A non-parametric Bayesian approach to decompounding from high frequency data

Shota Gugushvili  Frank van der Meulen  Peter Spreij

Workshop on Analysis, Geometry and Probability (and Statistics)

Ulm, October 1, 2015
Outline

1. Introduction
2. Non-parametric Bayes
3. Main result
4. Simulations
5. Proof
Compound Poisson process

- $N = (N_t, t \geq 0)$ is a Poisson process with a constant intensity $\lambda > 0$.
- $\{Y_j\}$ is a sequence of i.i.d. random variables, independent of $N$ and having a common distribution function $R$ with density $r$.
- A compound Poisson process (CPP) $X = (X_t, t \geq 0)$ is defined as
  $$X_t = \sum_{j=1}^{N_t} Y_j.$$  
- CPPs are basic models e.g. in risk theory and queueing.
Sample path
The ‘true’ parameters: $\lambda_0$ and $r_0$.

Observations: a discrete time sample $X_\Delta, \ldots, X_{n\Delta}$ is available, where $\Delta > 0$ is a sampling mesh.

Problem: (non-parametric) estimation of $\lambda_0$ and $r_0$. Recovering $\lambda_0$ and $r_0$ from the observations $X_{i\Delta}$’s is called decompounding.


More general context of Lévy processes: Comte and Genon-Catalot (2010), Comte and Genon-Catalot (2011), Neumann and Reiß (2009), and others.
Discrete observations

$Y_t$ vs $t$
Equivalent problem

The random variables $Z_i^\Delta = X_i \Delta - X_{(i-1)\Delta}$, $1 \leq i \leq n$, are i.i.d. Each $Z_i^\Delta$ is distributed as

$$Z^\Delta = \sum_{j=1}^{T^\Delta} Y_j,$$

where $T^\Delta$ is independent of the sequence $\{Y_j\}$ and has a Poisson distribution with parameter $\Delta \lambda$.

Equivalent problem: Estimate $\lambda_0$ and $r_0$ based on the sample $Z_n^\Delta = (Z_1^\Delta, Z_2^\Delta, \ldots, Z_n^\Delta)$. 
Introduction

Non-parametric Bayes

Main result

Simulations

Proof
To be or not to be?
The problem seems quite tough to me!
But don’t despair, my poor soul,
I’ve found the way to reach the goal

By thinking hard for many days
And pain depicted on my face:
I’ll use the rule by Thomas Bayes
To elegantly close the case!

Shota Gugushvili, September 2015
Non-parametric Bayes

- Non-parametric Bayesian approach to inference for Lévy processes has been considered in the literature only in Gugushvili et al. (2015a) and Gugushvili et al. (2015b).
- Advantages: automatic quantification of uncertainty in parameter estimates through Bayesian posterior credible sets; (Bayesian) adaptation.
- One may also think of a non-parametric Bayes approach as a means for obtaining a frequentist estimator.
Bayes’ theorem combines the likelihood and prior into the posterior. We start with the likelihood.

The law $Q_{\lambda,r}^{\Delta}$ of $Z_i^{\Delta}$ is not absolutely continuous with respect to the Lebesgue measure.

A specific choice of the dominating measure for $Q_{\lambda,r}^{\Delta}$ is not essential for inferential conclusions, but a clever choice may greatly simplify the theoretical analysis.
Jump measure

- For
  \[ B \in \mathcal{B}([0, \Delta]) \otimes \mathcal{B}(\mathbb{R} \setminus \{0\}) \]
  define the random measure \( \mu \) by
  \[
  \mu_X(B) = \{ \# t : (t, X_t - X_{t^-}) \in B \}.
  \]

- Let \( \mathbb{R}^{\Delta}_{\lambda, r} \) be the law of \((X_t : t \in [0, \Delta])\).

- Under \( \mathbb{R}^{\Delta}_{\lambda, r} \), the random measure \( \mu_X \) is a Poisson point process on \([0, \Delta] \times (\mathbb{R} \setminus \{0\})\) with intensity measure \( \Lambda(dt, dx) = \lambda dt r(x)dx \).
Jump measure
Fix $\tilde{\lambda}$ and $\tilde{r}$.

Provided $\lambda, \tilde{\lambda} > 0$, and $\tilde{r} > 0$,

$$
\frac{d\mathbb{R}_\Delta^{\lambda,r}}{d\mathbb{R}_\Delta^{\tilde{\lambda},\tilde{r}}}(X) = \exp \left( \int_0^\Delta \int_{\mathbb{R}} \log \left( \frac{\lambda r(x)}{\tilde{\lambda} \tilde{r}(x)} \right) \mu_X(dt, dx) - \Delta (\lambda - \tilde{\lambda}) \right).
$$
From the continuous to the discrete case

- The density $k_{\lambda, r}^\Delta$ of $Q_{\lambda, r}^\Delta$ with respect to $Q_{\tilde{\lambda}, \tilde{r}}^\Delta$ is given by the conditional expectation

$$k_{\lambda, r}^\Delta(x) = \mathbb{E}_{\tilde{\lambda}, \tilde{r}} \left( \frac{dR_{\lambda, r}^\Delta}{dR_{\tilde{\lambda}, \tilde{r}}^\Delta} (X) \middle| X_\Delta = x \right),$$

where the subscript in the conditional expectation operator signifies the fact that it is evaluated under $R_{\tilde{\lambda}, \tilde{r}}^\Delta$.

- The likelihood (in the parameter pair $(\lambda, r)$) associated with the sample $Z_n^\Delta$ is given by

$$L_n^\Delta(\lambda, r) = \prod_{i=1}^n k_{\lambda, r}^\Delta(Z_i^\Delta).$$
We will use the product prior $\Pi = \Pi_1 \times \Pi_2$ for $(\lambda_0, r_0)$.

The prior $\Pi_1$ for $\lambda_0$ will be assumed to be supported on the interval $[\underline{\lambda}, \bar{\lambda}]$ and to possess a density $\pi_1$ with respect to the Lebesgue measure.

The prior for $r_0$ will be specified as a Dirichlet process mixture of normal densities.
Digression

- Suppose we have an i.i.d. sample $V_1, \ldots, V_n \sim F$.
- An excellent nonparametric estimator of $F$ is the empirical distribution function:

$$
\hat{F}_n(x) = \frac{1}{n} \sum_{j=1}^{n} 1[V_j \leq x].
$$

- If we instead want to estimate the density $f$ of $F$, we can smooth $\hat{F}_n$ with a kernel $W$ to obtain a kernel estimator

$$
\hat{f}_{nh}(x) = \frac{1}{h} \int W \left( \frac{x - y}{h} \right) d\hat{F}_n(y) = \frac{1}{nh} \sum_{j=1}^{n} W \left( \frac{x - V_i}{h} \right).
$$

Here $h > 0$ is a smoothing parameter. Its choice is critical for a good statistical performance of $\hat{f}_{nh}$.
Dirichlet process prior

- Dirichlet process prior is a non-parametric prior on the set of distribution functions.
- Let \( \alpha \) be a finite measure on \( \mathbb{R} \) and let \( D_\alpha \) denote the Dirichlet process distribution with base measure \( \alpha \).
- By definition, if \( F \sim D_\alpha \), then for any Borel-measurable partition \( B_1, \ldots, B_k \) of \( \mathbb{R}^d \) the distribution of the vector \( (F(B_1), \ldots, F(B_k)) \) is the \( k \)-dimensional Dirichlet distribution with parameters \( \alpha(B_1), \ldots, \alpha(B_k) \).
- Dirichlet distribution has a density defined on the \( k \)-dimensional unit simplex \( \mathbb{S}_k \) that is proportional to \( x_1^{\alpha_1-1} \cdots x_k^{\alpha_k-1} \). It is a multivariate generalisation of the Beta distribution.
- Dirichlet prior has attractive properties: conjugacy, large topological support.
Dirichlet process location mixture of normals

- Introduce a convolution density

\[ r_{H,\sigma}(x) = \int \phi_{\sigma}(x - z)H(\,dz), \]

where \( H \) is a distribution function on \( \mathbb{R} \), \( \sigma > 0 \), and \( \phi_{\sigma} \) denotes the density of the centred normal distribution with variance \( \sigma^2 \).

- The Dirichlet process location mixture of normals prior \( \Pi_2 \) is obtained as the law of the random function \( r_{H,\sigma} \), where \( H \sim \mathcal{D}_\alpha \), and \( \sigma \sim G \) for some prior distribution function \( G \).

- In a rough sense the Dirichlet mixture prior resembles the kernel estimation approach (realisations of the Dirichlet process are discrete distributions with infinite number of atoms). But it is more ‘clever’ than a ‘naive’ kernel estimator.
By Bayes’ theorem, the posterior measure of any measurable set $A$ is given by

$$
\Pi(A|\mathcal{Z}_n^\Delta) = \frac{\iint_A L_n^\Delta(\lambda, r)d\Pi_1(\lambda)d\Pi_2(r)}{\iint L_n^\Delta(\lambda, r)d\Pi_1(\lambda)d\Pi_2(r)}.
$$
Our main result concerns study of asymptotic frequentist properties of Bayesian procedures.

We will establish the posterior contraction rate in a suitable metric around the true parameter pair \((\lambda_0, r_0)\). This will provide a frequentist justification of our approach.

The result also implies existence of Bayes point estimates converging in the frequentist sense to the true parameter pair \((\lambda_0, r_0)\) with the same rate.
Parametric example

Let $X_1, \ldots, X_n \sim \text{Bernoulli}(p)$. Take a uniform prior on $p$. Then

$$p|X_1, \ldots, X_n \sim \text{Beta} \left( \sum_{i=1}^{n} X_i + 1, n - \sum_{i=1}^{n} X_i + 1 \right).$$

Posterior mean and variance are

$$\frac{\sum_{i=1}^{n} X_i + 1}{n + 2}, \quad \frac{\left( \sum_{i=1}^{n} X_i + 1 \right) \left( n - \sum_{i=1}^{n} X_i + 1 \right)}{(n + 2)^2 (n + 3)}.$$

Let us view the posterior under the true parameter value $p_0$. As $n \to \infty$, the posterior puts more and more mass around the true parameter: it concentrates on the neighbourhoods of radius proportional to $\sqrt{n}$.

Bernstein-von Mises theorem (discovered by Laplace).
The Hellinger distance $h(Q_0, Q_1)$ between two probability laws $Q_0$ and $Q_1$ on a measurable space $(\Omega, \mathcal{F})$ is given by

$$h(Q_0, Q_1) = \left( \int \left( \frac{dQ_0}{2} - \frac{dQ_1}{2} \right)^2 \right)^{1/2}.$$

Define the complements of the Hellinger-type neighbourhoods of $(\lambda_0, r_0)$ by

$$A(\varepsilon_n, M) = \left\{ (\lambda, r) : \frac{1}{\sqrt{\Delta}} h(Q_{\lambda_0, r_0}^\Delta, Q_{\lambda, r}^\Delta) > M\varepsilon_n \right\},$$

where $\{\varepsilon_n\}$ is a sequence of positive numbers.

We say that $\varepsilon_n$ is a posterior contraction rate, if there exists a constant $M > 0$, such that

$$\Pi(A(\varepsilon_n, M)|Z_n^\Delta) \to 0$$

in $Q_{\lambda_0, r_0}^\Delta$-probability as $n \to \infty$. 
Lemma

The following holds:

\[
\lim_{\Delta \to 0} \frac{1}{\Delta} h^2(Q^\Delta_{\lambda,r}, Q^\Delta_{\lambda_0,r_0}) = h^2(\lambda r, \lambda_0 r_0)
= \int (\sqrt{\lambda r(x)} - \sqrt{\lambda_0 r_0(x)})^2 \, dx.
\]
Assumptions

Assumption

(i) \( \lambda_0 \) is in a compact set \([\underline{\lambda}, \bar{\lambda}]\) \( \subset (0, \infty) \);

(ii) The true density \( r_0 \) is a location mixture of normal densities, i.e.

\[
r_0(x) = r_{H_0, \sigma_0}(x) = \int \phi_{\sigma_0}(x - z) dH_0(z)
\]

for some fixed distribution \( H_0 \) and a constant \( \sigma_0 \in [\underline{\sigma}, \bar{\sigma}] \subset (0, \infty) \). Furthermore, for some \( 0 < \kappa_0 < \infty \), \( H_0[-\kappa_0, \kappa_0] = 1 \), i.e. \( H_0 \) has compact support.
The prior on $\lambda$, $\Pi_1$, has a density $\pi_1$ (with respect to the Lebesgue measure) that is supported on the finite interval $[\lambda, \bar{\lambda}] \subset (0, \infty)$ and is such that

$$0 < \underline{\pi}_1 \leq \pi_1(\lambda) \leq \bar{\pi}_1 < \infty, \quad \lambda \in [\lambda, \bar{\lambda}]$$

for some constants $\underline{\pi}_1$ and $\bar{\pi}_1$. 

Assumption on $\Pi_2$

**Assumption**

1. The base measure $\alpha$ of the Dirichlet process prior $D_\alpha$ has a continuous density on an interval $[-\kappa_0 - \zeta, \kappa_0 + \zeta]$, with $\kappa_0$ as in Assumption 1 (ii), for some $\zeta > 0$, is bounded away from zero there, and for all $t > 0$ satisfies the tail condition

$$\alpha(|z| > t) \lesssim e^{-b|t|^\delta}$$

with some constants $b > 0$ and $\delta > 0$;

2. The prior on $\sigma$, $\Pi_3$, is supported on the interval $[\underline{\sigma}, \overline{\sigma}] \subset (0, \infty)$ and is such that its density $\pi_3$ with respect to the Lebesgue measure satisfies

$$0 < \underline{\pi}_3 \leq \pi_3(\sigma) \leq \overline{\pi}_3 < \infty, \quad \sigma \in [\underline{\sigma}, \overline{\sigma}]$$

for some constants $\underline{\pi}_3$ and $\overline{\pi}_3$. 
Result

Theorem

Under previous assumptions, provided $n\Delta \to \infty$, there exists a constant $M > 0$, such that for

$$
\varepsilon_n = \frac{\log^{\kappa}(n\Delta)}{\sqrt{n\Delta}}, \quad \kappa = \max \left( \frac{2}{\delta}, \frac{1}{2} \right) + \frac{1}{2},
$$

we have

$$
\Pi \left( A(\varepsilon_n, M) \bigg| Z_n^\Delta \right) \to 0
$$

in $\mathbb{Q}_{\lambda_0, f_0}$-probability as $n \to \infty$. 
Discussion

- For fixed $\Delta$ (w.l.o.g. one may then assume $\Delta = 1$) the posterior contraction rate in Theorem 1 reduces to
  $$\varepsilon_n = \frac{\log^\kappa(n)}{\sqrt{n}}.$$  

- The posterior contraction rate is controlled by the parameter $\delta$. The stronger the decay rate in (1), the better the contraction rate, but all $\delta \geq 4$ give the same value $\kappa = 1$.

- Our theorem implies existence of Bayesian point estimates with the same frequentist convergence rates.

- The (frequentist) minimax convergence rate for estimation of $(\lambda_0, r_0)$, but some existing analogies suggest that up to a logarithmic factor it should be of order $\sqrt{n \Delta}$.

- Our result generalises to multivariate processes and Hölder smooth jump densities $r$. 
Implementation

- Implementation is work in progress.
- A fully non-parametric implementation appears quite a difficult task.
- Instead we concentrated on the case when the density $r$ is a discrete location mixture of normals.
- Keywords: MCMC, data augmentation, auxiliary variables.
- Interesting interplay between $\lambda$, $\Delta$, $n$ and $r$. 
Example 1

- Setup: $\Delta = 1$, $\lambda = 1$, $r = 0.8N(2, 1) + 0.2N(-1, 1)$, $n = 5000$.
- True density and posterior mean based on 15000 MCMC samples (the first 5000 are treated as burn-in).
Example 2

- Setup: $\Delta = 1$, $\lambda = 3$, $r = 0.8N(2, 1) + 0.2N(-1, 1)$, $n = 5000$.
- True density and posterior mean based on 25000 MCMC samples (the first 10000 are treated as burn-in).
Example 1: another plot
Example 2: another plot
Roadmap

- General results on posterior contraction rates (Ghosal et al (2000), Ghosal and van der Vaart (2001) and others) are not directly applicable in our case.
- However, key insights from the proofs are still valid with suitable modifications.
We start with the decomposition

\[ \Pi(A(\varepsilon_n, M) | Z_n^\Delta) = \Pi(A(\varepsilon_n, M) | Z_n^\Delta) \phi_n + \Pi(A(\varepsilon_n, M) | Z_n^\Delta)(1 - \phi_n), \]

where \( \phi_n \) is a sequence of tests.

The idea is to show that the terms on the right-hand side separately converge to zero in probability.

The tests \( \phi_n \) allow one to control the behaviour of the likelihood ratio

\[ \mathcal{L}_n^\Delta(\lambda, f) = \prod_{i=1}^{n} \frac{k^\Delta_{\lambda, f}(Z_i^\Delta)}{k^\Delta_{\lambda_0, f_0}(Z_i^\Delta)} \]

on the set where it is not well-behaved due to the fact that \((\lambda, f)\) is ‘far away’ from \((\lambda_0, f_0)\).
Tests

Sophisticated arguments show existence of tests, such that for any $\varepsilon > \varepsilon_n$,

$$
\mathbb{E}_{\lambda_0, f_0}[\phi_n] \leq 2 \exp \left( - (K M^2 - c_1) n \Delta \varepsilon_n^2 \right),
$$

$$
\sup \left\{ \mathbb{E}_{\lambda, f}[1 - \phi_n] \leq \exp \left( - K n \Delta M^2 \varepsilon_n^2 \right), \quad \right\}
$$

where $K > 0$ is a universal constant.
We have

\[
\mathbb{E}_{\lambda_0, f_0}[\prod(A(\varepsilon_n, M) | Z_n^\Delta) \phi_n] 
\leq \mathbb{E}_{\lambda_0, f_0}[\phi_n] \leq 2 \exp \left( - (KM^2 - c_1) n \Delta \varepsilon_n^2 \right),
\]

and so by the Chebyshev (Markov?) inequality \( \prod(A(\varepsilon_n, M) | Z_n^\Delta) \phi_n \) goes to zero in probability.
The second term is
\[
\frac{\int_{A(\varepsilon_n, M)} \mathcal{L}^n_{\triangle}(\lambda, f) d\Pi_1(\lambda) d\Pi_2(f)(1 - \phi_n)}{\int \mathcal{L}^n_{\triangle}(\lambda, f) d\Pi_1(\lambda) d\Pi_2(f)}.
\]

We will show that the numerator goes exponentially fast to zero in \( \mathbb{Q}^\Delta_{\lambda_0, f_0} \)-probability, while the denominator is bounded from below by an exponential function, with \( \mathbb{Q}^\Delta_{\lambda_0, f_0} \)-probability tending to one, in such a way that the ratio still goes to zero in \( \mathbb{Q}^\Delta_{\lambda_0, f_0} \)-probability.
By Fubini’s theorem one can show that

\[ \mathbb{E}_{\lambda_0, f_0}[\text{Num}_n] \leq \Pi(Q_n^c) + \iint_{Q_n \cap A(\varepsilon_n, M)} \mathbb{E}_{\lambda, f}[1 - \phi_n] d\Pi_1(\lambda) d\Pi_2(f). \]

The second term is bounded by \( \exp(-KM^2 n \Delta \varepsilon_n^2) \).
Furthermore,

\[ \Pi(Q_n^c) = \Pi_2(H[-a_n, a_n] < 1 - \eta_n, \sigma \in [\sigma, \bar{\sigma}]) \lesssim \frac{1}{\eta_n} e^{-b\alpha_n^2}. \]

Hence

\[ \mathbb{E}_{\lambda_0, f_0}[\text{Num}_n] \lesssim \frac{1}{\eta_n} e^{-b\alpha_n^2} + \exp(-KM^2 n \Delta \varepsilon_n^2). \]
Denominator

- Let

$$\tilde{\varepsilon}_n = \frac{\log(n\Delta)}{\sqrt{n\Delta}}.$$ 

- It can be shown that with $\mathbb{P}_{\lambda_0, f_0}^{\Delta, n}$-probability tending to one as $n \to \infty$, for any constant $C > 0$ we have

$$\text{Denom}_n > \exp \left( -(1 + C)n\Delta \tilde{\varepsilon}_n^2 - \bar{c} \log^2 \left( \frac{1}{\tilde{\varepsilon}_n} \right) \right).$$
The proof is completed by combining the previous bounds: provided the constant $M > 0$ is chosen large enough, the Chebyshev inequality yields the result.