, our proposal density function $q(a_i^\circ|\rho)$ is given by

$$\begin{aligned} q(a_i^\circ|\rho) &= p(n_i^\circ)p(a_i^\circ|n_i^\circ, \rho) \\ &= e^{-\lambda\Delta}\frac{(\lambda\Delta)^{n_i^\circ}}{n_i^\circ!}\binom{n_i^\circ}{a_{i1}^\circ, \ldots, a_{iJ}^\circ}\prod_{j=1}^{J}\rho_j^{a_{ij}^\circ} \\ &= \prod_{j=1}^{J} e^{-\lambda\rho_j\Delta}\frac{(\rho_j\lambda\Delta)^{a_{ij}^\circ}}{a_{ij}^\circ!} \end{aligned}$$

by the same calculation as in Proposition 3.1. Therefore, our acceptance probability A is

$$\begin{aligned} A &= \frac{p(a_i^\circ|\rho, \mu, \tau, Z_i)q(a_i|\rho)}{p(a_i|\rho, \mu, \tau, Z_i)q(a_i^\circ|\rho)} \\ &= \frac{\psi\left(Z_i; (a_i^\circ)^T\mu, \tau^{-1}(a_i^\circ)^T\mathbf{1}\right)}{\psi\left(Z_i; a_i^T\mu, \tau^{-1}a_i^T\mathbf{1}\right)}, \end{aligned}$$

and we accept a_i° with probability $1 \wedge A$.

Updating Mixture Weights. Let

$$s_j = \sum_{i=1}^{n} a_{ij} \tag{3.6}$$

be the number of jumps in component j. We use the following Lemma to update our mixture weights.

Lemma 3.1. *Conditional on a, we have that $\rho_1, \ldots, \rho_J$ are independent and*

$$\rho_j|a \overset{ind}{\sim} \mathcal{G}(\alpha + s_j, \lambda n\Delta).$$

Proof. We calculate $p(\rho|a)$ using Bayes Theorem:

$$\begin{aligned} p(\rho|a) &\propto p(a|\rho)\pi(\rho) \\ &= \prod_{j=1}^{J} \pi(\rho_j) \left(\prod_i p(a_{ij}|\rho) \right) \\ &\propto \prod_{j=1}^{J} \rho_j^{\alpha-1} \left(\prod_i e^{-\rho_j \lambda \Delta} (\rho_j \lambda \Delta)^{a_{ij}} \right) \\ &= \prod_{j=1}^{J} \rho_j^{\alpha-1} e^{-\rho_j \lambda n \Delta} (\rho_j \lambda \Delta)^{s_j} \\ &\propto \prod_{j=1}^{J} \rho_j^{s_j+\alpha-1} e^{-\rho_j \lambda n \Delta}. \end{aligned}$$

Since this is the probability density of independent $\mathcal{G}(s_j + \alpha, \lambda n \Delta)$ distributed random variables, the result follows. □

Updating Individual Component Parameters. Let

$$n_i = \sum_{j=1}^{J} a_{ij}$$

be the number of jumps in observation i. We use the following Lemma to update the parameters of each component.

Lemma 3.2. *[11, Lemma 2] Conditional on (Z, a), we have*

$$\begin{aligned} \tau|Z, a &\sim \mathcal{G}(\eta + n/2, \gamma + (R - q^T P^{-1} q)/2) \\ \mu|\tau, Z, a &\sim \mathcal{N}(P^{-1}q, \tau^{-1}P^{-1}) \end{aligned}$$

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

$$P = \kappa I_{J\times J} + \tilde{P}, \qquad \tilde{P}_{jk} = \sum_{i=1}^{n} n_i^{-1} a_{ij} a_{ik},$$

q is the J-dimensional vector with

$$q_j = \kappa \xi_j + \sum_{i=1}^{n} n_i^{-1} a_{ij} Z_i,$$

$R > 0$ is given by

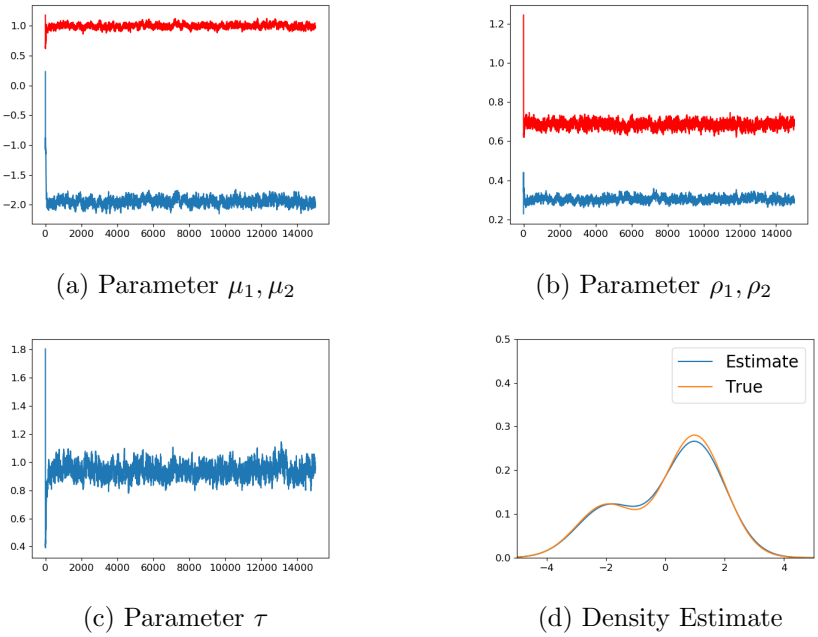
$$R = \kappa \sum_{j=1}^{J} \xi_j^2 + \sum_{i=1}^{n} n_i^{-1} Z_i^2,$$

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

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

Proof. See Appendix. □

3.2.4 Numerical Simulations



(a) Parameter μ_1, μ_2

(b) Parameter ρ_1, ρ_2

(c) Parameter τ

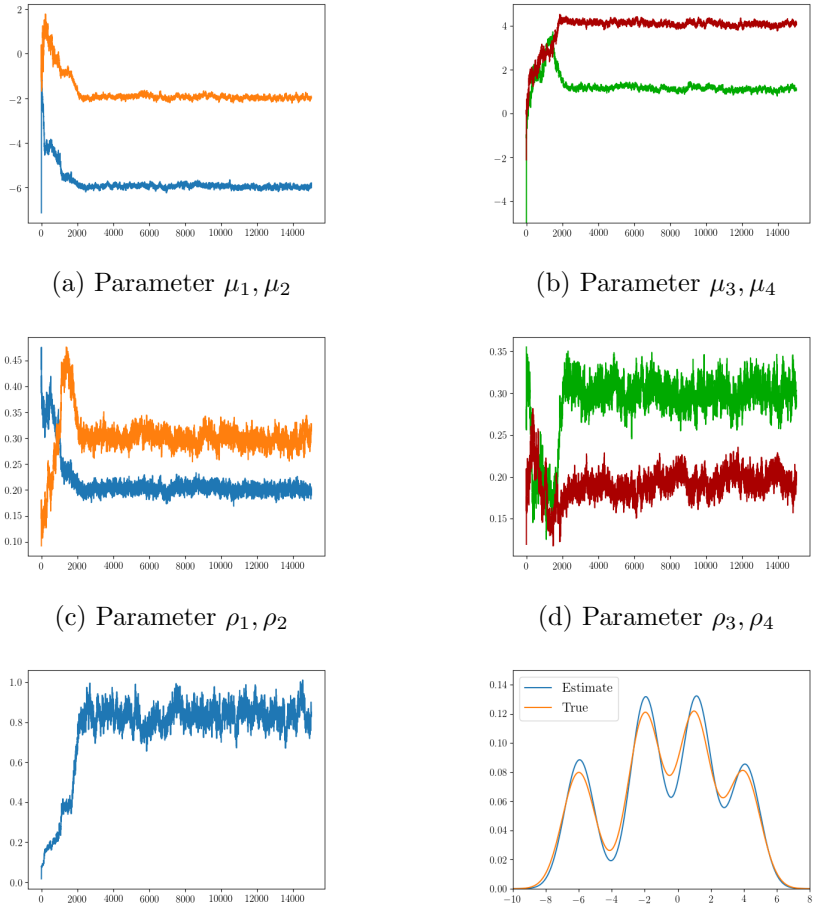
(d) Density Estimate

Figure 3.2: Run of MCMC algorithm for mixture of two Gaussians.

Example 1. We simulated a CPP with intensity $\lambda = 1$ and jump density function $f(x) = 0.3\psi(x; -2, 1) + 0.7\psi(x; 1, 1)$. Therefore,

$$(\mu_1, \mu_2) = (-2, 1), (\rho_1, \rho_2) = (0.3, 0.7), \tau = 1.$$

We obtained $n = 8000$ observations with spacing $\Delta = 1$ (giving roughly $5000 \approx (1 - e^{-\Delta\lambda})n$ non-zero observations) and performed 15000 MCMC iterations. We took hyperparameters $(\alpha, \eta, \gamma, \xi, \kappa) = (1, 1, 1, \mathbf{0}, 1)$. We obtained an average acceptance rate of 51% for the auxiliary variable Metropolis-Hastings proposals. The plots show that the chains move close to the true values of the mixture parameters in a satisfactory manner. A plot of posterior mean estimates of the parameters can be seen in Figure 3.2d.



(a) Parameter μ_1, μ_2

(b) Parameter μ_3, μ_4

(c) Parameter ρ_1, ρ_2

(d) Parameter ρ_3, ρ_4

(e) Parameter τ

(f) Density Estimate

Figure 3.3: Run of MCMC algorithm for mixture of four Gaussians.

Example 2. We simulated a CPP with intensity $\lambda = 1$ and jump density function $f(x) = 0.2\psi(x; -6, 1) + 0.3\psi(x; -2, 1) + 0.3\psi(x; 1, 1) + 0.2\psi(x; 4, 1)$. Therefore,

$$(\mu_1, \mu_2, \mu_3, \mu_4) = (-6, -2, 1, 4), (\rho_1, \rho_2, \rho_3, \rho_4) = (0.2, 0.3, 0.3, 0.2), \tau = 1.$$

As before, we obtained $n = 8000$ observations with spacing $\Delta = 1$, which consisted of roughly 5000 non-zero observations and we performed 15000 MCMC iterations. We took hyperparameters $(\alpha, \eta, \gamma, \xi, \kappa) = (1, 1, 1, \mathbf{0}, 1)$. We obtained an average acceptance rate of 29.7% for the auxiliary variable Metropolis-Hastings proposals.

The plots show good convergence around the true means and mixture weights, however, the chain for the precision parameter τ shows consistent underestimation and the sample variance is higher in the majority of parameters than in Example 1. A plot of posterior mean estimates of the

parameters can be seen in Figure 3.2d. Overall, we can see that the increase in number of components causes a significant reduction in the performance of the algorithm. This is due to the increased difficulty in moving around a higher dimensional space for the Gibbs sampler.

3.3 Non-Parametric Estimation via DPMM

In the previous section, we assumed that J was known, allowing us to reduce the problem to one of parametric estimation. However, such an assumption reduces the parameter space of densities significantly. Ideally we would like to have very little assumed about the space in which the jump probability density function lives.

We relax this assumption on J being known by using a Dirichlet Process Mixture Model (DPMM). The Dirichlet process is a distribution over the space of probability distributions. Therefore, the Dirichlet process allows us to treat Bayes' Theorem in the most general case as in Section 3.1, by specifying a Dirichlet Process prior on the space of probability distributions for the jump distribution F. We follow [9] to build the theory of Dirichlet processes.

3.3.1 Dirichlet Processes

Definition 3.1. (Random Measure. [5]) Let $(\Omega, \mathcal{F}, \mathcal{P})$ be some arbitrary probability space and $(\mathfrak{X}, \mathcal{X})$ be a measurable space. Then a map $P : \Omega \times \mathcal{X} \to \mathbb{R}$ is a random measure on $(\mathfrak{X}, \mathcal{X})$ if

1. For every $\omega \in \Omega$, the map $A \mapsto P(\omega, A)$ is a probability measure on $(\mathfrak{X}, \mathcal{X})$,
2. For every $A \in \mathcal{X}$, the map $\omega \mapsto P(\omega, A)$ is a random variable from $\Omega \to \mathbb{R}$.

Definition 3.2. (Dirichlet Process [7]). A random measure P on $(\mathfrak{X}, \mathcal{X})$ is said to possess a Dirichlet process distribution $\mathrm{DP}(\alpha)$ with base measure α on the measurable space $(\mathfrak{X}, \mathcal{X})$ if, for every finite measurable partition $A_1, \ldots, A_k$ of $\mathfrak{X}$, we have that the joint distribution of random variables $P(A_1), \ldots, P(A_k)$ satisfy

$$(P(A_1), \ldots, P(A_k)) \sim \mathrm{Dir}\left(k+1; \alpha(A_1), \ldots, \alpha(A_k)\right).$$

This definition does not provide much intuition about how a Dirichlet process could be used to deal with our problem of density estimation. Therefore, to build the intuition, we first simplify our discussion to a countable sample space.

Countable Dirichlet Process. [9] A probability measure on a countable sample space S (equipped with the σ-algebra $\mathfrak{S}$ generated by all finite subsets) can be represented as an infinite-length probability vector $s = (s_1, s_2, \dots)$ assigning probability weights to each element in the sample space. For example, a Poisson distribution on countable measurable space $(\mathbb{N}, \sigma(\mathbb{N}))$ can be represented as the probability vector

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

Thus, the space of probability measures on a countable space corresponds to the unit simplex

$$S_\infty = \left\{ s = (s_1, s_2, \dots) : s_j \geq 0, j \in \mathbb{N}, \sum_{j=1}^{\infty} s_j = 1 \right\}. \tag{3.7}$$

Consider the smallest σ-algebra $\mathfrak{S}_\infty$ on S_∞ that makes coordinate maps $s \mapsto s_i,\ i \in \mathbb{N}$ measurable. Consider some arbitrary probability space $(\Omega, \mathcal{F}, \mathcal{P})$ and a random measure $P : \Omega \times \mathfrak{S} \to \mathbb{R}$. Then, clearly for every $\omega \in \Omega$, P_ω is in the space S_∞, where $P_\omega(A) = P(\omega, A)$. Through this, we can see that P is a random element from $(\Omega, \mathcal{F})$ to $(S_\infty, \mathfrak{S}_\infty)$. It then makes sense to talk about

$$\mathcal{P}(P \in M) \qquad \text{for } M \in \mathfrak{S}_\infty.$$

As such, the distribution of P is a probability measure on $(S_\infty, \mathfrak{S}_\infty)$. Therefore, by constructing a random element that takes values in our space S_∞, we generate a distribution on the space S_∞.

Proposition 3.2. *Let $(S_\infty, \mathfrak{S}_\infty)$ be the measurable space defined in* (3.7) *and let $(\Omega, \mathcal{F}, \mathcal{P})$ be some arbitrary probability space. Then the map $p : \Omega \to S_\infty$ is a random element if and only if every coordinate p_i is a random variable.*

Proof. See Chapter 3 of [9]. □

In other words, Proposition 3.2 tells us that a distribution on $(S_\infty, \mathfrak{S}_\infty)$ corresponds to a sequence of random variables $(p_1, p_2, \dots)$ such that $\sum_{j=1}^{\infty} p_j = 1$ almost surely. From this and Definition 3.2, we see that random element $p = (p_1, p_2, \dots) \sim \mathrm{DP}(\alpha)$ if for every $k \in \mathbb{N}$,

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

where $\alpha = (\alpha_1, \alpha_2, \dots)$ is a (deterministic) sequence such that

$$\sum_{j=1}^{\infty} \alpha_j < \infty, \qquad \alpha_j \geq 0,\ j \in \mathbb{N}.$$

Construction through the Stick Breaking Process We perform the following algorithm to distribute the total probability mass 1, conceptually thought of as a stick of length 1, randomly to each coordinate $p_1, p_2, \dots$.

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 $V_2(1 - V_1)$ to the point p_2.
3. We repeat in this way so that point p_j has mass

$$p_j = V_j \prod_{l=1}^{j-1} (1 - V_l). \tag{3.8}$$

Proposition 3.3. *Let α be a probability distribution on $\mathbb{R}^d$ and let $M > 0$ be fixed. Suppose $\theta_1, \theta_2 \dots \overset{\text{i.i.d}}{\sim} \alpha$ and $V_1, V_2, \dots \overset{\text{i.i.d}}{\sim} \text{Beta}(1, M)$ are all mutually independent random variables. Then, for the random element $p = (p_1, p_2, \dots)$ defined in (3.8), we have that*

$$\sum_{j=1}^{\infty} p_j \delta_{\theta_j} \sim \text{DP}(M\alpha).$$

where $(M\alpha)(A) := M\alpha(A)$ for every $A \in \mathcal{B}(\mathbb{R}^d)$ is a measure on $\mathbb{R}^d$ of total mass M. We call $M > 0$ the concentration parameter.

Proof. See Theorem 4.12 of [9]. □

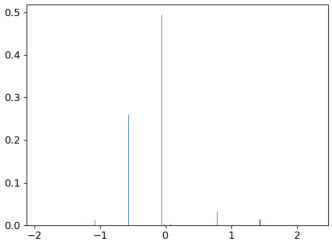


Figure 3.4: Realisation of Dirichlet process with $\alpha = \mathcal{N}(0, 1), M = 1$.

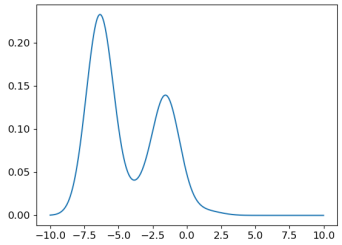


Figure 3.5: Realisation of Dirichlet process mixture.

From Proposition 3.3 and as we can see in Figure 3.4, it is clear that realisations of $\text{DP}(\alpha)$ are almost surely discrete. This does not seem suitable

for our purposes as we would like realisations of a DP to be distributions with continuous probability density functions. To solve this, we convolve the distribution generated from a Dirichlet process with a kernel, creating a Dirichlet process mixture.

Dirichlet Process Mixtures Let α be a probability measure on $\mathbb{R}^d$. Let Θ be some parameter set and, for $\theta \in \Theta \subset \mathbb{R}^d$, let the map $x \mapsto \psi(x, \theta)$ be a probability density function. For example, we could take $\theta = (\mu, \sigma^2) \in \mathbb{R} \times \mathbb{R}_{>0}$ and Gaussian probability density function

$$\psi(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

Let $F \sim \mathrm{DP}(\alpha)$. Then we define the Dirichlet process mixture to be the probability density function $p_{F,\psi}$ given by

$$p_{F,\psi}(x) = \int \psi(x, \theta) dF(\theta).$$

Since $F \overset{\mathcal{L}}{=} \sum_{j=1}^{\infty} p_j \delta_{\theta_j}$ for some $p = (p_1, p_2, \dots)$, $\theta = (\theta_1, \theta_2, \dots)$ constructed in Proposition 3.3, we get that

$$p_{F,\psi}(x) \overset{\mathcal{L}}{=} \sum_{j=1}^{\infty} p_j \psi(x, \theta).$$

Therefore, for Gaussian probability density function ψ, we have now extended (3.2) into a mixture of infinite number of Gaussians with a Dirichlet process prior on the mixture weights. In Figure 3.5, we have taken $\alpha_\mu = \mathcal{N}(0, 10), M = 1$ for mean parameter μ and

$$\psi(x; \mu) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-\mu)^2}.$$

We can see that a realisation of a Dirichlet process mixture is almost surely continuous, as desired. We our now back to estimation over an infinite number of parameters, therefore, such estimation is non-parametric. As we will see, even though we have an infinite number of parameters, the mixture weights will be significant in only a finite number of components.

3.3.2 Construction of Hierarchical Model

In similar fashion to Section 3.2, we will assume that the density f comes from a mixture of Gaussians as in (3.2). The notable difference is that we assume J is unknown and we hope that the Dirichlet process mixture will cause components weights to concentrate in only in a few components.

The stick breaking construction allows us to write down our Hierarchical model easily. We assume the same priors for (μ, τ) as in (3.5). We also put a prior on the concentration parameter α. Then we have the model:

$$\begin{aligned}
\alpha &\sim \mathcal{G}(\nu, \epsilon) \\
\tau &\sim \mathcal{G}(\eta, \gamma) \\
\mu_j|\tau &\overset{\text{ind}}{\sim} \mathcal{N}(\xi_j, \kappa^{-1}\tau^{-1}) \\
\beta_j|\alpha &\overset{\text{ind}}{\sim} \text{Beta}(1, \alpha) \\
\rho_j|\beta &= \beta_j \prod_{k=1}^{j-1}(1-\beta_k) \\
a_{ij}|\rho &\overset{\text{ind}}{\sim} \text{Po}(\lambda\rho_j\Delta) \\
Z_i|a, \mu, \tau &\overset{\text{ind}}{\sim} \mathcal{N}(a_i^T\mu, \tau^{-1}a_i^T\mathbf{1})
\end{aligned}$$

where $(\epsilon, \nu, \eta, \gamma, \xi, \kappa)$ are hyperparameters. We can illustrate this using the following probabilistic graphical model:

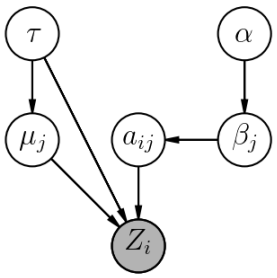


Figure 3.6: Probabilistic Graphical Model for the Hierarchical Model.

Truncating Mixture. Since we cannot generate infinite component mixtures in our simulations, we truncate our mixtures up to some sufficiently large number of components K. This can be justified intuitively, since with a finite number of observations, it seems quite likely that the number of mixture components that contribute non-negligible mass to the mixture will grow slower than the number of samples. This intuition can be formalized through the fact that the expected number of components that contribute non-negligible mass to the mixture approaches $\log n$, where n is the number of observations. Therefore, we take $K = \log n$.

3.3.3 Construction of MCMC Algorithm

We again use a Metropolis-Hastings-within-Gibbs algorithm to sample from the posterior distribution, but this time additionally performing a Metropolis-Hastings step to sample from conditional distribution $\beta|a, Z$.

Algorithm 2: Gibbs Sampler for DPMM Hierarchical Model

Result: Samples from the posterior distribution $p(a, \mu, \tau, \beta, \rho|Z)$.
Initialise $\alpha^{(0)}, \beta^{(0)}, \mu^{(0)}, \tau^{(0)}, a^{(0)}$;
for $k = 1, \ldots, K$ **do**
  Set $\rho_k^{(0)} \leftarrow \beta_k^{(0)} \prod_{j=1}^{k-1} (1 - \beta_j^{(0)})$
end
for $t = 1, \ldots, N$ **do**
  Update auxiliary variable a:

  1. Sample $a^{(t)} \sim p(a|\rho^{(t-1)}, \mu^{(t-1)}, \tau^{(t-1)}, Z)$ via Metropolis-Hastings step

  Update parameters α, β, τ, μ:

  1. Sample $\alpha^{(t)} \sim p(\alpha|\beta^{(t-1)})$
  2. Sample $\beta^{(t)} \sim p(\beta|a^{(t)}, \alpha^{(t)}, Z)$ via Metropolis-Hastings step
  3. Sample $\tau^{(t)} \sim p(\tau|a^{(t)}, Z)$
  4. Sample $\mu^{(t)} \sim p(\mu|a^{(t)}, \tau^{(t)}, Z)$

  Update parameter ρ:

  1. Set $\rho_k^{(t)} \leftarrow \beta_k^{(t)} \prod_{j=1}^{k-1} (1 - \beta_j^{(t)})$

end

Updating Concentration Parameter. We update auxiliary variable a using the same Metropolis-Hastings step as in Section 3.2.2. We also update parameters τ, μ using the same distributions in Lemma 3.2. To update α, we derive the conditional distribution $p(\alpha|\beta)$.

Lemma 3.3. *Conditional on β, we have*

$$\alpha|\beta \sim \mathcal{G}\left(\epsilon - K, \nu - \sum_{k=1}^{K} \log(1 - \beta_k)\right).$$

Proof. We have

$$\begin{aligned}
p(\alpha|\beta) &\propto p(\beta|\alpha)p(\alpha) \\
&= \left(\prod_{k=1}^{K}(1-\beta_k)^{\alpha-1}\frac{\Gamma(\alpha)}{\Gamma(\alpha+1)}\right)\alpha^{\epsilon-1}e^{-\nu\alpha} \\
&= \alpha^{\epsilon-K-1}\exp\left\{\sum_{k=1}^{K}(\alpha-1)\log(1-\beta_k)-\nu\alpha\right\} \\
&\propto \alpha^{\epsilon-K-1}\exp\left\{-\alpha\left(\nu-\sum_{k=1}^{K}\log(1-\beta_k)\right)\right\},
\end{aligned}$$

giving the required result. □

Updating Stick Breaking Weights. We update β using a Metropolis-Hastings step since its conditional distribution does not have a tractable form. For each $k = 1, \ldots, K$, we want to sample from $p(\beta_k|a, \beta_{-k}, \alpha) \propto p(a|\beta)p(\beta_k|\alpha)$, where

$$\beta_{-k} = (\beta_1, \ldots, \beta_{k-1}, \beta_{k+1}, \ldots, \beta_K).$$

Note that

$$\begin{aligned}
p(a|\beta)p(\beta_k) &\propto p(a|\rho)p(\beta_k) \\
&\propto \left(\prod_{i=1}^{n}\prod_{j=1}^{K}p(a_{ij}|\rho)\right)(1-\beta_k)^{\alpha-1} \\
&\propto (1-\beta_k)^{\alpha-1}\prod_{j=1}^{K}\prod_{i=1}^{n}e^{-\lambda\Delta\rho_j}\rho_j^{a_{ij}} \\
&= (1-\beta_k)^{\alpha-1}\prod_{j=1}^{K}e^{-n\lambda\Delta\rho_j}\rho_j^{s_j} \\
&= (1-\beta_k)^{\alpha-1}\prod_{j=1}^{K}\left(\exp\left\{-n\lambda\Delta\beta_j\prod_{l=1}^{j-1}(1-\beta_l)\right\}\beta_j^{s_j}\prod_{l=1}^{j-1}(1-\beta_l)^{s_j}\right) \\
&\propto (1-\beta_k)^{\alpha-1+\sum_{j=k+1}^{K}s_j}\beta_k^{s_k}\prod_{j=k}^{K}\exp\left\{-\lambda\Delta n\beta_j\prod_{l=1}^{j-1}(1-\beta_l)\right\}.
\end{aligned}$$

Therefore, we propose

$$\beta_k^{\circ} \sim \text{Beta}\left(s_k+1, \sum_{j=k+1}^{K}s_j+\alpha\right).$$

where the sum $\sum_{j=k+1}^{K} s_j$ is taken to be zero when $k = K$. Let

$$\beta^{\circ} = (\beta_1, \ldots, \beta_{k-1}, \beta_k^{\circ}, \beta_{k+1}, \ldots, \beta_K).$$

Our acceptance probability A is

$$A = \frac{\prod_{j=k}^{K} \exp\left\{-\lambda\Delta n \beta_j^{\circ} \prod_{l=1}^{j-1}(1-\beta_l^{\circ})\right\}}{\prod_{j=k}^{K} \exp\left\{-\lambda\Delta n \beta_j \prod_{l=1}^{j-1}(1-\beta_l)\right\}}$$

$$= \exp\left\{\lambda\Delta n \sum_{j=k}^{K}\left(\beta_j \prod_{l=1}^{j-1}(1-\beta_l) - \beta_j^{\circ}\prod_{l=1}^{j-1}(1-\beta_l^{\circ})\right)\right\}$$

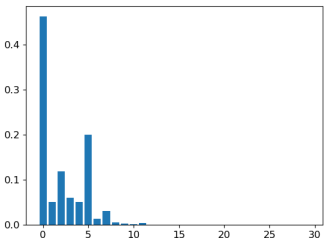
We accept β° with probability $1 \wedge A$. After we complete all Metropolis-Hastings steps, we update stick-breaking weights ρ accordingly with the new β.

3.3.4 Numerical Simulations

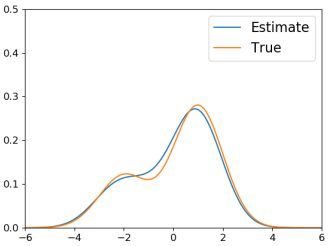
Pitfalls of this Method. Notice that for density function f given by a mixture of Gaussians, if we do not assume that each component is unique, we can split the mixture into a mixture with a greater number of components. More formally,

$$f(\cdot) \triangleq \sum_{j=1}^{J} \rho_j \psi(\cdot; \mu_j, 1/\tau) = \sum_{j=1}^{K} \rho_j^{'} \psi(\cdot; \mu_j; 1/\tau),$$

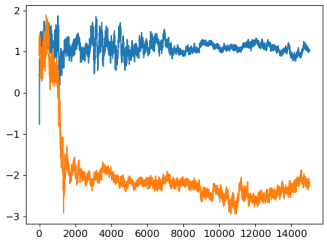
for $K \geq J$ and some possibly identical μ_j's. Therefore, for unknown number of components, we have non-identifiability in the parameters in question. Although we may obtain good estimates of the probability density functions, we may have estimates of parameters μ_j which are very close or identical to each other. Nonetheless, we hope that by placing a prior on the concentration parameter α, updating this in our Gibbs sampler conditional on the observations will cause our estimates to be concentrated in the same or almost the same number of components that the true density has.



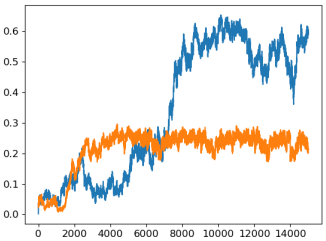
(a) Posterior mean of Mixing Weights



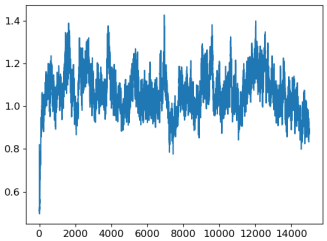
(b) Density Estimate



(c) Parameter μ_1, μ_2



(d) Parameter ρ_1, ρ_2



(e) Parameter τ

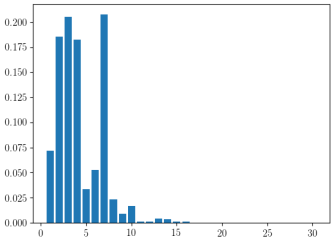
Figure 3.7: Run of MCMC algorithm for mixture of two Gaussians.

Example 1. We used the same CPP simulation of Example 1 of section 3.2 - we simulated a CPP with intensity $\lambda = 1$ and jump density function $f(x) = 0.3\psi(x; -2, 1) + 0.7\psi(x; 1, 1)$. Therefore,

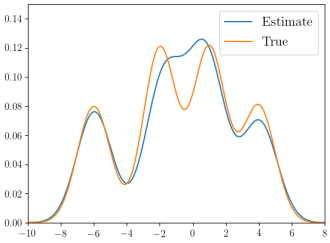
$$(\mu_1, \mu_2) = (-2, 1), (\rho_1, \rho_2) = (0.3, 0.7), \tau = 1.$$

We obtained $n = 8000$ observations which consisted of roughly 5000 non-zero observations and performed 15000 MCMC iterations, truncating the mixture to 30 components. We obtained an average acceptance rate of 46% for the auxiliary variable a Metropolis-Hastings proposals and an average acceptance rate of 63% for Metropolis-Hastings proposals of β. The density estimate in Figure 3.7b was obtained using the posterior mean estimates of the parameters. It shows an overall satisfactory fit, similar to that of section 3.2. This portrays the feasibility of this approach numerically.

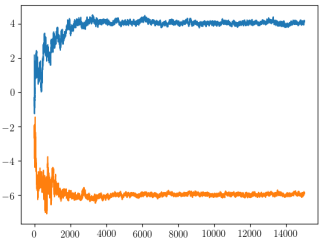
The mixture components converged much slower than in section 3.2. This may be due to the label switching problem, as discussed previously, as well as the fact that we have a much larger parameter space of 30 mixture components. We would need to test this for a greater number of iterations to see whether convergence is found.



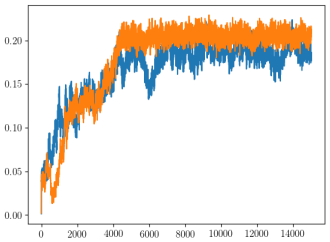
(a) Posterior mean of Mixing Weights



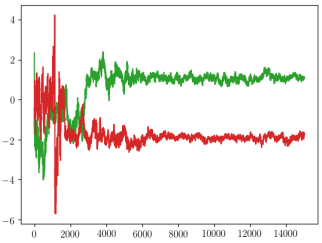
(b) Density Estimate



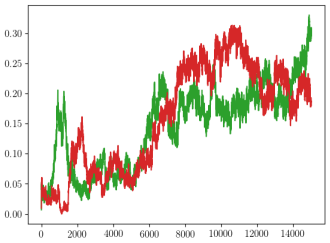
(c) Parameter μ_1, μ_2



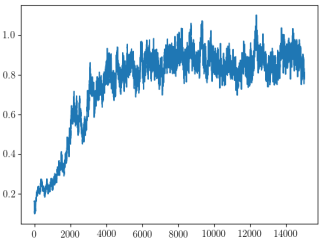
(d) Parameter ρ_1, ρ_2



(e) Parameter μ_3, μ_4



(f) Parameter ρ_3, ρ_4



(g) Parameter τ

Figure 3.8: Run of MCMC algorithm for mixture of four Gaussians.

Example 2. We again used the same CPP simulation of Example 2 of section 3.2 - we simulated a CPP with intensity $\lambda = 1$ and jump density function $f(x) = 0.2\psi(x; -6, 1) + 0.3\psi(x; -2, 1) + 0.3\psi(x; 1, 1) + 0.2\psi(x; 4, 1)$.

Therefore, in these plots,

$$(\mu_1, \mu_2, \mu_3, \mu_4) = (-6, 4, -2, 1), (\rho_1, \rho_2, \rho_3, \rho_4) = (0.2, 0.2, 0.3, 0.3), \tau = 1.$$

We obtained $n = 8000$ observations which consisted of roughly 5000 non-zero observations and performed 15000 MCMC iterations, truncating the mixture to 30 components. We obtained an average acceptance rate of 29.4% for the auxiliary variable Metropolis-Hastings proposals and an average acceptance rate of 45% for Metropolis-Hastings proposals of β. We obtained a density estimate using the posterior mean estimates of the parameters as seen in Figure 3.8b.

As we can see in the plots, the posterior mean of the mixing weights is significant in only four components. Therefore, we observe, as intended, that the Dirichlet process mixture will concentrate in the same number of components as the true density. Focusing the discussion to these components, we see good convergence in Figure 3.8c and 3.8d of the first two mixture components. However, convergence of the mixture weights in the last two mixture components is poor and this is reflected in the density estimate as we cannot see well-defined modes around -2 and 1.

4 Conclusion

We have provided four approaches to estimating the jump density of a compound Poisson process. We have rigorously constructed the estimators and have seen that the estimators show satisfactory fit on examples within the high frequency setting.

It is evident from the previous sections that the estimators have been defined on the basis of different conditions or assumptions. Therefore, it is somewhat difficult to make valid comparisons between the estimators since some assumptions are more strong than others. For example, the kernel density estimators do not assume that the probability density functions come from a mixture of Gaussians, whilst in the Bayesian density estimation setting, we have made this assumption in order to obtain a tractable likelihood.

The kernel density estimators provide the most general framework for density estimation out of the estimators shown. This can be seen in the numerical simulations where they show good fit on a range of distributions. However, we must tune the bandwidth h to obtain smooth estimates of the density. There are various methods to do this, for example, we could perform cross validation as suggested in [15].

The Bayesian approach is attractive since it provides a distribution rather than just a point estimate. However, in the estimators shown, we are limited to mixtures of Gaussians. Furthermore, computation is expensive and convergence can be slow, as shown in the numerical simulations in Section 3.3. If we were able to formulate a Bayesian approach without

having to assume the parametric family of the true density, then this would give the benefits of using a Bayesian framework, whilst also formulating the problem is a more general setting.

As discussed with Dr. Coca, we could instead work with the intractable likelihood

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

By truncating the series up to some K sufficiently large and using the Fourier Inversion Theorem, it is possible to approximate this likelihood for densities f.

Therefore, by specifying a Dirichlet process prior, defined in Section 3.3, over the space of probability densities, we can return to the Bayesian framework without having to bypass the intractable likelihood.

Statistical inference on discretely observed compound Poisson processes continues to prove itself a fascinating and active area of research, from both a computational and theoretical standpoint. We hope that this work clearly presents the different approaches for solving the statistical inverse problem and provides a good entry point towards exploring the field further.

A Appendix

Proof of Proposition 1.2.

$$\begin{aligned}
\phi_Z(t) &= \mathbb{E}e^{itZ} = \mathbb{E}e^{it\mathbb{1}(N>0)\sum_{i=1}^{N} Y_i} \\
&= \mathbb{E}[\mathbb{1}(N=0)] + \mathbb{E}\left[\mathbb{1}(N>0)\prod_{i=1}^{N} e^{itY_i}\right] \\
&= e^{-\lambda\Delta} + \mathbb{E}\left[\mathbb{1}(N>0)\mathbb{E}\left[\prod_{i=1}^{N} e^{itY_i}\middle| N\right]\right] && \text{(law of total expectation)} \\
&= e^{-\lambda\Delta} + \mathbb{E}\left[\mathbb{1}(N>0)\prod_{i=1}^{N} \mathbb{E}\left[e^{itY_1}\middle| N\right]\right] && \text{(i.i.d property of jumps)} \\
&= e^{-\lambda\Delta} + \mathbb{E}\left[\mathbb{1}(N>0)\prod_{i=1}^{N} \phi_f(t)\right] && \text{(independence of } Y_1 \text{ and } N) \\
&= e^{-\lambda\Delta} + \mathbb{E}\left[\mathbb{1}(N>0)e^{N\ln\phi_f(t)}\right] \\
&= \mathbb{E}\left[e^{N\ln\phi_f(t)}\right] \\
&= \exp(\lambda\Delta(e^{\ln\phi_f(t)}-1)) && \text{(MGF of } \mathcal{P}(\lambda\Delta)) \\
&= e^{-\lambda\Delta+\lambda\Delta\phi_f(t)}
\end{aligned}$$

Proof of Proposition 1.3. Suppose we have $A \in \mathcal{B}$ such that $\mu(A) = 0$. Then $0 \notin A$ and A has Lebesgue measure 0. Therefore, under event A, we have that $N > 0$ i.e.

$$\left\{ \mathbb{1}(N > 0) \sum_{j=1}^{N} Y_i \in A \right\} \subseteq \{N > 0\} \cap \left\{ \sum_{j=1}^{N} Y_i \in A \right\}.$$

Using this, we get that

$$\begin{aligned}
\mathbb{P}_Z(A) &= \mathbb{P}\left(\mathbb{1}(N > 0) \sum_{j=1}^{N} Y_i \in A \right) \\
&\leq \mathbb{P}\left(N > 0, \sum_{j=1}^{N} Y_i \in A \right) \\
&= \sum_{n=1}^{\infty} \mathbb{P}\left(\sum_{j=1}^{n} Y_i \in A, N = n \right) \\
&= \sum_{n=1}^{\infty} \mathbb{P}\left(\sum_{j=1}^{n} Y_i \in A \right) \mathbb{P}(N = n) \\
&= 0
\end{aligned}$$

since $\sum_{j=1}^{n} Y_i$ has a density by Lemma 1.1. Therefore, $\mathbb{P}_Z$ is absolutely continuous with respect to μ. Furthermore, for $A \in \mathcal{B}$,

$$\begin{aligned}
\mathbb{P}_Z(A) &= \mathbb{P}\left(0 \in A, N = 0\right) + \sum_{m=1}^{\infty} \mathbb{P}\left(\sum_{j=1}^{m} Y_i \in A \right) \mathbb{P}(N = m) \\
&= \mathbb{P}(N = 0) \int_A d\delta_0 + \sum_{m=1}^{\infty} e^{-\lambda\Delta} \frac{(\lambda\Delta)^m}{m!} \int_A f^{*m}(x) dx \\
&= \int_A e^{-\lambda\Delta} \mathbb{1}_{\{0\}}(x) + \sum_{m=1}^{\infty} e^{-\lambda\Delta} \frac{(\lambda\Delta)^m}{m!} f^{*m}(x) \mathbb{1}_{\mathbb{R}\setminus\{0\}}(x) d\mu(x)
\end{aligned}$$

giving the result.

Proof of Lemma 3.2. For (μ, τ) we get

$$\begin{aligned}
p(\mu,\tau|Z,a) &\propto p(Z|a,\mu,\tau)p(\mu,\tau) \\
&\propto p(Z|\mu,\tau,a)p(\mu|\tau)p(\tau) \\
&\propto \left(\prod_{i=1}^{n}\psi(Z_i;a_i^T\mu,n_i/\tau)\right)\left(\tau^{J/2}\exp\left\{-\frac{\tau\kappa}{2}\sum_{j=1}^{J}(\mu_j-\xi_j)^2\right\}\right)\left(\tau^{\eta-1}e^{-\gamma\tau}\right) \\
&\propto \tau^{n/2}\exp\left\{-\frac{\tau}{2}\sum_{i=1}^{n}n_i^{-1}(Z_i-a_i^T\mu)^2\right\}\left(\tau^{J/2}\exp\left\{-\frac{\tau\kappa}{2}\sum_{j=1}^{J}(\mu_j-\xi_j)^2\right\}\right)\left(\tau^{\eta-1}e^{-\gamma\tau}\right) \\
&\propto \tau^{\eta-1+(n+J)/2}\exp\left\{-\gamma\tau-\frac{D(\mu)}{2}\tau\right\}
\end{aligned}$$

where

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

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

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

Therefore,

$$p(\mu|\tau,a,Z)\propto\exp\left\{-\frac{\tau}{2}(\mu-P^{-1}q)^TP(\mu-P^{-1}q)\right\}$$

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

Also,

$$\begin{aligned}
\int\exp(-\frac{\tau}{2}D(\mu))d\mu &= \exp\left\{-\frac{\tau}{2}(R-q^TP^{-1}q)\right\}\int\exp\left(-\frac{\tau}{2}(\mu-P^{-1}q)^TP(\mu-P^{-1}q)\right)d\mu \\
&= \exp\left\{-\frac{\tau}{2}(R-q^TP^{-1}q)\right\}(2\pi)^{J/2}\sqrt{|\tau^{-1}P^{-1}|} \qquad \text{(A.1)}
\end{aligned}$$

Line (A.1) follows by the integral of a multivariate Gaussian distribution being equal to 1. Thus, we can write $p(\tau|Z,a)$ as

$$\begin{aligned}
p(\tau|Z,a) &= \int p(\tau,\mu|z,a)d\mu \\
&= \int\tau^{\eta-1+(n+J)/2}\exp\left\{-\gamma\tau-\frac{D(\mu)}{2}\tau\right\}d\mu \\
&\propto \tau^{\eta-1+(n+J)/2}e^{-\gamma\tau}(2\pi)^{J/2}\sqrt{|\tau^{-1}P^{-1}|}\exp\left\{-\frac{\tau}{2}(R-q^TP^{-1}q)\right\} \quad \text{using (A.1)} \\
&\propto \tau^{\eta+(n+J-J)/2-1}\exp\left\{-\tau\left(\gamma+\frac{1}{2}(R-q^TP^{-1}q)\right)\right\}
\end{aligned}$$

giving that $\tau|Z,a\sim\mathcal{G}(\eta+n/2,\gamma+(R-q^TP^{-1}q)/2)$

References

[1] Alexandre B. Tsybakov. Introduction to nonparametric estimation. 01 2009.

[2] Christophe Chesneau, Fabienne Comte, and Fabien Navarro. Fast nonparametric estimation for convolutions of densities. *Canadian Journal of Statistics*, 41(4):617–636, 2013.

[3] Kai Lai Chung. *A course in probability theory*. Academic press, 2001.

[4] Fabienne Comte, Cline Duval, and Valentine Genon-Catalot. Nonparametric density estimation in compound poisson process using convolution power estimators. *Metrika*, 77, 01 2014.

[5] Hans Crauel. *Random probability measures on Polish spaces*. CRC press, 2002.

[6] Céline Duval. Density estimation for compound poisson processes from discrete data. *Stochastic Processes and their Applications*, 123(11): 3963–3986, 2013.

[7] Thomas S. Ferguson. A bayesian analysis of some nonparametric problems. *Ann. Statist.*, 1(2):209–230, 03 1973.

[8] Gerald B Folland. *Fourier analysis and its applications*, volume 4. American Mathematical Soc., 2009.

[9] Subhashis Ghosal and Aad Van der Vaart. *Fundamentals of nonparametric Bayesian inference*, volume 44. Cambridge University Press, 2017.

[10] Evarist Gin and Richard Nickl. *Mathematical Foundations of Infinite-Dimensional Statistical Models*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 2015.

[11] Shota Gugushvili, Frank van der Meulen, and Peter Spreij. A nonparametric bayesian approach to decompounding from high frequency data. *Statistical Inference for Stochastic Processes*, 21(1):53–79, Apr 2018.

[12] Francis Begnaud Hildebrand. *Introduction to numerical analysis*. Courier Corporation, 1987.

[13] Robert C. Merton. Option pricing when underlying stock returns are discontinuous. *Journal of Financial Economics*, 3(1):125 – 144, 1976. ISSN 0304-405X.

[14] Thomas Mikosch. *Non-life insurance mathematics: an introduction with the Poisson process*. Springer Science & Business Media, 2009.

[15] Berwin A Turlach. Bandwidth selection in kernel density estimation: A review. In *CORE and Institut de Statistique*. Citeseer, 1993.

[16] Bert van Es, Shota Gugushvili, and Peter Spreij. A kernel type nonparametric density estimator for decompounding. *Bernoulli*, 13(3):672–694, 08 2007.

[17] D Vere-Jones and T Ozaki. Some examples of statistical estimation applied to earthquake data. *Annals of the Institute of Statistical Mathematics*, 34(1):189–207, 1982.

[18] MP Wand. Finite sample performance of deconvolving density estimators. *Statistics & Probability Letters*, 37(2):131–139, 1998.