Stat Comput (2018) 28:563–577
https://doi.org/10.1007/s11222-017-9748-4

Nonparametric estimation for compound Poisson process via variational analysis on measures

Alexey Lindo[1] · **Sergei Zuyev**[2] · **Serik Sagitov**[2]

Received: 2 February 2016 / Accepted: 17 April 2017 / Published online: 19 April 2017
© The Author(s) 2017. This article is an open access publication

Abstract The paper develops new methods of nonparametric estimation of a compound Poisson process. Our key estimator for the compounding (jump) measure is based on series decomposition of functionals of a measure and relies on the steepest descent technique. Our simulation studies for various examples of such measures demonstrate flexibility of our methods. They are particularly suited for discrete jump distributions, not necessarily concentrated on a grid nor on the positive or negative semi-axis. Our estimators also applicable for continuous jump distributions with an additional smoothing step.

Keywords Compound Poisson distribution · Decompounding · Measure optimisation · Gradient methods · Steepest descent algorithms

Mathematics Subject Classification Primary: 62G05; Secondary: 62M05 · 65C60

✉ Sergei Zuyev
sergei.zuyev@chalmers.se

Alexey Lindo
alexey.lindo@glasgow.ac.uk

Serik Sagitov
serik@chalmers.se

[1] School of Mathematics and Statistics, University of Glasgow, Glasgow, UK

[2] Department of Mathematical Sciences, Chalmers University of Technology and University of Gothenburg, Gothenburg, Sweden

1 Introduction

The paper develops new methods of nonparametric estimation of the distribution of compound Poisson data. A compound Poisson process $(W_t)_{t\geq 0}$ is a Markov jump process with $W_0 = 0$ characterised by a finite *compounding* measure Λ defined on the real line $\mathbb{R} = (-\infty, +\infty)$ such that

$$\Lambda(\{0\}) = 0, \quad \|\Lambda\| := \Lambda(\mathbb{R}) \in (0, \infty). \tag{1}$$

The jumps of this process occur at the constant rate $\|\Lambda\|$, and the jump sizes are independent random variables with a common distribution $\Lambda(\mathrm{d}x)/\|\Lambda\|$. In a more general context, the compound Poisson process is a particular case of a Lévy process with Λ being the corresponding integrable Lévy measure. Inference problems for such processes naturally arises in financial mathematics (Cont and Tankov 2003), queueing theory (Asmussen 2008), insurance (Mikosch 2009) and in many other situations modelled by compound Poisson and Lévy processes.

Suppose the compound Poisson process is observed at regularly spaced times $(W_h, W_{2h}, \ldots, W_{nh})$ for some time step $h > 0$. The consecutive increments $X_i = W_{ih} - W_{(i-1)h}$ then form a vector $(X_1, \ldots X_n)$ of independent random variables having a common compound Poisson distribution with the characteristic function

$$\varphi(\theta) = Ee^{i\theta W_h} = e^{h\psi(\theta)}, \quad \psi(\theta) = \int (e^{i\theta x} - 1)\Lambda(\mathrm{d}x). \tag{2}$$

Here and below the integrals are taken over the whole $\mathbb{R}$ unless specified otherwise. Estimation of the measure Λ in

terms of a sample $(X_1, \dots X_n)$ is usually called *decompounding* which is the main object of study in this paper.

We propose a combination of two nonparametric methods which we call characteristic function fitting (ChF) and convolution fitting (CoF). ChF may deal with a more general class of Lévy processes, while CoF explicitly targets the compound Poisson processes.

The ChF estimator for the jump measure Λ is obtained by minimisation of the loss functional

$$L_{\mathrm{ChF}}(\Lambda) = \int |e^{h\psi(\theta)} - \hat{\varphi}_n(\theta)|^2 \omega(\theta) \mathrm{d}\theta, \tag{3}$$

where $\psi(\theta) \equiv \psi(\theta, \Lambda)$ is given by (2),

$$\hat{\varphi}_n(\theta) = \frac{1}{n} \sum_{k=1}^{n} e^{i\theta X_k}.$$

is the empirical characteristic function and $\omega(\theta)$ is a weight function. It was shown in Neumann and Reiss (2009) in a more general Lévy process setting that minimising (3) leads to a consistent estimator of the Lévy triplet. Typically, $\omega(\theta)$ is a positive constant for $\theta \in [\theta_1, \theta_2]$ and zero otherwise, but it can also be chosen to grow as $\theta \to 0$, this would lead to boosting an agreement of the moments of a fitted jump distribution with the empirical moments.

We compute explicitly the derivative of the loss functional (3) with respect to the measure Λ, formula (18) in "Appendix", and perform the steepest descent directly on the cone of non-negative measures to a local minimiser, further developing the approach by Molchanov and Zuyev (2002). It must be noted that, as a simple example reveal, the functionals based on the empirical characteristic function usually have a very irregular structure, see Fig. 1. As a result, the steepest descent often fails to attend the global optimal solution, unless the starting point of the optimisation procedure is carefully chosen.

The CoF estimation method uses the fact that the convolution of $F(x) = \mathbf{P}(W_h \le x)$,

$$F^{*2}(x) = \int F(y)F(x-y)dy,$$

as a functional of Λ has an explicit form of an infinite Taylor series involving direct products of measures Λ, see Theorem 2 in Sect. 4. After truncating it to only the first k terms, we build a loss function $L^{(k)}_{\mathrm{CoF}}$ by comparing two estimates of F^{*2}: the one based on the truncated series and the other being the empirical convolution F_n^{2*}. CoF is able to produce nearly optimal estimates $\hat{\Lambda}_k$ when large values of k are taken, but at the expense a drastically increased computation time.

A practical combination of these methods recommended by this paper is to find $\hat{\Lambda}_k$ using CoF with a low value of k and then apply ChF with $\hat{\Lambda}_k$ as the starting value. The estimate for such a two-step procedure will be denoted by $\tilde{\Lambda}_k$ in the sequel.

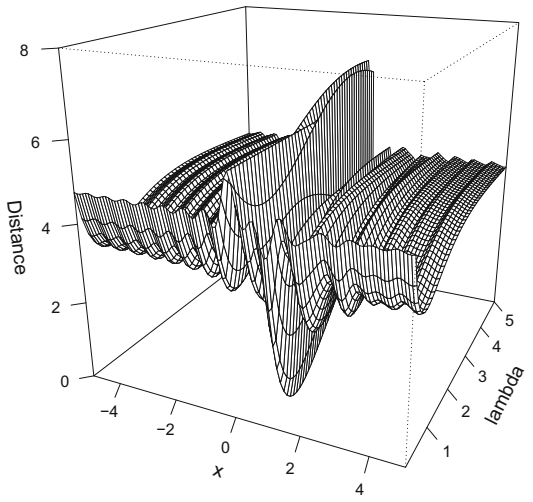

Fig. 1 Illustration of intrinsic difficulties faced by any characteristic function fitting procedure. Plotted is the integrated squared modulus of the difference between two characteristic functions with measures $\Lambda = \delta_1$ and $\Lambda' = \lambda\delta_x, x \in [-5, 5],\ \lambda \in (0, 5)$. Clearly, any algorithm based on closeness of characteristic functions, like (3), would have difficulties converging to the global minimum attained at point $x = 1, \lambda = 1$ even in this simple two-parameter model

To give an early impression of our approach, let us demonstrate the performance of our methods on the famous data by Ladislaus Bortkiewicz who collected the numbers of Prussian soldiers killed by a horse kick in 10 cavalry corps over a 20-year period (Bortkiewicz 1898). The counts 0, 1, 2, 3 and 4 were observed 109, 65, 22, 3 and 1 times, with 0.6100 deaths per year per cavalry unit. The author argues that the data are Poisson distributed which corresponds to the measure $\Lambda = \lambda\delta_1$ concentrated on the point $\{1\}$ (only jumps of size 1) and the mass λ being the parameter of the Poisson distribution estimated by the sample mean to be 0.61. Figure 2 on its top panel presents the estimated Lévy measures for the cut-off values $k = 1, 2, 3$ when using CoF method. For the values of $k = 1, 2$, the result is a measure having many atoms. This is explained by the fact that the accuracy of the convolution approximation is not enough for these data, but $k = 3$ already results in a measure $\hat{\Lambda}_3$ essentially concentrated at $\{1\}$, thus supporting the Poisson model with parameter $\|\hat{\Lambda}_3\| = 0.6098$. In Sect. 4, we return to this example and explain why the choice of $k = 3$ is reasonable here. Caused by a possibly very irregular behaviour of the score function L_{ChF} demonstrated above, we practically observed that the convergence of the ChF method depends critically on the choice of the initial measure, especially on its total mass. However, the proposed combination of CoF followed by ChF demonstrates (the bottom plot) that this two-step (faster) pro-

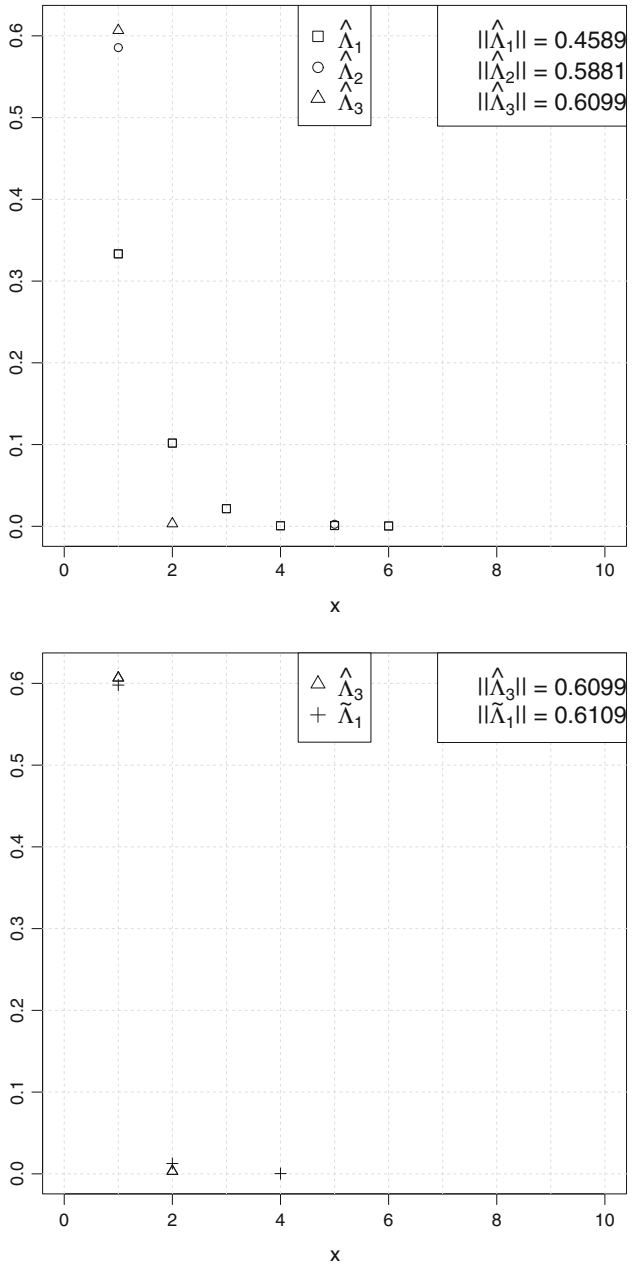

Fig. 2 The analysis of Bortkiewicz horse kick data. *Top panel* comparison of CoF estimates for $k = 1, 2, 3$. *Bottom panel* comparison of the estimate by CoF with $k = 3$ and a combination of CoF with $k = 1$ followed by ChF

cedure results in the estimate $\tilde{\Lambda}_1$, which is as good as more computationally demanding $\hat{\Lambda}_3$.

Previously developed methods include discrete decompounding approach based on the inversion of Panjer recursions as proposed in Buchmann and Grübel (2003). van Es et al. (2007) and, lately, Duval (2013), Comte et al. (2014) studied the continuous decompounding problem when the measure Λ is assumed to have a density. They apply Fourier inversion in combination with kernel smoothing techniques for estimating the unknown density of the measure Λ. In contrast, we do not distinguish between discrete and continuous Λ in that our algorithms, based on direct optimisation of functionals of a measure, work for both situations on a discretised phase space of Λ. However, if one sees many small atoms appearing in the solution, which fill a thin grid, this may indicate that the true measure is absolutely continuous and some kind of smoothing should yield its density.

In this paper, we do not address estimation of more general Lévy processes allowing for $\Lambda(-1, 1) = \infty$. In the Lévy process setting, the most straightforward approach for estimating the distribution $F(x) = \mathbf{P}(W_h \le x)$ is the moments fitting, see Feuerverger and McDunnough (1981b) and Carrasco and Florens (2000). Estimates of Λ can be obtained by maximising the likelihood ratio (see e.g. Quin and Lawless 1994) or by minimising some measure of proximity between F and the empirical distribution function $\hat{F}_n(x) = \frac{1}{n}\sum_{k=1}^{n} \mathbb{1}_{\{X_k \le x\}}$, where the dependence on Λ comes through F via the inversion formula of the characteristic function:

$$F(x) - F(x-0) = \frac{1}{2\pi} \lim_{y\to\infty} \int_{-y}^{y} \exp\{h\psi(\theta) - i\theta x\} \mathrm{d}\theta.$$

For the estimation, the characteristic function in the integral above is replaced by the empirical characteristic function.

Parametric inference procedures based on the empirical characteristic function have been known for some time, see Feuerverger and McDunnough (1981a) and Sueishi and Nishiyama (2005), and the references therein. Algorithms based on the inversion of the empirical characteristic function and on the relation between its derivatives were proposed in Watteel and Kulperger (2003). Note that the inversion of the empirical characteristic function, in contrast to the inversion of its theoretical counterpart, generally leads to a complex valued measure which needs to be dealt with.

One of the reviewers has drawn our attention to the recent preprint Coca (2015) which promises to be useful for testing the presence of discrete and/or continuous jump distribution components as well as for obtaining approximation accuracy bounds based on the central limit theorem. Practical implementations of these theoretical results are yet to be explored.

The rest of the paper has the following structure. Section 2 introduces the theoretical basis of our approach—a constraint optimisation technique in the space of measures. Section 3 provides an algorithmic implementation of the corresponding steepest descent method in R-language. Section 4 develops the necessary ingredients for the CoF method based on the main analytical result of the paper, Theorem 2. Section 5 contains a broad range of simulation results illustrating performance of our algorithms on simulated data with various compounding measures, both discrete and continuous. Section 6 presents an application of our approach to real currency exchange data. Section 7 summarises our approach

and gives some practical recommendations. We conclude by "Appendix" with proofs and explicit formulas for the gradients of the two loss functions used in our steepest descent algorithm.

2 Optimisation of functionals of a measure

In this section, we briefly present the main ingredients of the constrained optimisation of functionals of a measure. Theorem 1 gives necessary conditions for a local minimum of a strongly differentiable functional. This theorem justifies a major step in our optimisation algorithm described in Sect. 3. Further details of the underlying theory can be found in Molchanov and Zuyev (2000a, b).

Denote by $\mathbb{M}$ and $\mathbb{M}_+$ the class of signed and, respectively, non-negative measures with a finite total variation. The set $\mathbb{M}$ then becomes a Banach space with sum and multiplication by real numbers defined set-wise: $(\eta_1+\eta_2)(B) := \eta_1(B) + \eta_2(B)$ and $(t\eta)(B) := t\eta(B)$ for any Borel set B and any real t. The set $\mathbb{M}_+$ is a *pointed cone* in $\mathbb{M}$ meaning that the zero measure is in $\mathbb{M}_+$, $\mu_1+\mu_2 \in \mathbb{M}_+$, and $t\mu \in \mathbb{M}_+$ as long as $\mu_1, \mu_2, \mu \in \mathbb{M}_+$ and $t \geq 0$.

A functional $G : \mathbb{M} \mapsto \mathbb{R}$ is called *Fréchet* (or *strongly*) *differentiable* at $\eta \in \mathbb{M}$ if there exists a bounded linear operator (a *differential*) $DG(\eta)[\cdot] : \mathbb{M} \mapsto \mathbb{R}$ such that

$$G(\eta+\nu) - G(\eta) = DG(\eta)[\nu] + o(\|\nu\|), \quad \|\nu\| \to 0, \quad (4)$$

where $\|\nu\|$ is the total variation of a signed measure $\nu \in \mathbb{M}$. If for a given $\eta \in \mathbb{M}$ there exists a bounded function $\nabla G(\cdot\,;\eta) : \mathbb{R} \to \mathbb{R}$ such that

$$DG(\eta)[\nu] = \int \nabla G(x;\eta)\, \nu(\mathrm{d}x) \text{ for all } \nu \in \mathbb{M},$$

then $\nabla G(\cdot\,;\eta)$ is called the *gradient function* for G at η. Typically in applications, and it is indeed the case for the functionals of a measure considered in this paper, the gradient functions exist so that the differentials indeed have an integral form.

As a simple illustration, consider an integral of a bounded function $G(\eta) = \int f(x)\eta(dx)$. Since this is already a bounded linear functional of η, we get $\nabla G(x;\eta) = f(x)$ for any η. More generally, for a composition $G(\eta) = u(\int f(x)\eta(\mathrm{d}x))$, the gradient function can be obtained by the Chain rule:

$$\nabla G(x;\eta) = u'\Big(\int f(y)\,\eta(\mathrm{d}y)\Big)\, f(x). \quad (5)$$

The functional G in this example is strongly differentiable if both functions u' and f are bounded.

Taking into account condition (1), we aim to find a solution to the following constraint minimisation problem:

$$\Lambda = \arg\min\{L(\eta) : \eta \in \mathbb{M}_+,\ \eta(\{0\}) = 0\}, \quad (6)$$

where $L : \mathbb{M}_+ \mapsto \mathbb{R}$ is strongly differentiable functional of a measure. The following necessary condition of a minimum is proven in Appendix.

Theorem 1 *Suppose that a Λ solves (6), and the functional L possesses a gradient function $\nabla L(x;\Lambda)$ at this Λ. Then*

$$\begin{cases} \nabla L(x;\Lambda) \geq 0 \quad \text{for all } x \in \mathbb{R}\setminus\{0\}, \\ \nabla L(x;\Lambda) = 0 \quad \Lambda - \text{almost everywhere}. \end{cases} \quad (7)$$

Remark 1 It can be shown similarly that the necessary condition (7) also holds for optimisation over the class of Lévy measures satisfying, in addition to (1), the integrability condition $\int \min\{1, x^2\}\, \Lambda(\mathrm{d}x) < \infty$.

3 Steepest descent algorithm on the cone of positive measures

There is an extensive number of algorithms realising a *parametric* optimisation over a finite number of continuous variables, but optimisation algorithms over the cone of measures have been proposed only recently in Molchanov and Zuyev (2002) for the case of measures with a fixed total mass. The variation analysis of functionals of a measure outlined in the previous section allows us to develop a steepest descent type algorithm for minimisation of functionals of a compounding measure which we describe next. This algorithm has been used to obtain the simulation results presented in Sect. 5.

Recall that the principal optimisation problem has the form (6), where the functional $L(\Lambda)$ is minimised over the measures Λ subject to the constraint (1). For computational purposes, a measure $\Lambda \in \mathbb{M}_+$ is replaced by its discrete approximation which has a form of a linear combination $\boldsymbol{\Lambda} = \Sigma_{i=1}^{l} \lambda_i \delta_{x_i}$ of Dirac measures on a finite regular grid $x_1, \ldots, x_l \in \mathbb{R}$, $x_{i+1} = x_i + 2\Delta$. Specifically, for a given measure Λ, the atoms of $\boldsymbol{\Lambda}$ are given by

$$\begin{aligned} \lambda_1 &:= \Lambda((-\infty, x_1+\Delta)), \\ \lambda_i &:= \Lambda([x_i-\Delta, x_i+\Delta)), \quad \text{for } i = 2, \ldots, l-1, \\ \lambda_l &:= \Lambda([x_l-\Delta, \infty)). \end{aligned} \quad (8)$$

Clearly, the larger is l and the finer is the grid $\{x_1, \ldots, x_l\}$ the better is the approximation, however, at a higher computational cost.

Respectively, the discretised version of the gradient function $\nabla L(x;\Lambda)$ is the vector

$$\boldsymbol{g} = (g_1, \ldots, g_l), \quad g_i := \nabla L(x_i; \boldsymbol{\Lambda}), \quad i = 1, \ldots, l. \tag{9}$$

Our main optimisation algorithm has the following structure: In the master algorithm description above, the line 3

Steepest descent algorithm

Input: initial vector $\boldsymbol{\Lambda}$
1: **function** GoSteep($\boldsymbol{\Lambda}$)
2: initialise the discretised gradient
3: $\boldsymbol{g} \leftarrow (\nabla L(x_1; \boldsymbol{\Lambda}), \ldots, \nabla L(x_l; \boldsymbol{\Lambda}))$
4: **while** $\big(\min_i g_i < -\tau_2$ **or** $\max_{\{i:\,\lambda_i > \tau_1\}} g_i > \tau_2\big)$ **do**
5: choose a favourable step size ε depending on L and $\boldsymbol{\Lambda}$
6: compute new vector $\boldsymbol{\Lambda} \leftarrow$ MakeStep($\varepsilon, \boldsymbol{\Lambda}, \boldsymbol{g}$)
7: compute gradient at the new $\boldsymbol{\Lambda}$:
8: $\boldsymbol{g} \leftarrow (\nabla L(x_1; \boldsymbol{\Lambda}), \ldots, \nabla L(x_l; \boldsymbol{\Lambda}))$
9: **end while**
10: **return** $\boldsymbol{\Lambda}$
11: **end function**

uses the necessary condition (7) as a test condition for the main cycle. In the computer realisations, we usually want to discard the atoms of a negligible size: for this purpose, we use a zero-value threshold parameter τ_1. Another threshold parameter τ_2 decides when the coordinates of the gradient vector are sufficiently small to be discarded. For the examples considered in the next section, we typically used the following values: $\omega \equiv 1$, $\tau_1 = 10^{-2}$ and $\tau_2 = 10^{-6}$. The key MakeStep subroutine, mentioned on line 6, is described below. It calculates the admissible steepest direction $\boldsymbol{\nu}^*$ of size $\|\boldsymbol{\nu}^*\| \le \varepsilon$ and returns an updated vector $\boldsymbol{\Lambda} \leftarrow \boldsymbol{\Lambda} + \boldsymbol{\nu}^*$.

Algorithm for a steepest descent move

Input: maximal step size ε, current variable value $\boldsymbol{\Lambda}$ and current gradient value $\boldsymbol{g}$
function MakeStep($\varepsilon, \boldsymbol{\Lambda}, \boldsymbol{g}$)
initialise the optimal step $\boldsymbol{\nu^*} \leftarrow \boldsymbol{0}$
initialise the running coordinate $i \leftarrow 0$
initialise the total mass available $E \leftarrow \varepsilon$
while $((E > 0)$ **and** $(i \le l))$ **do**
if $g_i > |g_l|$ **then**
$\nu_i^* \leftarrow \max(-\lambda_i, -E)$
$E \leftarrow E - \nu_i^*$
else
$\nu_l^* \leftarrow E$
$E \leftarrow 0$
end if
$i \leftarrow i + 1$
end while
return $\boldsymbol{\Lambda} + \boldsymbol{\nu^*}$
end function

The MakeStep subroutine looks for a vector $\boldsymbol{\nu}^*$ which minimises the linear form $\sum_{i=1}^{l} g_i \nu_i$ appearing in the Taylor expansion

$$L(\boldsymbol{\Lambda} + \boldsymbol{\nu}) - L(\boldsymbol{\Lambda}) = \sum_{i=1}^{l} g_i \nu_i + o(|\boldsymbol{\nu}|).$$

This minimisation is subject to the following linear constraints

$$\sum_{i=1}^{l} |\nu_i| \le \varepsilon, \quad \nu_i \ge -\lambda_i, \quad i = 1, \ldots, l.$$

The just described linear programming task has a straightforward solution given below.

For simplicity, we assume that $g_1 \ge \cdots \ge g_l$. Note that this ordering can always be achieved by a permutation of the components of the vector $\boldsymbol{g}$ and respectively, $\boldsymbol{\Lambda}$. Assume also that the total mass of $\boldsymbol{\Lambda}$ is bigger than the given positive stepsize ε. Define two indices

$$i_g = \max\{i:\, g_i \ge |g_l|\}, \; i_\varepsilon = \max\{i:\, \sum_{j=1}^{i-1} \lambda_j < \varepsilon\}.$$

If $i_\varepsilon \le i_g$, then the coordinates of $\boldsymbol{\nu}^*$ are given by

$$\nu_i^* := \begin{cases} -\lambda_i & \text{for} \quad i \le i_\varepsilon, \\ \sum_{j=1}^{i_\varepsilon - 1} \lambda_j - \varepsilon & \text{for} \quad i = i_\varepsilon + 1, \\ 0 & \text{for} \quad i \ge i_\varepsilon + 2, \end{cases}$$

and if $i_\varepsilon > i_g$, then

$$\nu_i^* := \begin{cases} -\lambda_i, & \text{for} \quad i \le i_g, \\ 0 & \text{for} \quad i_g < i < l, \\ \varepsilon - \sum_{j=1}^{i_g} \lambda_j & \text{for} \quad i = l. \end{cases}$$

The presented algorithm is realised in the statistical computation environment R Core Team (2015) in the form of a library `mesop` which is freely downloadable from one of the authors' webpage.[1]

4 Description of the CoF method

As it was alluded in Introduction, the CoF method uses a representation of the convolution as a function of the compounding measure Λ. We now formulate the main theoretical result of the paper on which the CoF method is based. The proof is given in Appendix.

We will need the following notation. For a function F, denote $U_x F(y) = F(y - x) - F(x)$ and

[1] http://www.math.chalmers.se/~sergei/download.html.

$$U_{x_1,\dots,x_n}F(y) := U_{x_n}(U_{x_1,\dots,x_{n-1}}F(y)) = \sum_{J\subseteq\{1,2,\dots,n\}} (-1)^{n-|J|}F(y-\Sigma_{j\in J}x_j),$$

where the sum is taken over all the subsets J of $\{1, 2, \dots, n\}$ including the empty set. Denote $\Gamma_0(F, \Lambda, y) = F(y)$, and

$$\Gamma_i(F, \Lambda, y) = \frac{1}{i!}\int_{\mathbb{R}^i} U_{x_1,\dots,x_i}F(y)\Lambda(\mathrm{d}x_1)\dots\Lambda(\mathrm{d}x_i), \quad i \geq 1.$$

Theorem 2 *Let $(W_t)_{t\geq 0}$ be a compound Poisson process characterised by (2), and $F(y) = F_h(y)$ be the cumulative distribution function of W_h for a given positive h. Then for each real y, one has*

$$F^{*2}(y) = \sum_{i=0}^{\infty} h^i \Gamma_i(F, \Lambda, y). \tag{10}$$

Recall that the empirical convolution of a sample $(X_1, \dots, X_n)$,

$$\hat{F}_n^{*2}(y) := \frac{1}{\binom{n}{2}} \sum_{1\leq i<j\leq n} \mathbb{1}\{X_i + X_j \leq y\}. \tag{11}$$

is an unbiased and consistent estimator of $F^{*2}(x)$, see Frees (1986). The CoF method looks for a finite measure Λ that minimises the following loss function

$$L^{(k)}_{\mathrm{CoF}}(\Lambda) = \int \Big\{ \sum_{i=0}^{k} h^i \Gamma_i(\hat{F}_n, \Lambda, y) - \hat{F}_n^{*2}(y) \Big\}^2 \omega(y)\mathrm{d}y. \tag{12}$$

The infinite sum in (10) is truncated to k terms in (12) for computational reasons. The error introduced by the truncation can be accurately estimated by bounding the remainder term in the finite expansion formula (16) in the proof. Alternatively, turning to (10) and using $0 \leq F(y) \leq 1$, we obtain $\sup_{y\in\mathbb{R}} |U_{x_1,\dots,x_i}F(y)| \leq 2^{i-1}$, yielding a uniform bound

$$\sup_{y\in\mathbb{R}} \sum_{i=k+1}^{\infty} h^i \big|\Gamma_i(F, \Lambda, y)\big| \leq R_k(h\|\Lambda\|), \text{ where}$$

$$R_k(x) = \frac{1}{2}\sum_{n=k+1}^{\infty} \frac{(2x)^n}{n!}. \tag{13}$$

Thus, to have a good estimate with this method, the upper bound $R_k(h\|\Lambda\|)$ should be small, which could be achieved by reducing the time step h or/and increasing k. For instance, for the horse kick data considered in Introduction, we have $h = 1$ and the estimated value of $\|\Lambda\|$ is 0.61, giving the values $R_k(0.61) = 0.58, 0.21, 0.06$ for $k = 1, 2, 3$. This indicates that $k = 3$ is rather adequate cut-off for the data.

If the expected number of jumps, $h\|\Lambda\|$, in the time interval $[0, h]$, is large, the sample values X_i, in the case of a finite variances, would have approximately normal distribution. Since the normal distribution is determined by the first two moments only and not by the entire compounding distribution, an effective estimation of $\Lambda/\|\Lambda\|$ is hardly possible, see Duval (2014) for a related discussion. Indeed, to get the upper bound close to 0.2 given $h\|\Lambda\| = 8$, one would need to take $k = 41$ which is hardly computationally possible.

To summarise, if one has a control on the choice of h, it should be taken so that the estimated value of $h\|\Lambda\|$ is close to 1. For large values of this parameter, the central limit theorem prevents an effective estimation of λ, while the small values would result in almost always single jumps and the optimisation procesure giving basically the sample measure as a solution. Similarly to the problem of choice of a kernel estimator or the histogram's bin width, a compromise should be sought. A practical approach would be to try various values of h, as we demonstrate below in Sect. 6 on the real FX data.

5 Simulation results

To illustrate the performance of our estimation methods, we generated samples of size $n = 1000$ for compound Poisson processes driven by various kinds of measure Λ. In Sects. 5.1, 5.2 and 5.3, we considered examples of discrete jump size distributions. Note that lattice distributions with both positive and negative jumps are particularly challenging because of possible cancellations of jumps, the case barely considered in the literature so far.

In Sects. 5.4 and 5.5, we present simulation results for two cases of continuously distributed jumps: non-negative and general. The continuous measures are replaced in the simulations by their discretised versions given by (8). The grid size in these examples was $\Delta = 0.25$. Note that no special account is given to the fact that the measure is continuous and the algorithms work the same way as with genuine discrete measures. However, the presence of atoms filling the consecutive grid ranges should indicate that the true compounding measure is probably continuous. A separate analysis could be tried to formally check this hypothesis, for instance, by the methods proposed in Coca (2015). If confirmed, some kind of kernel smoothing could be used to produce an estimated density curve or specific estimation methods for continuously distributed jumps employed, like the ones mentioned in Introduction.

For all the considered examples, we applied three versions of the CoF with $h = 1, k = 1, 2, 3$ and $\omega \equiv 1$. We also apply

ChF using the estimate of CoF with $k = 1$. Observe that CoF with $k = 1$ can be made particularly fast because here we have a non-negative least squares optimisation problem. If the computation time is no concern, one can also implement CoF with higher values of k to use the resulting measure as a starting point to ChF. Given a complicated nature of the loss function, this may or may not lead to a better fit. In all these examples $\|\Lambda\| = 1$, which explains our particular choice of $h = 1$, see the discussion above in connection to the error estimate (13).

5.1 Degenerate jump measure

Consider first the simplest measure $\Lambda(dx) = \delta_1(dx)$ corresponding to a standard Poisson process with rate 1. Since all the jumps are integer valued and non-negative, it would be logical to take the non-negative integer grid for possible atom positions of the discretised Λ. This is the way we have done it for the horse kick data analysis in Introduction. However, to test the robustness of our methods, we took the grid $\{0, \pm 1/4, \pm 2/4, \ldots\}$. As a result, the estimated measures might place some mass on non-integer points or even on negative values of x to compensate for inaccurately fitted positive jumps. We have chosen to show on the graphs the discrepancies between the estimated and the true measure. An important indicator of the effectiveness of an estimation is the closeness of the total masses $\|\hat{\Lambda}\|$ and $\|\Lambda\|$. For $\Lambda = \delta_1$, the probability to have more than 3 jumps is approximately 0.02; therefore, with the CoF method we expect that $k = 3$ would give an adequate estimate for these data. Indeed, the top panel of Fig. 3 demonstrates that the CoF with $k = 3$ is much more effective in detecting the jumps of the Poisson process compared to $k = 2$ and, especially, to $k = 1$. The latter methods generate large discrepancies both in atom sizes and in the total mass of the obtained measure. Observe also the presence of artifactual small atoms at large x and even at some non-integer locations.

The bottom panel shows that a good alternative to a rather computationally demanding CoF method with $k = 3$ is a much faster combined CoF–ChF method when $\hat{\Lambda}_1$ measure is used as the initial measure in the ChF algorithm. The resulting measure $\tilde{\Lambda}_1$ is almost identical to $\hat{\Lambda}_3$, but also has the total mass closer to the target value 1. The total variation distances between the estimated measures $\hat{\Lambda}_k$ and the theoretical measure Λ are 0.435, 0.084 and 0.053 for $k = 1, 2, 3$, respectively. The best fit provides the combined CoF–ChF method which produces a measure $\tilde{\Lambda}_1$ within the distance of 0.043 from Λ.

5.2 Discrete positive and negative jumps

Consider now a jump measure $\Lambda = 0.2\delta_{-1} + 0.2\delta_1 + 0.6\delta_2$. This gives rise to a compound Poisson process with rate

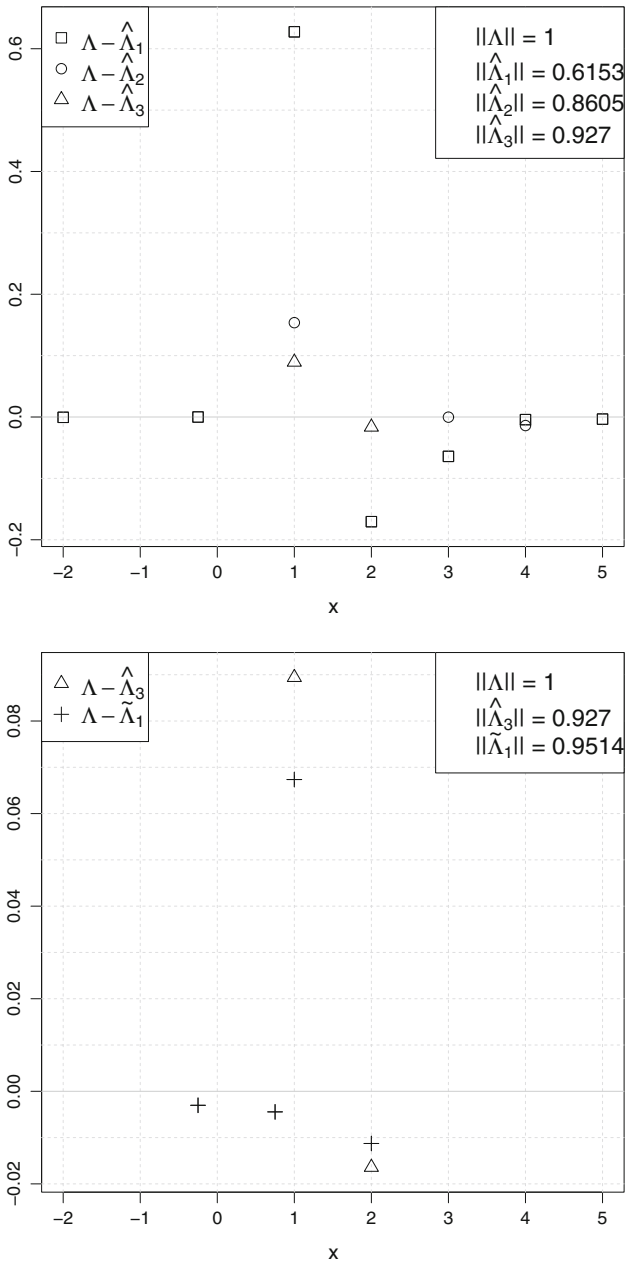

Fig. 3 Simulation results for $\Lambda = \delta_1$. *Top panel* the differences between $\Lambda(\{x\})$ and their estimates $\hat{\Lambda}_k(\{x\})$ obtained by CoF with $k = 1, 2, 3$. Zero values of the differences are not plotted. *Bottom panel* comparison of $\hat{\Lambda}_3$ with $\tilde{\Lambda}_1$ obtained by ChF initiated at $\hat{\Lambda}_1$. Notice a drastic change in the vertical axis scale as we go from the *top* to the *bottom panel*

$\|\Lambda\| = 1$ and jumps of sizes $-1, 1, 2$ having respective probabilities 0.2, 0.2 and 0.6. Figure 4 presents the results of our simulations. The presence of negative jumps cancelling positive jumps creates an additional difficulty for the estimation task. This phenomenon explains why the approximation obtained with $k = 2$ is worse than with $k = 1$ and $k = 3$: two jumps of sizes $+1$ and -1 sometimes cancel each other, which is indistinguishable from no jumps case, see the

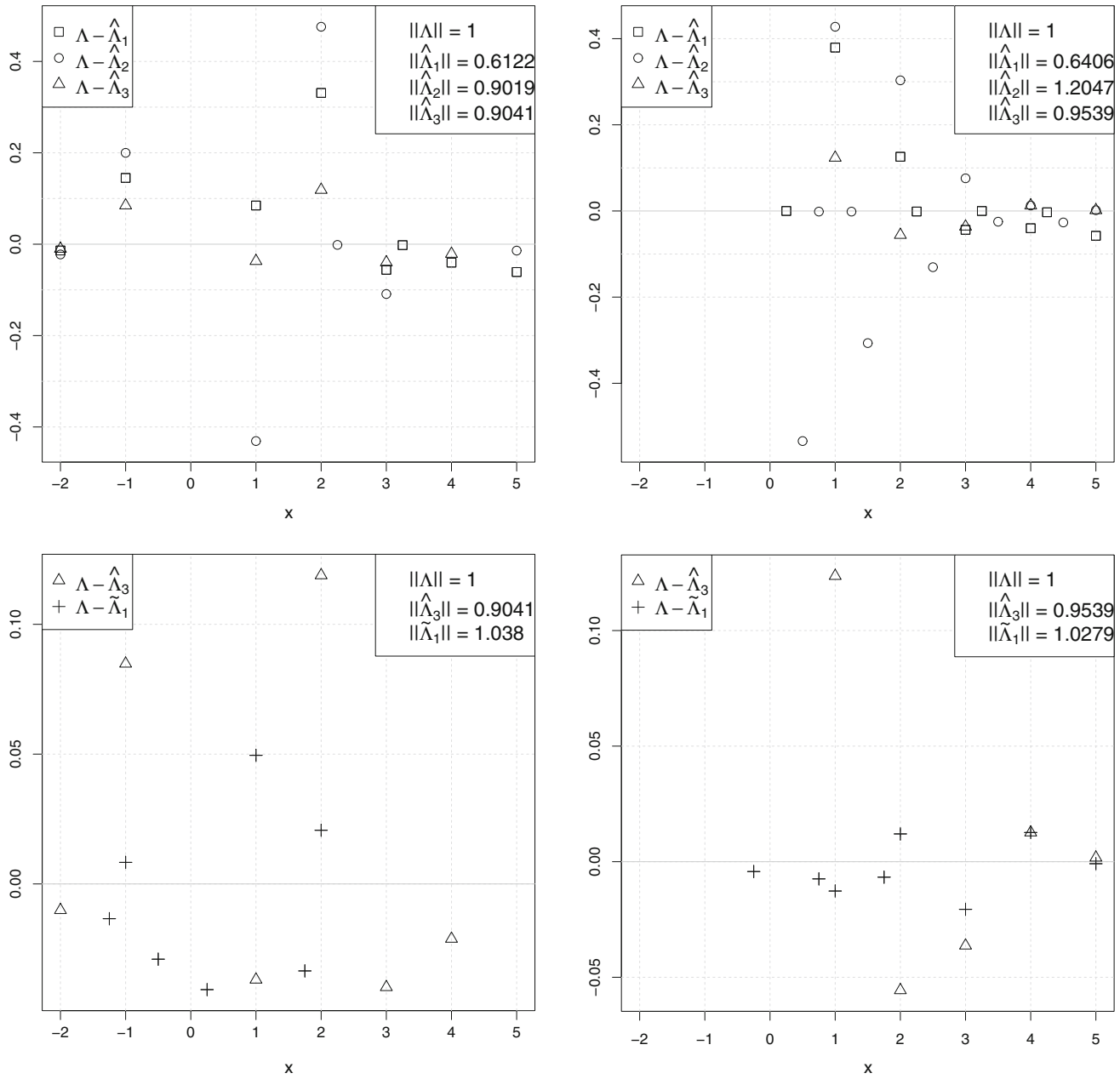

Fig. 4 Simulation results for $\Lambda = 0.2\delta_{-1} + 0.2\delta_1 + 0.6\delta_2$. *Top panel* the differences between $\Lambda(\{x\})$ and their estimates $\hat{\Lambda}_k(\{x\})$ obtained by CoF with $k = 1, 2, 3$. *Bottom panel* comparison of $\hat{\Lambda}_3$ with $\tilde{\Lambda}_1$

Fig. 5 Simulation results for a shifted Poisson distribution $\Lambda(\{x\}) = e^{-1}/(x-1)!$ for $x = 1, 2, \ldots$. *Top panel* the differences between $\Lambda(\{x\})$ and their estimates $\hat{\Lambda}_k(\{x\})$ obtained by CoF with $k = 1, 2, 3$. *Bottom panel* comparison of $\hat{\Lambda}_3$ with $\tilde{\Lambda}_1$ obtained by ChF initiated at $\hat{\Lambda}_1$

top panel of Fig. 4. Moreover, -1 and 2 added together are the same as having a single size 1 jump. The phenomenon still persists when we increased the sample size: $k = 1$ and $k = 3$ still perform better. Notice that going from $k = 1$ through $k = 2$ up to $k = 3$ improves the performance of CoF, although the computing time increases dramatically. The corresponding total variation distances of $\hat{\Lambda}_k$ to the theoretical distribution are 0.3669, 0.6268 and 0.1558. The combined method gives the distance 0.0975, and according to the bottom plot, it is again a clear winner in this case too. It is also much faster.

5.3 Unbounded compounding distribution

As an example of a measure Λ with unbounded support, we take a shifted Poisson distribution with parameter 1. Figure 5 presents our simulation results for this case; for

computation purposes, we took the interval $x \in [-2, 5]$ as the support range for the estimated measure. In practice, the support range should be enlarged if atoms start appearing on the boundaries of the chosen interval indicating a wider support of the estimated measure, see also Buchmann and Grübel (2003) for a related discussion. As the top panel reveals, also in this case the CoF method with $k = 3$ gives a better approximation than those with $k = 1$ or $k = 2$ (the total variation distance to the theoretical distribution is 0.1150 compared to 0.3256 and 0.9235, respectively) and the combined (faster) method gives an even better estimate with $d_{\mathrm{TV}}(\tilde{\Lambda}_1, \Lambda) = 0.0386$. Interestingly, the case of $k = 2$ was the worst in terms of the total variation distance to the original measure. We suspect that the 'pairing effect' may be responsible: the jumps are better fitted with a single integer- valued variable rather than with the sum of two. The algorithm may also got stuck in a local minimum producing small atoms at non-integer positions.

5.4 Continuous non-negative compounding distribution

Consider a compound Poisson process of rate 1 with the compounding distribution being exponential with parameter 1. The top plot of Fig. 6 shows that, as expected, the approximation accuracy increases with k. Observe that the total variation distance $d_{\mathrm{TV}}(\hat{\Lambda}_3, \boldsymbol{\Lambda}) = 0.0985$ is comparable with the discretisation error: $d_{\mathrm{TV}}(\Lambda, \boldsymbol{\Lambda}) = 0.075$. A Gaussian kernel smoothed version of $\hat{\Lambda}_3$ is presented at the bottom plot of Fig. 6. The visible discrepancy for small values of x is explained by the fact that there were no sufficiently many small jumps in the simulated sample for the algorithm to put more mass around 0.

Optimisation in the space of measures usually tends to produce atomic measures since these are boundary points of typical constraint sets in $\mathbb{M}$. Indeed, $\tilde{\Lambda}_1$ has smaller number of atoms than $\boldsymbol{\Lambda}$ does and still it approximates better the empirical characteristic function of the sample.

5.5 Continuous compounding distribution

Finally, Fig. 7 takes up the important example of compound Poisson processes with normally distributed jumps having both positive and negative values. Once again, the estimates $\hat{\Lambda}_k$ improve as k increases, and the combined method CoF–ChF gives an estimate similar to $\hat{\Lambda}_3$. Notice an inflection around 0 caused by the restraint on the estimated measure which imposes the origin to have a zero mass. This shows as a dip in the curve produced by the kernel smoother.

In the presented examples with continuous compounding distribution, when choosing the kernel smoother width, we were guided by a visual smoothness of the resulting curve.

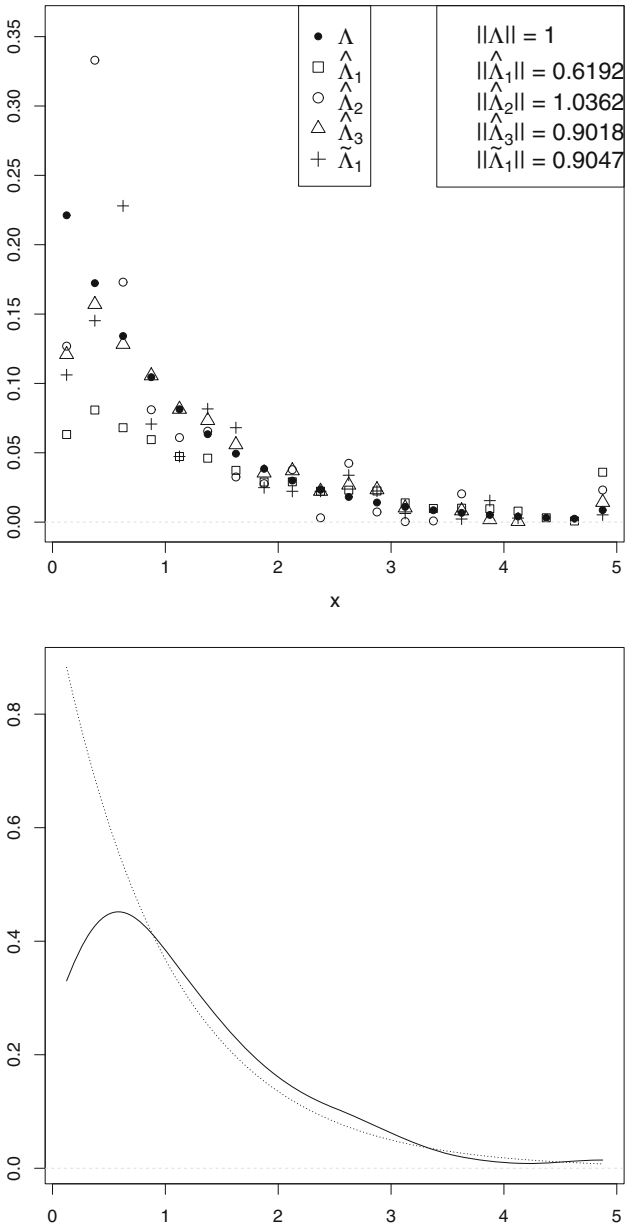

Fig. 6 Simulation results for a compound Poisson process with jump intensity 1 and jump sizes having an exponential distribution with parameter 1. *Top plot* obtained measures for various algorithms, the *bottom plot* the theoretical exponential density and the smoothed version of $\tilde{\Lambda}_1$ measure with a Gaussian kernel with the standard deviation 0.4

Similarly to a general smoothing procedure, optimisation of the kernel width requires additional criteria to be employed, like information criteria. It is also possible to add a specific term into the score function of the optimisation procedure depending on the kernel function which is responsible for the goodness of fit of the smoothed curve to the empirical data. We do not address, however, these issues here considering it a separate problem from a nonparametric measure fitting, see also Discussion section below.

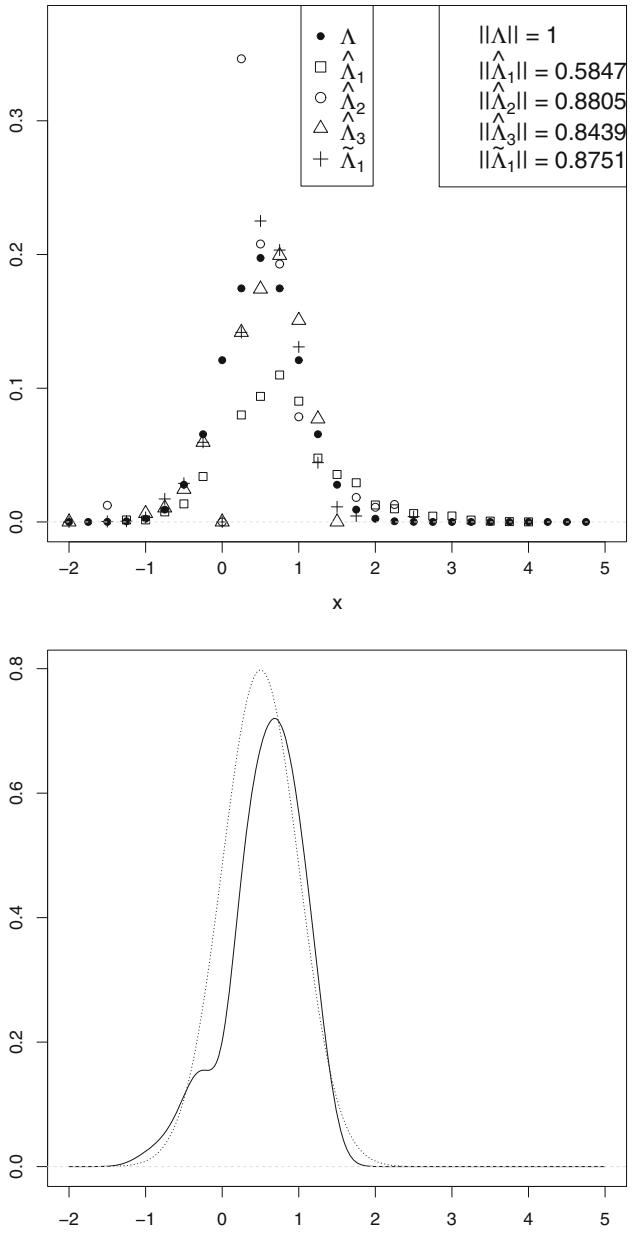

Fig. 7 *Top plot* estimated jump measure for a simulated sample with jump sizes having a normal distribution with the mean 0.5 and variance 0.25. *Bottom plot* the theoretical Gaussian density and the smoothed version of $\hat{\Lambda}_3$ measure with a Gaussian kernel with the standard deviation 0.2

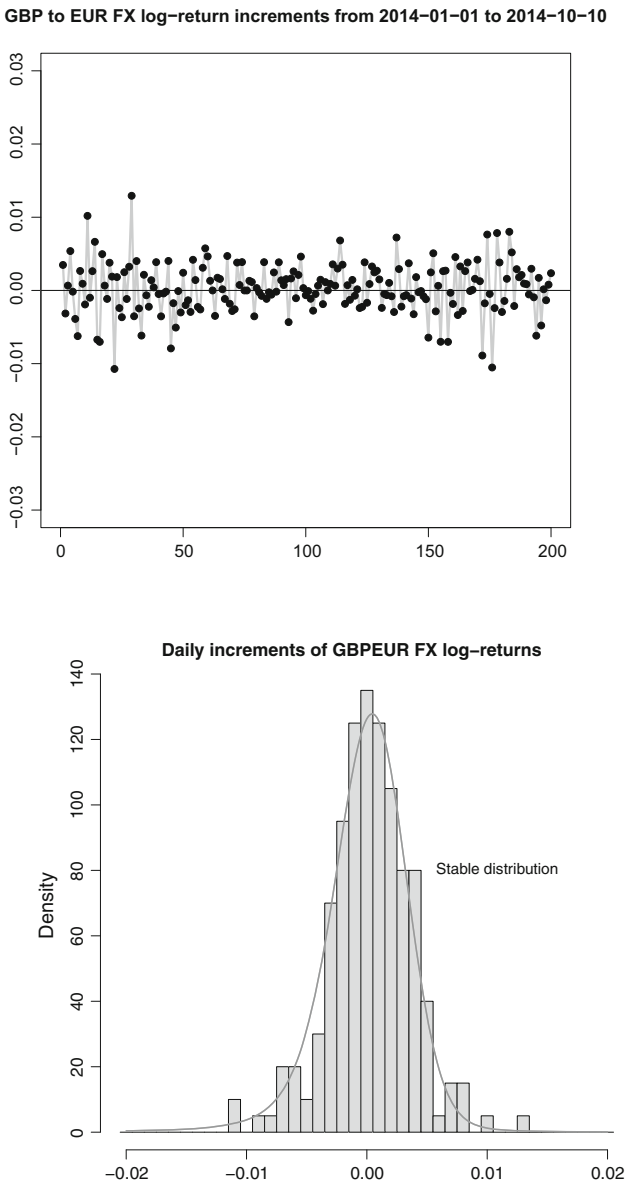

Fig. 8 *Top plot* Consecutive increments of the log-returns of GBP rate against EUR from 2014-01-02 to 2014-10-10. *Bottom plot* the fitted stable distribution to the increment data

6 Currency exchange data application

Lévy processes are widely used in financial mathematics to model the dynamics of the *log-returns* which for a commodity with price S_t and time t is defined to be $W_t = \log(S_t/S_0)$. For this model, the increments $W_h - W_0$, $W_{2h} - W_h, \ldots$ are independent and have a common infinitely divisible distribution. For example, many authors argue that the log-returns of the currency exchange rates in a stable market have indeed i.i.d. increments, see e.g. Cont (2001). We took FX data of the Great Britain Pound (GBP) against a few popular currencies and chose to work with GBP to EUR exchange rates in a period of 200 consecutive days of a relatively stable market from 2014-01-02 to 2014-10-10, see the top plot of Fig. 8. We fitted various, popular among financial analysts, distributions to the daily increments of the log-returns: Gaussian, GEV, Weibull and stable distributions. The best fit was obtained by the stable distribution. In order to have a consistent comparison with our methods, we used the loss function (3) to estimate the parameters of the stable distribution, such estimation method goes back to at least Paulson et al. (1975).

The fitted stable $S(1.882, -1, 0.002, 0; 0)$ distribution (in S0 parametrisation) is resented on the bottom plot of Fig. 8. A formal Chi-square test, however, rejected the null hypothesis that the data are coming from the fitted stable distribution due to the large discrepancies in the tails. The distance between the empirical characteristic function and the fitted stable distribution's characteristic function measured in terms of the score function L_{ChF} was 6.12×10^{-3}. We then ran our CoF algorithm with $k = 1$ and obtained a distribution within the distance 5.53×10^{-6} from the empirical characteristic function. Taking the resulting jump measure as a starting point to our ChF method, we arrived at the distribution within the distance 8.71×10^{-7}. The observed improvement is due to more accurate estimates of the large jumps of the exchange rates (which are not related to global economical or political events).

It may be expected that, as in the case of a linear regression, the agreement of the estimated model with the data could be "too good". To verify stability of our estimates, we ran our algorithms on the data with different time lags: every 2, 4 and 8 days records. It is interesting to note that even at 8 days lag our algorithms attained a distribution at the distance 2.19×10^{-4}, an order of magnitude closer to the empirical characteristic function than the fitted stable distribution despite that fact that 8 times less data were used, see Fig. 9.

The estimates of the measure Λ obtained for various lags are not that much different, apart from 8 days lag when only 25 observations are available, which reassures that our estimation methods give consistent results. These findings are illustrated on the bottom panel of Fig. 9.

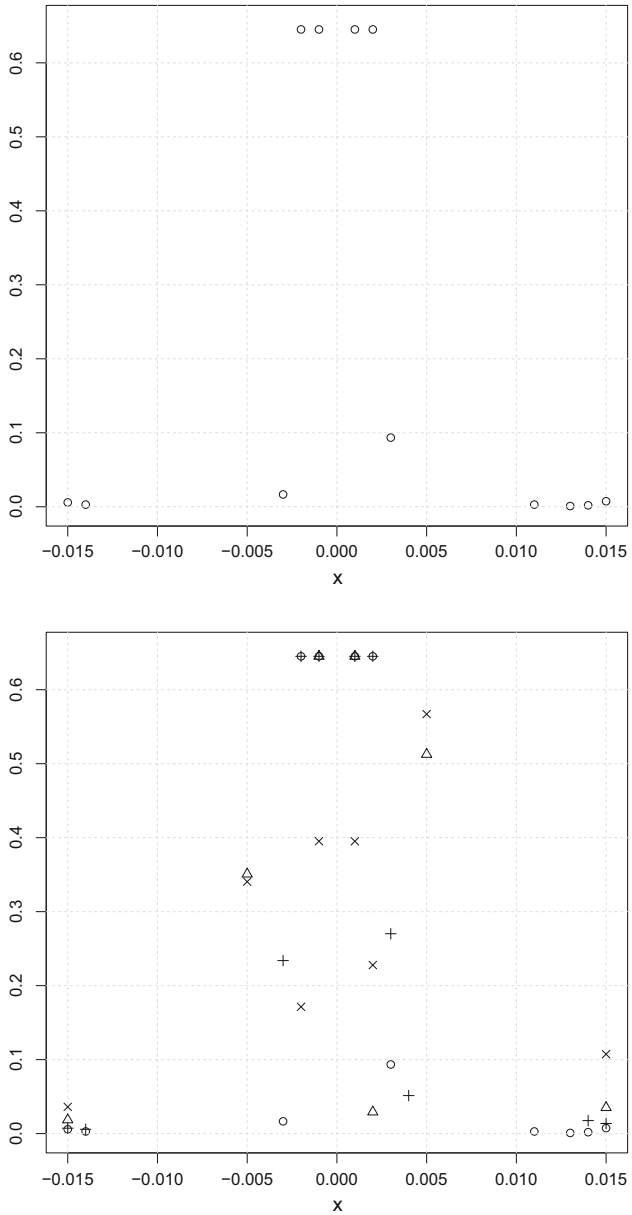

Fig. 9 *Top plot* estimated Lévy measure for a GBP/EUR rate log-return increments. *Bottom plot* estimated compounding measure for FX data recorded with various lags: 1 (*circle*), 2 (*plus*), 4 (*cross*) and 8 (*triangle*) day intervals

7 Discussion

This paper deals with nonparametric inference for compound Poisson processes. We proposed and analysed new algorithms based on the characteristic function fitting (ChF) and convoluted cumulative distribution function fitting (CoF). The algorithms are based on the recently developed variational analysis of functionals of measures and the corresponding steepest descent methods for constraint optimisation on the cone of measures. CoF methods are capable of producing very accurate estimates, but at the expense of growing computational complexity. The ChF method critically depends on the initial approximation measure due to highly irregular behaviour of the objective function. We have observed that the problems of convergence of the ChF algorithms can often be effectively overcome by choosing the sample measure (discretised to the grid) as the initial approximation measure. However, a better alternative, as we demonstrated in the paper, is to use the measure obtained by the simplest ($k = 1$) CoF algorithm. This combined CoF–ChF algorithm is fast and in majority of cases produces a measure which is closest in the total variation to the measure under estimation, and thus, this is our method of choice.

The practical experience we gained during various tests allows us to conclude that the suggested methods are especially well suited for estimation of discrete jump size distributions. They work well even with jumps that take both positive and negative values, not necessarily belonging to a regular lattice, demonstrating a clear advantage over the existing methods, see Buchmann and Grübel (2003), Buchmann and Grübel (2004). The use of our algorithms for continuous

compounding distributions requires more trial and error in choosing the right discretisation grid and smoothing procedures. In order to properly take into account the continuity of the compounding measure, one may apply direct methods of the density estimation proposed by van Es et al. (2007), Watteel and Kulperger (2003). Alternatively, one can try to develop an optimisation algorithm for the class of absolutely continuous measures by characterising their tangent cones. Additional conditions on the density may also be imposed, like Lipschitz kind of conditions, to make the feasible set closed in the corresponding measure topology.

Appendix

Proof of Theorem 1

First-order necessary criteria for constrained optimisation in a Banach space can be derived in terms of tangent cones. Let $\mathbb{A}$ be a subset of $\mathbb{M}$ and $\eta \in \mathbb{A}$. The *tangent cone* to $\mathbb{A}$ at η is the following subset of $\mathbb{M}$:

$$\mathbb{T}_{\mathbb{A}}(\eta) = \liminf_{t \downarrow 0} \; t^{-1}(\mathbb{A} - \eta).$$

Recall that the $\liminf_n A_n$ for a family of subsets (A_n) in a normed space is the set of the limits of all converging sequences $\{a_n\}$ such that $a_n \in A_n$ for all n. Equivalently, $\mathbb{T}_{\mathbb{A}}(\eta)$ is the closure of the set of such $\nu \in \mathbb{M}$ for which there exists an $\varepsilon = \varepsilon(\nu) > 0$ such that $\eta + t\nu \in \mathbb{A}$ for all $0 \le t \le \varepsilon$.

By the definition of the tangent cone, if η is a point of minimum of a strongly differentiable function L over a set $\mathbb{A}$, then one must have

$$DL(\eta)[\nu] \ge 0 \quad \text{for all } \nu \in \mathbb{T}_{\mathbb{A}}(\eta). \tag{14}$$

Indeed, assume that there exists $\nu \in \mathbb{T}_{\mathbb{A}}(\eta)$ such that $DL(\eta)[\nu] := -\varepsilon < 0$. Then, there is a sequence of positive numbers $t_n \downarrow 0$ and a sequence $\eta_n \in \mathbb{A}$ such that $\nu = \lim_n t_n^{-1}(\eta_n - \eta)$. Because $\|\eta - \eta_n\| = t_n(1 + o(1))\|\nu\| \to 0$, we obtain that $\eta_n \to \eta$. Since any bounded linear operator is continuous, we also have

$$\begin{aligned} DL(\eta)[\nu] &= DL(\eta)[\lim_n t_n^{-1}(\eta_n - \eta)] \\ &= \lim_n t_n^{-1} DL(\eta)[\eta_n - \eta] = -\varepsilon. \end{aligned}$$

Furthermore, by (4),

$$\begin{aligned} DL(\eta)[\eta_n - \eta] &= L(\eta_n) - L(\eta) + o(\|\eta - \eta_n\|) \\ &= L(\eta_n) - L(\eta) + o(t_n), \end{aligned}$$

implying

$$L(\eta_n) - L(\eta) = -t_n\varepsilon(1 + o(1)) < -t_n\varepsilon/2$$

for all sufficiently small t_n. Thus, in any ball around η there exists an $\eta_n \in \mathbb{A}$ such that $L(\eta_n) < L(\eta)$, so that η is not a point of a local minimum of L over $\mathbb{A}$. This finishes the proof of (14).

In our case, the constraint set $\mathbb{A}$ is the set $\mathbb{L} = \{\eta \in \mathbb{M}_+ : \eta(\{0\}) = 0\}$. Next step is to find a sufficiently rich class of measures belonging to the tangent cone $\mathbb{T}_{\mathbb{L}}(\Lambda)$ for s given $\Lambda \in \mathbb{L}$. For this, notice that for any such Λ, the Dirac measure δ_x belongs to $\mathbb{T}_{\mathbb{L}}(\Lambda)$ since $\Lambda + t\delta_x \in \mathbb{L}$ for any $t \ge 0$ as soon as $x \ne 0$. Similarly, given any Borel $B \subset \mathbb{R}$, the negative measure $-\Lambda|_B := -\Lambda(\cdot \cap B)$, which is the restriction of $-\Lambda$ onto B, is also in the tangent cone $\mathbb{T}_{\mathbb{L}}(\Lambda)$, because for any $0 \le t \le 1$ we have $\Lambda - t\Lambda|_B \in \mathbb{L}$.

Since, under the assumptions of the theorem, $\nabla L(x; \Lambda)$ is a gradient function, the necessary condition (14) becomes

$$\int \nabla L(x; \Lambda)\, \nu(\mathrm{d}x) \ge 0 \quad \text{for all } \nu \in \mathbb{T}_{\mathbb{L}}(\Lambda).$$

Substituting $\nu = \delta_x$ above we immediately obtain the inequality in (7). Finally, taking $\nu = -\Lambda|_B$ yields

$$\int_B \nabla L(x; \Lambda)\, \Lambda(\mathrm{d}x) \le 0.$$

Since this is true for any Borel B, we conclude that $\nabla L(x; \Lambda) \le 0$ Λ almost everywhere which, combined with the previous inequality, gives the second relation in (7).

Proof of Theorem 2

Let $\boldsymbol{N}$ be the space of locally finite counting measures φ on $\mathbb{R}$. Let $\mathcal{N}$ be the smallest σ-algebra which makes measurable all the mappings $\varphi \mapsto \varphi(B) \in \mathbb{Z}_+$ for $\varphi \in \boldsymbol{N}$ and compact sets B. A Poisson point process with the *intensity measure* μ is a measurable mapping Π from some probability space into $[\boldsymbol{N}, \mathcal{N}]$ such that for any finite family of disjoint compact sets $B_1, \dots, B_k$, the random variables $\Pi(B_1), \dots, \Pi(B_k)$ are independent and each $\Pi(B_i)$ has a Poisson distribution with parameter $\mu(B_i)$. Clearly $\mu(B) = \mathbf{E}\Pi(B)$ for any B. To emphasise the dependence of the distribution on μ, we write the expectation as $\mathbf{E}_\mu$ in the sequel.

Consider a measurable function $G\colon \boldsymbol{N} \to \mathbb{R}$, and for a given $z \in \mathbb{R}$ define the difference operator

$$D_z G(\varphi) := G(\varphi + \delta_z) - G(\varphi), \;\; \varphi \in \boldsymbol{N}.$$

For the iterations of such difference operators,

$$D_{z_1,\dots,z_n} G = D_{z_n}(D_{z_1,\dots,z_{n-1}} G), \quad (z_1, \dots, z_n) \in \mathbb{R}^n,$$

it can be checked that

$$D_{z_1,\dots,z_n}G(\nu) = \sum_{J\subseteq\{1,2,\dots,n\}} (-1)^{n-|J|} G\big(\nu + \Sigma_{j\in J}\delta_{z_j}\big),$$

where $|J|$ stands for the cardinality of J, so that if J is an empty set, then $|J| = 0$. Define

$$T_\mu G(z_1,\dots,z_n) := \mathbf{E}_\mu D_{z_1,\dots,z_n} G(\Pi).$$

Suppose that the functional G is such that there exists a constant $c > 0$ satisfying

$$|G\big(\Sigma_{j=1}^n \delta_{z_j}\big)| \le c^n \text{ for all } n \ge 1 \text{ and all } (z_1,\dots z_n).$$

It was proved in Molchanov and Zuyev (2000a, Theorem 2.1) that if μ, μ' are finite measures, the expectation $\mathbf{E}_{\mu+\mu'}G(\Pi)$ exists and

$$\mathbf{E}_{\mu+\mu'}G(\Pi) = \mathbf{E}_\mu G(\Pi) + \sum_{i=1}^\infty \frac{1}{i!}\int_{\mathbb{R}^i} T_\mu G(z_1,\dots,z_i)\mu'(\mathrm{d}z_1)\dots\mu'(\mathrm{d}z_i). \quad (15)$$

Generalisations of this formula to infinite and signed measures for square integrable functionals can be found in Last (2014). A finite-order expansion formula can be obtained by representing the expectation above in the form

$$\mathbf{E}_{\mu+\mu'}G(\Pi) = \mathbf{E}_\mu \mathbf{E}_{\mu'}[G(\Pi+\Pi') \;\big|\; \Pi],$$

where Π and Π' are independent Poisson processes with intensity measures μ and μ', respectively, and then applying the moment expansion formula by Błaszczyszyn et al. (1997, Theorem 3.1) to $G(\Pi+\Pi')$ viewed as a functional of Π' with a given Π. This gives us

$$\mathbf{E}_{\mu+\mu'}G(\Pi) = \mathbf{E}_\mu G(\Pi) + \sum_{i=1}^k \frac{1}{i!}\int_{\mathbb{R}^i} T_\mu G(z_1,\dots,z_i)\mu'(\mathrm{d}z_1)\dots\mu'(\mathrm{d}z_i) + \frac{1}{(k+1)!}\int_{\mathbb{R}^{k+1}} T_{\mu+\mu'}G(z_1,\dots,z_{k+1})\mu'(\mathrm{d}z_1)\dots \mu'(\mathrm{d}z_{k+1}). \quad (16)$$

To prove Theorem 2, we use a coupling of the compound Poisson process $(W_t)_{t\ge0}$ with a Poisson process Π on $\mathbb{R}_+\times\mathbb{R}$ driven by the intensity measure $\mu = \ell\times\Lambda$, where ℓ is the Lebesgue measure on $[0,+\infty)$. Clearly,

$$W_t = \Sigma_{(t_j,x_j)\in\Pi_t} x_j = \int_0^t\int_{\mathbb{R}} x\Pi(\mathrm{d}s\,\mathrm{d}x),$$

where for each realisation $\Sigma_j\delta_{z_j}$ of Π with $z_j = (t_j,x_j)$, we denote by Π_t the restriction of Π onto $[0,t]\times\mathbb{R}$. For a fixed arbitrary $y\in\mathbb{R}$ and a point configuration $\varphi = \Sigma_j\delta_{(t_i,x_j)}$, consider a functional G_y defined by

$$G_y(\varphi) = \mathbb{1}\Big\{\sum_{(t_j,x_j)\in\varphi} x_j \le y\Big\}$$

and notice that for any $z = (t,x)$,

$$G_y(\varphi+\delta_z) = \mathbb{1}\Big\{\sum_{(t_j,x_j)\in\varphi} x_j \le y - x\Big\} = G_{y-x}(\varphi). \quad (17)$$

Expressing the cumulative distribution function $F(y) = \mathbf{P}\{W_h\le y\}$ as an expectation

$$F(y) = \mathbf{P}_\mu\Big\{\sum_{(t_j,x_j)\in\Pi_h} x_j \le y\Big\} = \mathbf{E}_\mu G_y(\Pi_h),$$

and putting $\mu' = [0,h]\times\Lambda$, $\mu'' = [h,2h]\times\Lambda$, we find

$$\mathbf{E}_{\mu'+\mu''}G_y(\Pi) = \mathbf{P}\{W_{2h}\le y\} = \mathbf{P}\{W_h + W_h''\le y\} = F^{*2}(y),$$

where $W_h'' = W_{2h} - W_h$. Observe also that by iteration of (17),

$$\begin{aligned} T_{\mu'}G_y(z_1,\dots,z_n) &= \mathbf{E}_{\mu'}D_{z_1,\dots,z_n}G_y(\Pi) \\ &= \sum_{J\subseteq\{1,2,\dots,n\}} (-1)^{n-|J|}\mathbf{E}_{\mu'}G_y\big(\Pi + \Sigma_{j\in J}\delta_{z_j}\big) \\ &= \sum_{J\subseteq\{1,2,\dots,n\}} (-1)^{n-|J|}F(y - \Sigma_{j\in J}x_j) = U_{x_1,\dots,x_n}F(y). \end{aligned}$$

To finish the proof, it now remains to apply expansion (15):

$$\begin{aligned} F^{*2}(y) &= \mathbf{E}_{\mu'+\mu''}G_y(\Pi) = F(y) \\ &\quad + \sum_{i=1}^\infty \frac{1}{i!}\int_{(\mathbb{R}_+\times\mathbb{R})^i} U_{x_1,\dots,x_i}F(y)\,\mu''(\mathrm{d}t_1\,\mathrm{d}x_1)\dots\mu''(\mathrm{d}t_n\,\mathrm{d}x_i) \\ &= \sum_{i=0}^\infty h^i\Gamma_i(F,\Lambda,y). \end{aligned}$$

Gradient of ChF loss function

The ChF method is based on the loss function L_{ChF} given by (3), which is everywhere differentiable in Fréchet sense with respect to the measure Λ. Aiming at the steepest descent gradient method described in Sect. 3 for obtaining the minimum of the loss function, we compute here the gradient of L_{ChF} in terms of the following functions

$$q_1(\theta, x) := \cos(\theta x) - 1, \quad q_2(\theta, x) := \sin(\theta x) - \theta x \, \mathbb{I}_{\{|x|<\varepsilon\}},$$
$$Q_i(\theta, \Lambda) := \int q_2(\theta, x)\Lambda(\mathrm{d}x), \quad i = 1, 2.$$

Using this notation, the real and imaginary parts of an infinitely divisible distribution characteristic function $\varphi = \varphi_1 + i\varphi_2$ can be written down as

$$\varphi_1(\theta, \Lambda) = e^{hQ_1(\theta,\Lambda)} \cos\{hQ_2(\theta, \Lambda)\},$$
$$\varphi_2(\theta, \Lambda) = e^{hQ_1(\theta,\Lambda)} \sin\{hQ_2(\theta, \Lambda)\}.$$

After noticing that $\hat{\varphi}_n = \hat{\varphi}_{n,1} + i\hat{\varphi}_{n,2}$, with

$$\hat{\varphi}_{n,1}(\theta) = \frac{1}{n}\sum_{j=1}^{n} \cos(\theta X_j), \quad \hat{\varphi}_{n,2}(\theta) = \frac{1}{n}\sum_{j=1}^{n} \sin(\theta X_j),$$

the loss functional L_{ChF} can be written as

$$L_{\mathrm{ChF}}(\Lambda) = \int \left\{\varphi_1(\theta, \Lambda) - \hat{\varphi}_{n,1}(\theta)\right\}^2 \omega(\theta)\mathrm{d}\theta + \int \left\{\varphi_2(\theta, \Lambda) - \hat{\varphi}_{n,2}(\theta)\right\}^2 \omega(\theta)\mathrm{d}\theta.$$

From this representation, the gradient function corresponding to the Fréchet derivative with respect to the measure Λ is obtained using the Chain rule (5):

$$\begin{aligned}\nabla L_{\mathrm{ChF}}(x; \Lambda) &= 2\int \{\varphi_1(\theta, \Lambda) - \hat{\varphi}_{n,1}(\theta)]\}\nabla\varphi_1(\theta)[x, \Lambda]\omega(\theta)\mathrm{d}\theta \\ &\quad + 2\int \{\varphi_2(\theta, \Lambda) - \hat{\varphi}_{n,2}(\theta)\}\nabla\varphi_2(\theta)[x, \Lambda]\omega(\theta)\mathrm{d}\theta,\end{aligned} \tag{18}$$

where the gradients of $\varphi_i(\theta) := \varphi_i(\theta, \Lambda)$, $i = 1, 2$, with respect to the measure Λ, are given by

$$\begin{aligned}\nabla\varphi_1(\theta)(x; \Lambda) &= he^{hQ_1(\theta,\Lambda)} \\ &\quad \times \left\{\cos\left(hQ_2(\theta, \Lambda)\right)q_1(\theta, x) \right. \\ &\quad \left. - \sin\left(hQ_2(\theta, \Lambda)\right)q_2(\theta, x)\right\}, \\ \nabla\varphi_2(\theta)(x; \Lambda) &= he^{hQ_1(\theta,\Lambda)} \\ &\quad \times \left\{\sin\left(hQ_2(\theta, \Lambda)\right)q_1(\theta, x) \right. \\ &\quad \left. + \cos\left(hQ_2(\theta, \Lambda)\right)q_2(\theta, x)\right\}.\end{aligned}$$

Gradient of CoF loss function

As with the ChF method, the CoF algorithm relies on the steepest descent approach. The needed gradient function has the form

$$\nabla L^{(k)}_{\mathrm{CoF}}(x; \Lambda) = 2h\int \left\{\sum_{i=0}^{k} h^i \Gamma_i(\hat{F}_n, \Lambda, y) - \hat{F}_n^{*2}(y)\right\} \times \sum_{j=0}^{k-1} h^j \Xi_j(\hat{F}_n, \Lambda, y, x)\omega(y)\mathrm{d}y,$$

where

$$\Xi_j(F, \Lambda, y, x) = \frac{1}{j!}\int_{\mathbb{R}^j} U_{x,x_1,\ldots,x_j} F(y)\Lambda(\mathrm{d}x_1)\ldots\Lambda(\mathrm{d}x_j).$$

This formula follows from the Chain rule (5) and the equality

$$\begin{aligned}&\nabla\Big(\sum_{j=1}^{k} \frac{h^j}{j!}\int_{\mathbb{R}^j} U_{x_1,\ldots,x_j} F(y)\Lambda(\mathrm{d}x_1)\ldots\Lambda(\mathrm{d}x_j)\Big)(x; \Lambda) \\ &= h\sum_{j=0}^{k-1} \frac{h^j}{j!}\int_{\mathbb{R}^j} U_{x,x_1,\ldots,x_j} F(y)\Lambda(\mathrm{d}x_1)\ldots\Lambda(\mathrm{d}x_j).\end{aligned}$$

To justify the last identity, it suffices to see that for any integrable symmetric function $u(x_1, \ldots, x_j)$ of $j \geq 1$ variables,

$$\begin{aligned}&\nabla\Big(\int_{\mathbb{R}^j} u(x_1, \ldots, x_j)\Lambda(\mathrm{d}x_1)\ldots\Lambda(\mathrm{d}x_j)\Big)(x; \Lambda) \\ &= j\int_{\mathbb{R}^{j-1}} u(x, x_1, \ldots, x_{j-1})\Lambda(\mathrm{d}x_1)\ldots\Lambda(\mathrm{d}x_{j-1}).\end{aligned}$$

This is due to

$$\begin{aligned}&\int_{\mathbb{R}^j} u(x_1, \ldots, x_j)(\Lambda + \nu)(\mathrm{d}x_1)\ldots(\Lambda + \nu)(\mathrm{d}x_j) \\ &\quad - \int_{\mathbb{R}^j} u(x_1, \ldots, x_j)\Lambda(\mathrm{d}x_1)\ldots\Lambda(\mathrm{d}x_j) \\ &= \sum_{k=1}^{j}\int_{\mathbb{R}^j} u(x_1, \ldots, x_j)\Lambda(\mathrm{d}x_1)\ldots\Lambda(\mathrm{d}x_{k-1}) \\ &\quad \times \nu(\mathrm{d}x_k)\Lambda(\mathrm{d}x_{k+1})\ldots\Lambda(\mathrm{d}x_j) + o(\|\nu\|),\end{aligned}$$

where the last sum equals

$$j\int_{\mathbb{R}^j} u(x, x_1, \ldots, x_{j-1})\nu(\mathrm{d}x)\Lambda(\mathrm{d}x_1)\ldots\Lambda(\mathrm{d}x_{j-1}).$$

For example, the cost function (12) with $k = 1$ and $\omega(y) \equiv 1$ has the gradient

$$\begin{aligned}&\nabla L^{(1)}_{\mathrm{CoF}}(x; \Lambda) \\ &= 2h\int \left\{\hat{F}_n(y) - \hat{F}_n^{*2}(y) \right. \\ &\quad \left. + \int \hat{F}_n(y - z)\Lambda(\mathrm{d}z)\right\}\hat{F}_n(y - x)\,\mathrm{d}y.\end{aligned}$$

Respectively, the discretised gradient (9) used in the steepest descent algorithm is the vector $\boldsymbol{g}$ with the components

$$g_i = 2h \int \Big\{ \hat{F}_n(y) - \hat{F}_n^{*2}(y) + \sum_{j=1}^{l} \hat{F}_n(y - x_j)\lambda_j \Big\} \hat{F}_n(y - x_i)\, \mathrm{d}y, \quad i = 1, \ldots, l. \tag{19}$$

Acknowledgements This work was supported by Swedish Research Council Grant No. 11254331. The authors are grateful to Ilya Molchanov for fruitful discussions and to anonymous referees for valuable suggestions and stimulating criticism.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

Asmussen, S.: Applied Probability and Queues. Stochastic Modelling and Applied Probability. Springer, New York (2008)

Asmussen, S., Rosińsky, J.: Approximations of small jumps of Lévy process with a view towards simulation. J. Appl. Prob. **38**, 482–493 (2001)

Błaszczyszyn, B., Merzbach, E., Schmidt, V.: A note on expansion for functionals of spatial marked point processes. Stat. Probab. Lett. **36**(3), 299–306 (1997)

Bortkiewicz, L.: Das gesetz der kleinen zahlen. Monatshefte für Mathematik und Physik **9**(1), A39–A41 (1898)

Buchmann, B., Grübel, R.: Decompounding: an estimation problem for Poisson random sums. Ann. Stat. **31**(4), 1054–1074 (2003)

Buchmann, B., Grübel, R.: Decompounding Poisson random sums: recursively truncated estimates in the discrete case. Ann. Inst. Stat. Math. **56**(4), 743–756 (2004)

Carrasco, M., Florens, J.P.: Generalization of GMM to a continuum of moment conditions. Econ. Theory **16**, 797–834 (2000)

Coca, A.: Efficient nonparametric inference for discretely observed compound Poisson processes. Technical Report 1512.08472, ArXiv, December (2015)

Comte, F., Duval, C., Genon-Catalot, V.: Nonparametric density estimation in compound Poisson processes using convolution power estimators. Metrika **77**(1), 163–183 (2014)

Cont, R.: Empirical properties of asset returns: stylized facts and statistical issues. Quant. Financ. **1**, 223–236 (2001)

Cont, R., Tankov, P.: Financial Modelling with Jump Processes. Chapman & Hall/CRC, London (2003)

Duval, C.: Density estimation for compound Poisson processes from discrete data. Stoch. Process. Appl. **123**(11), 3963–3986 (2013)

Duval, C.: When is it no longer possible to estimate a compound Poisson process? Electron. J. Stat. **8**(1), 274–301 (2014)

Feuerverger, A., McDunnough, P.: On the efficiency of empirical characteristic function procedures. J. R. Stat. Soc. Ser. B **43**, 20–27 (1981a)

Feuerverger, A., McDunnough, P.: On some Fourier methods for inference. J. Am. Stat. Assoc. **76**, 379–387 (1981b)

Frees, E.W.: Nonparametric renewal function estimation. Ann. Stat. **14**(4), 1366–1378 (1986)

Last, G.: Perturbation analysis of Poisson processes. Bernoulli **20**(2), 486–513 (2014)

Mikosch, T.: Non-life Insurance Mathematics: An Introduction with the Poisson Process. Springer, New York (2009)

Molchanov, I., Zuyev, S.: Variational analysis of functionals of Poisson processes. Math. Oper. Res. **25**(3), 485–508 (2000a)

Molchanov, I., Zuyev, S.: Tangent sets in the space of measures: with applications to variational analysis. J. Math. Anal. Appl. **249**(2), 539–552 (2000b)

Molchanov, I., Zuyev, S.: Steepest descent algorithms in a space of measures. Stat. Comput. **12**, 115–123 (2002)

Neumann, M.H., Reiss, M.: Nonparametric estimation for Lévy process from low-frequency observations. Bernoulli **15**(1), 223–248 (2009)

Paulson, A.S., Holcomb, E.W., Leitch, R.A.: The estimation of the parameters of the stable laws. Biometrika **62**(1), 163–170 (1975)

Quin, J., Lawless, J.: Empirical likelihood and general estimating equations. Ann. Stat. **22**, 300–325 (1994)

R Core Team: R: a language and environment for statistical computing. R Foundation for Statistical Computing (2015)

Sueishi, N., Nishiyama, Y.: Estimation of Lévy processes in mathematical finance: a comparative study. In: Zerger, A., Argent, R.M. (eds.) MODSIM 2005 International Congress on Modelling and Simulation, pp. 953–959 (2005)

van Es, B., Gugushvili, S., Spreij, P.: A kernel type nonparametric density estimator for decompounding. Bernoulli **13**(3), 672–694 (2007)

Watteel, R.N., Kulperger, R.J.: Nonparametric estimation of the canonical measure for infinitely divisible distributions. J. Stat. Comput. Simul. **73**(7), 525–542 (2003)