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

Suraj Shah

April 15, 2019

Contents

Chapter 1

Introduction

Jump processes are a group of continuous-time stochastic processes whose values move in discrete amounts at random times. The compound Poisson process sits near the centre of this rich class of processes, and its versatility makes it an exceptional candidate for modelling phenomena in the world. However, its versatility comes at a price - statistical inference on them is difficult when we only observe values at discrete points in time because a significant amount of information is lost.

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 suitable kernel functions as shown in van Es et al. [4] and in Comte et al. [1]. We then visit non-parametric Bayesian inference. We simplify our problem to the parametric case in order to construct the density estimator laid out in Gugushvili et al. [3]. Finally, we generalise the Bayesian procedure usings Dirichlet processes and illustrate its performance via numerical simulations.

Since this essay is computational in flavour, we focus mainly on carefully deriving the forms of the estimators and analyse their performance on various example simulations. Therefore, we omit many theoretical guarantees of these estimators, but such results exist and are abundant in the literature referenced.

1.1 Compound Poisson Processes

Properties. Compound Poisson processes push the boundaries of Poisson processes, 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.

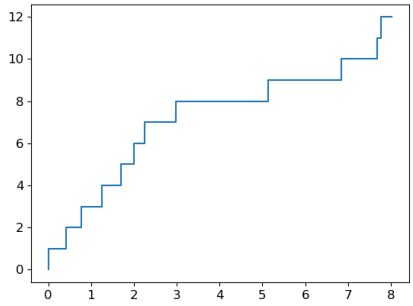


Figure 1.1.1: Poisson Process

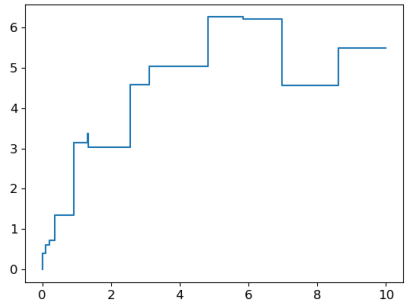


Figure 1.1.2: Compound Poisson Process

Definition 1.1.1 (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 $\lambda\Delta$:
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 **??**, the jump sizes are of unit length, whilst the compound Poisson process allows the jumps to vary in both magnitude and direction.

Definition 1.1.2 (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 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 Poison 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 us 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 probability density function f and assigns zero mass at the origin $\{0\}$, i.e.
$$F(\{0\}) = 0.$$
We assign zero mass at the origin for the jump distribution since, 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 $\{X_{i\Delta} : i = 1, \dots, n\}$. We approach this problem using non-parametric estimation, as we want to put little assumptions on density f. For example, we may assume only 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 $\{X_{i\Delta} : i = 1, \dots, n\}$, consider increments $\{Z_i : i = 1, \dots, n\}$ 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 1.2.1. *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 $X_{i\Delta}$. Therefore, henceforth, we will refer to our observations as $\{Z_i : i = 1, \dots, n\}$ and call such Z_i as 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 will be vital for the essay and we will refer to these properties throughout.

Poisson Random Sums. Let $Y_1, Y_2, \dots$ 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 1.3.1. *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.

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

□

For the next property, we require the following Lemma.

Lemma 1.3.1. *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 $*$ denotes the convolution i.e.

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

Proposition 1.3.2. *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. 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.

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

Therefore,

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

since $\sum_{j=1}^{n} Y_i$ has a density by Lemma 1.3.1. 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. □

In light of Proposition 1.3.2, 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

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

Chapter 2

Kernel Density Estimation

Kernel density estimation is a non-parametric way of 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 make it an ideal candidate for continuous probability density functions.

Motivation 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 , 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 to gives us an estimator $\hat{p}_n(x)$ of $p(x)$ where

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

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 2.0.1. (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

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

2.1 Estimation of Convolution Powers

Returning back to our problem, suppose we have non-zero observations $\{Z_i : i = 1, \ldots, n\}$ of the compound Poisson process. By (1.6) following Proposition 1.3.2, we have that each Z_i has density g given by

$$g = \frac{e^{-\lambda\Delta}}{1 - e^{-\lambda\Delta}} \sum_{m=1}^{\infty} \frac{(\lambda\Delta)^m}{m!} f^{*m} = \frac{1}{e^{\lambda\Delta} - 1} \sum_{m=1}^{\infty} \frac{(\lambda\Delta)^m}{m!} f^{*m} \tag{2.2}$$

We can see that density g is a weighted sum of convolution power 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.

Proposition 2.1.1. *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. 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 the same as the characteristic function of a random variable with probability density function f. We first show, using the Fourier Inversion Theorem, that the Fourier transform defined above is injective.

Fourier Inversion Theorem. Let $f \in L^1(\mathbb{R})$ be an integrable function. Suppose also that $\mathcal{F}[f]$ is integrable. Then

$$f(x) = \frac{1}{2\pi} \int_{\mathbb{R}} e^{-itx} \mathcal{F}[f](t)dt \tag{2.3}$$

To show that $\mathcal{F}$ is injective, suppose that f is in $L^1(\mathbb{R})$ such that $\mathcal{F}[f] = 0$. Then, we have that $\mathcal{F}[f]$ is in $L^1(\mathbb{R})$ since it is the zero function. Therefore, by Levy's Inversion formula, 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 that states that $\mathcal{F}[f * g] = \mathcal{F}[f]\mathcal{F}[g]$ for integrable functions f, g. From (2.2), 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

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

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 the previous section reduces to the principal branch of the logarithm. Thus, we get that

$$\mathcal{F}[f] = \frac{\log(1 + (e^{\lambda\Delta} - 1)\mathcal{F}[g])}{\lambda\Delta} = \sum_{m=1}^{\infty} \frac{(-1)^{m+1}}{m} \frac{(e^{\lambda\Delta} - 1)^m}{\lambda\Delta} \mathcal{F}[g]^m \tag{2.4}$$

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 a probability density function so $f \in L^1(\mathbb{R})$, gives the result. □

2.1.1 Construction

Consider the kernel density estimator, shown in Definition 2.0.1, given by

$$\hat{g}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} k\left(\frac{x - Z_i}{h}\right), \qquad \text{for some fixed } h > 0. \tag{2.5}$$

We take the kernel function k to be the sinus cardinal function

$$k(x) = \frac{\sin(\pi x)}{\pi x} \tag{2.6}$$

Proposition 2.1.2. *The sinus cardinal function given in* (2.6) *is a kernel function and has characteristic function*

$$\phi_k(t) = \int_{\mathbb{R}} k(x) e^{itx} dx = \mathbb{1}_{[-\pi,\pi]}(t)$$

We also have the following result:

Lemma 2.1.1. *Let k be a kernel function and let bandwidth $h > 0$. Let $\hat{g}_n$ be the kernel density estimator given in* (2.5) *and let ϕ_k be the characteristic function of kernel k. Also, let ϕ_{emp} be the empirical characteristic function given by*

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

Then for all $t \in \mathbb{R}$,

$$\phi_{\hat{g}_n}(t) = \phi_{emp}(t)\phi_k(ht)$$

Proof.

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

□

Therefore, using Levy's Inversion formula and Lemma 2.1.1 for our sinus cardinal kernel k, we see that the kernel density estimator can be rewritten as

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

Furthermore, using Lemma 1.3.1, which is equivalent to the Convolution Theorem, we note that $\phi_{g^{*m}} = (\phi_g)^m$. Therefore, the convolution power $\hat{g}_n^{*m}$ is given by

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

Therefore, using $\hat{g}_n^{*m}$ as our estimators 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. Especially, $\frac{(e^{\lambda\Delta}-1)^m}{m} \to 0$ as $m \to \infty$. Therefore, it is sensible to truncate the series up to some sufficiently large K to give

$$\hat{f}_n(x) = \sum_{m=1}^{K} \frac{(-1)^{m+1}}{m}\frac{(e^{\lambda\Delta}-1)^m}{\lambda\Delta}\hat{g}_n^{*m}(x) \tag{2.7}$$

2.1.2 Simulation Results

To compute estimators $\hat{g}_n^{*m}$, we decompose it into two terms, approximate the integral using the trapezoid rule and then perform the Fast Fourier Transform (FFT) to obtain a range of values for the density function estimator.

We can express $\hat{g}_n^{*m} = \hat{g}_{n,1}^{*m} + \hat{g}_{n,2}^{*m}$ where

$$\hat{g}_{n,1}^{(m)}(x) = \frac{1}{2\pi}\int_{0}^{\frac{\pi}{h}} \left(\phi_{\text{emp}}(t)\right)^m e^{-itx}dt$$

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

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.

Trapezoid Rule. 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

$$\int_a^b h(x)dx \approx \eta \sum_{j=0}^{N-1} h(t_j) \tag{2.8}$$

We take a grid $t_j = j\eta$ for $j = 0, 1, \ldots, N-1$ where N is some large power of 2 and $\eta = \frac{\pi}{(N-1)h}$.

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

We take $x_k = -\frac{N\delta}{2} + \delta k$ for $k = 0, 1, \ldots, N-1$ and δ is some constant to be defined later. Then

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

Similarly,

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

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

$$Y_j = \sum_{k=0}^{N-1} x_k e^{-ij\frac{2\pi k}{N}} \tag{2.9}$$

The inverse transform is given by

$$Y_j = \frac{1}{N} \sum_{k=0}^{N-1} x_k e^{ij\frac{2\pi k}{N}} \tag{2.10}$$

We choose $\delta = \frac{2h(N-1)}{N}$ 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 (2.7) gives an estimator $\hat{f}_n$ of f.

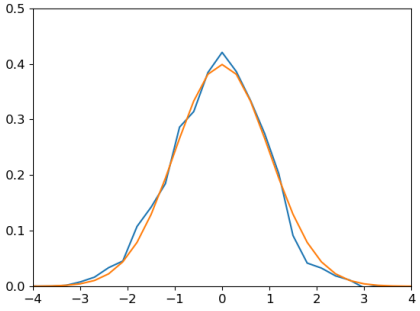


Figure 2.1.1: Standard Gaussian Jump Distribution

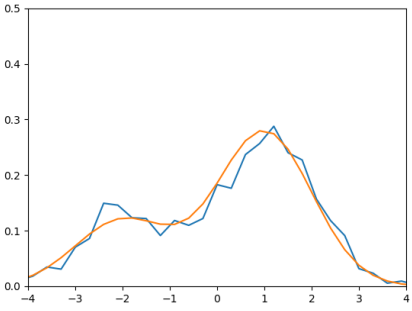


Figure 2.1.2: Mixture of Gaussians Jump Distribution

Examples. We took $\lambda = 0.3, \Delta = 0.4$, and kernel density estimators with $N = 16384$ number of points and bandwidth $h = 0.15$. The simulations show a satisfactory general fit, capturing the overall shape well.

2.2 Inversion of Characteristic Functions

Once again, suppose we have non-zero observations $\{Z_i : i = 1, \ldots, n\}$ of the compound Poisson process. Let

$$X = \mathbb{1}(N > 0) \sum_{j=1}^{N} Y_i$$

be a Poisson random sum with $N \sim \mathcal{P}(\lambda\Delta)$. Then

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

Using Proposition 1.3.1 and letting the probability density function of an observation Z_i be denoted by g (which exists by Proposition 1.3.2), we get that

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

Rearranging, we get

$$\phi_g(t) = \frac{1}{e^{\lambda\Delta} - 1}(e^{\lambda\Delta\phi_f(t)} - 1). \tag{2.11}$$

This suggests that if we could suitably invert the formula in (2.11) to get an expression in terms of ϕ_f, then an estimator of ϕ_g would induce an estimator for ϕ_f. Then, using Fourier Inversion Theorem, we would obtain an estimator for f.

2.2.1 Construction

The main issue to address is inverting relationship (2.11) to obtain an expression in terms of ϕ_f. However, 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 our the sense, since it is not injective: $e^{w+2\pi i} = e^w \ \forall w \in \mathbb{C}$.

Therefore, the following Lemmas are significant to obtain our desired expression.

Distinguished Logarithm.

Lemma 2.2.1. *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 2.2.2. *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 Appendix. □

For a function ϕ satisfying the assumptions of Lemma 2.2.2, 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 2.2.2:

Lemma 2.2.3. *For ϕ_1, ϕ_2 satisfying the assumptions of Lemma 2.2.2, we have for $\psi(t) = \phi_1(t)\phi_2(t)$,*

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

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

$$\begin{aligned}
\lambda\Delta(\phi_f(t) - 1) &= \text{Log}\left(e^{\lambda\Delta(\phi_f-1)}\right)(t) && \text{(Lemma 2.2.1)} \\
&= \text{Log}\left(e^{-\lambda\Delta}\left[(e^{\lambda\Delta} - 1)\phi_g + 1\right]\right)(t) && \text{(using (2.11))} \\
&= -\lambda\Delta + \text{Log}\left((e^{\lambda\Delta} - 1)\phi_g + 1\right)(t)
\end{aligned}$$

(we need that Log of constant function is constant). Therefore,

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

By Levy's Inversion formula, for integrable ϕ_f we have

$$f(x) = \frac{1}{2\pi\lambda\Delta}\int_{-\infty}^{\infty} e^{-itx}\text{Log}\left((e^{\lambda\Delta} - 1)\phi_g + 1\right)(t)dt \tag{2.12}$$

Using our theory of kernel density estimation, 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

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

Technical Issues. Using Lemma 2.1.1, it is tempting to introduce an estimator $\hat{f}_n$ of f

$$\hat{f}_n(x) = \frac{1}{2\pi\lambda\Delta}\int_{-\infty}^{\infty} e^{-itx}\text{Log}\left((e^{\lambda\Delta} - 1)\phi_{\text{emp}}\phi_k(h\ \cdot) + 1\right)(t)dt \tag{2.13}$$

but this brings two main issues:

1. In light of Lemma 2.2.2, we may have some measurable set A with non-zero Lebesgue measure such that $(e^{\lambda\Delta} - 1)\phi_{\text{emp}}(t)\phi_w(ht) + 1$ 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,

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

would give $\hat{f}_n(1)$ 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.

2.2.2 Simulation Results

We note that for $\lambda\Delta < \log 2$, the distinguished logarithm in (2.13) reduces to the principal branch of the logarithm. Therefore, we can directly use the built-in logarithm from scientific computing packages, and bounding $\hat{f}_n$ by a suitable sequence is not needed. Thus, we use (2.13) directly to compute our estimator on cases where $\lambda\Delta < \log 2$.

Kernel. We use the following kernel function k given by

$$k(t) = \frac{48t(t^2-1)\cos t - 144(2t^2-5)\sin t}{\pi t^7} \tag{2.14}$$

This expression is fairly non-trivial, but has a much simpler expression for the characteristic function.

Proposition 2.2.1. *The function k defined in* (2.14) *is a kernel and has characteristic function*

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

In a similar fashion to Section 2.1.2, we rewrite (2.13) as $\hat{f}_n(x) = \hat{f}_{n,1}(x) + \hat{f}_{n,2}(x)$ where

$$\hat{f}_{n,1}(x) = \frac{1}{2\pi\lambda\Delta} \int_0^\infty e^{-itx} \log\Big((e^{\lambda\Delta}-1)\phi_{\text{emp}}(t)\phi_k(ht)+1\Big)dt \tag{2.15}$$

$$\begin{aligned}\hat{f}_{n,2}(x) &= \frac{1}{2\pi\lambda\Delta} \int_{-\infty}^0 e^{-itx} \log\Big((e^{\lambda\Delta}-1)\phi_{\text{emp}}(t)\phi_k(ht)+1\Big)dt \\ &= \frac{1}{2\pi\lambda\Delta} \int_0^\infty e^{itx} \log\Big((e^{\lambda\Delta}-1)\phi_{\text{emp}}(-t)\phi_k(ht)+1\Big)dt \end{aligned} \tag{2.16}$$

Line (2.16) follows since ϕ_k is symmetric. We again deal with $\hat{f}_{n,1}$ - the case of $\hat{f}_{n,2}$ is very similar.

Approximating (2.15) using the Trapezoid rule (2.8), we get, for spacing parameter $\eta > 0$ and $t_j = j\eta$, that

$$\hat{f}_{n,1}(x) \approx \frac{\eta}{2\pi\lambda\Delta} \sum_{k=0}^{N-1} e^{-it_j x} \log\Big((e^{\lambda\Delta}-1)\phi_{\text{emp}}(t_j)\phi_k(ht_j)+1\Big), \tag{2.17}$$

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

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

and $\delta > 0$ is some constant to be defined later. Thus we have

$$\hat{f}_{n,1}(x_k) \approx \frac{1}{2\pi\lambda} \sum_{j=0}^{N-1} e^{-ijk\eta\delta} e^{it_j \frac{N\delta}{2}} \psi_1(t_j)\eta, \tag{2.18}$$

where $\psi_1(t) = \log\big((e^{\lambda\Delta}-1)\phi_{\text{emp}}(t)\phi_k(ht)+1\big)$. Similarly,

$$\hat{f}_{n,2}(x_k) \approx \frac{1}{2\pi\lambda} \sum_{j=0}^{N-1} e^{ijk\eta\delta} e^{-it_j \frac{N\delta}{2}} \psi_2(t_j)\eta, \tag{2.19}$$

where $\psi_2(t) = \log\big((e^{\lambda\Delta}-1)\phi_{\text{emp}}(-t)\phi_k(ht)+1\big)$.

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 smaller so that δ can be relatively larger and thus, points at which we evaluate our density estimator are relatively separate from one another.

Examples. We took $\lambda = 0.3$, $\Delta = 0.4$, $N = 16384$, $\eta = 0.01$ and bandwidth $h = 0.14$.

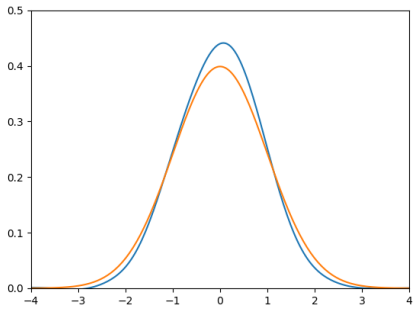


Figure 2.2.1: Standard Gaussian Jump Distribution

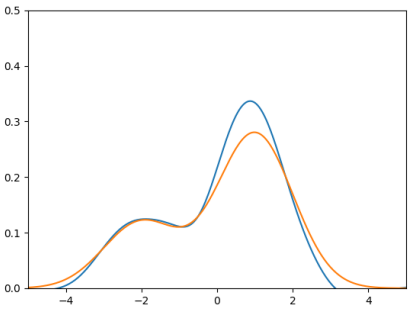


Figure 2.2.2: Mixture of Gaussians Jump Distribution

Chapter 3

Bayesian Density Estimation

The kernel density estimators provide a satisfactory attempt in recovering jump 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 [3], 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 as

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

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.

(Add some additional details).

3.2 Parametric Estimation using a Data Augmentation Scheme

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. We then use a data augmentation scheme to make our likelihood tractable and implement 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:

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

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 (1.6) following Proposition 1.3.2 that for non-zero observation Z, the likelihood of Z given f is given by

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

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 $Z_i = \sum_{j=1}^{N} Y_j$ 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

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

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 MCMC Algorithm

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 some hyperparameters.

Proposition 3.2.1. *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 our prior distribution on θ specified above. Then, we can write our model as:

$$\begin{aligned} \theta &\sim \pi(\theta) \\ a_{ij}|\ \rho &\overset{\text{ind}}{\sim} \mathcal{P}(\lambda\rho_j\Delta) \\ Z_i|a_i, \theta &\overset{\text{ind}}{\sim} \mathcal{N}(a_i^T\mu, a_i^T\tau^{-1}) \end{aligned}$$

This model gives us the following joint distribution decomposition:

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

This suggests performing a Gibbs sampling algorithm 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 $q(n_i^\circ, a_i^\circ|\rho)$ 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 3.2.1. Therefore, our acceptance probability A is

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

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

Updating parameters. Let

$$n_i = \sum_{j=1}^{J} a_{ij}, \qquad s_j = \sum_{i=1}^{n} a_{ij} \tag{3.6}$$

be the number of jumps in observation i and the number of jumps in component j respectively. We use the following Lemma to update our parameters. (Need to define s_j and n_i).

Lemma 3.2.1. *Conditional on a, we have that $\rho_1, \ldots, \rho_J$ are independent and the following holds:*

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

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

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

q is the J-dimensional vector with

$$q_j = \kappa \xi_j + \sum_i n_i^{-1} a_{ij} Z_i$$

$R > 0$ 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. (Move this to Appendix).

$$\begin{aligned} p(\rho|a) &\propto p(a|\rho)\pi(\rho) \\ &= \prod_{j=1}^{J} \pi(\rho_j) \left(\prod_i p(a_{ij}|\rho) \right) \\ &\propto \prod_{j=1}^{J} \rho_j^{\alpha-1} \left(\prod_i e^{-\rho_j \lambda \Delta} (\rho_j \lambda \Delta)^{a_{ij}} \right) \\ &= \prod_{j=1}^{J} \rho_j^{\alpha-1} e^{-\rho_j \lambda n \Delta} (\rho_j \lambda \Delta)^{s_j} \\ &\propto \prod_{j=1}^{J} \rho_j^{s_j + \alpha_0 - 1} e^{-\rho_j \lambda n \Delta} \end{aligned}$$

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

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

where

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

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

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

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

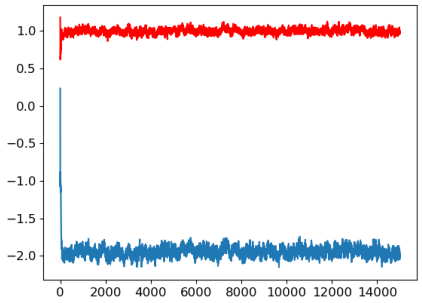
$$\int \exp(-\frac{\tau}{2} D(\mu)) d\mu \propto e^{-\frac{\tau R}{2}} \int \exp \left(-\frac{\tau}{2} (\mu - P^{-1} q)^T P (\mu - P^{-1} q) \right) d\mu \tag{3.7}$$

Thus, we get that

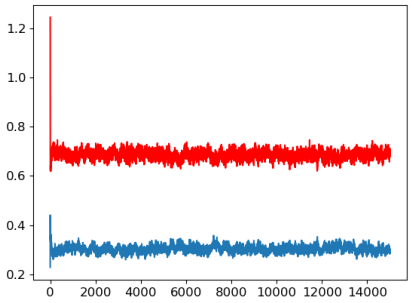
$$\begin{aligned}
p(\tau | z, a) &= \int p(\tau, \mu | z, a) d\mu \\
&\propto \tau^{\eta - 1 + (n+J)/2} \exp(-\gamma \tau) e^{-\tau R/2} (2\pi)^{J/2} \sqrt{|\tau^{-1} P^{-1}|} \exp \left(\frac{1}{2} \tau q^T P^{-1} q \right) \\
&\propto \tau^{\eta + (n+J)/2 - 1} \exp \left\{ -\tau \left(\gamma + \frac{1}{2}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

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

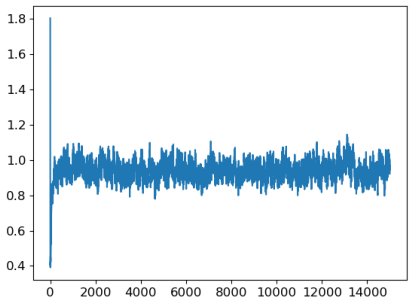
3.2.3 Simulation Results



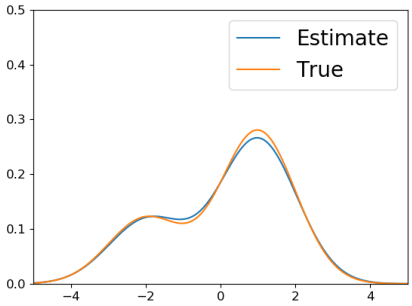
(a) Parameter μ



(b) Parameter ρ



(c) Parameter τ



(d) Density Estimate

Figure 3.2.1: Run of MCMC for mixture of two Gaussians, $\lambda = 1, \Delta = 1$

3.3 Non-Parametric Estimation via Dirichlet Process Mixture Model

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 of probability density functions that our jump density comes from.

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.

3.3.1 Dirichlet Processes

Definition 3.3.1. (Random Measure). Let $(\Omega, \mathcal{F}, \mathcal{P})$ be a 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 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 3.3.2. (Dirichlet Process). 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 $A_1, \dots, A_k$ of $\mathfrak{X}$, we have that the joint distribution of random variables $P(A_1), \dots, P(A_k)$ satisfy

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

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

Countable Dirichlet Process. 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

$$S_\infty = \left\{ s = (s_1, s_2, \dots) : s_j \geq 0, j \in \mathbb{N}, \sum_{j=1}^{\infty} s_j = 1 \right\}. \tag{3.8}$$

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 3.3.1. *Let $(S_\infty, \mathfrak{S}_\infty)$ be the measurable space defined in* (3.8) *let $(\Omega, \mathcal{F}, \mathcal{P})$ be some arbitrary probability space. Then a map $p : \Omega \to S_\infty$ is a random element if and only if every coordinate p_i is a random variable.*

In other words, Proposition 3.3.1 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 3.3.2, we see that random element $p = (p_1, p_2, \dots) \sim \text{DP}(\alpha)$ if for every $k \in \mathbb{N}$,

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

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

$$p_j = V_j \prod_{l=1}^{j-1} (1 - V_l) \tag{3.9}$$

Proposition 3.3.2. *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 (3.9), 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.

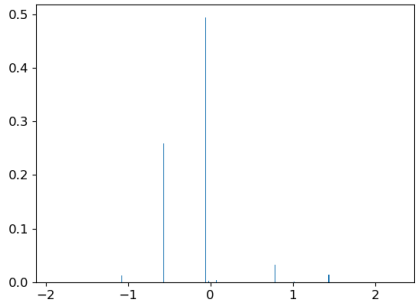


Figure 3.3.1: Realisation of Dirichlet process with $\alpha = \mathcal{N}(0, 1), M = 1$.

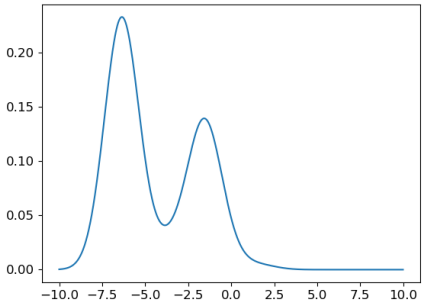


Figure 3.3.2: Realisation of Dirichlet process mixture.

From Proposition 3.3.2 and as we can see in Figure 3.3.1, 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 $F \sim \text{DP}(\alpha)$. 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 3.3.2, 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 3.3.2, 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.

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 try to infer it from our observations.

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). Then we have the model:

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

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

(Picture of diagram)

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 $K = \log n$.

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 $\beta^{(0)}, \mu^{(0)}, \tau^{(0)}, a^{(0)}$;
for $j = 1, \ldots, K$ **do**
  Set $\rho_j^{(0)} \leftarrow \beta_j^{(0)} \prod_{k=1}^{j-1} (1 - \beta_k^{(0)})$
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 β, τ, μ:
    1. Sample $\beta^{(t)} \sim p(\beta|a^{(t)}, Z)$ via Metropolis-Hastings step
    2. Sample $\tau^{(t)} \sim p(\tau|a^{(t)}, Z)$
    3. Sample $\mu^{(t)} \sim p(\mu|a^{(t)}, \tau^{(t)}, Z)$
  Update parameter ρ:
    1. Set $\rho_j^{(t)} \leftarrow \beta_j^{(t)} \prod_{k=1}^{j-1} (1 - \beta_k^{(t)})$
end

Updating Parameters. 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 3.2.1. However, to update β we use a Metropolis-Hastings step using we cannot write down its conditional distribution in a 'nice' form.

We want to sample from $p(\beta|a) \propto p(a|\beta)p(\beta)$. Note that

$$\begin{aligned}
p(a|\beta)p(\beta) &\propto p(a|\rho)p(\beta) \\
&\propto \left(\prod_{i=1}^{n} \prod_{j=1}^{K} p(a_{ij}|\rho) \right) \prod_{j=1}^{K} (1 - \beta_j)^{\alpha - 1} \\
&\propto \prod_{j=1}^{K} \left((1 - \beta_j)^{\alpha - 1} \prod_{i=1}^{n} e^{-\lambda \Delta \rho_j} \rho_j^{a_{ij}} \right) \\
&= \prod_{j=1}^{K} (1 - \beta_j)^{\alpha - 1} e^{-n\lambda \Delta \rho_j} \rho_j^{s_j} && (s_j \text{ defined in (3.6)}) \\
&= \prod_{j=1}^{K} \beta_j^{s_j} (1 - \beta_j)^{\alpha - 1} e^{-n\lambda \Delta \beta_j \prod_{k=1}^{j-1} (1 - \beta_k)} \prod_{k=1}^{j-1} (1 - \beta_k)^{s_j}
\end{aligned}$$

We can decompose this into

$$\prod_{j=1}^{K} \nu_j \exp \left\{ -n\lambda \Delta \beta_j \prod_{k=1}^{j-1} (1 - \beta_k) \right\}$$

, where

$$\begin{aligned} v_1 &= \beta_1^{s_1}(1-\beta_1)^{\alpha-1+\sum_{k=2}^K s_k} \\ v_2 &= \beta_2^{s_2}(1-\beta_2)^{\alpha-1+\sum_{k=3}^K s_k} \\ \vdots &\quad \vdots \\ v_{K-1} &= \beta_{K-1}^{s_{k-1}}(1-\beta_{K-1})^{\alpha-1+s_K} \\ v_K &= \beta_K^{s_K}(1-\beta_K)^{\alpha-1} \end{aligned}$$

Therefore, for $k = 1, \ldots, K$ we propose

$$\begin{aligned} \beta_1^\circ &\sim \text{Beta}(s_1+1, \textstyle\sum_{k=2}^K s_k + \alpha) \\ \beta_2^\circ &\sim \text{Beta}(s_2+1, \textstyle\sum_{k=3}^K s_k + \alpha) \\ \vdots &\quad \vdots \\ \beta_{K-1}^\circ &\sim \text{Beta}(s_{K-1}+1, s_K + \alpha) \\ \beta_K^\circ &\sim \text{Beta}(s_K+1, \alpha) \end{aligned}$$

independently. Thus, our acceptance probability A is

$$\begin{aligned} A &= \frac{p(\beta^\circ|a)q(\beta|a)}{p(\beta|a)q(\beta^\circ|a)} \\ &= \prod_{j=1}^K \exp\left\{-n\lambda\Delta\beta_j^\circ \prod_{k=1}^{j-1}(1-\beta_k^\circ) + n\lambda\Delta\beta_j \prod_{k=1}^{j-1}(1-\beta_k)\right\} \end{aligned}$$

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

3.3.3 Simulation Results

Chapter 4

Comparison of Estimators

Bibliography

[1] Fabienne Comte, Cline Duval, and Valentine Genon-Catalot. Nonparametric density estimation in compound poisson process using convolution power estimators. *Metrika*, 77, 01 2014. doi: 10.1007/s00184-013-0475-3.

[2] Cline Duval. Density estimation for compound poisson processes from discrete data. *Stochastic Processes and their Applications*, 123(11):3963–3986, 2013. URL `https://doi.org/10.1016/j.spa.2013.06.006`.

[3] Shota Gugushvili, Frank van der Meulen, and Peter Spreij. A non-parametric bayesian approach to decompounding from high frequency data. *Statistical Inference for Stochastic Processes*, 21(1):53–79, Apr 2018. URL `https://doi.org/10.1007/s11203-016-9153-1`.

[4] Bert van Es, Shota Gugushvili, and Peter Spreij. A kernel type nonparametric density estimator for decompounding. *Bernoulli*, 13(3):672–694, 08 2007. doi: 10.3150/07-BEJ6091. URL `https://doi.org/10.3150/07-BEJ6091`.