TU Delft

Delft University of Technology

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

Gugushvili, Shota; van der Meulen, Frank; Spreij, Peter

DOI
10.1007/s11203-016-9153-1

Publication date
2016

Document Version
Final published version

Published in
Statistical Inference for Stochastic Processes

Citation (APA)
Gugushvili, S., van der Meulen, F., & Spreij, P. (2016). A non-parametric Bayesian approach to decompounding from high frequency data. *Statistical Inference for Stochastic Processes*, 1-27. https://doi.org/10.1007/s11203-016-9153-1

Important note
To cite this publication, please use the final published version (if applicable).
Please check the document version above.

Copyright
Other than for strictly personal use, it is not permitted to download, forward or distribute the text or part of it, without the consent of the author(s) and/or copyright holder(s), unless the work is under an open content license such as Creative Commons.

Takedown policy
Please contact us and provide details if you believe this document breaches copyrights.
We will remove access to the work immediately and investigate your claim.

This work is downloaded from Delft University of Technology.
For technical reasons the number of authors shown on this cover page is limited to a maximum of 10.

Stat Inference Stoch Process
DOI 10.1007/s11203-016-9153-1

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

Shota Gugushvili[1] · **Frank van der Meulen**[2] · **Peter Spreij**[3,4]

Received: 1 June 2016 / Accepted: 29 November 2016
© The Author(s) 2016. This article is published with open access at Springerlink.com

Abstract Given a sample from a discretely observed compound Poisson process, we consider non-parametric estimation of the density f_0 of its jump sizes, as well as of its intensity λ_0. We take a Bayesian approach to the problem and specify the prior on f_0 as the Dirichlet location mixture of normal densities. An independent prior for λ_0 is assumed to be compactly supported and to possess a positive density with respect to the Lebesgue measure. We show that under suitable assumptions the posterior contracts around the pair (λ_0, f_0) at essentially (up to a logarithmic factor) the $\sqrt{n\Delta}$-rate, where n is the number of observations and Δ is the mesh size at which the process is sampled. The emphasis is on high frequency data, $\Delta \to 0$, but the obtained results are also valid for fixed Δ. In either case we assume that $n\Delta \to \infty$. Our main result implies existence of Bayesian point estimates converging (in the frequentist sense, in probability) to (λ_0, f_0) at the same rate. We also discuss a practical implementation of our approach. The computational problem is dealt with by inclusion of auxiliary variables and we develop a Markov chain Monte Carlo algorithm that samples from the joint distribution of the unknown parameters in the mixture density and the introduced auxiliary variables. Numerical examples illustrate the feasibility of this approach.

✉ Frank van der Meulen
f.h.vandermeulen@tudelft.nl

Shota Gugushvili
shota.gugushvili@math.leidenuniv.nl

Peter Spreij
spreij@uva.nl

1 Mathematical Institute, Leiden University, P.O. Box 9512, 2300 RA Leiden, The Netherlands

2 Faculty of Electrical Engineering, Mathematics and Computer Science, Delft Institute of Applied Mathematics, Delft University of Technology, Mekelweg 4, 2628 CD Delft, The Netherlands

3 Korteweg-de Vries Institute for Mathematics, University of Amsterdam, P.O. Box 94248, 1090 GE Amsterdam, The Netherlands

4 Radboud University Nijmegen, Nijmegen, The Netherlands

Published online: 16 December 2016

Keywords Compound Poisson process · Non-parametric Bayesian estimation · Posterior contraction rate · High frequency observations

Mathematics Subject Classification Primary: 62G20 · Secondary: 62M30

1 Introduction

1.1 Problem formulation

Let $N = (N_t,\, t \geq 0)$ be a Poisson process with a constant intensity $\lambda > 0$ and let $Y_1,\, Y_2,\, Y_3, \ldots$ be a sequence of independent random variables independent of N and having a common distribution function F with density f (with respect to the Lebesgue measure). A compound Poisson process (abbreviated CPP) $X = (X_t,\, t \geq 0)$ is defined as

$$X_t = \sum_{j=1}^{N_t} Y_j, \tag{1}$$

where the sum over an empty set is by definition equal to zero. CPPs form a basic model in a variety of applied fields, most notably in e.g., queueing and risk theory, see Embrechts et al. (1997) and Prabhu (1998) and the references therein, but also in other fields of science, see, e.g., Alexandersson (1985) and Burlando and Rosso (1993) for stochastic models for precipitation, Katz (2002) on modelling of hurricane damage, or Scalas (2006) for applications in economics and finance.

Suppose that corresponding to the 'true' parameter values $\lambda = \lambda_0$ and $f = f_0$, a discrete time sample $X_\Delta,\, X_{2\Delta}, \ldots, X_{n\Delta}$ is available from (1), where $\Delta > 0$. Such a discrete time observation scheme is common in a number of applications of CPP, e.g., in the precipitation models of the above references. Based on the sample $\mathcal{X}_n^\Delta = (X_\Delta,\, X_{2\Delta}, \ldots, X_{n\Delta})$, we are interested in (non-parametric) estimation of λ_0 and f_0. Before proceeding further, we notice that by the stationary independent increments property of a CPP, the random variables $Z_i^\Delta = X_{i\Delta} - X_{(i-1)\Delta},\ 1 \leq i \leq n$, are independent and identically distributed. Each Z_i^Δ has the same distribution as the random variable

$$Z^\Delta = \sum_{j=1}^{T^\Delta} Y_j, \tag{2}$$

where T^Δ is independent of the sequence $Y_1,\, Y_2, \ldots$ and has a Poisson distribution with parameter $\Delta\lambda$. Hence, our problem is equivalent to estimating (non-parametrically) λ_0 and f_0 based on the sample $\mathcal{Z}_n^\Delta = (Z_1^\Delta,\, Z_2^\Delta, \ldots, Z_n^\Delta)$. We will henceforth use this alternative formulation of the problem. Our emphasis is on *high frequency* data, $\Delta = \Delta_n \to 0$ as $n \to \infty$, but the obtained results are also valid for *low frequency* observations, i.e., for fixed Δ.

Our main result is on the contraction rate of the posterior distribution, which we show to be, up to a logarithmic factor, $(n\Delta)^{-1/2}$. A by now standard approach to obtain contraction rates in an IID setting is to verify the assumptions of the fundamental Theorem 2.1 in Ghosal et al. (2000). It should be noted that in the present high frequency setting, this theorem is not applicable. One of the model assumptions underlying this theorem, which is satisfied in Gugushvili et al. (2015), is that one deals with samples of a *fixed* distribution, whereas in our present high frequency observation regime the distribution of Z^Δ is *varying*, with the Dirac

distribution concentrated at zero as its limit for $\Delta \to 0$. Therefore we propose an alternative approach, circumventing the use of the cited Theorem 2.1. The theoretical contribution of the present paper is therefore not only the statement of the main result itself, but also its proof. Next to this we also discuss a practical implementation of our non-parametric Bayesian approach, a Markov chain Monte Carlo algorithm that samples from the joint distribution of the unknown parameters in the mixture density and certain introduced auxiliary variables.

1.2 Literature review and present approach

Because adding a Poisson number of Y_j's amounts to compounding their distributions, the problem of recovering the intensity λ_0 and the density f_0 from the observations Z_i's can be referred to as decompounding. Decompounding already has some history: the early contributions (Buchmann and Grübel 2003, 2004) dealt with estimation of the distribution function F_0, paying particular attention to the case when F_0 is discrete, while the later contributions (Comte et al. 2014; Duval 2013; Es et al. 2007) concentrated on estimation of the density f_0 instead. More (frequentist) theory on statistical inference on CPPs (and more generally on Lévy processes) can be found in the volume (Belomestny et al. 2015), with the survey paper (Comte et al. 2015) devoted to statistical methods for high frequency discrete observations, with a special section on CPPs. Other references on statistics for Lévy processes in the high frequency data setting are Comte and Genon-Catalot (2011), Comte and Genon-Catalot (2010), Comte et al. (2010), Figueroa-López (2008), Figueroa-Lopez (2009), Nickl and Reiß (2012), Nickl et al. (2016), and Ueltzhöfer and Klüppelberg (2011). All these approaches are frequentist in nature. On the other hand, theoretical and computational advances made over the recent years have shown that a non-parametric Bayesian approach is feasible in various statistical settings; see e.g., Hjort et al. (2010) for an overview. This is the approach we will take in this work to estimate λ_0 and f_0.

To the best of our knowledge, non-parametric Bayesian approach to inference for (a class of) Lévy processes was first considered in Gugushvili et al. (2015). That paper, contrary to the present context, dealt with observations at fixed equidistant times, and was strongly based on an application of Theorem 2.1 of Ghosal et al. (2000), as already alluded to in the problem formulation of Sect. 1.1. The present work complements the results from Gugushvili et al. (2015), in the sense that we now allow high frequency observations, which requires a substantially different route to prove our results, as we will explain in more detail in Sect. 1.3.

We will study the non-parametric Bayesian approach to decompounding from a frequentist point of view (in the sense specified below), so that one may also think of it as a means for obtaining a frequentist estimator. Advantages of the non-parametric Bayesian approach include automatic quantification of uncertainty in parameter estimates through Bayesian posterior credible sets and automatic selection of the degree of smoothing required in non-parametric inferential procedures.

1.3 Results

The non-parametric class $\mathcal{F}$ of densities f that we consider is that of location mixtures of normal densities. So we consider densities specified by

$$f(x) = f_{H,\sigma}(x) = \int \phi_\sigma(x - z)\mathrm{d}H(z), \tag{3}$$

where ϕ_σ denotes the density of the normal distribution with mean zero and variance σ^2 and H is a mixing measure. These mixtures form a rich and flexible class of densities, see Marron and Wand (1992) and McLachlan and Peel (2000), that are capable of closely

approximating many densities that themselves are not representable in this way. The resulting mixture densities will be infinitely smooth, which is arguably the case in many, if not most, practical applications.

Bayesian estimation requires specification of prior distributions on λ and f. We propose independent priors on λ and f that we denote by Π_1 and Π_2, respectively. For f, we take a Dirichlet mixture of normal densities as a prior. This type of prior in the context of Bayesian density estimation has been introduced in Ferguson (1983) and Lo (1984); for recent references see, e.g., Ghosal and Vaart (2001). The prior for f is defined as the law of the function $f_{H,\sigma}$ as in (3), with H assumed to follow a Dirichlet process prior D_α with base measure α and σ a priori independent with distribution Π_3. Recall that a Dirichlet process D_α on $\mathbb{R}$ with the base measure α defined on the Borel σ-algebra $\mathcal{B}(\mathbb{R})$ (we assume α to be non-negative and σ-additive) is a random probability measure G on $\mathbb{R}$, such that for every finite and measurable partition $B_1, B_2, \ldots, B_k$ of $\mathbb{R}$, the probability vector $(G(B_1), G(B_2), \ldots, G(B_k))$ possesses the Dirichlet distribution on the k-dimensional simplex with parameters $(\alpha(B_1), \alpha(B_2), \ldots, \alpha(B_k))$. See, e.g., the original paper (Ferguson 1973), or the overview article (Ghosal 2010) for more information on Dirichlet process priors.

A nonparametric Bayesian approach to density estimation employing a Dirichlet mixture of normal densities as a prior can in very rough sense be thought of as a Bayesian counterpart of kernel density estimation (with a Gaussian kernel), cf. Ghosal and van der Vaart (2007, p. 697).

With the sample size n tending to infinity, the Bayesian approach should be able to discern the true parameter pair (λ_0, f_0) with increasing accuracy. We can formalise this by requiring, for instance, that for any fixed neighbourhood A (in an appropriate topology) of (λ_0, f_0), $\Pi(A^c|\mathcal{Z}_n^\Delta) \to 0$ in $\mathbb{Q}_{\lambda_0,f_0}^{\Delta,n}$-probability. Here Π is used as a shorthand notation for the posterior distribution of (λ, f) and we use $\mathbb{Q}_{\lambda_0,f_0}^{\Delta}$ to denote the law of the random variable Z^Δ in (2) and $\mathbb{Q}_{\lambda_0,f_0}^{\Delta,n}$ the law of $\mathcal{Z}_n^\Delta$. More generally, one may take a sequence of shrinking neighbourhoods A_n of (λ_0, f_0) and try to determine the rate at which the neighbourhoods A_n are allowed to shrink, while still capturing most of the posterior mass. This rate is referred to as a posterior convergence rate (we will give the precise definition in Sect. 3). Two fundamental references dealing with establishing it in various statistical settings are Ghosal et al. (2000) and Ghosal and Vaart (2001). This convergence rate can be thought of as an analogue of the convergence rate of a frequentist estimator. The analogy can be made precise: contraction of the posterior distribution at a certain rate implies existence of a Bayes point estimate with the same convergence rate (in the frequentist sense); see Theorem 2.5 in Ghosal et al. (2000) and the discussion on pp. 506–507 there.

Obviously, for our programme to be successful, Δ has to satisfy the assumption $n\Delta \to \infty$, which is a necessary condition for consistent estimation of (λ_0, f_0), as it ensures that asymptotically we observe an infinite number of jumps in the process. We cover both the case of so called high frequency observation schemes ($\Delta \to 0$) as well as low frequency observations (fixed Δ). A sufficient condition, which covers both observation regimes and which relates Δ to n, is $\Delta = n^{-\alpha}$, where $0 \leq \alpha < 1$.

We note that in Ghosal and Tang (2006) and Tang and Ghosal (2007) non-parametric Bayesian inference for Markov processes is studied, of which CPPs form a particular class, but these papers deal with estimation of the transition density of a discretely observed Markov process, which is different from the problem we consider here. A parametric Bayesian approach to inference for CPPs is studied in Insua et al. (2012, Sects. 5.5 and 10.3).

The main result of our paper is Theorem 1, in which we state sufficient conditions on the prior that yield a posterior rate of contraction of the order $(\log^\kappa(n\Delta))/\sqrt{n\Delta}$, for some constant $\kappa > 0$. We argue that this rate is a nearly (up to a logarithmic factor) optimal posterior

contraction rate in our problem. Our main result complements the one in Gugushvili et al. (2015), in that it treats both the low and high frequency observation schemes simultaneously, with emphasis on the latter. We note (again) a fundamental difference between the present paper and Gugushvili et al. (2015), when it comes down to the techniques to prove the main result. As Theorem 2.1 of Ghosal et al. (2000) cannot immediately be used, we take an alternative tour that avoids this theorem, but instead refines a number of technical results involving properties of statistical tests that form essential ingredients of the proof in Ghosal et al. (2000). These refined results are then used as key technical steps in a direct proof of our Theorem 1. Furthermore, it establishes the posterior contraction rate for infinitely smooth jump size densities f_0, which is not covered by Gugushvili et al. (2015). On the other hand, Gugushvili et al. (2015) deals with multi-dimensional CPPs, while in this paper we consider only the one-dimensional case. Finally, in this work we also discuss a practical implementation of our non-parametric Bayesian approach. The computational problem is dealt with by inclusion of auxiliary variables. More precisely, we show how a Markov chain Monte Carlo algorithm can be devised that samples from the joint distribution of the unknown parameters in the mixture density and the introduced auxiliary variables. Numerical examples illustrate the feasibility of this approach.

1.4 Organisation

The remainder of the paper is organised as follows. In the next section we state some preliminaries on the likelihood, prior and notation. In Sect. 3 we first motivate the use of the scaled Hellinger metric to define neighbourhoods for which posterior contraction rate is derived in case the observations are sampled at high frequency. Then we present the main result on the posterior contraction rate (Theorem 1), whose proof is given in Sect. 5. We discuss the numerical implementation of our results in Sect. 4. Technical lemmas and their proofs used to prove the main theorem are gathered in the Appendix.

2 Preliminaries and notation

2.1 Likelihood, prior and posterior

We are interested in Bayesian inference with Bayes' formula. Therefore we need to specify the likelihood in our model. We use the following notation:

$\mathbb{P}_f$	law of Y_1 (law of the jumps of the CPP)
$\mathbb{Q}^{\Delta}_{\lambda,f}$	law of Z^{Δ}_1 (law of the increments of the discretely observed CPP)
$\mathbb{Q}^{\Delta,n}_{\lambda,f}$	law of $\mathcal{Z}^{\Delta}_n$ (joint law of the increments of the discretely observed CPP)
$\mathbb{R}^{\Delta}_{\lambda,f}$	law of $(X_t,\ t \in [0,\ \Delta])$ (law of the CPP on $[0,\ \Delta]$)

The characteristic function of the Poisson sum Z^{Δ} defined in (2) is given by

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

where ϕ_f is the characteristic function of f. This can be rewritten as

$$\phi(t) = e^{-\lambda\Delta} + \left(1 - e^{-\lambda\Delta}\right)\frac{1}{e^{\lambda\Delta} - 1}\left(e^{\lambda\Delta\phi_f(t)} - 1\right),$$

which, using the fact that ϕ_f vanishes at infinity, shows that the distribution of Z^{Δ} is a mixture of a point mass at zero and an absolutely continuous distribution. Letting $t \to \infty$, we get

that $\phi(t) \to e^{-\lambda\Delta}$. Hence λ is identifiable from the law of Z^Δ, and then so is f. The density of the law $\mathbb{Q}^\Delta_{\lambda,f}$ of Z^Δ with respect to the measure μ, which is the sum of Lebesgue measure and the Dirac measure concentrated at zero, can in fact be written explicitly as (cf. van Es et al. 2007, p. 681 and Proposition 2.1 in Duval 2013)

$$\frac{\mathrm{d}\mathbb{Q}^\Delta_{\lambda,f}}{\mathrm{d}\mu}(x) = e^{-\lambda\Delta}\mathbf{1}_{\{0\}}(x) + \left(1 - e^{-\lambda\Delta}\right)\sum_{m=1}^{\infty} a_m(\lambda\Delta) f^{*m}(x)\mathbf{1}_{\mathbb{R}\setminus\{0\}}(x), \tag{4}$$

where $\mathbf{1}_A$ denotes the indicator of a set A,

$$a_m(\lambda\Delta) = \frac{1}{e^{\lambda\Delta} - 1}\frac{(\lambda\Delta)^m}{m!}, \tag{5}$$

and f^{*m} denotes the m-fold convolution of f with itself. However, the expression (4) is useless for Bayesian computations. To work around this problem, we will employ a different dominating measure. Consider the law $\mathbb{R}^\Delta_{\lambda,f}$ of $(X_t,\ t \in [0,\ \Delta])$. By the Theorem in Skorohod (1964, p. 261) $\mathbb{R}^\Delta_{\lambda,f}$ is absolutely continuous with respect to $\mathbb{R}^\Delta_{\widetilde{\lambda},\widetilde{f}}$ if and only if $\mathbb{P}_f$ is absolutely continuous with respect to $\mathbb{P}_{\widetilde{f}}$ (we of course assume that $\lambda,\ \widetilde{\lambda} > 0$). A simple condition to ensure the latter is to assume that $\widetilde{f}$ is continuous and does not take the value zero on $\mathbb{R}$.

Define the random measure μ by

$$\mu(B) = \{\#t\colon (t,\ X_t - X_{t-}) \in B\}, \quad B \in \mathcal{B}([0,\ \Delta]) \otimes \mathcal{B}(\mathbb{R}\setminus\{0\}).$$

Under $\mathbb{R}_{\lambda,f}$, the random measure μ is a Poisson point process on $[0,\ \Delta] \times (\mathbb{R}\setminus\{0\})$ with intensity measure $\Lambda(dt,\ dx) = \lambda dt f(x)dx$, which follows, e.g., from Theorem 1 on p. 69 and Corollary on p. 64 in Skorohod (1964). By formula (46.1) on p. 262 in Skorohod (1964), we have

$$\frac{\mathrm{d}\mathbb{R}^\Delta_{\lambda,f}}{\mathrm{d}\mathbb{R}^\Delta_{\widetilde{\lambda},\widetilde{f}}}(X) = \exp\left(\int_0^\Delta\int_{\mathbb{R}} \log\left(\frac{\lambda f(x)}{\widetilde{\lambda}\widetilde{f}(x)}\right)\mu(dt,\ dx) - \Delta(\lambda - \widetilde{\lambda})\right). \tag{6}$$

By Theorem 2 on p. 245 in Skorohod (1964) and Corollary 2 on p. 246 there, the density $k^\Delta_{\lambda,f}$ of $\mathbb{Q}^\Delta_{\lambda,f}$ with respect to $\mathbb{Q}^\Delta_{\widetilde{\lambda},\widetilde{f}}$ is given by the conditional expectation

$$k^\Delta_{\lambda,f}(x) = \mathbb{E}_{\widetilde{\lambda},\widetilde{f}}\left(\frac{\mathrm{d}\mathbb{R}^\Delta_{\lambda,f}}{\mathrm{d}\mathbb{R}^\Delta_{\widetilde{\lambda},\widetilde{f}}}(X)\middle| X_\Delta = x\right), \tag{7}$$

where the subscript in the conditional expectation operator signifies the fact that it is evaluated under the probability $\mathbb{R}^\Delta_{\widetilde{\lambda},\widetilde{f}}$. Hence the likelihood [in the parameter pair $(\lambda,\ f)$] associated with the sample $\mathcal{Z}^\Delta_n$ is given by the product

$$L^\Delta_n(\lambda,\ f) = \prod_{i=1}^n k^\Delta_{\lambda,f}\left(Z^\Delta_i\right). \tag{8}$$

An advantage of specifying the likelihood in this manner is that it allows one to reduce some of the difficult computations for the laws $\mathbb{Q}^\Delta_{\lambda,f}$ to those for the laws $\mathbb{R}^\Delta_{\lambda,f}$, which are simpler.

Observe that the priors on λ and f indirectly induce the prior $\Pi = \Pi_1 \times \Pi_2$ on the collection of densities $k^\Delta_{\lambda,f}$. We will indiscriminately use the symbol Π to signify both the prior on $(\lambda,\ f)$, but also on the density $k^\Delta_{\lambda,f}$. The posterior in the first case will be understood as the posterior for the pair $(\lambda,\ f)$, while in the second case as the posterior for the density

$k^{\Delta}_{\lambda,f}$. We will often use the same symbol Π to denote the posterior distribution of $(\lambda,\ f)$ and on the density $k^{\Delta}_{\lambda,f}$. This simplifies notationally some of the formulations below.

By Bayes' theorem, the posterior measure of any measurable set $A \subset (0,\ \infty) \times \mathcal{F}$ is given by

$$\Pi\left(A|\mathcal{Z}_n^{\Delta}\right) = \frac{\iint_A L_n^{\Delta}(\lambda,\ f)\mathrm{d}\Pi_1(\lambda)\mathrm{d}\Pi_2(f)}{\iint L_n^{\Delta}(\lambda,\ f)\mathrm{d}\Pi_1(\lambda)\mathrm{d}\Pi_2(f)}.$$

Upon setting $\overline{A} = \{k_{\lambda,f}\colon (k,\ \lambda) \in A\}$ and recalling our conventions above, this can also be written as

$$\Pi\left(\overline{A}|\mathcal{Z}_n^{\Delta}\right) = \frac{\int_{\overline{A}} L_n^{\Delta}(k)\mathrm{d}\Pi(k)}{\int L_n^{\Delta}(k)\mathrm{d}\Pi(k)}.$$

Once the posterior is available, one can next proceed with computation of other quantities of interest in Bayesian statistics, such as Bayes point estimates or credible sets.

2.2 Notation

Throughout the paper we will use the following notation to compare two sequences $\{a_n\}$ and $\{b_n\}$ of positive real numbers: $a_n \lesssim b_n$ will mean that there exists a constant $C > 0$ that is independent of n and is such that $a_n \leq Cb_n$, while $a_n \gtrsim b_n$ will signify the fact that $a_n \geq Cb_n$.

Next we introduce various notions of distances between probability measures. The Hellinger distance $h(\mathbb{Q}_0,\ \mathbb{Q}_1)$ between two probability laws $\mathbb{Q}_0$ and $\mathbb{Q}_1$ on a measurable space $(\Omega,\ \mathfrak{F})$ is defined as

$$h\left(\mathbb{Q}_0,\ \mathbb{Q}_1\right) = \left(\int \left(\mathrm{d}\mathbb{Q}_0^{1/2} - \mathrm{d}\mathbb{Q}_1^{1/2}\right)^2\right)^{1/2}.$$

Assume further $\mathbb{Q}_0 \ll \mathbb{Q}_1$. The Kullback–Leibler (or informational) divergence $\mathrm{K}(\mathbb{Q}_0,\ \mathbb{Q}_1)$ is defined as

$$\mathrm{K}\left(\mathbb{Q}_0,\ \mathbb{Q}_1\right) = \int \log\left(\frac{\mathrm{d}\mathbb{Q}_0}{\mathrm{d}\mathbb{Q}_1}\right)\mathrm{d}\mathbb{Q}_0,$$

while the V-discrepancy is defined through

$$\mathrm{V}\left(\mathbb{Q}_0,\ \mathbb{Q}_1\right) = \int \log^2\left(\frac{\mathrm{d}\mathbb{Q}_0}{\mathrm{d}\mathbb{Q}_1}\right)\mathrm{d}\mathbb{Q}_0.$$

Here is some additional notation. For $f,\ g$ nonnegative integrable functions, not necessarily densities, we write

$$h^2(f,\ g) = \int (\sqrt{f} - \sqrt{g})^2,$$

$$\mathrm{K}(f,\ g) = \int \log\frac{f}{g} f - \int f + \int g,$$

$$\mathrm{V}(f,\ g) = \int \log^2\frac{f}{g} f.$$

Note that these 'distances' are all nonnegative and only zero if $f = g$ a.e. If f and g are densities of probability measures $\mathbb{Q}_0$ and $\mathbb{Q}_1$ on $(\mathbb{R},\ \mathcal{B})$, respectively, then the above 'distances' reduce to the previously introduced ones.

We will also use $\mathrm{K}(x,\ y) = x\log\frac{x}{y} - x + y$ for $x,\ y > 0$. Note that also $\mathrm{K}(x,\ y) \geq 0$ and $\mathrm{K}(x,\ y) = 0$ if and only if $x = y$.

3 Main result on posterior contraction rate

Denote the true parameter values for the CPP by $(\lambda_0,\ f_0)$. Recall that the problem is to estimate f_0 and λ_0 based on the observations $\mathcal{Z}_n^\Delta$ and that $\Delta \to 0$ in a high frequency regime. To say that a pair $(f,\ \lambda)$ lies in a neighbourhood of $(f_0,\ \lambda_0)$, one needs a notion of distance on the corresponding measures $\mathbb{Q}^\Delta_{\lambda,f}$ and $\mathbb{Q}^\Delta_{\lambda_0,f_0}$, the two possible induced laws of $Z_i^\Delta = X_{i\Delta} - X_{(i-1)\Delta}$. The Hellinger distance is a popular and rather reasonable choice to that end in non-parametric Bayesian statistics. However, for $\Delta \to 0$ the Hellinger metric h between those laws automatically tends to 0. The first assertion of Lemma 1 below states that $h(\mathbb{Q}^\Delta_{\lambda,f},\ \mathbb{Q}^\Delta_{\lambda_0,f_0})$ is of order $\sqrt{\Delta}$ when $\Delta \to 0$. This motivates to replace the ordinary Hellinger metric h with the scaled metric $h^\Delta = h/\sqrt{\Delta}$ in our asymptotic analysis for high frequency data. Of course, for fixed Δ (in which case one can take $\Delta = 1$ w.l.o.g.), nothing changes with this replacement. The lemma also shows that the Kullback–Leibler divergence and the V-discrepancy are of order Δ for $\Delta \to 0$. Therefore we will also use the scaled distances $\mathrm{K}^\Delta = \mathrm{K}/\Delta$ and $\mathrm{V}^\Delta = \mathrm{V}/\Delta$

Lemma 1 *The following expressions hold true*:

$$\lim_{\Delta\to 0}\frac{1}{\Delta}h^2\left(\mathbb{Q}^\Delta_{\lambda,f},\ \mathbb{Q}^\Delta_{\lambda_0,f_0}\right) = h^2\left(\lambda f,\ \lambda_0 f_0\right) = \int\left(\sqrt{\lambda f(x)} - \sqrt{\lambda_0 f_0(x)}\right)^2 \mathrm{d}x, \tag{9}$$

$$\lim_{\Delta\to 0}\frac{1}{\Delta}\mathrm{K}\left(\mathbb{Q}^\Delta_{\lambda,f},\ \mathbb{Q}^\Delta_{\lambda_0,f_0}\right) = \mathrm{K}\left(\lambda f,\ \lambda_0 f_0\right) = \lambda\mathrm{K}\left(f,\ f_0\right) + \mathrm{K}\left(\lambda,\ \lambda_0\right), \tag{10}$$

$$\lim_{\Delta\to 0}\frac{1}{\Delta}\mathrm{V}\left(\mathbb{Q}^\Delta_{\lambda,f},\ \mathbb{Q}^\Delta_{\lambda_0,f_0}\right) = \mathrm{V}\left(\lambda f,\ \lambda_0 f_0\right) = \int \log^2\frac{\lambda f(x)}{\lambda_0 f_0(x)}\lambda f(x)\mathrm{d}x. \tag{11}$$

The proof will be presented in the appendix.

Remark 1 The Hellinger process (here deterministic) of order $\frac{1}{2}$ for *continuous* observations of X on an interval $[0,\ t]$ is given by Jacod and Shiryaev (2003, Sects. IV.3 and IV.4a)

$$h_t = \frac{t}{2}\int\left(\sqrt{\lambda f(x)} - \sqrt{\lambda_0 f_0(x)}\right)^2 \mathrm{d}x = h_1 t,$$

from which it follows that $h^2(\mathbb{R}^t_{\lambda,f},\ \mathbb{R}^t_{\lambda_0,f_0}) = 2 - 2\exp(-h_t)$, whose derivative in $t = 0$ is the same as in (9) and thus equal to $2h_1$. For the Kullback–Leibler divergence and the discrepancy V similar assertions hold. These observations have the following heuristic explanation. For $\Delta \to 0$, there is no big difference between observing the path of X over the interval $[0,\ \Delta]$ and X_Δ, as the probability of $\{N_\Delta \geq 2\}$ is small (of order Δ^2).

In order to determine the posterior contraction rate in our problem, we now specify suitable neighbourhoods A_n of $(\lambda_0,\ f_0)$, for which this will be done. Let $M > 0$ be a constant and let $\{\varepsilon_n\}$ be a sequence of positive numbers, such that $\varepsilon_n \to 0$ as $n \to \infty$. Let

$$h^\Delta\left(\mathbb{Q}_0,\ \mathbb{Q}_1\right) = \frac{1}{\sqrt{\Delta}}h\left(\mathbb{Q}_0,\ \mathbb{Q}_1\right),$$

be a rescaled Hellinger distance. Lemma 1 suggests that this is the right scaling to use. Introduce the complements of the Hellinger-type neighbourhoods of $(\lambda_0,\ f_0)$,

$$A\left(\varepsilon_n,\ M\right) = \left\{(\lambda,\ f)\colon h^\Delta\left(\mathbb{Q}^\Delta_{\lambda_0,f_0},\ \mathbb{Q}^\Delta_{\lambda,f}\right) > M\varepsilon_n\right\}.$$

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

$$\Pi\left(A\left(\varepsilon_n,\ M\right)|\mathcal{Z}_n^\Delta\right) \to 0, \tag{12}$$

in $\mathbb{Q}^{\Delta,n}_{\lambda_0,f_0}$-probability as $n \to \infty$. Our goal in this section is to determine the 'fastest' rate at which ε_n is allowed to tend to zero, while not violating (12).

We will assume that the observations are generated from a CPP that satisfies the following assumption.

Assumption 1 (i) λ_0 is in a compact set $[\underline{\lambda}, \overline{\lambda}] \subset (0, \infty)$;
(ii) The true density f_0 is a location mixture of normal densities, i.e.,

$$f_0(x) = f_{H_0,\sigma_0}(x) = \int \phi_{\sigma_0}(x - z)\mathrm{d}H_0(z),$$

for some fixed distribution H_0 and a constant $\sigma_0 \in [\underline{\sigma}, \overline{\sigma}] \subset (0, \infty)$. Furthermore, for some $0 < \kappa_0 < \infty$, $H_0[-\kappa_0, \kappa_0] = 1$, i.e., H_0 has compact support.

The more general location-scale mixtures of normal densities,

$$f_0(x) = f_{H_0,K_0}(x) = \iint \phi_\sigma(x - z)\mathrm{d}H_0(z)\mathrm{d}K_0(\sigma),$$

possess even better approximation properties than the location mixtures of the normals (here H_0 and K_0 are distributions) and could also be considered in our setup. However, this would lead to additional technical complications, which could obscure essential contributions of our work.

For obtaining posterior contraction rates we need to make some assumptions on the prior.

Assumption 2 (i) The prior on λ, Π_1, has a density π_1 (with respect to the Lebesgue measure) that is supported on the finite interval $[\underline{\lambda}, \overline{\lambda}] \subset (0, \infty)$ and is such that

$$0 < \underline{\pi}_1 \le \pi_1(\lambda) \le \overline{\pi}_1 < \infty, \quad \lambda \in [\underline{\lambda}, \overline{\lambda}], \tag{13}$$

for some constants $\underline{\pi}_1$ and $\overline{\pi}_1$;
(ii) The base measure α of the Dirichlet process prior D_α has a continuous density on an interval $[-\kappa_0 - \zeta, \kappa_0 + \zeta]$, with κ_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

$$\alpha(|z| > t) \lesssim e^{-b|t|^\delta}, \tag{14}$$

with some constants $b > 0$ and $\delta > 0$;
(iii) The prior on σ, Π_3, is supported on the interval $[\underline{\sigma}, \overline{\sigma}] \subset (0, \infty)$ and is such that its density π_3 with respect to the Lebesgue measure satisfies

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

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

Assumptions 1 and 2 parallel those given in Ghosal and Vaart (2001) in the context of non-parametric Bayesian density estimation using the Dirichlet location mixture of normal densities as a prior. We refer to that paper for an additional discussion.

The following is our main result. Note that it covers both the case of high frequency observations ($\Delta \to 0$) and observations with fixed intersampling intervals. We use Π to denote the posterior on (λ, f).

Theorem 1 *Under Assumptions 1 and 2, 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

$$\Pi\left(A\left(\varepsilon_n,\ M\right)\Big|\mathcal{Z}_n^{\Delta}\right)\to 0,$$

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

For fixed Δ (w.l.o.g. one may then assume $\Delta = 1$) the posterior contraction rate in Theorem 1 reduces to $\varepsilon_n = \frac{\log^{\kappa}(n)}{\sqrt{n}}$. We also see that the posterior contraction rate is controlled by the parameter δ of the tail behaviour in (14). Note that if (14) is satisfied for some $\delta > 4$, it is also automatically satisfied for all $0 < \delta \le 4$. The stronger the decay rate in (14), the better the contraction rate, but all $\delta \ge 4$ give the same value $\kappa = 1$. The best possible posterior contraction rate in Theorem 1 for minimal δ is obtained for $\delta = 4$. In the proof in Sect. 5 we can therefore assume that $\delta \le 4$.

As on p. 1239 in Ghosal and Vaart (2001) and similar Corollary 5.1 there, Theorem 1 implies existence of a point estimate of $(\lambda_0,\ f_0)$ with a frequentist convergence rate ε_n. The (frequentist) minimax convergence rate for estimation of $k_{\lambda,f}^{\Delta}$ relative to the Hellinger distance is unknown in our problem, but an analogy to Ibragimov and Khas'minskiĭ (1982) suggests that up to a logarithmic factor it should be of order $\sqrt{n\Delta}$ (cf. Ghosal and Vaart 2001, p. 1236). The logarithmic factor is insignificant for all practical purposes. The convergence rate of an estimator of the Lévy density with loss measured in the L_2-metric in a more general Lévy model than the CPP model is $(n\Delta)^{-\beta/(2\beta+1)}$, whenever the target density is Sobolev smooth of order β (cf. Comte and Genon-Catalot 2011). Our contraction rate is hence, roughly speaking, a limiting case of the convergence in Comte and Genon-Catalot (2011) for $\beta\to\infty$.

4 Algorithms for drawing from the posterior

In this section we discuss computational methods for drawing from the distribution of the pair $(\lambda,\ f)$, conditional on $\mathcal{X}_n^{\Delta}$ (or equivalently: conditional on $\mathcal{Z}_n^{\Delta}$). In the following there is no specific need that the observational times are equidistant. We will assume observations at times $0 < t_1 < \cdots < t_n$ and set $\Delta_i = t_i - t_{i-1}$ $(1 \le i \le n)$. Further, for consistency with notation following shortly, we set $z_i = X_{t_i} - X_{t_{i-1}}$ and $z = (z_1, \ldots, z_n)$. We will use "Bayesian notation" throughout and write p for a probability density of mass function and use π similarly for a prior density or mass function.

In general, it is infeasible to generate independent realisations of the posterior distribution of $(\lambda,\ f)$. To see this: from (4) one obtains that the conditional density of a nonzero increment z on a time interval of length Δ is given by

$$p(z\mid\lambda,\ f) = \frac{e^{-\lambda\Delta}}{1-e^{-\lambda\Delta}}\sum_{k=1}^{\infty}\frac{(\lambda\Delta)^k}{k!}f^{*k}(z), \tag{15}$$

which generally is rather intractable due to the infinite weighted sum of convolutions. We specialise to the case where the jump size distribution is a mixture of $J \ge 1$ Gaussians. The richness and versatility of the class of finite normal mixtures is convincingly demonstrated in Marron and Wand (1992).

Hence, we assume

$$f(\cdot) = \sum_{j=1}^{J}\rho_j\phi\left(\cdot;\ \mu_j,\ 1/\tau\right),\quad \sum_{j=1}^{J}\rho_j = 1, \tag{16}$$

where $\phi(\cdot;\ \mu,\ \sigma^2)$ denotes the density of a random variable with $\mathcal{N}(\mu,\ \sigma^2)$ distribution. Note that in (16) we parametrise the density with the *precision* τ. In the "simple" case $J = 2$ the convolution density of k independent jumps is given by

$$f^{*k}(\cdot) = \sum_{\ell=0}^{k} \binom{k}{\ell} \rho_1^{\ell} \rho_2^{k-\ell} \phi\left(\cdot;\ \ell\mu_1 + (k-\ell)\mu_2;\ k/\tau\right).$$

Plugging this expression into Eq. (15) confirms the intractable form of $p(z \mid \lambda,\ f)$.

We will introduce auxiliary variables to circumvent the intractable form of the likelihood. In case the CPP is observed *continuously*, the problem is much easier as now the continuous time likelihood on an interval $[0,\ T]$ is known to be (Shreve 2008, Theorem 11.6.7)

$$\lambda^{|V|} e^{-\lambda T} \prod_{i \in V} f\left(J_i\right),$$

where the T_i are the jump times of the CPP, J_i the corresponding jump sizes and $V = \{i : T_i \leq T\}$. The tractability of the continuous time likelihood naturally suggests the construction of a data augmentation scheme. Denote the values of the CPP in between times t_{i-1} and t_i by $x_{(i-1,i)}$. We will refer to $x_{(i-1,i)}$ as the missing values on the ith segment. Set

$$x^{mis} = \left\{x_{(i-1,i)},\ 1 \leq i \leq n\right\}.$$

A data augmentation scheme now consists of augmenting auxiliary variables x^{mis} to $(\lambda,\ f)$ and constructing a Markov chain that has $p(x^{mis},\ \lambda,\ f \mid z)$ as invariant distribution. More specifically, a standard implementation of this algorithm consists of the following steps:

(1) Initialise x^{mis}.
(2) Draw $(\lambda,\ f) \mid (x^{mis},\ z)$.
(3) Draw $x^{mis} \mid (\lambda,\ f,\ z)$.
(4) Repeat steps 2 and 3 many times.

Under weak conditions, the iterates for $(\lambda,\ f)$ are (dependent) draws from the posterior distribution. Step 3 entails generating compound Poisson bridges. By the Markov property, bridges on different segments can be drawn independently. Data augmentation has been used in many Bayesian computational problems, see, e.g., Tanner and Wong (1987). The outlined scheme can be applied to the problem at hand, but we explain shortly that imputation of complete CPP-bridges (which is nontrivial) is unnecessary and we can do with less imputation, thereby effectively reducing the state space of the Markov chain.

As we assume that the jumps are drawn from a non-atomic distribution, imputation is only necessary on segments with nonzero increments. For this reason we let

$$\mathcal{I} = \{i \in \{1, \ldots, n\} : z_i \neq 0\},$$

denote the set of observations with nonzero jump sizes and define the number of segments with nonzero jumps to be $I = |\mathcal{I}|$.

4.1 Auxiliary variables

Note that if $Y \sim f$ with f as in (16), then Y can be simulated by first drawing its label L, which equals j with probability ρ_j, and next drawing from the $N(\mu_L,\ 1/\tau)$ distribution. Knowing the labels, sampling the jumps conditional on their sum being z is much easier compared to the case with unknown labels. Adding auxiliary variables as labels is a standard trick used for inference in mixture models (see, e.g., Diebolt and Robert 1994; Richardsen

and Green 1997). For the problem at hand, we can do with even less imputation: all we need to know is the number of jumps of each type on every segment with nonzero jump size. For $i \in \mathcal{I}$ and $j \in \{1, \ldots, J\}$, let n_{ij} denote the number of jumps of type j on segment i. Denote the set of all auxiliary variables by $\mathbf{a} = \{a_i,\ i \in I\}$, where

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

In the following we will use the following additional notation: for $i = 1, \ldots, n,\ j = 1, \ldots, J$ we set

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

These are the number of jumps on the i-th segment, the total number of jumps of type j (summed over all segments) and the total number of jumps of all types, respectively.

4.2 Reparametrisation and prior specification

Instead of parametrising with $(\lambda,\ \rho_1, \ldots, \rho_J)$, we define

$$\psi_j = \lambda\rho_j, \quad j = 1, \ldots, J.$$

Then

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

The background of this reparametrisation is the observation that a compound Poisson random variable Z whose jumps are of J types can be decomposed as $Z = \sum_{j=1}^{J} Z_j$, where the Z_j are independent, compound Poisson random variables whose jumps are of type j only, and where the parameter of the Poisson random variable is ψ_j. In what follows we use $\theta = (\psi,\ \mu,\ \tau)$ with $\psi = (\psi_1, \ldots, \psi_J)$ and $\mu = (\mu_1, \ldots, \mu_J)$.

Denote the Gamma distribution with shape parameter α and rate β by $\mathcal{G}(\alpha,\ \beta)$. We take priors

$$\begin{aligned} \psi_1, \ldots, \psi_J &\overset{\text{iid}}{\sim} \mathcal{G}\,(\alpha_0,\ \beta_0)\,, \\ \mu \mid \tau &\sim \mathcal{N}\left([\xi_1, \ldots, \xi_J]',\ I_{J\times J}(\tau\kappa)^{-1}\right), \\ \tau &\sim \mathcal{G}\,(\alpha_1,\ \beta_1)\,, \end{aligned}$$

with positive hyperparameters $(\alpha_0,\ \beta_0,\ \alpha_1,\ \beta_1,\ \kappa)$ fixed.

4.3 Hierarchical model and data augmentation scheme

We construct a Metropolis–Hastings algorithm to draw from

$$p(\theta,\ \mathbf{a} \mid z) = \frac{p(\theta,\ z,\ \mathbf{a})}{p(z)}.$$

For an index $i \in \mathcal{I}$ we set $\mathbf{a}_{-i} = \{a_j,\ j \in \mathcal{I} \setminus \{i\}\}$. The two main steps of the algorithm are:

(i) *Update segments* for each segment $i \in \mathcal{I}$, draw a_i conditional on $(\theta,\ z,\ \mathbf{a}_{-i})$;
(ii) *Update parameters* draw θ conditional on $(z,\ \mathbf{a})$.

Compared to the full data augmentation scheme discussed previously, the present approach is computationally much cheaper as the amount of imputation scales with the number of segments that need imputation. If the time in between observations is fixed and equal to Δ, then the expected number of segments for imputation equals $n(1-e^{-\lambda\Delta})$, which is for small Δ approximately proportional to $n\Delta\lambda$.

Denote the Poisson distribution with mean λ by $\mathcal{P}(\lambda)$. Including the auxiliary variables, we can write the observation model as a *hierarchical model*

$$\begin{aligned} z_i \mid a_i,\ \mu,\ \tau &\overset{\text{ind}}{\sim} N\left(a_i'\mu,\ n_i/\tau\right), \\ n_{ij} \mid \psi &\overset{\text{ind}}{\sim} \mathcal{P}\left(\psi_j\Delta_i\right), \\ (\psi,\ \mu,\ \tau) &\sim \pi(\psi,\ \mu,\ \tau) \end{aligned} \tag{17}$$

(with $i \in \{1,\ldots,n\}$ and $j \in \{1,\ldots,J\}$). This implies

$$p(\theta,\ z,\ \mathbf{a}) = \pi(\theta) \times \prod_{i=1}^{n}\left(\phi\left(z_i;\ a_i'\mu,\ n_i/\tau\right)\prod_{j=1}^{J} e^{-\psi_j\Delta_i}\frac{(\psi_j\Delta_i)^{n_{ij}}}{n_{ij}!}\right).$$

4.4 Updating segments

Updating the ith segment requires drawing from

$$p\left(a_i \mid \theta,\ z,\ \mathbf{a}_{-i}\right) \propto \phi\left(z_i;\ a_i'\mu,\ n_i/\tau\right)\prod_{j=1}^{J}\frac{(\psi_j\Delta_i)^{n_{ij}}}{n_{ij}!}.$$

We do this with a Metropolis–Hastings step. First we draw a proposal n_i° (for n_i) from a $\mathcal{P}(\lambda\Delta_i)$ distribution, conditioned to have nonzero outcome. Next, we draw

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

where $\mathcal{MN}$ denotes the multinomial distribution. Hence the proposal density equals

$$\begin{aligned} q\left(n_{i1}^\circ,\ldots,n_{iJ}^\circ \mid \theta\right) &= \frac{e^{-\lambda\Delta_i}}{1-e^{-\lambda\Delta_i}}\frac{(\lambda\Delta_i)^{n_i^\circ}}{n_i^\circ!}\binom{n_i^\circ}{n_{i1}^\circ\ldots n_{iJ}^\circ}\prod_{j=1}^{J}\left(\psi_j/\lambda\right)^{n_{ij}^\circ} \\ &= \frac{e^{-\lambda\Delta_i}}{1-e^{-\lambda\Delta_i}}\prod_{j=1}^{J}\frac{(\psi_j\Delta_i)^{n_{ij}^\circ}}{n_{ij}^\circ!}. \end{aligned}$$

The acceptance probability for the proposal n° equals $1 \wedge A$, with

$$A = \frac{\phi(z_i;\ (a_i^\circ)'\mu,\ n_i^\circ/\tau)}{\phi(z_i;\ a_i'\mu,\ n_i/\tau)}.$$

4.5 Updating parameters

The proof of the following lemma is given in Appendix 1.

Lemma 2 *Conditional on* $\mathbf{a}$, $\psi_1,\ldots,\psi_J$ *are independent and*

$$\psi_j \mid \mathbf{a} \sim \mathcal{G}\left(\alpha_0+s_j,\ \beta_0+T\right).$$

Furthermore,

$$\begin{aligned} \mu \mid \tau, z, \mathbf{a} &\sim \mathcal{N}\left(P^{-1}q,\ \tau^{-1}P^{-1}\right), \\ \tau \mid z, \mathbf{a} &\sim \mathcal{G}\left(\alpha_1 + I/2,\ \beta_1 + \left(R - q'P^{-1}q\right)/2\right), \end{aligned} \tag{18}$$

where P is the symmetric $J \times J$ matrix with elements

$$P = \kappa I_{J\times J} + \tilde{P} \quad \tilde{P}_{j,k} = \sum_{i\in\mathcal{I}} n_i^{-1} n_{ij} n_{ik}, \quad j, k \in \{1, \ldots, J\}, \tag{19}$$

q is the J-dimensional vector with

$$q_j = \kappa\xi_j + \sum_{i\in\mathcal{I}} n_i^{-1} n_{ij} z_i, \tag{20}$$

$R > 0$ is given by

$$R = \kappa \sum_{j=1}^{J} \xi_j^2 + \sum_{i\in\mathcal{I}} n_i^{-1} z_i^2, \tag{21}$$

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

Remark 2 If for some $j \in \{1, \ldots, J\}$ we have $s_j = 0$ (no jumps of type j), then the matrix $\tilde{P}$ is singular. However, adding $\kappa I_{J\times J}$ ensures invertibility of P.

4.6 Numerical illustrations

The first two examples concern mixtures of two normal distributions We simulated $n = 5.000$ segments with $\Delta = 1$, $\mu_1 = 2$, $\mu_2 = -1$ and $\tau = 1$. For the prior-hyperparameters we took $\alpha_0 = \beta_0 = \alpha_1 = \beta_1 = 1$, $\xi_1 = \xi_2 = 0$ and $\kappa = 1$.

The results for $\lambda\Delta = 1$, $\rho_1 = 0.8$, $\rho_2 = 0.2$ and hence $\psi_1 = 0.8$ and $\psi_2 = 0.2$ are shown in Fig. 1. The densities obtained from the posterior mean of the parameter estimates and the true density are shown in Fig. 2. The average acceptance probability for updating the segments was 51%.

The results for $\lambda\Delta = 3$, $\rho_1 = 0.8$, $\rho_2 = 0.2$ and hence $\psi_1 = 2.4$ and $\psi_2 = 0.6$ are shown in Fig. 3. The densities obtained from the posterior mean of the parameter estimates and the true density are shown in Fig. 4. The average acceptance probability for updating the segments was 41%. Observe that the autocorrelation functions of the iterations of the ψ_i in the second case display a much slower decay.

We also assessed the performance of our method on a more complicated example where we took a mixture of four normals. Here $\Delta = 1$, $(\mu_1, \mu_2, \mu_3, \mu_4) = (-1, 0, 0.8, 2)$, $(\psi_1, \psi_2, \psi_3, \psi_4) = (0.3, 0.4, 0.2, 0.1)$ (hence $\lambda = 1$) and $\tau^{-1} = 0.09$. The results obtained after simulating $n = 10.000$ segments are shown in Figs. 5 and 6.

Mixtures of normals need not be multimodal and can also yield skew densities. As an example, we consider the case where $(\mu_1, \mu_2) = (0, 2)$, $(\psi_1, \psi_2) = (1.5, 0.5)$ (hence $\lambda = 2$) and $\tau = 1$. Data were generated and discretely sampled with $\Delta = 1$ and $n = 5.000$ segments. A plot of the posterior mean is shown in Fig. 7.

4.7 Discussion

As can be seen from the autocorrelation plots, mixing of the chain deteriorates when $\lambda\Delta$ increases. As the focus in this article is on high frequency data, where there are on average only

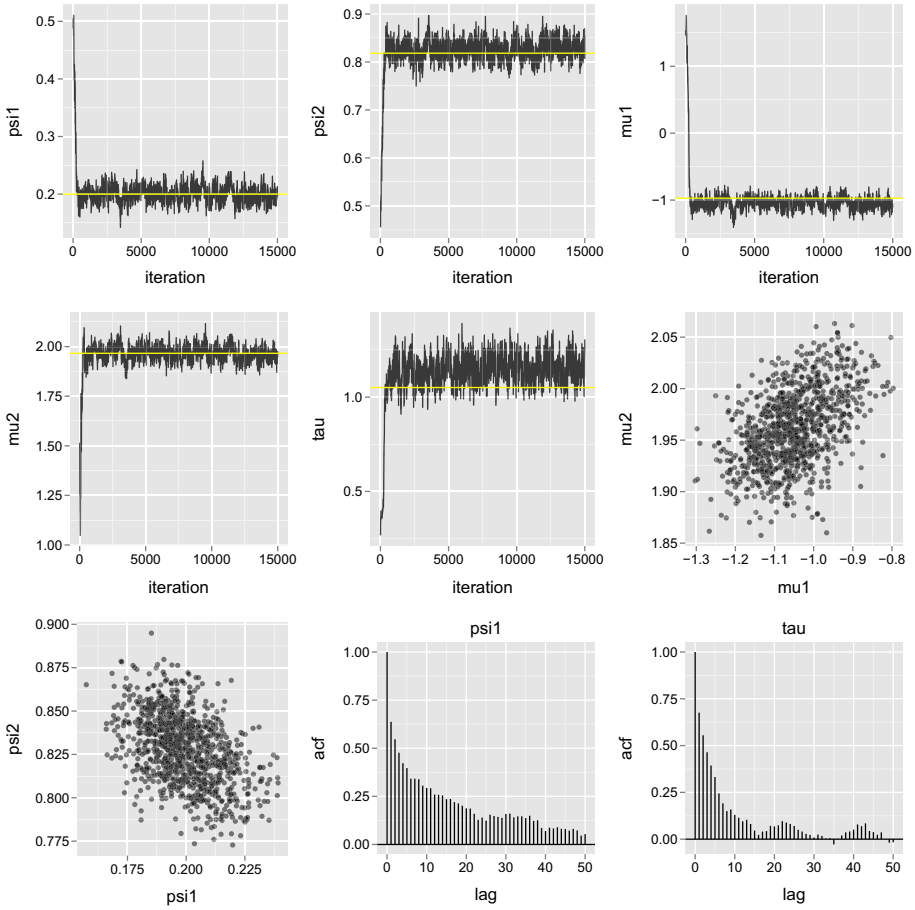

Fig. 1 Results for $\lambda = 1$ using 15.000 MCMC iterations. The trace plots show all iterations; in the other plots the first 5.000 iterations are treated as burnin. The figures are obtained after subsampling the iterates, where only each fifth iterate was saved. The *horizontal yellow lines* are obtained from computing the posterior mean of θ based on the true auxiliary variables on all segments

a few jumps in between observations, we do not go into details on improving the algorithm. We remark that a non-centred parametrisation (see for instance Papaspiliopoulos et al. 2007) may give more satisfactory results when $\lambda\Delta$ is large. A non centred parametrisation can be obtained by changing the hierarchical model in (17). Denote by F_λ^{-1} the inverse cumulative distribution function of the $\mathcal{P}(\lambda)$ distribution. Let u_{ij} $(i = 1, \ldots, n$ and $j = 1, \ldots, J)$ be a sequence of independent $U(0,\ 1)$ random variables and set $u = \{u_{ij},\ i = 1, \ldots, n,\ j = 1, \ldots, J\}$. By considering the hierarchical model

$$
\begin{aligned}
z_i \mid u,\ \mu,\ \tau &\overset{\text{ind}}{\sim} N\left(\sum_{j=1}^{J} \mu_j F^{-1}_{\psi_j \Delta_i}\left(u_{ij}\right),\ \tau^{-1} \sum_{j=1}^{J} F^{-1}_{\psi_j \Delta_i}\left(u_{ij}\right)\right), \\
u_{ij} &\overset{\text{iid}}{\sim} U(0,\ 1), \\
(\psi,\ \mu,\ \tau) &\sim \pi(\psi,\ \mu,\ \tau),
\end{aligned}
\tag{22}
$$

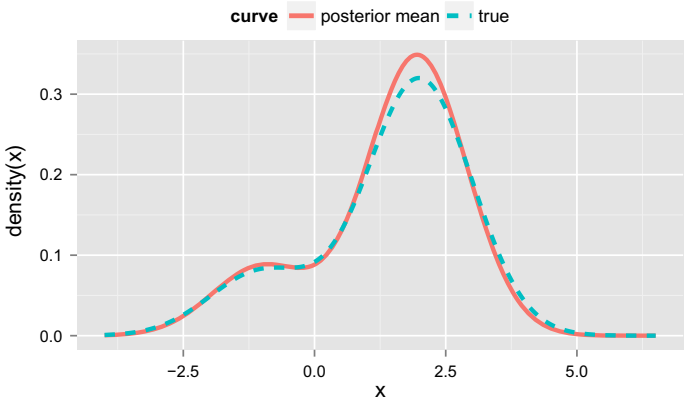

Fig. 2 Results for $\lambda = 1$; the first 5.000 iterations are treated as burnin. Shown are the true jump size density and the density obtained from the posterior mean of the non-burnin iterates

($i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, J\}$), ψ can be updated using a Metropolis–Hastings step. In this way $\{n_{ij}\}$ and ψ are updated simultaneously.

Another option is to integrate out (μ, τ) from $p(\theta, z, \mathbf{a})$. In this model it is even possible to integrate out ψ as well. In that case only the auxiliary variables $\mathbf{a}$ have to be updated. Yet another method to improve the efficiency of the algorithm is to use ideas from parallel tempering (Cf. Brooks et al. 2011, Chap. 11).

5 Proof of Theorem 1

There are a number of general results in Bayesian nonparametric statistics, such as the fundamental Theorem 2.1 in Ghosal et al. (2000) and Theorem 2.1 in Ghosal and Vaart (2001), which allow determination of the posterior contraction rates through checking certain conditions, but none of these results is easily and directly applicable in our case. The principle bottleneck is that a main assumption underlying these theorems is sampling from a fixed distribution, whereas in our high frequency setting, the distributions vary with Δ. Therefore, for the clarity of exposition in the proof of our main theorem we will choose an alternative path, which consists in mimicking the main steps of the proof of Theorem 2.1, involving judiciously chosen statistical tests, as in Ghosal et al. (2000), while also employing some results on the Dirichlet location mixtures of normal densities from Ghosal and Vaart (2001). However, a significant part of technicalities we will encounter are characteristic of the decompounding problem only.

Throughout this section we assume that Assumptions 1 and 2 hold. Furthermore, in view of the discussion that followed Theorem 1 we will without loss of generality assume that $0 < \delta \leq 4$. All the technical lemmas used in this section are collected in the appendices.

We start with the decomposition

$$\Pi\left(A\left(\varepsilon_n, M\right)|\mathcal{Z}_n^{\Delta}\right) = \Pi\left(A\left(\varepsilon_n, M\right)|\mathcal{Z}_n^{\Delta}\right)\phi_n + \Pi\left(A\left(\varepsilon_n, M\right)|\mathcal{Z}_n^{\Delta}\right)(1-\phi_n) =: \mathrm{I}_n + \mathrm{II}_n, \tag{23}$$

where $0 \leq \phi_n \leq 1$ is a sequence of tests based on observations $\mathcal{Z}_n^{\Delta}$ and with properties to be specified below. The idea is to show that the terms on the right-hand side of the above display separately converge to zero in probability. The tests ϕ_n allow one to control the behaviour of

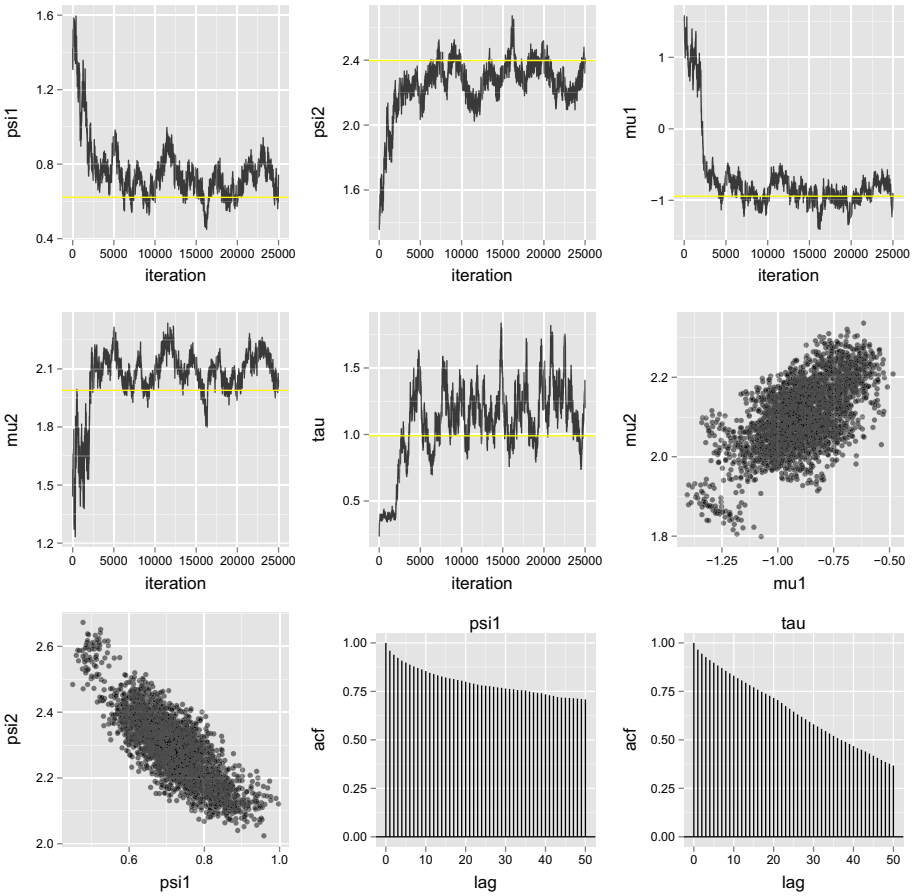

Fig. 3 Results for $\lambda = 3$ using 25.000 MCMC iterations. The trace plots show all iterations; in the other plots the first 10.000 iterations are treated as burnin. The figures are obtained after subsampling the iterates, where only each fifth iterate was saved. The *horizontal yellow lines* are obtained from computing the posterior mean of θ based on the true auxiliary variables on all segments

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 $(\lambda_0,\ f_0)$.

5.1 Construction of tests

The next lemma is an adaptation of Theorem 7.1 from Ghosal et al. (2000) to decompounding. A proof is given in the appendix. We use the notation $D(\varepsilon,\ A,\ d)$ to denote the ε-packing number of a set A in a metric space with metric d, applied in our case with d the scaled Hellinger metric h^{Δ}.

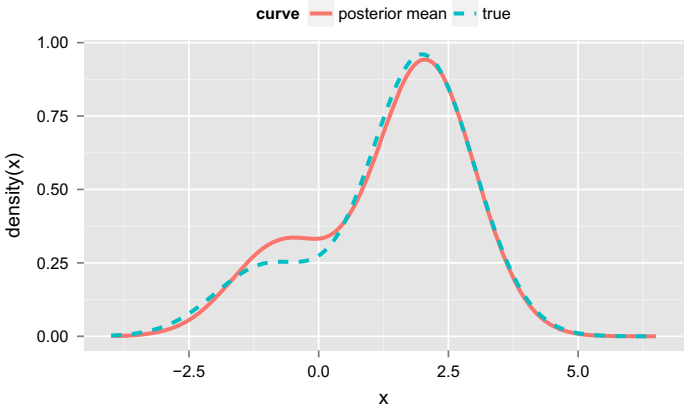

Fig. 4 Results for $\lambda = 3$; the first 10.000 iterations are treated as burnin. Shown are the true jump size density and the density obtained from the posterior mean of the non-burnin iterates

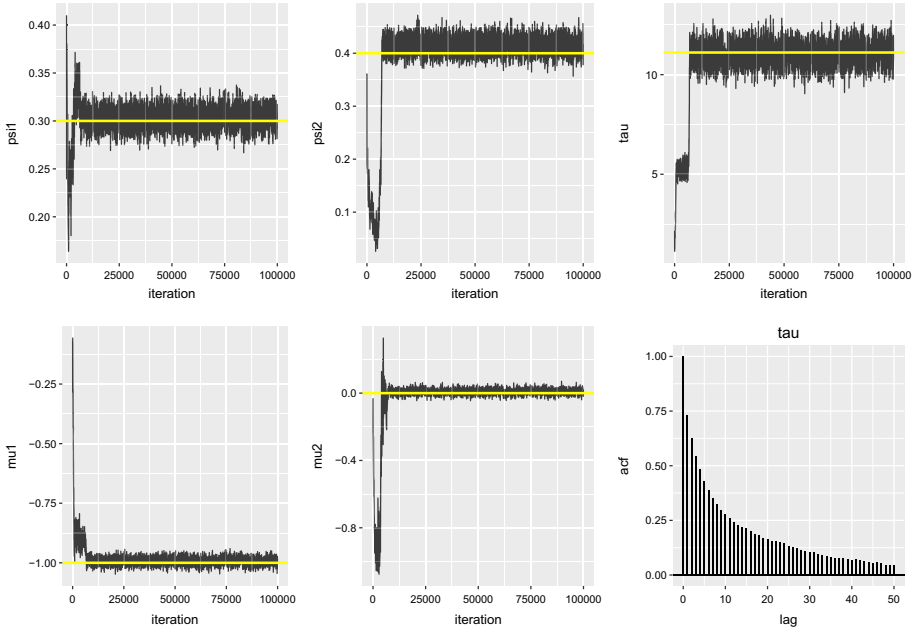

Fig. 5 Results for the example with a mixture of four normals using 100.000 MCMC iterations. The trace plots show all iterations, in the autocorrelation plot the first 20.000 iterations are treated as burnin. The figures are obtained after subsampling the iterates, where only each fifth iterate was saved. The *horizontal yellow lines* indicate true values. The results for the other parameters are similar and therefore not displayed

Lemma 3 *Let $\mathcal{Q}$ be an arbitrary set of probability measures $\mathbb{Q}^{\Delta}_{\lambda,f}$. Suppose for some non-increasing function $D(\varepsilon)$, some sequence $\{\varepsilon_n\}$ of positive numbers and every $\varepsilon > \varepsilon_n$,*

$$D\left(\frac{\varepsilon}{2},\ \left\{\mathbb{Q}^{\Delta}_{\lambda,f} \in \mathcal{Q}\colon \varepsilon \le h^{\Delta}\left(\mathbb{Q}^{\Delta}_{\lambda_0,f_0},\ \mathbb{Q}^{\Delta}_{\lambda,f}\right) \le 2\varepsilon\right\},\ h^{\Delta}\right) \le D(\varepsilon). \tag{24}$$

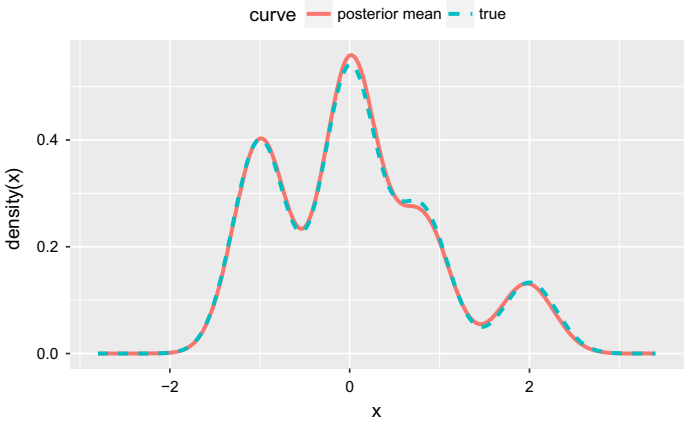

Fig. 6 Results for the example with a mixture of four normals; the first 20.000 iterations are treated as burnin. Shown are the true jump size density and the density obtained from the posterior mean of the non-burnin iterates

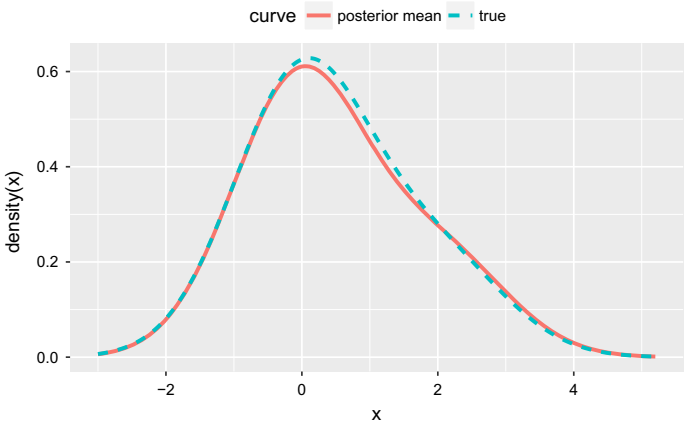

Fig. 7 Results for the example with a skew density; the first 20.000 iterations are treated as burnin. Shown are the true jump size density and the density obtained from the posterior mean of the non-burnin iterates

Then for every $\varepsilon > \varepsilon_n$ there exists a sequence of tests $\{\phi_n\}$ (depending on $\varepsilon > 0$), such that

$$\mathbb{E}_{\lambda_0, f_0}[\phi_n] \leq D(\varepsilon) \exp\left(-Kn\Delta\varepsilon^2\right) \frac{1}{1 - \exp(-Kn\Delta\varepsilon^2)},$$
$$\sup_{\{\mathbb{Q}^{\Delta}_{\lambda,f} \in \mathcal{Q}: h^{\Delta}(\mathbb{Q}^{\Delta}_{\lambda_0,f_0}, \mathbb{Q}^{\Delta}_{\lambda,f}) > \varepsilon\}} \mathbb{E}_{\lambda, f}[1 - \phi_n] \leq \exp\left(-Kn\Delta\varepsilon^2\right),$$

where $K > 0$ is a universal constant.

In the proofs of Propositions 1 and 2 we need the inequalities below. There exists a constant $\overline{C} \in (0, \infty)$ depending on $\underline{\lambda}$ and $\overline{\lambda}$ only, such that for all $\lambda_1, \lambda_2 \in [\underline{\lambda}, \overline{\lambda}]$ and f_1, f_2 it holds that

$$\mathrm{K}\left(\mathbb{Q}^{\Delta}_{\lambda_1,f_1},\,\mathbb{Q}^{\Delta}_{\lambda_2,f_2}\right)\le \overline{C}\Delta\left(\mathrm{K}\left(\mathbb{P}_{f_1},\,\mathbb{P}_{f_2}\right)+|\lambda_1-\lambda_2|^2\right), \tag{25}$$

$$\mathrm{V}\left(\mathbb{Q}^{\Delta}_{\lambda_1,f_1},\,\mathbb{Q}^{\Delta}_{\lambda_2,f_2}\right)\le \overline{C}\Delta\left(\mathrm{V}\left(\mathbb{P}_{f_1},\,\mathbb{P}_{f_2}\right)+\mathrm{K}\left(\mathbb{P}_{f_1},\,\mathbb{P}_{f_2}\right)+|\lambda_1-\lambda_2|^2\right), \tag{26}$$

$$h\left(\mathbb{Q}^{\Delta}_{\lambda_1,f_1},\,\mathbb{Q}^{\Delta}_{\lambda_2,f_2}\right)\le \overline{C}\sqrt{\Delta}\left(|\lambda_1-\lambda_2|+h\left(\mathbb{P}_{f_1},\,\mathbb{P}_{f_2}\right)\right). \tag{27}$$

These inequalities can be proven in the same way as Lemma 1 in Gugushvili et al. (2015).

Let ε_n be as in Theorem 1. Throughout, $\overline{C}$ denotes the above constant. For a constant $L>0$ define the sequences $\{a_n\}$ and $\{\eta_n\}$ by

$$a_n=L\log^{2/\delta}\left(\frac{1}{\eta_n}\right),\quad \eta_n=\frac{\varepsilon_n}{4\overline{C}}.$$

We will show that inequality (24) holds true for every $\varepsilon=M\varepsilon_n$ with $M>2$ and the set of measures $\mathcal{Q}$ equal to

$$\mathcal{Q}_n=\left\{\mathbb{Q}^{\Delta}_{\lambda,f_{H,\sigma}}:\lambda\in[\underline{\lambda},\,\overline{\lambda}],\ H\left[-a_n,\,a_n\right]\ge 1-\eta_n,\ \sigma\in[\underline{\sigma},\,\overline{\sigma}]\right\}.$$

As a first step, note that we have

$$\begin{aligned}\log D\left(\frac{\varepsilon}{2},\,\mathcal{Q}_n,\,h^{\Delta}\right)&\le \log D\left(\varepsilon_n,\,\mathcal{Q}_n,\,h^{\Delta}\right)\\&\le \log N\left(\frac{\varepsilon_n}{2},\,\mathcal{Q}_n,\,h^{\Delta}\right)=\log N\left(\frac{\varepsilon_n\sqrt{\Delta}}{2},\,\mathcal{Q}_n,\,h\right),\end{aligned} \tag{28}$$

where $N\left(\frac{\varepsilon_n\sqrt{\Delta}}{2},\,\mathcal{Q}_n,\,h\right)$ is the covering number of the set $\mathcal{Q}_n$ with h-balls of size $\varepsilon_n\sqrt{\Delta}/2$. The first inequality in (28) follows from assuming $M>2$. For bounding the righthand side in (28), we have the following proposition.

Proposition 1 *We have*

$$\log N\left(\frac{\varepsilon_n\sqrt{\Delta}}{2},\,\mathcal{Q}_n,\,h\right)\lesssim\log^{4/\delta+1}\left(\frac{1}{\varepsilon_n}\right). \tag{29}$$

Proof Define

$$\mathcal{F}_n=\left\{f_{H,\sigma}:H\left[-a_n,\,a_n\right]\ge 1-\eta_n,\ \sigma\in[\underline{\sigma},\,\overline{\sigma}]\right\}.$$

Let $\{\lambda_i\}$ be centres of the balls from a minimal covering of $[\underline{\lambda},\,\overline{\lambda}]$ with $|\cdot|$-balls of size η_n. Let $\{f_j\}$ be centres of the balls from a minimal covering of $\mathcal{F}_n$ with h-balls of size η_n. For any $\mathbb{Q}_{\lambda,f_{H,\sigma}}\in\mathcal{Q}_n$, by (27) we have

$$h\left(\mathbb{Q}_{\lambda,f_{H,\sigma}},\,\mathbb{Q}_{\lambda_i,f_j}\right)\le\frac{\varepsilon_n\sqrt{\Delta}}{2},$$

by appropriate choices of i and j. It follows that

$$\log N\left(\frac{\varepsilon_n\sqrt{\Delta}}{2},\,\mathcal{Q}_n,\,h\right)\le\log N\left(\eta_n,\,[\underline{\lambda},\,\overline{\lambda}],\,|\cdot|\right)+\log N\left(\eta_n,\,\mathcal{F}_n,\,h\right).$$

Evidently,

$$\log N\left(\eta_n,\,[\underline{\lambda},\,\overline{\lambda}],\,|\cdot|\right)\lesssim\log\left(\frac{1}{\varepsilon_n}\right).$$

As we assume $\delta \leq 4$, we can apply the arguments in Ghosal and van der Vaart (2001, pp. 1251–1252) see in particular formulae (5.8)–(5.10) (cf. also Theorem 3.1 and Lemma A.3 there), which yield

$$\log N\left(\eta_n,\ \mathcal{F}_n,\ h\right) \lesssim \log^{4/\delta+1}\left(\frac{1}{\varepsilon_n}\right).$$

Combination of the above three inequalities implies the statement of the proposition.

An application of Proposition 1 to (28) gives

$$\log D\left(\frac{\varepsilon}{2},\ \mathcal{Q}_n,\ h^{\Delta}\right) \lesssim \log^{4/\delta+1}\left(\frac{1}{\varepsilon_n}\right) \leq c_1 n\Delta\varepsilon_n^2,$$

for some positive constant c_1. Here, the final inequality follows from our choice for ε_n. Hence, (24) is satisfied for

$$D(\varepsilon) = \exp\left(\left(c_1/M^2 - K\right) n\Delta\varepsilon^2\right).$$

By Lemma 3 there exist tests ϕ_n such that for all n large enough

$$\mathbb{E}_{\lambda_0,f_0}\left[\phi_n\right] \leq 2\exp\left(-\left(KM^2 - c_1\right) n\Delta\varepsilon_n^2\right), \tag{30}$$

$$\sup_{\{\mathbb{Q}^{\Delta}_{\lambda,f}\in\mathcal{Q}_n:\, h^{\Delta}(\mathbb{Q}^{\Delta}_{\lambda_0,f_0},\, \mathbb{Q}^{\Delta}_{\lambda,f})>\varepsilon\}} \mathbb{E}_{\lambda,f}\left[1-\phi_n\right] \leq \exp\left(-Kn\Delta M^2\varepsilon_n^2\right). \tag{31}$$

5.2 Bound on I_n in (23)

First note that by Eq. (30)

$$\mathbb{E}_{\lambda_0,f_0}\left[\mathrm{I}_n\right] \leq \mathbb{E}_{\lambda_0,f_0}\left[\phi_n\right] \leq 2\exp\left(-\left(KM^2 - c_1\right) n\Delta\varepsilon_n^2\right).$$

Chebyshev's inequality implies that I_n converges to zero in $\mathbb{Q}^{\Delta,n}_{\lambda_0,f_0}$-probability as $n\to\infty$, as soon as M is chosen so large that $KM^2 - c_1 > 0$. □

5.3 Bound on II_n

Now we consider II_n. We have

$$\mathrm{II}_n = \frac{\iint_{A(\varepsilon_n,M)} \mathcal{L}_n^{\Delta}(\lambda,\ f)\mathrm{d}\Pi_1(\lambda)\mathrm{d}\Pi_2(f)(1-\phi_n)}{\iint \mathcal{L}_n^{\Delta}(\lambda,\ f)\mathrm{d}\Pi_1(\lambda)\mathrm{d}\Pi_2(f)} =: \frac{\mathrm{III}_n}{\mathrm{IV}_n}.$$

We will show that the numerator III_n goes exponentially fast to zero, in $\mathbb{Q}^{\Delta,n}_{\lambda_0,f_0}$-probability, while the denominator IV_n is bounded from below by an exponential function, with $\mathbb{Q}^{\Delta,n}_{\lambda_0,f_0}$-probability tending to one, in such a way that the ratio of III_n and IV_n still goes to zero in $\mathbb{Q}^{\Delta,n}_{\lambda_0,f_0}$-probability.

5.3.1 Bounding III_n

As $1_{\{A(\varepsilon_n,M)\}} \leq 1_{\mathcal{Q}_n^c} + 1_{\{A(\varepsilon_n,M)\cap\mathcal{Q}_n\}}$ we have

$$\mathbb{E}_{\lambda_0,f_0}\left[\mathrm{III}_n\right] \leq \Pi\left(\mathcal{Q}_n^c\right) + \iint_{\mathcal{Q}_n\cap A(\varepsilon_n,M)} \mathbb{E}_{\lambda,f}\left[1-\phi_n\right]\mathrm{d}\Pi_1(\lambda)\mathrm{d}\Pi_2(f).$$

Here we applied Fubini's theorem to obtain the second term on the right-hand-side, which by (31) is bounded by $\exp(-KM^2n\Delta\varepsilon_n^2)$. Furthermore,

$$\Pi\left(\mathcal{Q}_n^c\right) = \Pi_2\left(H\left[-a_n,\ a_n\right] < 1-\eta_n,\ \sigma \in [\underline{\sigma},\ \overline{\sigma}]\right) \lesssim \frac{1}{\eta_n} e^{-ba_n^{\delta}},$$

where the last inequality is formula (5.11) in Ghosal and Vaart (2001). Hence

$$\mathbb{E}_{\lambda_0, f_0}\left[\mathrm{III}_n\right] \lesssim \frac{1}{\eta_n} e^{-ba_n^{\delta}} + \exp\left(-KM^2n\Delta\varepsilon_n^2\right). \tag{32}$$

5.3.2 Bounding IV_n

Recall $\mathrm{K}_\Delta = \mathrm{K}/\Delta$ and $\mathrm{V}_\Delta = \mathrm{V}/\Delta$. Let

$$B^\Delta\left(\varepsilon,\ (\lambda_0,\ f_0)\right) = \left\{(\lambda,\ f)\colon K^\Delta\left(\mathbb{Q}^\Delta_{\lambda_0,f_0},\ \mathbb{Q}^\Delta_{\lambda,f}\right) \le \varepsilon^2,\ V^\Delta\left(\mathbb{Q}^\Delta_{\lambda_0,f_0},\ \mathbb{Q}^\Delta_{\lambda,f}\right) \le \varepsilon^2\right\},$$

and

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

Note that $n\Delta\widetilde{\varepsilon}_n^2 \to \infty$ when $n \to \infty$.

We will use the following bound, an adaptation of Lemma 8.1 in Ghosal et al. (2000) to our setting, valid for every $\varepsilon > 0$ and $C > 0$,

$$\mathbb{Q}^{\Delta,n}_{\lambda_0,f_0}\left(\iint_{B^\Delta(\varepsilon,(\lambda_0,f_0))} \mathcal{L}_n(\lambda,\ f)\mathrm{d}\widetilde{\Pi}(\lambda,\ f) \le \exp\left(-(1+C)n\Delta\varepsilon^2\right)\right) \le \frac{1}{C^2n\Delta\varepsilon^2}, \tag{33}$$

where

$$\widetilde{\Pi}(\cdot) = \frac{\Pi(\cdot)}{\Pi(B^\Delta(\varepsilon,\ (\lambda_0,\ f_0)))},$$

is a normalised restriction of $\Pi(\cdot)$ to $B^\Delta(\varepsilon,\ (\lambda_0,\ f_0))$.

By virtue of (33), with $\mathbb{Q}^{\Delta,n}_{\lambda_0,f_0}$-probability tending to one, for any constant $C > 0$ we have

$$\begin{aligned}\mathrm{IV}_n &\ge \iint_{B^\Delta(\widetilde{\varepsilon}_n,(\lambda_0,f_0))} \mathcal{L}_n^\Delta(\lambda,\ f)\mathrm{d}\Pi_1(\lambda) \times \mathrm{d}\Pi_2(f) \\ &> \Pi\left(B^\Delta\left(\widetilde{\varepsilon}_n,\ (\lambda_0,\ f_0)\right)\right)\exp\left(-(1+C)n\Delta\widetilde{\varepsilon}_n^2\right).\end{aligned} \tag{34}$$

We will now work out the product probability on the right-hand side of this inequality.

Proposition 2 *It holds that*

$$\Pi\left(B^\Delta\left(\widetilde{\varepsilon}_n,\ \mathbb{Q}_{\lambda_0,f_0}\right)\right) \gtrsim \exp\left(-\bar{c}\log^2\left(\frac{1}{\widetilde{\varepsilon}_n}\right)\right),$$

for some constant $\bar{c}$.

Proof Let $0 < c \le 1/\sqrt{5\overline{C}}$ be a constant. Here $\overline{C}$ is the constant in (25) and (26). By these inequalities it is readily seen that

$$\left\{(\lambda,\ f)\colon K\left(\mathbb{P}_{f_0},\ \mathbb{P}_f\right) \le c^2\widetilde{\varepsilon}_n^2,\ V\left(\mathbb{P}_{f_0},\ \mathbb{P}_f\right) \le c^2\widetilde{\varepsilon}_n^2,\ |\lambda_0-\lambda|^2 \le c^2\widetilde{\varepsilon}_n^2\right\} \subset B^\Delta\left(\widetilde{\varepsilon}_n,\ \mathbb{Q}^\Delta_{\lambda_0,f_0}\right).$$

It then follows by the independence assumption on Π_1 and Π_2 that

$$\Pi\left(B^{\Delta}\left(\widetilde{\varepsilon}_n,\, \mathbb{Q}^{\Delta}_{\lambda_0,f_0}\right)\right) \geq \Pi_1\left(|\lambda_0-\lambda| \leq c\widetilde{\varepsilon}_n\right) \times \Pi_2\left(f\colon \mathrm{K}\left(\mathbb{P}_{f_0},\, \mathbb{P}_f\right) \leq c^2\widetilde{\varepsilon}_n^2,\, \mathrm{V}\left(\mathbb{P}_{f_0},\, \mathbb{P}_f\right) \leq c^2\widetilde{\varepsilon}_n^2\right).$$

For the first factor on the right-hand side we have by (13) that

$$\Pi_1\left(|\lambda_0-\lambda| \leq c\widetilde{\varepsilon}_n\right) \gtrsim \widetilde{\varepsilon}_n.$$

As far as the second factor is concerned, for some constants $\bar{c}_1,\ \bar{c}_2$ it is bounded from below by

$$\bar{c}_1 \exp\left(-\bar{c}_2 \log^2\left(\frac{1}{\widetilde{\varepsilon}_n}\right)\right),$$

by the same arguments as in inequality (5.17) in Ghosal and Vaart (2001). The result now follows by combining the two lower bounds.

Combining (34) with Proposition 2, with $\mathbb{Q}^{\Delta,n}_{\lambda_0,f_0}$-probability tending to one as $n \to \infty$, for any constant $C > 0$ we have

$$\mathrm{IV}_n > \exp\left(-(1+C)n\Delta\widetilde{\varepsilon}_n^2 - \bar{c}\log^2\left(\frac{1}{\widetilde{\varepsilon}_n}\right)\right). \tag{35}$$

We are now ready for showing the final steps of proving that II_n tends to zero in $\mathbb{Q}^{\Delta,n}_{\lambda_0,f_0}$-probability. Let G_n denote the set on which inequality (35) is true. Then by (32) we obtain

$$\mathbb{E}_{\lambda_0,f_0}\left[\mathrm{II}_n 1_{G_n}\right] \lesssim \exp\left((1+C)n\Delta\widetilde{\varepsilon}_n^2 + \bar{c}\log^2\left(\frac{1}{\widetilde{\varepsilon}_n}\right)\right) \times \left[\frac{1}{\eta_n}e^{-ba_n^{\delta}} + \exp\left(-KM^2n\Delta\varepsilon_n^2\right)\right].$$

Recall that $n\Delta\widetilde{\varepsilon}_n^2 = \log^2(n\Delta)$. Hence, the exponent in the first factor of this display is of order $\log^2(n\Delta)$. Furthermore $a_n^{\delta} = L^{\delta}\log^2(4\overline{C}/\varepsilon_n)$, which is of order $\log^2(n\Delta)$ as well. It follows that, provided the constants L and M are chosen large enough, the right-hand side of the above display converges to zero as $n \to \infty$. Chebyshev's inequality then implies that II_n converges to zero in probability as $n \to \infty$. This completes the proof of Theorem 1. □

Acknowledgements We wish to thank Wikash Sewlal from Delft University of Technology for the simulation results of the example with a mixture of four normals and the skewed density. The research leading to these results has received funding from the European Research Council under ERC Grant Agreement 320637.

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.

Additional lemmas and proofs

Proof of Lemma 1

We give a detailed proof of equality (9). As we are interested in small values of Δ, we make some necessary approximations. Starting point is the expansion for the 'density' of $\mathbb{Q}^{\Delta}_{\lambda,f}$

with respect to the Lebesgue measure,

$$e^{-\lambda\Delta}\delta_0(x) + \left(1 - e^{-\lambda\Delta}\right) \sum_{m=1}^{\infty} a_m(\lambda\Delta) f^{*m}(x),$$

see (4), with coefficients a_m defined in (5). It follows that we have the likelihood ratio

$$\begin{aligned}\frac{\mathrm{d}\mathbb{Q}^{\Delta}_{\lambda,f}}{\mathrm{d}\mathbb{Q}^{\Delta}_{\lambda_0,f_0}}(x) &= \mathbf{1}_{x=0} e^{-(\lambda-\lambda_0)\Delta} + \mathbf{1}_{x\neq 0} \frac{(1-e^{-\lambda\Delta}) \sum_{m=1}^{\infty} a_m(\lambda\Delta) f^{*m}(x)}{(1-e^{-\lambda_0\Delta}) \sum_{m=1}^{\infty} a_m(\lambda_0\Delta) f_0^{*m}(x)} \\ &= e^{-(\lambda-\lambda_0)\Delta} \left(\mathbf{1}_{x=0} + \mathbf{1}_{x\neq 0} \frac{\lambda f(x)}{\lambda_0 f_0(x)} + o(\Delta)\right),\end{aligned}$$

where we collected terms of order Δ^m for $m \geq 2$ as $o(\Delta)$. Hence we get for the Hellinger affinity

$$H\left(\mathbb{Q}^{\Delta}_{\lambda,f},\, \mathbb{Q}^{\Delta}_{\lambda_0,f_0}\right) = \int \sqrt{\mathrm{d}\mathbb{Q}^{\Delta}_{\lambda,f} \mathrm{d}\mathbb{Q}^{\Delta}_{\lambda_0,f_0}},$$

the approximating expression

$$H\left(\mathbb{Q}^{\Delta}_{\lambda,f},\, \mathbb{Q}^{\Delta}_{\lambda_0,f_0}\right) = e^{-(\lambda+\lambda_0)\Delta/2} \left(1 + \Delta\sqrt{\lambda_0\lambda} H\left(f,\ f_0\right) + o(\Delta)\right).$$

It follows that for $\Delta \to 0$,

$$\begin{aligned}h^2\left(\mathbb{Q}^{\Delta}_{\lambda,f},\, \mathbb{Q}^{\Delta}_{\lambda_0,f_0}\right) &= 2 - 2H\left(\mathbb{Q}^{\Delta}_{\lambda,f},\, \mathbb{Q}^{\Delta}_{\lambda_0,f_0}\right) \\ &= 2 - 2e^{-(\lambda+\lambda_0)\Delta/2} \left(1 + \Delta\sqrt{\lambda_0\lambda} H\left(f,\ f_0\right) + o(\Delta)\right) \\ &= 2\left(1 - e^{-(\lambda+\lambda_0)\Delta/2}\right) - 2e^{-(\lambda+\lambda_0))\Delta/2} \left(\Delta\sqrt{\lambda_0\lambda} H\left(f,\ f_0\right) + o(\Delta)\right).\end{aligned}$$

Hence, for $\Delta \to 0$,

$$\begin{aligned}\frac{1}{\Delta} h^2\left(\mathbb{Q}^{\Delta}_{\lambda,f},\, \mathbb{Q}^{\Delta}_{\lambda_0,f_0}\right) &\to \lambda + \lambda_0 - 2\sqrt{\lambda_0\lambda} H\left(f,\ f_0\right) \\ &= \int \left(\sqrt{\lambda f(x)} - \sqrt{\lambda_0 f_0(x)}\right)^2 \mathrm{d}x.\end{aligned}$$

Equality (9) follows. The proofs of the equalities (10) and (11) follow a similar line of reasoning.

Proof of Lemma 3

The proof is an adaptation of Theorem 7.1 from Ghosal et al. (2000) to decompounding. In all what follows it is assumed that $\mathbb{Q}^{\Delta}_{\lambda,f} \in \mathcal{Q}$, but we suppress this assumption in the notation. Observe that

$$\begin{aligned}&D\left(\frac{\varepsilon}{2},\, \left\{\mathbb{Q}^{\Delta}_{\lambda,f} : \varepsilon \leq h^{\Delta}\left(\mathbb{Q}^{\Delta}_{\lambda_0,f_0},\, \mathbb{Q}^{\Delta}_{\lambda,f}\right) \leq 2\varepsilon\right\},\, h^{\Delta}\right) \\ &= D\left(\frac{\varepsilon\sqrt{\Delta}}{2},\, \left\{\mathbb{Q}^{\Delta}_{\lambda,f} : \varepsilon\sqrt{\Delta} \leq h\left(\mathbb{Q}^{\Delta}_{\lambda_0,f_0},\, \mathbb{Q}^{\Delta}_{\lambda,f}\right) \leq 2\varepsilon\sqrt{\Delta}\right\},\, h\right).\end{aligned}$$

From this point on the arguments from the proof of Theorem 7.1 in Ghosal et al. (2000) are applicable (with ε replaced by $\varepsilon\sqrt{\Delta}$) and eventually lead to the desired result. The role of formulae (7.1)–(7.2) in that proof are played in the present context by (36) and (37) below.

For a given $(\lambda_1,\ f_1)$ there exists a sequence of tests ϕ_n based on $\mathcal{Z}_n^{\Delta}$, such that

$$\mathbb{E}_{\lambda_0, f_0}[\phi_n] \leq \exp\left(-\frac{1}{2} n \Delta h^{\Delta}\left(\mathbb{Q}_{\lambda_0, f_0}^{\Delta},\ \mathbb{Q}_{\lambda, f}^{\Delta}\right)^2\right), \tag{36}$$

$$\sup_{h^{\Delta}(\mathbb{Q}_{\lambda, f}^{\Delta},\ \mathbb{Q}_{\lambda_1, f_1}^{\Delta}) < h^{\Delta}(\mathbb{Q}_{\lambda_0, f_0}^{\Delta},\ \mathbb{Q}_{\lambda_1, f_1}^{\Delta})} \mathbb{E}_{\lambda, f}[1-\phi_n] \leq \exp\left(-\frac{1}{2} n \Delta h^{\Delta}\left(\mathbb{Q}_{\lambda_0, f_0},\ \mathbb{Q}_{\lambda, f}\right)^2\right). \tag{37}$$

These two inequalities simply follow by rewriting the inequalities

$$\mathbb{E}_{\lambda_0, f_0}[\phi_n] \leq \exp\left(-\frac{1}{2} n h^2\left(\mathbb{Q}_{\lambda_0, f_0}^{\Delta},\ \mathbb{Q}_{\lambda, f}^{\Delta}\right)\right),$$

$$\sup_{h(\mathbb{Q}_{\lambda, f}^{\Delta},\ \mathbb{Q}_{\lambda_1, f_1}^{\Delta}) < h(\mathbb{Q}_{\lambda_0, f_0}^{\Delta},\ \mathbb{Q}_{\lambda_1, f_1}^{\Delta})} \mathbb{E}_{\lambda, f}[1-\phi_n] \leq \exp\left(-\frac{1}{2} n h^2\left(\mathbb{Q}_{\lambda_0, f_0}^{\Delta},\ \mathbb{Q}_{\lambda, f}^{\Delta}\right)\right),$$

which are proved in Ghosal et al. (2000, pp. 520–521) and rely upon the results in Birgé (1984) and Cam (1986).

Proof of Lemma 2

As the priors for $\psi_1, \ldots, \psi_J$ are independent, we obtain that

$$\begin{aligned} p(\psi \mid \mu,\ \tau,\ z,\ \mathbf{a}) = p(\psi \mid \mathbf{a}) &\propto \prod_{j=1}^{J}\left(e^{-\psi_j T}\psi_j^{s_j}\pi\left(\psi_j\right)\right) \\ &= \prod_{j=1}^{J}\left(e^{-(\psi_j T+\beta_0)}\psi_j^{s_j+\alpha_0-1}\right), \end{aligned}$$

which proves the first statement of the lemma.

For $(\mu,\ \tau)$ we get

$$\begin{aligned} p(\mu,\ \tau \mid z,\ \mathbf{a}) \propto &\prod_{i\in\mathcal{I}} \phi\left(z_i;\ a_i'\mu,\ n_i/\tau\right) \\ &\times \tau^{\alpha_1-1} e^{-\beta_1\tau}\tau^{J/2}\exp\left(-\frac{\tau\kappa}{2}\sum_{j=1}^{J}\left(\mu_j-\xi_j\right)^2\right). \end{aligned}$$

This is proportional to

$$\tau^{\alpha_1-1+(I+J)/2}\exp\left(-\beta_1\tau-\frac{D(\mu)}{2}\tau\right),$$

where

$$D(\mu) = \kappa\sum_{j=1}^{J}\left(\mu_j-\xi_j\right)^2 + \sum_{i\in\mathcal{I}} n_i^{-1}\left(z_i-a_i'\mu\right)^2.$$

From this expression it is easily seen that we can integrate out μ to obtain the distribution of τ, conditional on $(z,\ \mathbf{a})$. To get this right, write $D(\mu)$ as a quadratic form of μ:

$$D(\mu) = \mu' P\mu - 2q'\mu + R.$$

By completing the square, we find that

$$\int \exp\left(-\frac{\tau}{2}D(\mu)\right) d\mu = e^{-\tau R/2} \int \exp\left(-\frac{1}{2}\mu\tau P\mu + \tau q'\mu\right) d\mu.$$

The integrand is (up to a proportionality constant), the density of a bivariate normal random vector with mean vector $P^{-1}q$ and covariance matrix $\tau^{-1}P^{-1}$ evaluated in μ. This implies that the preceding display equals

$$e^{-\tau R/2}(2\pi)^{J/2}\sqrt{\left|\tau^{-1}P^{-1}\right|}\exp\left(\frac{1}{2}\tau q'P^{-1}q\right).$$

We conclude that

$$p(\tau \mid z, \mathbf{a}) \propto \tau^{\alpha_1 + I/2 - 1} \exp\left(-\left(\beta_1 + \frac{1}{2}\left(R - q'P^{-1}q\right)\right)\tau\right),$$

which proves the asserted Gamma distribution of τ. This computation also immediately leads to the assertion on the distribution of μ. We finally show that the rate parameter appearing for τ is positive. By definition $D(\mu) \geq 0$ for all μ. This implies that $D(P^{-1}q) = q'P^{-1}q - 2q'P^{-1}q + R = R - q'P^{-1}q \geq 0$.

References

Alexandersson H (1985) A simple stochastic model of the precipitation process. J Clim Appl Meteorol 24(12):1282–1295

Belomestny D, Comte F, Genon-Catalot V, Masuki H, Reiß M (eds) (2015) Lévy matters IV, estimation for discretely observed Lévy processes. Lecture notes in mathematics 2128. Springer, Cham

Birgé L (1984) Sur un théorème de minimax et son application aux tests. Probab Math Stat 3:259–282

Brooks S, Gelman A, Jones GL, Meng XL (2011) Handbook of Markov chain Monte Carlo. Chapman and Hall/CRC, Hoboken

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

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

Burlando P, Rosso R. (1993) Stochastic Models of Temporal Rainfall: Reproducibility, Estimation and Prediction of Extreme Events. In: *Stochastic Hydrology and its Use in Water Resources Systems, Simulation and Optimization*, Marco JB, Harboe R, Salas JD (eds.), NATO ASI Series 237: 137–173. Springer

Comte F, Genon-Catalot V (2010) Non-parametric estimation for pure jump irregularly sampled or noisy Lévy processes. Stat Neerl 64(3):290–313

Comte F, Genon-Catalot V (2010) Nonparametric adaptive estimation for pure jump Lévy processes. *Annales de Institut Henri Poincare (B), Probability and Statistics* 46(3): 595–617

Comte F, Genon-Catalot V (2015) Adaptive estimation for Lévy processes. In: Belomestny D, Comte F, Genon-Catalot V, Masuki H, Reiß M (eds.). *Lévy matters IV, Estimation for discretely observed Lévy processes*. Lecture Notes in Mathematics 2128: 77–177. Springer, Cham

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

Comte F, Genon-Catalot V (2011) Estimation for Lévy processes from high frequency data within a long time interval. Ann Stat 39:803–837

Diebolt J, Robert CP (1994) Estimation of finite mixture distributions through Bayesian sampling. J R Stat Soc B 56:363–375

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

Embrechts P, Klüppelberg C, Mikosch T (1997) Modelling Extremal Events for Insurance and Finance. Applications of Mathematics (New York), 33. Springer-Verlag, Berlin

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

Ferguson TS (1973) A Bayesian analysis of some nonparametric problems. Ann Stat 1:209–230
Ferguson TS (1983) Bayesian density estimation by mixtures of normal distributions. In: Recent advances in statistics. Academic, New York, p 287–302
Figueroa-López JE (2008) Small-time moment asymptotics for Lévy processes. Stat Probab Lett 78(18):3355–3365
Figueroa-López JE (2009) Nonparametric estimation of Lévy models based on discrete-sampling. In: Optimality. IMS lecture notes monograph series, vol 57. Institute of Statistical Mathematics, Beachwood, p 117–146
Ghosal S (2010) The Dirichlet process, related priors and posterior asymptotics. In: Bayesian nonparametrics. Cambridge series in statistical and probabilistic mathematics. Cambridge University Press, Cambridge, p 35–79
Ghosal S, Ghosh JK, van der Vaart AW (2000) Convergence rates of posterior distributions. Ann Stat 28:500–531
Ghosal S, Tang Y (2006) Bayesian consistency for Markov processes. Sankhyā 68:227–239
Ghosal S, van der Vaart AW (2001) Entropies and rates of convergence for maximum likelihood and Bayes estimation for mixtures of normal densities. Ann Stat 29:1233–1263
Ghosal S, van der Vaart AW (2007) Posterior convergence rates of Dirichlet mixtures at smooth densities. Ann Stat 35:697–723
Gugushvili S, van der Meulen F, Spreij P (2015) Non-parametric Bayesian inference for multi-dimensional compound Poisson processes. Mod Stoch Theory Appl 2:1–15
Hjort NL, Holmes C, Müller P, Walker SG (2010) Bayesian nonparametrics. Cambridge series in statistical and probabilistic mathematics, 28. Cambridge University Press, Cambridge
Ibragimov IA, Khas'minskiĭ RZ (1982) An estimate of the density of a distribution belonging to a class of entire functions (Russian). Teor Veroyatnost i Primenen 27:514–524
Insua DR, Ruggeri F, Wiper MP (2012) Bayesian analysis of stochastic process models. Wiley, Chichester
Jacod J, Shiryaev AN (2003) Limit theorems for stochastic processes, 2nd edn. Grundlehren der Mathematischen Wissenschaften, vol 288. Springer, Berlin
Katz RW (2002) Stochastic modeling of hurricane damage. J Appl Meteorol 41(7):754–762
Le Cam LM (1986) Asymptotic methods in statistical decision theory. Springer, New York
Lo AY (1984) On a class of Bayesian nonparametric estimates: I. Density estimates. Ann Stat 12:351–357
McLachlan G, Peel D (2000) Finite mixture models. Wiley series in probability and statistics: applied probability and statistics. Wiley-Interscience, New York
Marron JS, Wand MP (1992) Exact mean integrated squared error. Ann Stat 20(2):712–736
Nickl R, Reiß M (2012) A Donsker theorem for Lévy measures. J Funct Anal 263(10):3306–3332
Nickl R, Reiß M, Söhl J, Trabs M (2016) High-frequency Donsker theorems for Lévy measures. Probab Theory Relat Fields 164(1):61–108
Papaspiliopoulos O, Roberts GO, Sköld M (2007) A general framework for the parametrization of hierarchical models. Stat Sci 22(1):59–73
Prabhu NU (1998) Stochastic storage processes. Queues, insurance risk, dams, and data communication, 2nd edn. Applications of mathematics (New York), vol 15. Springer, New York
Richardsen S, Green PJ (1997) On Bayesian analysis of mixtures with an unknown number of components (with discussion). J R Stat Soc B 59:731–792
Scalas E (2006) The application of continuous time random walks in finance and economics. Physica A 362(2):225–239
Shreve SE (2008) Stochastic calculus for finance II, 2nd edn. Springer, New York
Skorohod AV (1964) Random processes with independent increments. Izdat. "Nauka", Moscow
Tang Y, Ghosal S (2007) Posterior consistency of Dirichlet mixtures for estimating a transition density. J Stat Plan Inference 137:1711–1726
Tanner MA, Wong WH (1987) The calculation of posterior distributions by data augmentation. J Am Stat Assoc 82:528–540
Ueltzhöfer FAJ, Klüppelberg C (2011) An Oracle inequality for penalised projection estimation of Lévy densities from high frequency observations. J Nonparametr Stat 23(4):967–989