Compound multivariate Hawkes processes: Large deviations and rare event simulation

RAVIAR S. KARIM[1,a], ROGER J. A. LAEVEN[2,b] and MICHEL MANDJES[3,c]

[1]*Department of Quantitative Economics, University of Amsterdam, P.O. Box 15867, 1001 NJ Amsterdam, The Netherlands.* [a]R.S.Karim@uva.nl

[2]*Department of Quantitative Economics, University of Amsterdam, EURANDOM and CentER, P.O. Box 15867, 1001 NJ Amsterdam, The Netherlands.* [b]R.J.A.Laeven@uva.nl

[3]*Korteweg-de Vries Institute for Mathematics, University of Amsterdam, EURANDOM and Amsterdam Business School, P.O. Box 94248, 1090 GE Amsterdam, The Netherlands.* [c]M.R.H.Mandjes@uva.nl

In this paper, we establish a large deviations principle for a multivariate compound process induced by a multivariate Hawkes process with random marks. Our proof hinges on showing essential smoothness of the limiting cumulant of the multivariate compound process, resolving the inherent complication that this cumulant is implicitly characterized through a fixed-point representation. We employ the large deviations principle to derive logarithmic asymptotic results on the marginal ruin probabilities of the associated multivariate risk process. We also show how to conduct rare event simulation in this multivariate setting using importance sampling and prove the asymptotic efficiency of our importance sampling based estimators.

MSC2020 subject classifications: Primary 60G55, 60F10; secondary 60E10, 62E20, 91G05

Keywords: Compound processes; Hawkes processes; large deviation; mutual excitation; rare event simulation; ruin probability

1. Introduction

Mutually exciting processes, or multivariate Hawkes processes ([22,23]), constitute an important class of point processes, particularly suitable to describe stochastic dependences among occurrences of events across time and space. Due to their built-in feedback mechanism, they are a natural contender to model contagious phenomena where clusters of events occur in both the temporal and spatial dimensions. Over the past decades, Hawkes processes have been increasingly applied across a broad variety of fields, such as finance ([1,2,4,24]), social interaction ([11]), neuroscience ([37]), seismology ([25,35]), and many others.

The key property of a Hawkes process is that it exhibits 'self-exciting' behavior: informally, any event instantaneously increases the likelihood of, hence potentially triggers, additional future events. A crucial element in its definition is the so-called *decay function* that quantifies how quickly the effect of an initial event on future events vanishes. Choosing this function to be exponential renders the model Markovian, which facilitates the explicit evaluation of various relevant risk and performance metrics (e.g., transient and stationary moments). In practical applications, however, it can be more natural to allow for other, i.e., not necessarily exponential, decay functions admitting non-Markovian, and long-memory, properties but making the analysis substantially more challenging; see, e.g., [16,31] who indicate the relevance of non-Markovian models in describing contagious phenomena.

In applied probability and mathematical risk theory, Hawkes processes have been used to model the claim arrival process, and, likewise, *compound* Hawkes processes to model the associated cumulative claim process that an insurance firm is facing; see the related literature discussed in detail below. In collective risk theory, *multivariate* Hawkes processes provide an appealing candidate for modeling, for example, the claim arrival process associated with technological risks.[1] Bearing in mind that ruin and exceedance probabilities ought to be kept small, a primary research goal concerns their analysis in the asymptotic regime in which the initial reserve level of the insurer or the time horizon of aggregation grows large, such that the event of interest becomes increasingly rare. This explains the interest in deriving large deviations principles for (compound) Hawkes processes, providing a formal tool to assess their rare event behavior, and facilitating in particular the identification of the asymptotics of ruin and exceedance probabilities. At the same time, it is noted, however, that large deviations results usually yield rough, logarithmic asymptotics only, in that they focus on identifying the associated decay rate. To remedy this, one could attempt to develop 'large deviations informed' simulation techniques by which rare events can be evaluated fast and accurately. This is particularly useful when the probabilities of rare events are too small to be estimated with reasonable accuracy using regular Monte Carlo simulations.

Whereas large deviations for *univariate* Hawkes processes are well understood, their *multivariate* counterpart is to a large extent unexplored. In this context, we mention [42], which focuses on the broad class of d-dimensional affine processes of Markovian type, covering the special case of the multivariate Hawkes process with an exponential decay function. Importantly, however, [42] presents large deviations results pertaining to the *sum* of the d components, rendering the result essentially single-dimensional; see also the refinements in [19]. In addition, a moderate deviations result has been derived in [41]. To the best of our knowledge, multi-dimensional large deviations principles for multivariate compound Hawkes processes allowing for general decay functions have not been established, and in addition, no rare event simulation techniques have been developed in this setting. These are the main subjects of this paper.

In the univariate case, large deviations results for compound Hawkes processes with general decay function have been derived in [39] building on [7]. The underlying argumentation relies on the *cluster representation* of the driving Hawkes process, as developed in the seminal work [23], from which it is concluded that the cluster size follows a so-called *Borel* distribution. A crucial element in proving the large deviations principle lies in showing that the limiting cumulant of the random object under study is *steep*, entailing that its derivative grows to infinity when approaching the boundary of its domain, such that the Gärtner-Ellis Theorem can be invoked. In the univariate (compound) Hawkes case considered in [39], steepness could be established by using the explicit expression for the cluster size distribution. When studying large deviations for *multivariate* (compound) Hawkes processes, however, a main technical difficulty that arises is that the cluster size distribution of the process is not known in closed form. In the multivariate setting, all one has is a vector-valued fixed-point representation for the limiting cumulant of the (multivariate) cluster size, as was derived in [29]. Importantly, this relation does not allow the closed-form identification of the limiting cumulant (let alone that one can find the distribution itself), entailing that one cannot explicitly characterize the boundary of its domain.

In light of the gaps in the literature described above, the contributions of this paper are the following:

∘ First, we establish a large deviations principle for multivariate compound Hawkes processes allowing for general decay functions. We also allow the Hawkes process to be *marked*, such that the intensity process experiences jumps of random size constituting another potential source of rare, atypical behavior. We have succeeded in establishing steepness based on an implicit fixed-point representation for the limiting cumulant of the joint cluster size distribution. Specifically, without having an explicit expression for the limiting cumulant, and without having an explicit characterization of its domain, we prove that the derivative of the limiting cumulant grows to infinity when approaching the boundary of its domain. This steepness property facilitates the use of the Gärtner-Ellis Theorem, so as to establish the desired large deviations principle. The mathematical details of the required multivariate analysis are involved.

[1] See, e.g., [5] who apply the contagion model of [1] to model cyber attacks and also [6].

∘ Based on these results, we characterize the asymptotic behavior of the ruin probability for the marginal ruin processes in the regime that the initial reserve level grows large. We prove that this ruin probability decays essentially exponentially, with the corresponding decay rate being equal to the unique zero of the limiting cumulant pertaining to that marginal. The proof exploits the discrete-time results of [34]; see also [10,15].

∘ Finally, we develop an importance sampling algorithm for estimating rare event probabilities in our multivariate setting. More precisely, we first derive the parameters of the exponentially twisted multivariate Hawkes process. The identification of the exponential change of measure is non-trivial, as some of the relevant functions pertaining to the model under the original measure are only known as solutions of a vector-valued fixed-point representation, and of potential independent interest. The twisted marginal ruin process has positive drift, yielding ruin with probability one under the new measure, but with the likelihood ratio being bounded by a function that decays exponentially in the initial reserve, thus leading to a considerable speedup in importance sampling relative to regular Monte Carlo simulation. We prove that this estimator is, in fact, asymptotically efficient in the sense of Siegmund ([38]), in passing also establishing a Lundberg-type upper bound on the ruin probability. In addition, we devise an asymptotically efficient importance sampling algorithm for estimating the probability of the multivariate compound Hawkes process (at a given point in time, that is) attaining a rare large value. The attainable speedup, relative to regular Monte Carlo simulation, is quantified through a series of simulation experiments.

Without attempting to provide an exhaustive overview, we now review a few important related papers, all of which focus on the *univariate* setting. We already mentioned [39], which analyzes the asymptotic behavior of ruin probabilities, under the assumption of light-tailed claims, drawing upon earlier large deviations results derived in [7] for more general Poisson cluster processes. Furthermore, in [39], an importance sampling based algorithm is proposed that is capable of efficiently generating estimates of the rare event probabilities of interest. In [28], the limiting cumulant of the cluster size distribution is implicitly characterized for the setting with random marks using a fixed-point argument, while proving a large deviations principle using the Gärtner-Ellis theorem for the upper bound and an exponential tilting method for the lower bound. Where the contributions above focus primarily on the case of light-tailed claims, subexponentially distributed claims are studied in [43], in the context of a non-stationary version of the Hawkes process. For (non-compound) Hawkes processes (i.e., not involving claims), 'precise' large deviations results, providing asymptotics beyond the leading order term, are obtained in [20]. The setting of a large initial intensity is studied in [17] and [18]. For the more general class of non-linear Hawkes processes, [44] proves the process-level large deviations, and [45] derives large deviations in the Markovian setting. The present paper can be viewed as belonging to the broader stream of papers in which one attempts to replace the usual Poissonian arrivals assumption by more realistic assumptions; see also e.g., [13,36] on shot-noise driven arrival rates, and [8] on models with a fluctuating client population.

The rest of this paper is organized as follows. In Section 2, we introduce the relevant processes and discuss some basic properties that are used throughout the paper. Section 3 derives results on the transform of the joint cluster size distribution, and provides an implicit characterization of the domain of the limiting cumulant. Section 4 establishes the large deviations principle for the multivariate compound Hawkes process with general decay function and random marks. In Section 5, we exploit the large deviations principle to develop importance sampling algorithms to efficiently estimate rare event probabilities; this includes an analysis of the decay rate of the marginal ruin probability. Concluding remarks are in Section 6. Some proofs and all simulation experiments, which numerically demonstrate the performance of our importance sampling based estimators, are relegated to the Appendix provided as supplementary material ([30]).

2. Multivariate compound Hawkes processes

In this section, we first provide the definitions of multivariate Hawkes and compound Hawkes processes, and next introduce some objects and discuss some properties that are relevant in the context of this paper. Throughout, we use the boldface notation $\boldsymbol{x} = (x_1, \ldots, x_d)^\top$ to denote a d-dimensional vector, for a given dimension $d \in \mathbb{N}$. Inequalities between vectors are understood componentwise, e.g., $\boldsymbol{x} > \boldsymbol{y}$ means $x_i > y_i$ for all $i = 1, \ldots, d$.

Consider a d-dimensional càdlàg counting process $\boldsymbol{N}(\cdot) \equiv (\boldsymbol{N}(t))_{t\in\mathbb{R}_+}$, where each increment $N_i(t) - N_i(s)$ records the number of points in component $i \in [d] := \{1, \ldots, d\}$ in the time interval $(s, t]$, with $s < t$. We label the points by considering, for each $j \in [d]$, a sequence of a.s. increasing positive random variables $\boldsymbol{T}_j = \{T_{j,r}\}_{r\in\mathbb{N}} = \{T_{j,1}, T_{j,2}, \ldots\}$ representing event times. We associate to this sequence the one-dimensional counting process $N_j(\cdot)$ by setting

$$N_j(t) := N_{\boldsymbol{T}_j}(0, t] = \sum_{r=1}^{\infty} \mathbf{1}_{\{T_{j,r} \leq t\}}.$$

The process $\boldsymbol{N}(\cdot) = (N_1(\cdot), \ldots, N_d(\cdot))^\top$ is then the d-dimensional counting process associated with the sequences of event times in all components, $\boldsymbol{T}_1, \ldots, \boldsymbol{T}_d$, compactly denoted by $\boldsymbol{N}(t) = \boldsymbol{N}_{\boldsymbol{T}}(0, t]$. Throughout, the points will be referred to as *events* and the terms point process and counting process are used interchangeably for $\boldsymbol{N}(\cdot)$. We assume that the point process starts empty, i.e., $\boldsymbol{N}(0) = \mathbf{0} = (0, \ldots, 0)^\top$.

In the original work [22], the Hawkes process is defined by relying on the concept of the conditional intensity function. An alternative, equivalent definition, known as the cluster process representation, can be given by representing the Hawkes process as a Poisson cluster process; it was first described in [23] in the setting of the conventional univariate Hawkes process, see also [12, Example 6.3(c)] and [32, Ch. IV]. The cluster process representation distinguishes between two types of events: first, there are *immigrant* events generated according to a homogeneous Poisson process with a given rate; and second, there are *offspring* events generated by an inhomogeneous Poisson process with rates that account for self-excitation and, in the multivariate context also, cross-excitation. In the following, we introduce the relevant terminology and provide the formal definitions of the process.

For $j \in [d]$, we consider *base rates* $\overline{\lambda}_j \geqslant 0$, with at least one of the base rates being strictly positive. For each combination $i, j \in [d]$, we let the *decay function* $g_{ij}(\cdot) : \mathbb{R}_+ \to \mathbb{R}_+$ be non-negative and integrable. Also, for $j \in [d]$, we define the *random marks* through the generic non-negative, non-degenerate random vector $\boldsymbol{B}_j = (B_{1j}, \ldots, B_{dj})$, asserting that the sequence of random marks $\{\boldsymbol{B}_{j,r}\}_{r\in\mathbb{N}}$ consists of i.i.d. random vectors that are distributed as $\boldsymbol{B}_j$. We allow the random variables $B_{ij,r}$ to be dependent for fixed j and r. Finally, we let $K_{ij}(\cdot)$ be an inhomogeneous Poisson process with intensity $B_{ij,r}\, g_{ij}(\cdot)$, given the value of $B_{ij,r}$. With these elements in place, we provide the following two equivalent definitions of a multivariate Hawkes process.

Definition 1 (conditional intensity function). *A multivariate Hawkes process is a d-dimensional point process $\boldsymbol{N}(\cdot)$ of which the components satisfy*

$$\begin{cases} \mathbb{P}(N_j(t+\Delta) - N_j(t) = 0 \,|\, \mathcal{F}_t) = 1 - \lambda_j(t)\Delta + o(\Delta), \\ \mathbb{P}(N_j(t+\Delta) - N_j(t) = 1 \,|\, \mathcal{F}_t) = \lambda_j(t)\Delta + o(\Delta), \\ \mathbb{P}(N_j(t+\Delta) - N_j(t) > 1 \,|\, \mathcal{F}_t) = o(\Delta), \end{cases} \tag{1}$$

for $j \in [d]$ as $\Delta \downarrow 0$, where $\mathcal{F}_t = \sigma(\boldsymbol{N}(s) : s \leqslant t)$ is the natural filtration generated by $\boldsymbol{N}(\cdot)$,

$$\lambda_i(t) = \overline{\lambda}_i + \sum_{j=1}^{d} \int_0^t B_{ij}(s)\, g_{ij}(t-s)\, \mathrm{d}N_j(s), \tag{2}$$

for $i \in [d]$, $\boldsymbol{N}(0) = \boldsymbol{0}$ and the integral in (2) is understood as $\int_{(0,t)}$, i.e., excluding t.

The conditional intensity function $\boldsymbol{\lambda}(\cdot)$ is taken left-continuous and is predictable; see also [12, Example 7.2(b) and Ch. 14] for further details.

Definition 2 (cluster process representation). *Define a d-dimensional point process $\boldsymbol{N}(\cdot)$ componentwise by $N_j(t) = N_{\boldsymbol{T}_j}(0,t]$ for $j \in [d]$ and $t > 0$, where the sequences of event times $\boldsymbol{T}_1, \ldots, \boldsymbol{T}_d$ are generated as follows:*

(i) First, for each $j \in [d]$, let there be a sequence of immigrant event times $\{T_{j,r}^{(0)}\}_{r\in\mathbb{N}}$ on the interval $(0,\infty)$ generated by a homogeneous Poisson process $I_j(\cdot)$ with rate $\overline{\lambda}_j$.
(ii) Second, let each immigrant event independently generate a d-dimensional cluster $\boldsymbol{C}_j \equiv \boldsymbol{C}_{T_{j,r}^{(0)}}$, consisting of event times associated with generations of events:
  (a) The immigrant with event time $T_{j,r}^{(0)}$ is labeled to be of generation 0 and into each component $m \in [d]$, it generates a sequence of first-generation event times $\{T_{m,r}^{(1)}\}_{r\in\mathbb{N}}$ on the interval $(T_{j,r}^{(0)}, \infty)$, according to $K_{mj}(\cdot - T_{j,r}^{(0)})$ with $B_{mj,r}$ the random mark associated to $T_{j,r}^{(0)}$.
  (b) Iterating (a) *above, with $T_{m,r}^{(n-1)}$ designating the r-th event time of generation $n-1$ in component $m \in [d]$, yields generation n event times $\{T_{l,r}^{(n)}\}_{r\in\mathbb{N}}$ in component $l \in [d]$ on the interval $(T_{m,r}^{(n-1)}, \infty)$, generated according to $K_{lm}(\cdot - T_{m,r}^{(n-1)})$.*

Upon taking the union over all generations, we obtain, for each component $j \in [d]$,

$$\boldsymbol{T}_j = \{T_{j,r}\}_{r\in\mathbb{N}} = \bigcup_{n=0}^{\infty} \{T_{j,r}^{(n)}\}_{r\in\mathbb{N}}.$$

The process $\boldsymbol{N}(\cdot)$ defined above for $t > 0$ and with $\boldsymbol{N}(0) = \boldsymbol{0}$ constitutes a multivariate Hawkes process.

To ensure that the clusters described in part (ii) of Definition 2 are a.s. finite, we assume that a *stability condition* applies throughout this paper. It is shown in [22] that this stability condition guarantees non-explosiveness of $\boldsymbol{N}(\cdot)$, see also [12, Example 8.3(c)].

Assumption 1. *Assume that the matrix $\boldsymbol{H} := (h_{mj})_{m,j\in[d]}$ with elements*

$$h_{mj} := \mathbb{E}[B_{mj}]c_{mj}, \tag{3}$$

with $c_{mj} = \int_0^\infty g_{mj}(v)\,\mathrm{d}v$, has spectral radius strictly smaller than 1.

We next define the multivariate compound Hawkes process as follows. Let $d^\star \in \mathbb{N}$ be fixed and note that we allow $d \neq d^\star$. For each $j \in [d]$, let $\{\boldsymbol{U}_{j,r}\}_{r\in\mathbb{N}} = \{(U_{1j,r}, \ldots, U_{d^\star j,r})^\top\}_{r\in\mathbb{N}}$ be a sequence of non-negative, non-degenerate i.i.d. random vectors of length $d^\star$. We allow the random variables $U_{ij,r}$ to be dependent for fixed j and r.

Definition 3 (multivariate compound Hawkes process). *Define* $\boldsymbol{Z}(\cdot) := (Z_1(\cdot), \ldots, Z_{d^\star}(\cdot))^\top$ *for each component* $Z_i(\cdot)$ *with* $i \in [d^\star]$ *by*

$$Z_i(t) := \sum_{j=1}^{d} \sum_{r=1}^{N_j(t)} U_{ij,r}, \quad t > 0, \tag{4}$$

where $\boldsymbol{U}_{j,r} = (U_{1j,r}, \ldots, U_{d^\star j,r})^\top$ *is drawn independently for every event in* $N_j(t)$*, with* $j \in [d]$*. The process* $\boldsymbol{Z}(\cdot)$ *defined above for any* $t > 0$ *constitutes a multivariate compound Hawkes process.*

If we define the random matrix $\boldsymbol{U} \in \mathbb{R}_+^{d^\star \times d}$ as

$$\boldsymbol{U} \equiv \left[\boldsymbol{U}_1, \ldots, \boldsymbol{U}_d\right] := \begin{bmatrix} U_{11} & U_{12} & \ldots & U_{1d} \\ U_{21} & U_{22} & \ldots & U_{2d} \\ \vdots & \vdots & \ddots & \vdots \\ U_{d^\star 1} & U_{d^\star 2} & \cdots & U_{d^\star d} \end{bmatrix}, \tag{5}$$

we can represent Eqn. (4) in vector-matrix form by

$$\boldsymbol{Z}(t) = \boldsymbol{N}(t) \star \boldsymbol{U} := \sum_{j=1}^{d} \sum_{r=1}^{N_j(t)} \boldsymbol{U}_{j,r}, \tag{6}$$

where $\star$ denotes the compound sum operation.

We proceed with a brief discussion of the dimensionality of the objects appearing in Definition 3. The Hawkes process $\boldsymbol{N}(\cdot)$ is of dimension d and the random vectors $\boldsymbol{U}_j$ are of dimension $d^\star$, which results in the compound Hawkes process $\boldsymbol{Z}(\cdot)$ being also of dimension $d^\star$. This reflects that the random variables of the type U_{ij}, with $i \in [d^\star]$ and $j \in [d]$, can be interpreted as the effect that an event in the j-th component of the Hawkes process $\boldsymbol{N}(\cdot)$ has on the i-the component of the compound Hawkes process $\boldsymbol{Z}(\cdot)$. As stated before, we allow $d \neq d^\star$. Intuitively, for instance, in the context of insurance, this means that the number of risk drivers may be larger ($d > d^\star$) or smaller ($d < d^\star$) than the number of insurance product categories.

We now introduce some objects related to the cluster process representation that are relevant for later analysis. Recall that for each immigrant in component $j \in [d]$, the d-dimensional cluster $\boldsymbol{C}_j$ from Definition 2 contains the sequences of event times in each component that have the immigrant with event time $T_{j,r}^{(0)}$ as oldest ancestor. We associate to $\boldsymbol{C}_j$ the d-dimensional *cluster point process* $\boldsymbol{S}_j(\cdot)$, by setting

$$\boldsymbol{S}_j(u) := \boldsymbol{S}_{\boldsymbol{C}_j}(0, u], \tag{7}$$

such that it counts the number of events of $\boldsymbol{C}_j$ on the interval $(0, u]$, where $u = t - T_{j,r}^{(0)} > 0$ is the remaining time after the arrival of the immigrant event. Concretely, we have

$$\boldsymbol{S}_j(u) := \begin{bmatrix} S_{1\leftarrow j}(u) \\ \vdots \\ S_{d\leftarrow j}(u) \end{bmatrix}, \tag{8}$$

where each entry $S_{i\leftarrow j}(u)$ records the number of events generated into component $i \in [d]$ in the cluster $\boldsymbol{C}_j$, its oldest ancestor being the immigrant event in component j that generated the cluster. To avoid double counts, we let the immigrant itself be included in the cluster (only) when $i = j$.

If we let u tend to infinity, the entries of the random vector $\boldsymbol{S}_j(u)$ ultimately count the *total* number of events within the cluster $\boldsymbol{C}_j$ generated into each component $i \in [d]$. Observe that $u \mapsto \boldsymbol{S}_j(u)$ is increasing componentwise and $\sup_{u\in\mathbb{R}_+} \|\boldsymbol{S}_j(u)\|_{\mathbb{R}^d} < \infty$ with probability 1 due to Assumption 1. Hence, we can define a random vector that counts the total number of events in all components, or simply *cluster size*, by setting $\boldsymbol{S}_j := \lim_{u\to\infty} \boldsymbol{S}_j(u)$, where convergence is understood in the a.s. sense.

One can interpret these clusters in terms of d-type Galton-Watson processes, where the total progeny equals the sum of all generations of offspring that descend from one individual ([27]). Suppose the Galton-Watson process starts with an individual of type $j \in [d]$, and let $\boldsymbol{S}_j^{(k)}$ denote the k-th generation of descendants. Then one can write

$$\boldsymbol{S}_j = \sum_{k=0}^{\infty} \boldsymbol{S}_j^{(k)}, \tag{9}$$

where $\boldsymbol{S}_j^{(0)} = \boldsymbol{e}_j$, the unit vector (i.e., with j-th entry equal to 1, and other entries equal to 0). In [27], the total progeny, i.e. cluster size, of such a process is analyzed in the one-dimensional setting and shown to have a so-called Borel distribution. For higher-dimensional Hawkes processes, by using [9, Theorem 1.2], it is, in principle, also possible to derive a representation of the multivariate cluster size distribution. However, the resulting expression is neither explicit nor workable for the goal at hand due to (highly) convoluted sums that arise in the derivation. More specifically, the multiplicity of the different possible sample paths to generate a certain number of events in each component yields a complex combinatorial problem.

We conclude this section by stating two convergence results that will be needed later in this paper. Under the stability condition, we have that the Hawkes process $\boldsymbol{N}(\cdot)$ satisfies the following strong law of large numbers, as shown in [4]: as $t \to \infty$, we have

$$\frac{\boldsymbol{N}(t)}{t} \to (\boldsymbol{I} - \boldsymbol{H})^{-1}\overline{\boldsymbol{\lambda}}, \tag{10}$$

a.s., where $\overline{\boldsymbol{\lambda}} = (\overline{\lambda}_1, \dots, \overline{\lambda}_d)^\top$. This result naturally extends to the corresponding compound Hawkes process $\boldsymbol{Z}(\cdot)$ (see [21, Theorem 1]): as $t \to \infty$, a.s.,

$$\frac{\boldsymbol{Z}(t)}{t} \to \mathbb{E}[\boldsymbol{U}](\boldsymbol{I} - \boldsymbol{H})^{-1}\overline{\boldsymbol{\lambda}}. \tag{11}$$

3. Transform analysis

In this section, we discuss probability generating functions and moment generating functions pertaining to the processes introduced in the previous section, viz. the multivariate Hawkes and compound Hawkes processes. These functions will play a pivotal role when deriving large deviations results later in this paper.

It is directly seen from the definition of $\boldsymbol{Z}(t)$ that, for fixed $t > 0$, its moment generating function satisfies the following composite expression in terms of the probability generating function of $\boldsymbol{N}(t)$:

$$m_{\boldsymbol{Z}(t)}(\boldsymbol{\theta}) \equiv \mathbb{E}\left[e^{\boldsymbol{\theta}^\top \boldsymbol{Z}(t)}\right] = \mathbb{E}\left[\prod_{l=1}^{d} \left(m_{\boldsymbol{U}_l}(\boldsymbol{\theta})\right)^{N_l(t)}\right], \tag{12}$$

where

$$m_{\boldsymbol{U}_l}(\boldsymbol{\theta}) \equiv \mathbb{E}[e^{\boldsymbol{\theta}^\top \boldsymbol{U}_l}] = \mathbb{E}\Big[\prod_{i=1}^{d^\star} e^{\theta_i U_{il}}\Big].$$

For now, we assume $\boldsymbol{\theta} \in \mathbb{R}^{d^\star}$ is chosen such that (12) exists—we will further discuss the domain of convergence below. We are interested in the limiting cumulant of $\boldsymbol{Z}(t)$ as $t \to \infty$, that is, we wish to analyze

$$\lim_{t\to\infty} \frac{1}{t} \log m_{\boldsymbol{Z}(t)}(\boldsymbol{\theta}). \tag{13}$$

To derive an expression for (13), we use a characterization of the probability generating function of $\boldsymbol{N}(t)$ in terms of the cluster point processes $\boldsymbol{S}_j(u)$, obtained in [29, Theorem 1]:

$$\mathbb{E}\Big[\prod_{l=1}^{d} z_l^{N_l(t)}\Big] = \prod_{j=1}^{d} \exp\Big(\overline{\lambda}_j \int_0^t \big(\mathbb{E}\Big[\prod_{l=1}^{d} z_l^{S_{l\leftarrow j}(u)}\Big] - 1\big)\, \mathrm{d}u\Big), \tag{14}$$

where, for each $j \in [d]$, the probability generating function of $\boldsymbol{S}_j(u)$ appearing on the right-hand side of (14) satisfies the fixed-point representation

$$f_j(z,u) := \mathbb{E}\Big[\prod_{l=1}^{d} z_l^{S_{l\leftarrow j}(u)}\Big] = z_j \mathbb{E}\Big[\exp\Big(\sum_{m=1}^{d} B_{mj} \int_0^u g_{mj}(v)\Big(f_m(z,u-v) - 1\Big)\,\mathrm{d}v\Big)\Big]; \tag{15}$$

see [29, Theorem 2].

In order to exploit this characterization to establish our large deviations result in the multivariate compound Hawkes setting, we need to analyze the domain of $z = (z_1, \ldots, z_d)^\top$ for which Eqns. (14) and (15) are valid, i.e., where the probability generating functions of $\boldsymbol{S}_j(u)$ exist. More precisely, since we focus on the regime $t \to \infty$ in (13), we need to consider the probability generating function of the total cluster size $\boldsymbol{S}_j$ instead of $\boldsymbol{S}_j(u)$, which in the sequel is denoted by

$$f_j(z) := \lim_{u\to\infty} f_j(z,u) = \mathbb{E}\Big[\prod_{l=1}^{d} z_l^{S_{l\leftarrow j}}\Big], \tag{16}$$

and uncover its domain. Let $\boldsymbol{f} : \mathbb{R}_+^d \to \overline{\mathbb{R}}^d$ be given by $\boldsymbol{f}(z) = (f_1(z), \ldots, f_d(z))^\top$ and denote its effective domain by $\mathscr{D}_f := \{z \in \mathbb{R}_+^d : \|\boldsymbol{f}(z)\|_{\mathbb{R}^d} < \infty\}$. For the set $\mathscr{D}_f$, denote the interior by $\mathscr{D}_f^\circ$ and the boundary by $\partial\mathscr{D}_f$.

Observe that the right-hand side of Eqn. (15) is expressed in terms of the moment generating function of the random vector $\boldsymbol{B}_j$. We assume the following to hold throughout the paper.

Assumption 2. *Assume that for some* $\boldsymbol{\vartheta} \in \mathbb{R}_+^d$,

$$m_{\boldsymbol{B}_j}(\boldsymbol{\vartheta}) = \mathbb{E}\Big[\exp\Big(\sum_{m=1}^{d} B_{mj}\vartheta_m\Big)\Big] < \infty, \tag{17}$$

for all $j \in [d]$.

The following result gives an implicit characterization of $\boldsymbol{f}(\cdot)$ and its domain $\mathscr{D}_f$ in terms of a fixed-point representation. Its proof is lengthy and involved.

Proposition 1. *The vector-valued function $\boldsymbol{f}(\boldsymbol{z})$ is the unique increasing function that satisfies*

$$f_j(\boldsymbol{z}) = z_j\,\mathbb{E}\Big[\exp\Big(\sum_{m=1}^{d} B_{mj}c_{mj}(f_m(\boldsymbol{z})-1)\Big)\Big], \tag{18}$$

for $\boldsymbol{z}\in\mathbb{R}_+^d$ such that $\boldsymbol{z}\leqslant\widehat{\boldsymbol{z}}_{\boldsymbol{r}}\equiv\widehat{\boldsymbol{z}}$. Here, for an arbitrarily given positive vector $\boldsymbol{r}\in\mathbb{R}_+^d$, $\widehat{\boldsymbol{z}}=(\widehat{z}_1,\dots,\widehat{z}_d)^\top$ is given for each $j\in[d]$ by

$$\widehat{z}_j = r_j\left(\sum_{k=1}^{d} r_k\mathbb{E}\Big[B_{kj}c_{kj}\exp\Big(\sum_{m=1}^{d} B_{mj}c_{mj}(\widehat{x}_m-1)\Big)\Big]\right)^{-1}, \tag{19}$$

where $\widehat{\boldsymbol{x}}=(\widehat{x}_1,\dots,\widehat{x}_d)^\top$ is the solution of

$$\begin{aligned} &x_j\left(\sum_{k=1}^{d} r_k\mathbb{E}\Big[B_{kj}c_{kj}\exp\Big(\sum_{m=1}^{d} B_{mj}c_{mj}(x_m-1)\Big)\Big]\right) \\ &= r_j\mathbb{E}\Big[\exp\Big(\sum_{m=1}^{d} B_{mj}c_{mj}(x_m-1)\Big)\Big]. \end{aligned} \tag{20}$$

Proof of Proposition 1. The proof consists of three parts: *(i)* identifying the limit of $f_j(\boldsymbol{z},u)$ as $u\to\infty$; *(ii)* implicit characterization of the domain $\mathscr{D}_f$; *(iii)* proving uniqueness of $\boldsymbol{f}(\cdot)$.

— *Proof of (i).* We show that for $\boldsymbol{z}\in\mathscr{D}_f$, we have that $f_j(\boldsymbol{z},u)\to f_j(\boldsymbol{z})$ for all $j\in[d]$. At this point, we do not yet know the precise domain $\mathscr{D}_f$, but we do know it is a convex subset of $\mathbb{R}_+^d$ and we implicitly derive it later in the proof.

When $\boldsymbol{z}=\boldsymbol{1}$, we have $\boldsymbol{f}(\boldsymbol{1},u)\equiv\boldsymbol{1}\equiv\boldsymbol{f}(\boldsymbol{1})$ and convergence follows trivially. Observe that when $\boldsymbol{0}\leqslant\boldsymbol{z}<\boldsymbol{1}$, $f_j(\boldsymbol{z},u)$ is decreasing in u and $0\leqslant f_j(\boldsymbol{z},u)<1$, hence, $f_j(\boldsymbol{z},u)$ converges by the monotone convergence theorem to a finite limit as $u\to\infty$, satisfying the limit of (15). When $\boldsymbol{z}>\boldsymbol{1}$, $f_j(\boldsymbol{z},u)$ is increasing in u and either diverges to ∞ or converges to a finite limit, satisfying the limit of (15). In the intermediate case, where for some $k,m\in[d]$ one has $z_k\leqslant 1$ and $z_m>1$, we proceed as follows. Recall that for each $j\in[d]$, the map $u\mapsto\boldsymbol{S}_j(u)$ is a.s. increasing in all components. We obtain the following upper bound:

$$\begin{aligned} \limsup_{u\to\infty} f_j(\boldsymbol{z},u) &= \limsup_{u\to\infty}\mathbb{E}\Big[\prod_{i=1}^{d} z_i^{S_{i\leftarrow j}(u)}\Big] \\ &= \limsup_{u\to\infty}\mathbb{E}\Big[\prod_{k:z_k\leqslant 1} z_k^{S_{k\leftarrow j}(u)}\prod_{m:z_m>1} z_m^{S_{m\leftarrow j}(u)}\Big] \\ &\leqslant \limsup_{u\to\infty}\mathbb{E}\Big[\prod_{k:z_k\leqslant 1} z_k^{S_{k\leftarrow j}(u)}\prod_{m:z_m>1} z_m^{S_{m\leftarrow j}}\Big] \\ &\overset{(*)}{=} \mathbb{E}\Big[\prod_{i=1}^{d} z_i^{S_{i\leftarrow j}}\Big] = f_j(\boldsymbol{z}), \end{aligned}$$

if the limit is finite, where in $(*)$ we have used the monotone convergence on the product over k, as this product is decreasing. Similarly, we obtain the lower bound

$$\liminf_{u\to\infty} f_j(z,u) \geqslant \liminf_{u\to\infty} \mathbb{E}\Big[\prod_{k:z_k\leqslant 1} z_k^{S_{k\leftarrow j}} \prod_{m:z_m>1} z_m^{S_{m\leftarrow j}(u)}\Big]$$

$$\overset{(*)}{=} \mathbb{E}\Big[\prod_{i=1}^{d} z_i^{S_{i\leftarrow j}}\Big] = f_j(z),$$

which implies that the lim inf and lim sup coincide, and so $f_j(z,u) \to f_j(z)$ for all $j \in [d]$. Provided all components converge, we have convergence of the vector $\boldsymbol{f}(z,u) \to \boldsymbol{f}(z)$. Hence, from (15) we obtain for each $j \in [d]$ and with $\boldsymbol{z} \in \mathscr{D}_f$ that

$$f_j(z) = z_j\,\mathbb{E}\Big[\exp\Big(\sum_{m=1}^{d} B_{mj}c_{mj}(f_m(z)-1)\Big)\Big], \tag{21}$$

yielding a vector-valued fixed-point representation for $\boldsymbol{f}(z)$.

— *Proof of (ii)*. We now implicitly characterize the domain $\mathscr{D}_f$. To that end, consider the function $\boldsymbol{G} : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}^d$, where each entry $G_j : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$, with $j \in [d]$, is given by

$$G_j(z,\boldsymbol{x}) = z_j\,\mathbb{E}\Big[\exp\Big(\sum_{m=1}^{d} B_{mj}c_{mj}(x_m-1)\Big)\Big] - x_j. \tag{22}$$

Note that $\boldsymbol{G}(\cdot)$ is continuously differentiable for all $\boldsymbol{z} \in \mathbb{R}^d$ and $\boldsymbol{x} \in \mathbb{R}^d$ for which the respective moment generating functions of $\boldsymbol{B}_j$ are defined. To obtain a characterization of $\mathscr{D}_f$, we need to find the set of $\boldsymbol{z}$ on the boundary of $\mathscr{D}_f$, for which $\boldsymbol{f}(z)$ exists and satisfies (18). Since Eqn. (18) is analogous to solving $\boldsymbol{G}(z,\boldsymbol{f}(z)) = \boldsymbol{0}$, we can find the domain of $\boldsymbol{f}(\cdot)$ by investigating the set $\boldsymbol{G}^{-1}(\boldsymbol{0}) = \{(z,\boldsymbol{x}) : \boldsymbol{G}(z,\boldsymbol{x}) = \boldsymbol{0}\}$. Since the preimage $\boldsymbol{G}^{-1}(\boldsymbol{0})$ can be a complicated set, we resort to the preimage theorem, a variation of the implicit function theorem also known as the regular level set theorem, see e.g., [40, Theorem 9.9], which states two results. First, the preimage has codimension equal to the dimension of the image, and second, the tangent space at a point of the preimage coincides with the kernel of the Jacobian at that point, provided that the Jacobian is of full-rank.

We proceed by providing a specification of the preimage theorem in our setting. The first part of the preimage theorem states that $\boldsymbol{G}^{-1}(\boldsymbol{0})$ is a d-dimensional space. Note that in the univariate ($d = 1$) setting, $\boldsymbol{G}^{-1}(0)$ would be a curve embedded in $\mathbb{R}\times\mathbb{R}$, and the tangent space would be a line. In our multivariate ($d > 1$) setting, $\boldsymbol{G}^{-1}(\boldsymbol{0})$ is a d-dimensional manifold embedded in $\mathbb{R}^d\times\mathbb{R}^d$, and the tangent space is again d-dimensional. The second part concerns tangent spaces, defined as follows: for any $(z,\boldsymbol{x}) \in \boldsymbol{G}^{-1}(\boldsymbol{0})$, the tangent space $T_{z,\boldsymbol{x}}(\boldsymbol{G}^{-1}(\boldsymbol{0}))$ consists of the set of vectors $(\boldsymbol{q},\boldsymbol{r}) \in \mathbb{R}_+^d \times \mathbb{R}_+^d$ for which there exists a curve $\gamma \subseteq \boldsymbol{G}^{-1}(\boldsymbol{0})$ with $\gamma(0) = (z,\boldsymbol{x})$ and $\gamma'(0) = (\boldsymbol{q},\boldsymbol{r})$. The second part of the preimage theorem then states

$$\operatorname{Ker}\big(\boldsymbol{J}_{\boldsymbol{G}}(z,\boldsymbol{x})\big) = T_{z,\boldsymbol{x}}(\boldsymbol{G}^{-1}(\boldsymbol{0})), \tag{23}$$

with $\boldsymbol{J}_{\boldsymbol{G}}(z,\boldsymbol{x}) \in \mathbb{R}^{d\times d}$ denoting the full Jacobian of $\boldsymbol{G}$ evaluated at $(z,\boldsymbol{x})$. We compute the $d\times d$-dimensional Jacobian matrices of partial derivatives of $\boldsymbol{G}$ w.r.t. z and $\boldsymbol{x}$ separately by

$$\boldsymbol{J}_{\boldsymbol{G},z} := \left[\frac{\partial G_j}{\partial z_k}(z,\boldsymbol{x})\right]_{j,k\in[d]} = \left[\boldsymbol{1}_{\{j=k\}}\mathbb{E}\Big[\exp\Big(\sum_{m=1}^{d} B_{mj}c_{mj}(x_m-1)\Big)\Big]\right]_{j,k\in[d]},$$

and

$$\boldsymbol{J}_{\boldsymbol{G},\boldsymbol{x}} := \left[\frac{\partial G_j}{\partial x_k}(z,\boldsymbol{x})\right]_{j,k\in[d]} = \left[z_j\mathbb{E}\Big[B_{kj}c_{kj}\exp\Big(\sum_{m=1}^{d}B_{mj}c_{mj}(x_m-1)\Big)\Big]-\mathbf{1}_{\{j=k\}}\right]_{j,k\in[d]},$$

such that $\boldsymbol{J}_{\boldsymbol{G}} = (\boldsymbol{J}_{\boldsymbol{G},z}\,|\,\boldsymbol{J}_{\boldsymbol{G},\boldsymbol{x}})$. To utilize Eqn. (23), we need to look for vectors $(\boldsymbol{q},\boldsymbol{r}) \in \mathbb{R}_+^d\times\mathbb{R}_+^d$ such that $\boldsymbol{J}_{\boldsymbol{G}}\cdot(\boldsymbol{q},\boldsymbol{r}) = \boldsymbol{J}_{\boldsymbol{G},z}\cdot\boldsymbol{q}+\boldsymbol{J}_{\boldsymbol{G},\boldsymbol{x}}\cdot\boldsymbol{r} = \mathbf{0}$, where we denote $\boldsymbol{J}_{\boldsymbol{G}} \equiv \boldsymbol{J}_{\boldsymbol{G}}(z,\boldsymbol{x})$ for brevity. Note that $\boldsymbol{J}_{\boldsymbol{G},z}$ is a diagonal matrix with positive entries, such that $\boldsymbol{J}_{\boldsymbol{G},z}\cdot\boldsymbol{q} = \mathbf{0}$ only if $\boldsymbol{q} = \mathbf{0}$; so we can focus on $\boldsymbol{r}$. Observe that the set of points where the determinant of the Jacobian $\boldsymbol{J}_{\boldsymbol{G},\boldsymbol{x}}$ has the same sign is a connected set, due to strict convexity in each entry of the functions $G_j(z,\boldsymbol{x})$, with $j\in[d]$, and by continuity of the determinant and partial derivatives. For a fixed vector $\boldsymbol{r}\in\mathbb{R}_+^d$, we can establish systems of equations for z and $\boldsymbol{x}$ such that $\boldsymbol{J}_{\boldsymbol{G},\boldsymbol{x}}\cdot\boldsymbol{r} = \mathbf{0}$. As will become clear later in the proof, convexity will play a crucial role in determining uniqueness of these solutions.

With the objective of substantiating the claim in (23), we compute the solution to the systems of equations $\boldsymbol{G}(z,\boldsymbol{x}) = \mathbf{0}$ and $\boldsymbol{J}_{\boldsymbol{G},\boldsymbol{x}}\cdot\boldsymbol{r} = \mathbf{0}$ by using the expression for $\boldsymbol{J}_{\boldsymbol{G},\boldsymbol{x}}$ and solving for z and $\boldsymbol{x}$. This yields the solutions $\widehat{z} = (\widehat{z}_1,\dots,\widehat{z}_d)$ and $\widehat{\boldsymbol{x}} = (\widehat{x}_1,\dots\widehat{x}_d)$ given in Eqns. (19) and (20), with $(\widehat{z},\widehat{\boldsymbol{x}}) \equiv (\widehat{z}_{\boldsymbol{r}},\widehat{\boldsymbol{x}}_{\boldsymbol{r}})$ parameterized by vectors $\boldsymbol{r}\in\mathbb{R}_+^d$, such that $\boldsymbol{G}(\widehat{z},\widehat{\boldsymbol{x}}) = \mathbf{0}$ and $\boldsymbol{J}_{\boldsymbol{G},\boldsymbol{x}}(\widehat{z},\widehat{\boldsymbol{x}})\cdot\boldsymbol{r} = \mathbf{0}$. Moreover, for any given $\boldsymbol{r}\in\mathbb{R}_+^d$, we show that the associated pair $(\widehat{z}_{\boldsymbol{r}},\widehat{\boldsymbol{x}}_{\boldsymbol{r}})$ is unique. The condition $\boldsymbol{J}_{\boldsymbol{G},\boldsymbol{x}}(\widehat{z}_{\boldsymbol{r}},\widehat{\boldsymbol{x}}_{\boldsymbol{r}})\cdot\boldsymbol{r} = \mathbf{0}$ stated for each row yields $\nabla_{\boldsymbol{x}}G_j(\widehat{z}_{\boldsymbol{r}},\widehat{\boldsymbol{x}}_{\boldsymbol{r}})\cdot\boldsymbol{r} = 0$ for all $j\in[d]$, with $\nabla_x G_j(\cdot)$ the j-th row of the Jacobian. Note that each $G_j(\cdot)$ only depends on z_j and $\boldsymbol{x}$. Due to strict convexity of $G_j(z,\boldsymbol{x})$ in each entry, we have that the sub-level set $G_j^{-1}(\leqslant 0) := \{(z,\boldsymbol{x}) : G_j(z,\boldsymbol{x}) \leqslant 0\}$ is a strictly convex set, and the level set $G_j^{-1}(0)$ is the boundary of $G_j^{-1}(\leqslant 0)$. This implies that

$$\boldsymbol{G}^{-1}(\mathbf{0}) = \{(z,\boldsymbol{x}) : G_j(z,\boldsymbol{x}) = 0, \forall j\in[d]\} = \bigcap_{j=1}^{d} G_j^{-1}(0),$$

is the boundary of a strictly convex set, namely $\boldsymbol{G}^{-1}(\leqslant\mathbf{0})$, as the latter is the intersection of strictly convex sets. Since $\boldsymbol{J}_{\boldsymbol{G},\boldsymbol{x}}(\widehat{z}_{\boldsymbol{r}},\widehat{\boldsymbol{x}}_{\boldsymbol{r}})\cdot\boldsymbol{r} = \mathbf{0}$ means $\boldsymbol{r}\in T_{\widehat{z}_{\boldsymbol{r}},\widehat{\boldsymbol{x}}_{\boldsymbol{r}}}(\boldsymbol{G}^{-1}(\mathbf{0}))$ by Eqn. (23), and since $\boldsymbol{G}^{-1}(\mathbf{0})$ is the boundary of a strictly convex set, we have that $\boldsymbol{r}$ uniquely determines the point $(\widehat{z}_{\boldsymbol{r}},\widehat{\boldsymbol{x}}_{\boldsymbol{r}})$.

The next step amounts to relating what we found so far to the domain $\mathscr{D}_f$. A given value of $z\in\mathbb{R}_+^d$ determines whether one can find $\boldsymbol{x}\in\mathbb{R}_+^d$ for which $\boldsymbol{G}(z,\boldsymbol{x}) = \mathbf{0}$, such that $(z,\boldsymbol{x})\in\boldsymbol{G}^{-1}(\mathbf{0})$. Observe that the set $R_z := \{z\in\mathbb{R}_+^d : z = \widehat{z}_{\boldsymbol{r}}, \boldsymbol{r}\in\mathbb{R}_+^d\}$ divides the positive quadrant $\mathbb{R}_+^d$ into two disjoint sets. The first set is the inner (convex) region, defined as the set of $z\in\mathbb{R}_+^d$ enclosed by the origin, the z axes and the set R_z, with R_z included; denote this set by $\mathcal{Z}$. The second set is the outer region, denoted by $\mathcal{Z}^c$, and it is the complement of $\mathcal{Z}$, such that $\mathcal{Z}\cup\mathcal{Z}^c = \mathbb{R}_+^d$. Note that when $z\in\mathcal{Z}^c$, then $\boldsymbol{G}(z,\boldsymbol{x}) \neq \mathbf{0}$ for any $\boldsymbol{x}\in\mathbb{R}_+^d$, since $\mathcal{Z}^c\times\mathbb{R}_+^d$ does not intersect $\boldsymbol{G}^{-1}(\mathbf{0})$. This yields $\boldsymbol{G}^{-1}(\mathbf{0}) = \{(z,\boldsymbol{x}) : z\in\mathcal{Z}, \boldsymbol{G}(z,\boldsymbol{x}) = \mathbf{0}\}$, and using that $\boldsymbol{G}(z,\boldsymbol{f}(z)) = \mathbf{0}$ for all $z\in\mathscr{D}_f$, we find $\mathscr{D}_f\subseteq\mathcal{Z}$, which proves *(ii)*. However, for $z\in\mathcal{Z}$, we may have multiple $\boldsymbol{x}\in\mathbb{R}_+^d$ such that $\boldsymbol{G}(z,\boldsymbol{x}) = \mathbf{0}$, so we investigate this further.

— *Proof of (iii)*. We are left with proving uniqueness of $\boldsymbol{f}(\cdot)$. We prove this by considering points $(z,\boldsymbol{x})\in\boldsymbol{G}^{-1}(\mathbf{0})$ and relating them to $\boldsymbol{f}(\cdot)$. From the preimage theorem, we know that $\boldsymbol{G}^{-1}(\mathbf{0})$ is d-dimensional, so we need only d parameters to describe this set. We can use the implicit function theorem to describe the $\boldsymbol{x}$ coordinate of $(z,\boldsymbol{x})\in\boldsymbol{G}^{-1}(\mathbf{0})$ in terms of an implicit function of z. We consider a particular point in this set and then show how the argument extends to other points.

Consider the point $(z,\boldsymbol{x}) = (\mathbf{1},\mathbf{1})\in\boldsymbol{G}^{-1}(\mathbf{0})$ since it satisfies $\boldsymbol{G}(\mathbf{1},\mathbf{1}) = \mathbf{0}$, where we use Assumption 2 to ensure existence of the moment generating functions of $\boldsymbol{B}_j$ around this point. Evaluated at the point $(\mathbf{1},\mathbf{1})$, the Jacobian of $\boldsymbol{G}$ with respect to $\boldsymbol{x}$ is given by

$$\boldsymbol{J}_{\boldsymbol{G},\boldsymbol{x}}(\mathbf{1},\mathbf{1}) = \boldsymbol{H}^\top - \boldsymbol{I}, \tag{24}$$

which is invertible due to Assumption 1. Then by the implicit function theorem, there exist open sets $V, W \subseteq \mathbb{R}_+^d$ both containing $\mathbf{1}$, and a unique continuously differentiable function $\tilde{\boldsymbol{f}} : V \to W$ such that $\tilde{\boldsymbol{f}}(\mathbf{1}) = \mathbf{1}$ and $\boldsymbol{G}(z, \tilde{\boldsymbol{f}}(z)) = \mathbf{0}$ for all $z \in V$. Note that this implies $V \subseteq \mathcal{Z}$ and that $\tilde{\boldsymbol{f}}(\cdot)$ satisfies the fixed-point equation in (18). Moreover, since $\tilde{\boldsymbol{f}}(\cdot)$ is unique, we have $\tilde{\boldsymbol{f}}(\cdot) = \boldsymbol{f}(\cdot)$ on V and $V \subseteq \mathscr{D}_f^\circ$, provided that $\tilde{\boldsymbol{f}}(\cdot)$ is increasing in all entries, as by definition $\boldsymbol{f}(\cdot)$ is increasing in all entries.

The point $(\mathbf{1}, \mathbf{1})$ is not particularly special; if we take another point $(z_0, \boldsymbol{x}_0) \in \boldsymbol{G}^{-1}(\mathbf{0})$, we find that the Jacobian $\boldsymbol{J}_{\boldsymbol{G},\boldsymbol{x}}(z_0, \boldsymbol{x}_0)$ is invertible provided $(z_0, \boldsymbol{x}_0) \neq (\widehat{z}, \widehat{\boldsymbol{x}})$. We can then apply the implicit function theorem to obtain open sets $z_0 \in V_0 \subseteq \mathbb{R}_+^d$, $\boldsymbol{x}_0 \in W_0 \subseteq \mathbb{R}_+^d$ and a unique map $\tilde{\boldsymbol{f}}_0 : V_0 \to W_0$ that satisfies $\boldsymbol{G}(z, \tilde{\boldsymbol{f}}_0(z)) = \mathbf{0}$ for all $z \in V_0$, again with $V_0 \subseteq \mathcal{Z}$. As before, we obtain $\tilde{\boldsymbol{f}}(\cdot) = \boldsymbol{f}(\cdot)$ on V_0 and $V_0 \subseteq \mathscr{D}_f^\circ$, due to uniqueness of $\tilde{\boldsymbol{f}}_0$, provided $\tilde{\boldsymbol{f}}_0(\cdot)$ is increasing in all entries. Since we can do this for arbitrary points, we obtain uniqueness of $\boldsymbol{f}(\cdot)$ on all of $\mathcal{Z}$, such that $\mathcal{Z} \subseteq \mathscr{D}_f^\circ$. Finally, for any pair of solutions $(\widehat{z}, \widehat{\boldsymbol{x}})$ to Eqns. (19) and (20), we have by monotonicity of $\boldsymbol{f}(\cdot)$ that $\lim_{z \nearrow \widehat{z}} \boldsymbol{f}(z) = \widehat{\boldsymbol{x}}$, which yields the characterization $\mathcal{Z} = \mathscr{D}_f$. □

We remark that taking $d = 1$ in Proposition 1 yields agreement with the results obtained in [28, Theorem 3.1.1], where our condition that the implicit function is increasing, is equivalent to the condition in [28] where they take the minimal solution of the equation $G(z, x) = 0$ for fixed $z < \widehat{z}$, with $G(\cdot, \cdot)$ defined in (22). Next, we focus on the limiting cumulant of $\boldsymbol{Z}(\cdot)$ given in (13). Note that the moment generating function of $\boldsymbol{Z}(t)$ in (12), and hence also in (13), is expressed in terms of the moment generating functions of the random vectors $\boldsymbol{U}_1, \ldots, \boldsymbol{U}_d$. Denote $\boldsymbol{m}_{\boldsymbol{U}}(\boldsymbol{\theta}) = (m_{\boldsymbol{U}_1}(\boldsymbol{\theta}), \ldots, m_{\boldsymbol{U}_d}(\boldsymbol{\theta}))^\top$ as the vector of moment generating functions of $\boldsymbol{U}_1, \ldots, \boldsymbol{U}_d$. We impose the following condition, assumed to hold throughout the paper.

Assumption 3. *Assume that for any $\widehat{z}$ in (19), there exists a vector $\widehat{\boldsymbol{\theta}} \in \mathbb{R}^{d^\star}$ such that*

$$\boldsymbol{m}_{\boldsymbol{U}}(\widehat{\boldsymbol{\theta}}) = \widehat{z}. \tag{25}$$

Define the function $\Lambda : \mathbb{R}^{d^\star} \to \mathbb{R}$ by

$$\Lambda(\boldsymbol{\theta}) = \sum_{j=1}^{d} \overline{\lambda}_j \big(f_j(\boldsymbol{m}_{\boldsymbol{U}}(\boldsymbol{\theta})) - 1 \big). \tag{26}$$

Also define the domain of convergence $\mathscr{D}_\Lambda := \{\boldsymbol{\theta} \in \mathbb{R}^{d^\star} : \Lambda(\boldsymbol{\theta}) < \infty\}$, denote its interior by $\mathscr{D}_\Lambda^\circ$ and denote by $\partial \mathscr{D}_\Lambda$ its boundary. We note that the dimension of elements in $\mathscr{D}_\Lambda$ is $d^\star$, whereas that of elements in $\mathscr{D}_f$ is d. We now characterize the limiting cumulant of $\boldsymbol{Z}(\cdot)$ in (13).

Lemma 1. *We have $\mathbf{0} \in \mathscr{D}_\Lambda^\circ$, and for $\boldsymbol{\theta} \in \mathbb{R}^{d^\star}$ such that $\boldsymbol{\theta} \leqslant \widehat{\boldsymbol{\theta}}$, where $\boldsymbol{m}_{\boldsymbol{U}}(\widehat{\boldsymbol{\theta}}) = \widehat{z}$ and with $\widehat{z}$ the solution to (19), we have*

$$\lim_{t \to \infty} \frac{1}{t} \log m_{\boldsymbol{Z}(t)}(\boldsymbol{\theta}) = \Lambda(\boldsymbol{\theta}). \tag{27}$$

Proof of Lemma 1. We showed that $\mathbf{1} \in \mathscr{D}_f^\circ$ in the proof of Proposition 1. We then immediately have by Assumption 3 that the vector of moment generating functions $\boldsymbol{m}_{\boldsymbol{U}}(\cdot)$ is defined in a neighborhood of the origin. Taking $\boldsymbol{\theta} = \mathbf{0}$, we have $\boldsymbol{m}_{\boldsymbol{U}}(\mathbf{0}) = \mathbf{1}$, which implies $\mathbf{0} \in \mathscr{D}_\Lambda^\circ$.

We now prove that Eqn. (27) holds. Combining Eqns. (12) and (14), we obtain

$$m_{\boldsymbol{Z}(t)}(\boldsymbol{\theta}) = \prod_{j=1}^{d} \exp\Big(\overline{\lambda}_j \int_0^t \Big(\mathbb{E}\Big[\prod_{l=1}^{d} m_{\boldsymbol{U}_l}(\boldsymbol{\theta})^{S_{l\leftarrow j}(u)}\Big] - 1\Big)\,\mathrm{d}u\Big),$$

or, equivalently,

$$\frac{1}{t}\log m_{\boldsymbol{Z}(t)}(\boldsymbol{\theta}) = \sum_{j=1}^{d} \overline{\lambda}_j \int_0^1 \Big(\mathbb{E}\Big[\prod_{l=1}^{d} m_{\boldsymbol{U}_l}(\boldsymbol{\theta})^{S_{l\leftarrow j}(vt)}\Big] - 1\Big)\,\mathrm{d}v, \tag{28}$$

performing an elementary change of variables. We want to take limits $t \to \infty$ in Eqn. (28); to that end we first focus on the expectation in the integrand. By mimicking Part (i) of the proof of Proposition 1, we can apply the monotone convergence theorem on the integrand to find

$$\lim_{t\to\infty} \mathbb{E}\Big[\prod_{l=1}^{d} m_{\boldsymbol{U}_l}(\boldsymbol{\theta})^{S_{l\leftarrow j}(vt)}\Big] = \mathbb{E}\Big[\prod_{l=1}^{d} m_{\boldsymbol{U}_l}(\boldsymbol{\theta})^{S_{l\leftarrow j}}\Big],$$

provided $\boldsymbol{\theta} \in \mathscr{D}_\Lambda$. Now since we took $\boldsymbol{\theta} \in \mathscr{D}_\Lambda$, we have by Assumption 3 that there exists $\widehat{\boldsymbol{\theta}} \in \mathbb{R}^{d^\star}$ such that $\boldsymbol{\theta} \leqslant \widehat{\boldsymbol{\theta}}$ and $\boldsymbol{m}_{\boldsymbol{U}}(\widehat{\boldsymbol{\theta}}) = \widehat{\boldsymbol{z}}$. Using again a similar argument as in Part (i) in the proof of Proposition 1, distinguishing between indices $k, n \in [d]$ for which $m_{\boldsymbol{U}_k}(\boldsymbol{\theta}) \leqslant 1$ and $m_{\boldsymbol{U}_n}(\boldsymbol{\theta}) > 1$, we can apply the dominated convergence theorem to obtain

$$\begin{aligned}
&\lim_{t\to\infty} \sum_{j=1}^{d} \overline{\lambda}_j \int_0^1 \Big(\mathbb{E}\Big[\prod_{l=1}^{d} m_{\boldsymbol{U}_l}(\boldsymbol{\theta})^{S_{l\leftarrow j}(vt)}\Big] - 1\Big)\,\mathrm{d}v \\
&= \sum_{j=1}^{d} \overline{\lambda}_j \int_0^1 \lim_{t\to\infty} \Big(\mathbb{E}\Big[\prod_{l=1}^{d} m_{\boldsymbol{U}_l}(\boldsymbol{\theta})^{S_{l\leftarrow j}(vt)}\Big] - 1\Big)\,\mathrm{d}v \\
&= \sum_{j=1}^{d} \overline{\lambda}_j \big(f_j(\boldsymbol{m}_{\boldsymbol{U}}(\boldsymbol{\theta})) - 1\big),
\end{aligned}$$

which proves Eqn. (27). □

4. Large deviations

In this section, we show that the multivariate compound Hawkes process satisfies a large deviations principle (LDP). The proof proceeds by establishing the required conditions on the limiting cumulant $\Lambda(\boldsymbol{\theta})$—*essential smoothness*, most notably—such that the Gärtner-Ellis theorem (see e.g., [14, Theorem 2.3.6]) can be invoked.

First recall the definition of an LDP for $\mathbb{R}^d$-valued random vectors; see [14, Section 1.2] for more background. Let $\mathcal{B}(\mathbb{R}^d)$ be the Borel σ-field on $\mathbb{R}^d$. Consider a family of random vectors $\{\boldsymbol{X}_\epsilon\}_{\epsilon\in\mathbb{R}_+}$ taking values in $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. We say that $\{\boldsymbol{X}_\epsilon\}_{\epsilon\in\mathbb{R}_+}$ satisfies the LDP with rate function $I(\cdot)$ if $I : \mathbb{R}^d \to [0,\infty]$ is a lower semicontinuous mapping, and if for every Borel set $A \in \mathcal{B}(\mathbb{R}^d)$,

$$-\inf_{\boldsymbol{x}\in A^\circ} I(\boldsymbol{x}) \leqslant \liminf_{\epsilon\to 0} \epsilon \log \mathbb{P}(\boldsymbol{X}_\epsilon \in A) \leqslant \limsup_{\epsilon\to 0} \epsilon \log \mathbb{P}(\boldsymbol{X}_\epsilon \in A) \leqslant -\inf_{\boldsymbol{x}\in \overline{A}} I(\boldsymbol{x}), \tag{29}$$

where A° and $\overline{A}$ denote the interior and closure of A. Also, recall that $I(\cdot)$ is lower semicontinuous if, for all $\alpha \geqslant 0$, the level sets $\{\boldsymbol{x} \in \mathbb{R}^d : I(\boldsymbol{x}) \leqslant \alpha\}$ are closed; we call $I(\cdot)$ a *good* rate function if the level sets are compact.

We proceed by establish that $(\boldsymbol{Z}(t)/t)_{t\in\mathbb{R}_+}$ satisfies an LDP on $(\mathbb{R}^{d^\star}, \mathcal{B}(\mathbb{R}^{d^\star}))$, as stated in the following theorem. A distinguishing feature of our proof is that, due to the fact that the distribution of $\boldsymbol{S}_j$ is not explicitly known, we prove *steepness*—a key step in proving essential smoothness—*implicitly*, i.e., through the fixed-point representation (18) that the probability generating functions $f_j(\cdot)$ satisfy. In particular, we cannot mimic the proof that was developed in [39] for the univariate case, as that proof heavily rests on explicit expressions for the univariate cluster size distribution.

Our steepness proof, as given below, may be somewhat obscured by the involved notation and complex objects needed due to the fact that we work in a multivariate setting. To remedy this, we have also included in Appendix A a separate proof for the univariate setting that is based on the same reasoning, but is considerably more transparent.

Theorem 1. *The process $(\mathbf{Z}(t)/t)_{t\in\mathbb{R}_+}$ satisfies on $(\mathbb{R}^{d^\star}, \mathcal{B}(\mathbb{R}^{d^\star}))$ the LDP with good rate function*

$$\Lambda^*(\boldsymbol{x}) = \sup_{\boldsymbol{\theta}\in\mathbb{R}^{d^\star}} (\boldsymbol{\theta}^\top \boldsymbol{x} - \Lambda(\boldsymbol{\theta})). \tag{30}$$

Proof of Theorem 1. The proof relies on an application of the Gärtner-Ellis theorem, for which we need to show that the limiting cumulant $\Lambda(\boldsymbol{\theta})$ is an essentially smooth, lower semicontinuous function. For essential smoothness, we need to show that $\mathscr{D}_\Lambda^\circ$ is non-empty and that $\mathbf{0} \in \mathscr{D}_\Lambda^\circ$, that $\Lambda(\cdot)$ is differentiable on $\mathscr{D}_\Lambda^\circ$, and finally that $\Lambda(\cdot)$ is steep; see [14, Section 2.3] for further details.

Lemma 1 shows that $\mathscr{D}_\Lambda^\circ$ is non-empty and $\mathbf{0} \in \mathscr{D}_\Lambda^\circ$. To show that $\Lambda(\cdot)$ is differentiable on $\mathscr{D}_\Lambda^\circ$, recall from the proof of Proposition 1 that $\boldsymbol{f}(\cdot)$ is continuously differentiable on $\mathscr{D}_f^\circ$, exploiting Assumptions 1 and 2. Using this property, in combination with the fact that the moment generating functions $m_{\boldsymbol{U}_l}(\boldsymbol{\theta})$ are differentiable for $\boldsymbol{\theta} \in \mathscr{D}_\Lambda^\circ$ by invoking Assumption 3, we conclude differentiability of $\Lambda(\cdot)$ on $\mathscr{D}_\Lambda^\circ$.

Next, we prove that $\Lambda(\cdot)$ is steep, i.e., for any $\bar{\boldsymbol{\theta}} \in \partial\mathscr{D}_\Lambda^\circ$ and a sequence $\boldsymbol{\theta}_n \nearrow \bar{\boldsymbol{\theta}}$ as $n \to \infty$, we have that $\lim_{n\to\infty} \|\nabla\Lambda(\boldsymbol{\theta}_n)\|_{\mathbb{R}^{d^\star}} = \infty$. For any $i \in [d^\star]$, we first observe that

$$\begin{aligned}\frac{\partial}{\partial\theta_i}\Lambda(\boldsymbol{\theta}) &= \sum_{j=1}^{d} \bar{\lambda}_j \frac{\partial}{\partial\theta_i} f_j(\boldsymbol{m}_{\boldsymbol{U}}(\boldsymbol{\theta})) \\ &= \sum_{j=1}^{d} \bar{\lambda}_j \sum_{k=1}^{d} \Big(\frac{\partial}{\partial\theta_i} m_{\boldsymbol{U}_k}(\boldsymbol{\theta})\Big)\mathbb{E}\Big[S_{k\leftarrow j} \prod_{l=1}^{d} m_{\boldsymbol{U}_l}(\boldsymbol{\theta})^{S_{l\leftarrow j} - \mathbf{1}_{\{k=l\}}}\Big].\end{aligned} \tag{31}$$

This identity entails that entries of $\nabla\Lambda(\cdot)$ are given in terms of the partial derivatives of the probability generating function of $\boldsymbol{S}_j$, for all $j \in [d]$.

To establish steepness of $\Lambda(\cdot)$, it suffices to show that the partial derivatives of $f_j(\cdot)$ diverge on the boundary of $\mathscr{D}_\Lambda^\circ$. Recall that the input for the probability generating function $\boldsymbol{f}(\cdot)$ in Eqn. (26) is the vector $\boldsymbol{m}_{\boldsymbol{U}}(\boldsymbol{\theta}) \in \mathbb{R}_+^d$. In the remainder of the proof, we first derive steepness of $\boldsymbol{f}(\cdot)$ at a specific $\boldsymbol{z} \in \mathbb{R}_+^d$, after which we consider the setting in which $\boldsymbol{f}(\cdot)$ is evaluated in the vector $\boldsymbol{m}_{\boldsymbol{U}}(\boldsymbol{\theta})$.

Define the matrix $\widehat{\boldsymbol{B}}(z) = (\widehat{B}_{mj}(z))_{m,j\in[d]}$ by

$$\widehat{B}_{mj}(z) := z_j \mathbb{E}\Big[B_{mj} c_{mj} \exp\Big(\sum_{i=1}^{d} B_{ij} c_{ij} (f_i(z) - 1)\Big)\Big]. \tag{32}$$

Taking the partial derivative of the fixed-point representation (18) with respect to z_k, for $k \in [d]$, yields

$$\begin{aligned}\frac{\partial f_j(z)}{\partial z_k} &= \mathbb{E}\Big[\exp\Big(\sum_{m=1}^{d} B_{mj} c_{mj}(f_m(z)-1)\Big)\Big]\mathbf{1}_{\{k=j\}} + \sum_{m=1}^{d}\frac{\partial f_m(z)}{\partial z_k}\widehat{B}_{mj}(z)\\ &= \frac{f_j(z)}{z_j}\mathbf{1}_{\{k=j\}} + \sum_{m=1}^{d}\frac{\partial f_m(z)}{\partial z_k}\widehat{B}_{mj}(z),\end{aligned} \tag{33}$$

where the second equality is due to the fixed-point representation (18) itself. We can write (33) compactly in matrix-vector form by considering the Jacobian $\boldsymbol{J_f}$ of $\boldsymbol{f}(\cdot)$, which yields

$$\boldsymbol{J_f}(z) = (\boldsymbol{I} - \widehat{\boldsymbol{B}}(z)^\top)^{-1}\mathrm{diag}(\boldsymbol{f}(z)/z), \tag{34}$$

where the division $\boldsymbol{f}(z)/z$ is to be understood componentwise, provided the inverse exists. We now explore for which values of z the inverse appearing in (34) fails to exist, i.e., when the associated determinant equals 0. Consider an element $\widehat{z}$ on the boundary of $\mathscr{D}_f$. Recall from Eqn. (19) that this $\widehat{z}$ is parametrized by some positive vector $\boldsymbol{r} \in \mathbb{R}_+^d$. Moreover, in this point we have $\boldsymbol{f}(\widehat{z}) = \widehat{\boldsymbol{x}}$, with $\widehat{\boldsymbol{x}}$ the solution to (20), and hence

$$\widehat{B}_{mj}(\widehat{z}) = \widehat{z}_j\mathbb{E}\Big[B_{mj}c_{mj}\exp\Big(\sum_{i=1}^{d} B_{ij}c_{ij}(\widehat{x}_i - 1)\Big)\Big]. \tag{35}$$

Combining this with Eqns. (19) and (20), we obtain

$$\widehat{\boldsymbol{B}}(\widehat{z})\cdot\boldsymbol{r} = \boldsymbol{r} \iff (\boldsymbol{I} - \widehat{\boldsymbol{B}}(\widehat{z}))\cdot\boldsymbol{r} = \boldsymbol{0}, \tag{36}$$

implying that $\boldsymbol{r}$ is in the kernel of $\boldsymbol{I} - \widehat{\boldsymbol{B}}(\widehat{z})$. Since $\boldsymbol{r}$ is a positive (non-zero) vector, we obtain that $\boldsymbol{I} - \widehat{\boldsymbol{B}}(\widehat{z})$ is not invertible and so

$$\det(\boldsymbol{I} - \widehat{\boldsymbol{B}}(\widehat{z})^\top) = 0. \tag{37}$$

Then by Eqn. (34), we find that the directional derivative into the positive quadrant diverges, i.e., for any $\boldsymbol{q} \in \mathbb{R}_+^d$ and sequence $\{z_n\} \subseteq \mathscr{D}_f^\circ$ such that $z_n \nearrow \widehat{z}$, we have

$$\lim_{z_n\nearrow\widehat{z}} \|\boldsymbol{J_f}(z_n)\cdot\boldsymbol{q}\|_{\mathbb{R}^d} = \infty,$$

since each element of $\mathrm{diag}(\boldsymbol{f}(\widehat{z})/\widehat{z}) = \mathrm{diag}(\widehat{\boldsymbol{x}}/\widehat{z})$ is positive and bounded. This proves that $\boldsymbol{f}(\cdot)$ is steep in each argument.

We now use the above observations to prove steepness of $\Lambda(\cdot)$. By Assumption 3, there exists $\widehat{\boldsymbol{\theta}}$ on the boundary of $\mathscr{D}_\Lambda$ such that $\boldsymbol{m_U}(\widehat{\boldsymbol{\theta}}) = \widehat{z}$. With the same argument as above, we find $\det(\boldsymbol{I} - \widehat{\boldsymbol{B}}(\boldsymbol{m_U}(\widehat{\boldsymbol{\theta}}))^\top) = 0$, such that $\boldsymbol{I} - \widehat{\boldsymbol{B}}(\boldsymbol{m_U}(\widehat{\boldsymbol{\theta}}))^\top$ is not invertible at the boundary of $\mathscr{D}_\Lambda$. Hence, for any positive vector $\boldsymbol{q} \in \mathbb{R}^d$ and a sequence $\{\boldsymbol{\theta}_n\} \subseteq \mathscr{D}_\Lambda^\circ$ such that $\boldsymbol{\theta}_n \nearrow \widehat{\boldsymbol{\theta}}$, we have

$$\liminf_{\boldsymbol{\theta}_n\nearrow\widehat{\boldsymbol{\theta}}} \|\boldsymbol{J_f}(\boldsymbol{m_U}(\boldsymbol{\theta}_n))\cdot\boldsymbol{q}\|_{\mathbb{R}^d} = \infty. \tag{38}$$

If we denote the entries of $\boldsymbol{J_f}$ by $\boldsymbol{J_f}^{(jk)} = \partial f_j / \partial z_k$, then from Eqn. (31), we have

$$\begin{aligned}
&\liminf_{\boldsymbol{\theta}_n \nearrow \widehat{\boldsymbol{\theta}}} \|\nabla\Lambda(\boldsymbol{\theta}_n)\|_{\mathbb{R}^{d^\star}} = \liminf_{\boldsymbol{\theta}_n \nearrow \widehat{\boldsymbol{\theta}}} \Big\| \Big(\frac{\partial}{\partial \theta_1}\Lambda(\boldsymbol{\theta}_n), \dots, \frac{\partial}{\partial \theta_{d^\star}}\Lambda(\boldsymbol{\theta}_n) \Big) \Big\|_{\mathbb{R}^{d^\star}} \\
&= \Big\| \Big(\sum_{j=1}^{d} \overline{\lambda}_j \sum_{k=1}^{d} \liminf_{\boldsymbol{\theta}_n \nearrow \widehat{\boldsymbol{\theta}}} \frac{\partial}{\partial \theta_1} m_{\boldsymbol{U}_k}(\boldsymbol{\theta}_n) \mathbb{E}\Big[S_{k\leftarrow j} \prod_{l=1}^{d} m_{\boldsymbol{U}_l}(\boldsymbol{\theta}_n)^{S_{l\leftarrow j} - \mathbf{1}_{\{k=l\}}} \Big], \\
&\qquad \dots, \sum_{j=1}^{d} \overline{\lambda}_j \sum_{k=1}^{d} \liminf_{\boldsymbol{\theta}_n \nearrow \widehat{\boldsymbol{\theta}}} \frac{\partial}{\partial \theta_{d^\star}} m_{\boldsymbol{U}_k}(\boldsymbol{\theta}_n) \mathbb{E}\Big[S_{k\leftarrow j} \prod_{l=1}^{d} m_{\boldsymbol{U}_l}(\boldsymbol{\theta}_n)^{S_{l\leftarrow j} - \mathbf{1}_{\{k=l\}}} \Big] \Big) \Big\|_{\mathbb{R}^{d^\star}} \\
&\geqslant \Big\| \Big(\sum_{j=1}^{d} \overline{\lambda}_j \sum_{k=1}^{d} \frac{\partial}{\partial \theta_1} m_{\boldsymbol{U}_k}(\widehat{\boldsymbol{\theta}}) \mathbb{E}\Big[\liminf_{\boldsymbol{\theta}_n \nearrow \widehat{\boldsymbol{\theta}}} S_{k\leftarrow j} \prod_{l=1}^{d} m_{\boldsymbol{U}_l}(\boldsymbol{\theta}_n)^{S_{l\leftarrow j} - \mathbf{1}_{\{k=l\}}} \Big], \qquad (39) \\
&\qquad \dots, \sum_{j=1}^{d} \overline{\lambda}_j \sum_{k=1}^{d} \frac{\partial}{\partial \theta_{d^\star}} m_{\boldsymbol{U}_k}(\widehat{\boldsymbol{\theta}}) \mathbb{E}\Big[\liminf_{\boldsymbol{\theta}_n \nearrow \widehat{\boldsymbol{\theta}}} S_{k\leftarrow j} \prod_{l=1}^{d} m_{\boldsymbol{U}_l}(\boldsymbol{\theta}_n)^{S_{l\leftarrow j} - \mathbf{1}_{\{k=l\}}} \Big] \Big) \Big\|_{\mathbb{R}^{d^\star}} \\
&= \Big\| \Big(\sum_{j=1}^{d} \overline{\lambda}_j \sum_{k=1}^{d} \frac{\partial}{\partial \theta_1} m_{\boldsymbol{U}_k}(\widehat{\boldsymbol{\theta}}) \boldsymbol{J}_{\boldsymbol{f}}^{jk}(\boldsymbol{m}_{\boldsymbol{U}}(\widehat{\boldsymbol{\theta}})), \dots, \sum_{j=1}^{d} \overline{\lambda}_j \sum_{k=1}^{d} \frac{\partial}{\partial \theta_{d^\star}} m_{\boldsymbol{U}_k}(\widehat{\boldsymbol{\theta}}) \boldsymbol{J}_{\boldsymbol{f}}^{jk}(\boldsymbol{m}_{\boldsymbol{U}}(\widehat{\boldsymbol{\theta}})) \Big) \Big\|_{\mathbb{R}^{d^\star}} \\
&= \infty,
\end{aligned}$$

where the inequality is an application of Fatou's lemma and the last equality is a consequence of (38).

We finally prove lower semicontinuity of $\Lambda(\cdot)$. Since we consider a metric space $\mathbb{R}^{d^\star}$, it suffices to show lower semicontinuity through sequences. Consider $\boldsymbol{\theta}_n \nearrow \boldsymbol{\theta} \in \mathscr{D}_\Lambda^\circ$ and observe that by Fatou's lemma, we have

$$\liminf_{\boldsymbol{\theta}_n \nearrow \boldsymbol{\theta}} \mathbb{E}\Big[\prod_{l=1}^{d} m_{\boldsymbol{U}_l}(\boldsymbol{\theta}_n)^{S_{l\leftarrow j}} \Big] \geqslant \mathbb{E}\Big[\prod_{l=1}^{d} \liminf_{\boldsymbol{\theta}_n \nearrow \boldsymbol{\theta}} m_{\boldsymbol{U}_l}(\boldsymbol{\theta}_n)^{S_{l\leftarrow j}} \Big], \qquad (40)$$

for any $j \in [d]$. Furthermore, it is easily shown that, for any integer $k \in \mathbb{N}$, another application of Fatou's lemma yields

$$\liminf_{\boldsymbol{\theta}_n \nearrow \boldsymbol{\theta}} m_{\boldsymbol{U}_l}(\boldsymbol{\theta}_n)^k = \liminf_{\boldsymbol{\theta}_n \nearrow \boldsymbol{\theta}} \mathbb{E}\big[\exp(\boldsymbol{\theta}_n^\top \boldsymbol{U}_l) \big]^k \geqslant \mathbb{E}\big[\exp(\boldsymbol{\theta}^\top \boldsymbol{U}_l) \big]^k. \qquad (41)$$

Since the random variables $S_{l\leftarrow j}$ are non-negative, we obtain

$$\begin{aligned}
\liminf_{\boldsymbol{\theta}_n \nearrow \boldsymbol{\theta}} \Lambda(\boldsymbol{\theta}_n) &= \liminf_{\boldsymbol{\theta}_n \nearrow \boldsymbol{\theta}} \sum_{j=1}^{d} \overline{\lambda}_j \Big(\mathbb{E}\Big[\prod_{l=1}^{d} m_{\boldsymbol{U}_l}(\boldsymbol{\theta}_n)^{S_{l\leftarrow j}} \Big] - 1 \Big) \\
&\geqslant \sum_{j=1}^{d} \overline{\lambda}_j \Big(\mathbb{E}\Big[\liminf_{\boldsymbol{\theta}_n \nearrow \boldsymbol{\theta}} \prod_{l=1}^{d} m_{\boldsymbol{U}_l}(\boldsymbol{\theta}_n)^{S_{l\leftarrow j}} \Big] - 1 \Big) \geqslant \Lambda(\boldsymbol{\theta}).
\end{aligned} \qquad (42)$$

We have now verified that the limiting cumulant $\Lambda(\cdot)$ satisfies all conditions for the Gärtner-Ellis theorem [14, Theorem 2.3.6] to apply. This concludes the proof of the LDP. □

The consequence of this LDP is that, for any Borel set $A \in \mathcal{B}(\mathbb{R}^{d^\star})$, we have that the measure $\nu_t : \mathcal{B}(\mathbb{R}^{d^\star}) \to [0,1]$ defined by $\nu_t(A) := \mathbb{P}(\boldsymbol{Z}(t)/t \in A)$ satisfies

$$-\inf_{\boldsymbol{x} \in A^\circ} \Lambda^*(\boldsymbol{x}) \leqslant \liminf_{t\to\infty} \frac{1}{t} \log \nu_t(A) \leqslant \limsup_{t\to\infty} \frac{1}{t} \log \nu_t(A) \leqslant -\inf_{\boldsymbol{x} \in \overline{A}} \Lambda^*(\boldsymbol{x}). \tag{43}$$

We note that when substituting the limiting value of $\boldsymbol{Z}(t)/t$ (as $t \to \infty$), given by (11), into $\Lambda^*(\cdot)$, one obtains 0, as should be. This is verified by taking the partial derivatives of $\Lambda(\cdot)$ w.r.t. $\boldsymbol{\theta}$ and evaluating the resulting expressions at $\boldsymbol{\theta} = \boldsymbol{0}$, using (31)–(33), which indeed yields the limiting value of $\boldsymbol{Z}(t)/t$.

5. Rare event simulation

In this section, we show how to use importance sampling to efficiently estimate rare event probabilities. This is accomplished by first exponentially twisting the underlying probability measure $\mathbb{P}$. In Section 5.1, as a contribution of potential independent interest, we describe how to identify the model primitives under this new measure, which we throughout refer to as $\mathbb{Q}$. In the two subsequent subsections, we specifically consider the probability of ruin in component $i \in [d]$ (of which the logarithmic asymptotics are derived in Proposition 2), and the probability of the multivariate compound Hawkes process attaining rare values (of which the logarithmic asymptotics have been established in Theorem 1). An assessment of the performance of the rare event simulation procedures is in Appendix E.

5.1. Identification of the alternative distribution

In this subsection, we describe how to construct the exponentially twisted version of the multivariate compound Hawkes process, which we associate with the probability measure $\mathbb{Q}$, without having explicit expressions for the moment and probability generating functions of the original process. More specifically, we identify a stochastic process of which the limiting cumulant equals, for a vector $\boldsymbol{\theta}^\star \in \mathbb{R}^{d^\star}$,

$$\Psi^{\mathbb{Q}}(\boldsymbol{\theta}) := \Psi(\boldsymbol{\theta} + \boldsymbol{\theta}^\star) - \Psi(\boldsymbol{\theta}^\star),$$

with $\Psi(\boldsymbol{\theta}) = \Lambda(\boldsymbol{\theta}) - \boldsymbol{r}^\top \boldsymbol{\theta}$, and, by virtue of Lemma 1,

$$\Lambda(\boldsymbol{\theta}) = \sum_{j=1}^{d} \overline{\lambda}_j \big(f_j(\boldsymbol{m}_{\boldsymbol{U}}(\boldsymbol{\theta})) - 1\big). \tag{44}$$

To this end, it is first verified that

$$\Lambda(\boldsymbol{\theta} + \boldsymbol{\theta}^\star) - \Lambda(\boldsymbol{\theta}^\star) = \sum_{j=1}^{d} \overline{\lambda}_j f_j(\boldsymbol{m}_{\boldsymbol{U}}(\boldsymbol{\theta}^\star)) \left(\frac{f_j(\boldsymbol{m}_{\boldsymbol{U}}(\boldsymbol{\theta} + \boldsymbol{\theta}^\star))}{f_j(\boldsymbol{m}_{\boldsymbol{U}}(\boldsymbol{\theta}^\star))} - 1 \right).$$

Next, for $j \in [d]$,

$$\begin{aligned} \frac{f_j(\boldsymbol{m}_{\boldsymbol{U}}(\boldsymbol{\theta} + \boldsymbol{\theta}^\star))}{f_j(\boldsymbol{m}_{\boldsymbol{U}}(\boldsymbol{\theta}^\star))} &= \frac{1}{f_j(\boldsymbol{m}_{\boldsymbol{U}}(\boldsymbol{\theta}^\star))} \sum_{\boldsymbol{n} \in \mathbb{N}_0^d} \mathbb{P}(\boldsymbol{S}_j = \boldsymbol{n}) \prod_{l=1}^{d} (m_{\boldsymbol{U}_l}(\boldsymbol{\theta} + \boldsymbol{\theta}^\star))^{n_l} \\ &= \sum_{\boldsymbol{n} \in \mathbb{N}_0^d} \frac{\mathbb{P}(\boldsymbol{S}_j = \boldsymbol{n})}{f_j(\boldsymbol{m}_{\boldsymbol{U}}(\boldsymbol{\theta}^\star))} \prod_{l=1}^{d} (m_{\boldsymbol{U}_l}(\boldsymbol{\theta}^\star))^{n_l} \left(\frac{m_{\boldsymbol{U}_l}(\boldsymbol{\theta} + \boldsymbol{\theta}^\star)}{m_{\boldsymbol{U}_l}(\boldsymbol{\theta}^\star)} \right)^{n_l}. \end{aligned}$$

Now define, for $j \in [d]$,

$$\mathbb{Q}(\boldsymbol{S}_j = \boldsymbol{n}) := \frac{\mathbb{P}(\boldsymbol{S}_j = \boldsymbol{n})}{f_j(\boldsymbol{m}_{\boldsymbol{U}}(\boldsymbol{\theta}^\star))} \prod_{l=1}^{d} (m_{\boldsymbol{U}_l}(\boldsymbol{\theta}^\star))^{n_l},$$

which induces a probability distribution (i.e., non-negative and summing to 1) by its very construction. Define by $f_j^{\mathbb{Q}}(z)$ the corresponding probability generating function, which is the counterpart of $f_j(z)$ under $\mathbb{Q}$: for $j \in [d]$,

$$f_j^{\mathbb{Q}}(\boldsymbol{z}) = \sum_{\boldsymbol{n} \in \mathbb{N}_0^d} \mathbb{Q}(\boldsymbol{S}_j = \boldsymbol{n}) \prod_{l=1}^{d} z_l^{n_l} = \frac{f_j(m_{\boldsymbol{U}_1}(\boldsymbol{\theta}^\star)z_1, \ldots, m_{\boldsymbol{U}_d}(\boldsymbol{\theta}^\star)z_d)}{f_j(\boldsymbol{m}_{\boldsymbol{U}}(\boldsymbol{\theta}^\star))}. \tag{45}$$

In addition, define, for $l \in [d]$,

$$\mathbb{Q}(U_{1l} \in \mathrm{d}x_1, \ldots, U_{d^\star l} \in \mathrm{d}x_{d^\star}) := \frac{\mathbb{P}(U_{1l} \in \mathrm{d}x_1, \ldots, U_{d^\star l} \in \mathrm{d}x_{d^\star})}{m_{\boldsymbol{U}_l}(\boldsymbol{\theta}^\star)} \prod_{k=1}^{d^\star} e^{\theta_k^\star x_k}, \tag{46}$$

which generates a probability distribution (i.e., non-negative and integrating to 1); let $m_{\boldsymbol{U}_l}^{\mathbb{Q}}(\boldsymbol{\theta})$ be the associated moment generating function, given by

$$m_{\boldsymbol{U}_l}^{\mathbb{Q}}(\boldsymbol{\theta}) = \frac{m_{\boldsymbol{U}_l}(\boldsymbol{\theta} + \boldsymbol{\theta}^\star)}{m_{\boldsymbol{U}_l}(\boldsymbol{\theta}^\star)}.$$

We finally define the base rates under $\mathbb{Q}$ via

$$\overline{\lambda}_j^{\mathbb{Q}} := \overline{\lambda}_j f_j(\boldsymbol{m}_{\boldsymbol{U}}(\boldsymbol{\theta}^\star)), \tag{47}$$

for $j \in [d]$.

Upon combining the objects defined above, it now requires an elementary verification to conclude that

$$\Lambda(\boldsymbol{\theta} + \boldsymbol{\theta}^\star) - \Lambda(\boldsymbol{\theta}^\star) = \sum_{j=1}^{d} \overline{\lambda}_j^{\mathbb{Q}} \left(f_j^{\mathbb{Q}}(\boldsymbol{m}_{\boldsymbol{U}}^{\mathbb{Q}}(\boldsymbol{\theta})) - 1 \right),$$

as desired; cf. (44). This means that we have uniquely characterized the joint distribution of the cluster sizes $\boldsymbol{S}_j$ (for $j \in [d]$), the joint distribution of the claim sizes $\boldsymbol{U}_l$ (for $l \in [d]$), and the base rates under the alternative measure $\mathbb{Q}$.

The only question left is: How does one sample a cluster size $\boldsymbol{S}_j$ under $\mathbb{Q}$? More concretely: What is the distribution of the marks B_{lj} under the alternative measure $\mathbb{Q}$, and how should the corresponding decay functions $g_{lj}(\cdot)$ be adapted? To this end, we revisit (18), which we rewrite to

$$f_j(z) = z_j\, m_{\boldsymbol{B}_j}(c_{1j}(f_1(z) - 1), \ldots, c_{dj}(f_d(z) - 1)), \tag{48}$$

using the self-evident notation

$$m_{\boldsymbol{B}_j}(\boldsymbol{\theta}) := \mathbb{E} \exp\left(\sum_{m=1}^{d} \theta_m B_{mj} \right).$$

Introduce the compact notation $\boldsymbol{y}_{\boldsymbol{\theta}^\star}(z) := (m_{\boldsymbol{U}_1}(\boldsymbol{\theta}^\star)z_1, \dots, m_{\boldsymbol{U}_d}(\boldsymbol{\theta}^\star)z_d)^\top$. Hence, as an immediate consequence of (45), we obtain

$$f_j^{\mathbb{Q}}(z) = \frac{f_j(\boldsymbol{y}_{\boldsymbol{\theta}^\star}(z))}{f_j(\boldsymbol{y}_{\boldsymbol{\theta}^\star}(\mathbf{1}))}.$$

Upon combining the two previous displays, we conclude that we can rewrite $f_j^{\mathbb{Q}}(z)$ as

$$\begin{aligned} f_j^{\mathbb{Q}}(z) &= \frac{m_{\boldsymbol{U}_j}(\boldsymbol{\theta}^\star)z_j \cdot m_{\boldsymbol{B}_j}(c_{1j}(f_1(\boldsymbol{y}_{\boldsymbol{\theta}^\star}(z)) - 1), \dots, c_{dj}(f_d(\boldsymbol{y}_{\boldsymbol{\theta}^\star}(z)) - 1))}{m_{\boldsymbol{U}_j}(\boldsymbol{\theta}^\star) \cdot m_{\boldsymbol{B}_j}(c_{1j}(f_1(\boldsymbol{y}_{\boldsymbol{\theta}^\star}(\mathbf{1})) - 1), \dots, c_{dj}(f_d(\boldsymbol{y}_{\boldsymbol{\theta}^\star}(\mathbf{1})) - 1))} \\ &= z_j \frac{m_{\boldsymbol{B}_j}(c_{1j}(f_1(\boldsymbol{y}_{\boldsymbol{\theta}^\star}(z)) - 1), \dots, c_{dj}(f_d(\boldsymbol{y}_{\boldsymbol{\theta}^\star}(z)) - 1))}{m_{\boldsymbol{B}_j}(c_{1j}(f_1(\boldsymbol{y}_{\boldsymbol{\theta}^\star}(\mathbf{1})) - 1), \dots, c_{dj}(f_d(\boldsymbol{y}_{\boldsymbol{\theta}^\star}(\mathbf{1})) - 1))}. \end{aligned}$$

To simplify this further, we write

$$c_{lj}^{\mathbb{Q}} = c_{lj} f_l(\boldsymbol{y}_{\boldsymbol{\theta}^\star}(\mathbf{1})), \quad \bar{c}_{lj}^{\mathbb{Q}} := c_{lj}^{\mathbb{Q}} - c_{lj}, \tag{49}$$

such that

$$f_j^{\mathbb{Q}}(z) = z_j \frac{m_{\boldsymbol{B}_j}\left(c_{1j}^{\mathbb{Q}}\left(\frac{f_1(\boldsymbol{y}_{\boldsymbol{\theta}^\star}(z))}{f_1(\boldsymbol{y}_{\boldsymbol{\theta}^\star}(\mathbf{1}))} - 1\right) + \bar{c}_{1j}^{\mathbb{Q}}, \dots, c_{dj}^{\mathbb{Q}}\left(\frac{f_d(\boldsymbol{y}_{\boldsymbol{\theta}^\star}(z))}{f_d(\boldsymbol{y}_{\boldsymbol{\theta}^\star}(\mathbf{1}))} - 1\right) + \bar{c}_{dj}^{\mathbb{Q}}\right)}{m_{\boldsymbol{B}_j}(\bar{c}_{1j}^{\mathbb{Q}}, \dots, \bar{c}_{dj}^{\mathbb{Q}})}.$$

We now focus on the distribution of the marks under the alternative measure $\mathbb{Q}$. Denoting $\bar{\boldsymbol{c}}_j^{\mathbb{Q}} = (\bar{c}_{1j}^{\mathbb{Q}}, \dots, \bar{c}_{dj}^{\mathbb{Q}})^\top$, we define

$$\mathbb{Q}(B_{1j} \in \mathrm{d}x_1, \dots, B_{dj} \in \mathrm{d}x_d) := \frac{\mathbb{P}(B_{1j} \in \mathrm{d}x_1, \dots, B_{dj} \in \mathrm{d}x_d)}{m_{\boldsymbol{B}_j}(\bar{\boldsymbol{c}}_j^{\mathbb{Q}})} \prod_{l=1}^{d} e^{\bar{c}_{lj}^{\mathbb{Q}} x_l}, \tag{50}$$

so that

$$m_{\boldsymbol{B}_j}^{\mathbb{Q}}(\boldsymbol{\theta}) = \frac{m_{\boldsymbol{B}_j}(\boldsymbol{\theta} + \bar{\boldsymbol{c}}_j^{\mathbb{Q}})}{m_{\boldsymbol{B}_j}(\bar{\boldsymbol{c}}_j^{\mathbb{Q}})}.$$

Combining the above relations, we thus conclude that

$$f_j^{\mathbb{Q}}(z) = z_j \, m_{\boldsymbol{B}_j}^{\mathbb{Q}}\left(c_{1j}^{\mathbb{Q}}(f_1^{\mathbb{Q}}(z) - 1), \dots, c_{dj}^{\mathbb{Q}}(f_d^{\mathbb{Q}}(z) - 1)\right),$$

which has, appealing to Eqn. (48), the right structure. This means that we have identified the distribution of the marks and the decay functions under $\mathbb{Q}$.

The following summarizes the above findings. Most importantly, the exponentially twisted version of the multivariate compound Hawkes process is again a multivariate compound Hawkes process, but (evidently) with different model primitives. Specifically, the $\boldsymbol{\theta}^\star$-twisted version of the multivariate compound Hawkes process can be constructed as follows:

- the base rate $\overline{\lambda}_j$ is replaced by $\overline{\lambda}_j^{\mathbb{Q}} = \overline{\lambda}_j \, f_j(\boldsymbol{m}_{\boldsymbol{U}}(\boldsymbol{\theta}^\star))$; cf. (47).
- the density of $\boldsymbol{U}_l$ is replaced by $\mathbb{Q}(U_{1l} \in \mathrm{d}x_1, \dots, U_{d^\star l} \in \mathrm{d}x_{d^\star})$, as given by (46);

- the density of $\boldsymbol{B}_j$ is replaced by $\mathbb{Q}(B_{1j} \in \mathrm{d}x_1, \ldots, B_{dj} \in \mathrm{d}x_d)$, as given by (50);
- the decay function $g_{lj}(\cdot)$ is replaced by $g_{lj}^{\mathbb{Q}}(\cdot) := g_{lj}(\cdot)\, f_l(\boldsymbol{y}_{\boldsymbol{\theta}^\star}(\mathbf{1})) = g_{lj}(\cdot)\, f_l(\boldsymbol{m}_{\boldsymbol{U}}(\boldsymbol{\theta}^\star))$; cf. (49).

This exponentially twisting mechanism generalizes the one identified for the *univariate* compound Hawkes process with *unit* marks, featuring in the statement of [39, Theorem 2.2].

5.2. Ruin probabilities

In this subsection, we consider, for a given $i \in [d^\star]$, the following net cumulative claim process (or: risk process):

$$Y_i(t) := Z_i(t) - rt, \tag{51}$$

in which claims are generated by a multivariate compound Hawkes process $\boldsymbol{Z}(t)$ and $r > 0$ is a constant premium rate per unit of time. Our first objective is to find the asymptotics of the associated ruin probability, i.e., the probability that this net cumulative process ever exceeds level u, for some $u > 0$.

From the LLN result for $\boldsymbol{Z}(t)/t$ given in Eqn. (11), we have that

$$\frac{Z_i(t)}{t} \to \mathbb{E}[\boldsymbol{U}_{(i)}](\boldsymbol{I} - \boldsymbol{H})^{-1}\overline{\boldsymbol{\lambda}}, \tag{52}$$

a.s., with $\mathbb{E}[\boldsymbol{U}_{(i)}] = (\mathbb{E}[U_{i1}], \ldots, \mathbb{E}[U_{id}])$. To make sure that ruin is rare, we impose throughout the *net profit condition*:

$$r > \mathbb{E}[\boldsymbol{U}_{(i)}](\boldsymbol{I} - \boldsymbol{H})^{-1}\overline{\boldsymbol{\lambda}}, \tag{53}$$

such that the process $Y_i(t)$ drifts towards $-\infty$. For some initial capital $u > 0$, the time of ruin is defined as

$$\tau_u := \inf\{t > 0 : u + rt - Z_i(t) < 0\} = \inf\{t > 0 : Y_i(t) > u\},$$

and the associated infinite horizon ruin probability is defined as

$$p(u) := \mathbb{P}(\tau_u < \infty). \tag{54}$$

We study the behavior of $p(u)$ for u large.

From Lemma 1, it immediately follows that the limiting cumulant function of $Y_i(\cdot)$ satisfies

$$\Psi_i(\theta) := \lim_{t\to\infty} \frac{1}{t} \log \mathbb{E}\big[e^{\theta Y_i(t)}\big] = \Lambda_i(\theta) - r\theta, \tag{55}$$

with

$$\Lambda_i(\theta) := \lim_{t\to\infty} \frac{1}{t} \log \mathbb{E}[e^{\theta Z_i(t)}] = \Lambda(0, \ldots, \theta, \ldots, 0).$$

By [14, Lemma 2.3.9], we know that $\Lambda_i(\cdot)$ is a convex function, which implies that $\Psi_i(\cdot)$ is also convex. We assume throughout the paper that we are in the light-tailed regime, in the sense that there exists $\theta^\star > 0$ such that

$$\Psi_i(\theta^\star) = 0. \tag{56}$$

We prove that $p(u)$ decays essentially exponentially as u increases, as made precise in the following proposition, the proof of which is in Appendix B.

Proposition 2. *For fixed $i \in [d^\star]$, the ruin probability $p(u)$ associated to the risk process $Y_i(\cdot)$ has logarithmic decay rate $-\theta^\star$, i.e.,*

$$\lim_{u\to\infty} \frac{1}{u} \log p(u) = -\theta^\star, \tag{57}$$

where $\theta^\star$ is the unique positive solution of (56).

We emphasize that although we consider just one out of the $d^\star$ net cumulative claim processes, the setting is still intrinsically multi-dimensional, as it involves the multivariate compound Hawkes process (of which the components cannot be described in isolation).

In our importance sampling algorithm presented in what follows, we will twist $\boldsymbol{Y}(t)$ by $\boldsymbol{\theta}^\star = (0,\dots,0,\theta^\star,0,\dots,0)^\top$, with the $\theta^\star$ corresponding to the i-th entry and solving $\Psi_i(\theta^\star) = 0$. We show that this leads to an estimator that is asymptotically efficient (also sometimes referred to as logarithmically efficient, or asymptotically optimal); in the remainder of this subsection, we refer to the alternative measure induced by this specific twist for fixed and given i by $\mathbb{Q}$. For more background on optimality notions of importance sampling procedures, such as asymptotic efficiency, we refer to [3, Section VI.1]. Our proof is given in Appendix C. In principle, it follows the same structure as the one given in [39, Section 4]; therefore, we focus on the main innovations in this general, multivariate setting with random marks.

Recall that $p(u) = \mathbb{P}(\tau_u < \infty)$, with τ_u the first time that $Y_i(\cdot)$ exceeds level u. First note that $p(u) = \mathbb{E}_{\mathbb{Q}}[L_{\tau_u} I]$, where I is the indicator function of the event $\{\tau_u < \infty\}$ and L_{τ_u} is the appropriate likelihood ratio, which quantifies the likelihood of the sampled path under $\mathbb{P}$ relative to $\mathbb{Q}$. More precisely, L_{τ_u} is the Radon-Nikodym derivative of the sampled path under the measure $\mathbb{P}$ relative to the measure $\mathbb{Q}$, evaluated at the ruin time τ_u. As in [39, Lemma 4.3], it can be concluded that—essentially due to the fact that we changed the drift of the risk process from a negative value (under $\mathbb{P}$) into a positive value (under $\mathbb{Q}$)—under $\mathbb{Q}$ eventually any positive value is reached by the process $Y_i(\cdot)$. Thus, $I \equiv 1$ with $\mathbb{Q}$-probability 1, and hence $p(u) = \mathbb{E}_{\mathbb{Q}}[L_{\tau_u}]$.

Following the reasoning in [39] (i.e., effectively relying on a general result in [26]), we can express the likelihood ratio in terms of the various quantities pertaining to the original measure $\mathbb{P}$ and their counterparts under $\mathbb{Q}$. Indeed, the likelihood ratio at time t equals

$$\begin{aligned}\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\Big|_{\mathcal{F}_t} = L_t = \exp\left(-\sum_{j=1}^{d}\int_0^t (\lambda_j(s) - \lambda_j^{\mathbb{Q}}(s))\,\mathrm{d}s\right)\exp\left(\sum_{j=1}^{d}\int_0^t \log\frac{\lambda_j(s)}{\lambda_j^{\mathbb{Q}}(s)}\mathrm{d}N_j(s)\right)\\ \times\exp\left(\sum_{j=1}^{d}\sum_{r=1}^{N_j(t)}\log \ell_j(\boldsymbol{B}_{j,r})\right)\prod_{j=1}^{d}\frac{m_{\boldsymbol{U}_j}(\boldsymbol{\theta}^\star)^{N_j(t)}}{e^{\theta_j^\star Z_j(t)}},\end{aligned} \tag{58}$$

where

$$\lambda_j(s) = \overline{\lambda}_j + \sum_{l=1}^{d}\sum_{r=1}^{N_l(s)} B_{jl,r}\, g_{jl}(s - T_{l,r}), \quad \lambda_j^{\mathbb{Q}}(s) = \overline{\lambda}_j^{\mathbb{Q}} + \sum_{l=1}^{d}\sum_{r=1}^{N_l(s)} B_{jl,r}\, g_{jl}^{\mathbb{Q}}(s - T_{l,r}), \tag{59}$$

with $\overline{\lambda}_j^{\mathbb{Q}}$ and $g_{jl}^{\mathbb{Q}}(\cdot)$ as defined in Section 5.1 and all random objects sampled under $\mathbb{Q}$, and with $\ell_j(\boldsymbol{x})$ denoting the ratio of the density of the random marks $\boldsymbol{B}_j$ under $\mathbb{P}$ and its counterpart under $\mathbb{Q}$ evaluated in the argument $\boldsymbol{x}$. It is directly seen, from the construction of the measure $\mathbb{Q}$, that

$$\lambda_j^{\mathbb{Q}}(s) = \lambda_j(s) f_j(\boldsymbol{m_U}(\boldsymbol{\theta}^\star)) > \lambda_j(s).$$

Note that the relation between $\lambda_j^{\mathbb{Q}}(s)$ and $\lambda_j(s)$, and the fact that $\boldsymbol{\theta}^\star$ is non-zero in the i-th entry, allows us to express the likelihood ratio as

$$L_t = \exp\left(-\sum_{j=1}^{d}\left(1 - f_j(\boldsymbol{m}_{\boldsymbol{U}}(\boldsymbol{\theta}^\star))\right)\int_0^t \lambda_j(s)\mathrm{d}s\right) e^{-\theta^\star Z_i(t)}$$
$$\times \exp\left(\sum_{j=1}^{d}\sum_{r=1}^{N_j(t)} \log \ell_j(\boldsymbol{B}_{j,r})\right)\prod_{j=1}^{d}\left(\frac{m_{\boldsymbol{U}_j}(\boldsymbol{\theta}^\star)}{f_j(\boldsymbol{m}_{\boldsymbol{U}}(\boldsymbol{\theta}^\star))}\right)^{N_j(t)}. \tag{60}$$

We now introduce the importance sampling estimator and establish its efficiency. With $n \in \mathbb{N}$, we define the importance sampling estimator of $p(u)$ by

$$p_n(u) := \frac{1}{n}\sum_{m=1}^{n} L_{\tau_u}^{(m)}, \tag{61}$$

where $L_{\tau_u}^{(m)}$ (for $m = 1, \ldots, n$) are independent replications of L_{τ_u}, sampled under $\mathbb{Q}$. In our context, asymptotic efficiency is to be understood as

$$\lim_{u\to\infty} \frac{1}{u} \log \sqrt{\mathbb{V}\mathrm{ar}_{\mathbb{Q}} L_{\tau_u}} \leqslant \lim_{u\to\infty} \frac{1}{u} \log p(u),$$

that is, the measure $\mathbb{Q}$ is asymptotically efficient for simulations; see Siegmund's criterion [38].

Theorem 2. *The importance sampling estimator $p_n(u)$ in (61), which relies on the alternative measure $\mathbb{Q}$ that corresponds to the exponential twist $\boldsymbol{\theta}^\star = (0, \ldots, 0, \theta^\star, 0, \ldots, 0)^\top$, is asymptotically efficient.*

Remark 1. In the final part of the proof of Theorem 2, contained in Appendix C, we have in passing derived a Lundberg-type inequality for this non-standard ruin model. Indeed, we have that the ruin probability $p(u)$, corresponding to the net cumulative claim process $Y_i(\cdot)$, satisfies the upper bound

$$p(u) \leqslant e^{-\theta^\star u}, \tag{62}$$

uniformly in $u > 0$.

The immediate consequence of the above theorem, which substantially generalizes [39, Theorem 4.5], is the following. Suppose that we wish to obtain an estimate with a certain *precision*, defined as the ratio of the confidence interval's half-width (which is proportional to the standard deviation of the estimate) and the estimate itself. Using simulation under the actual measure $\mathbb{P}$, the number of runs required to obtain a given precision is inversely proportional to the probability to be estimated. In our specific case, this means that under $\mathbb{P}$, due to Proposition 2, this number grows exponentially in u (roughly like $e^{\theta^\star u}$, that is). Under the alternative measure $\mathbb{Q}$, however, Theorem 2 entails that the number of runs to achieve this precision grows *subexponentially*, thus yielding a substantial variance reduction. This means that, despite the fact that the ruin probability decays very rapidly as u grows, the simulation effort required to estimate it grows at a relatively modest pace.

5.3. Exceedance probabilities

In this subsection, we consider the estimation of multivariate exceedance probabilities of the type

$$q_t(\boldsymbol{a}) := \mathbb{P}\left(\frac{Z_1(t)}{t} \geqslant a_1, \ldots, \frac{Z_{d^\star}(t)}{t} \geqslant a_{d^\star}\right),$$

where the set $A := [a_1, \infty) \times \cdots \times [a_{d^\star}, \infty)$ does not contain the vector $\boldsymbol{\mu}$, with, as before

$$\mu_i = \mathbb{E}[\boldsymbol{U}_{(i)}](\boldsymbol{I} - \boldsymbol{H})^{-1}\overline{\boldsymbol{\lambda}},$$

the asymptotic value of the process $Z_i(t)/t$. We consider the regime that t grows large, in which the event of interest becomes increasingly rare by Theorem 1. We show that the associated importance sampling estimator is asymptotically efficient.

Let $I \equiv I_{\boldsymbol{a}}$ be the indicator for the rare event, i.e., set $I_{\boldsymbol{a}} := \{Z_1(t) \geqslant a_1 t, \ldots, Z_{d^\star}(t) \geqslant a_{d^\star} t\}$ for any given $t > 0$. We define the importance sampling estimator for the probability of this event by

$$q_{t,n}(\boldsymbol{a}) := \frac{1}{n}\sum_{m=1}^{n} L_t^{(m)} I_{\boldsymbol{a}}^{(m)}, \tag{63}$$

where $L_t^{(m)}$ are independent replications of L_t, sampled under $\mathbb{Q}$ with a twist parameter depending on $\boldsymbol{a}$. Here, $I_{\boldsymbol{a}}^{(m)}$ are the associated indicators. We state the following theorem, the proof of which is contained in Appendix D.

Theorem 3. *The importance sampling estimator $q_{t,n}(\boldsymbol{a})$ in (63), when using the alternative measure $\mathbb{Q}$ that corresponds to the exponential twist*

$$\boldsymbol{\theta}(\boldsymbol{a}^\star) = \arg\sup_{\boldsymbol{\theta}}(\boldsymbol{\theta}^\top \boldsymbol{a}^\star - \Lambda(\boldsymbol{\theta})),$$

with $\boldsymbol{a}^\star := \arg\inf_{\boldsymbol{x}\in A} \Lambda^\star(\boldsymbol{x})$, is asymptotically efficient.

The analysis becomes significantly more complicated when considering rare event sets that are *unions* rather than *intersections*, e.g., $\{Z_1(t) \geqslant a_1 t\} \cup \{Z_2(t) \geqslant a_2 t\}$, in which we do not have a uniform bound on the likelihood ratio $L_t\, I_{\boldsymbol{a}}$ (as opposed to the case of an intersection of events; see the proof of Theorem 3). There are various ways to deal with this inherent complication; see e.g., the discussions on this issue in [33].

6. Concluding remarks

This paper has established a large deviations principle for multivariate compound Hawkes processes, with the underlying Hawkes processes admitting general decay functions and random marks. In order to prove the LDP, the main technical hurdle concerned proving that the limiting cumulant is steep. Our steepness proof is methodologically novel, in that we manage to show that the derivative of the cumulant grows to infinity when approaching the boundary of its domain, but, remarkably, without having an explicit characterization of this domain. Using the LDP, the logarithmic asymptotics of the corresponding ruin probability are identified. The final contribution concerns the development of rare event simulation procedures, based on importance sampling, and proven to be asymptotically efficient.

An interesting topic for future research is to consider other types of deviations for multivariate compound Hawkes processes, such as precise or process-level large deviations. Furthermore, recalling that the steepness proof does not require knowledge of the boundary of the domain, it could also be explored whether our approach carries over to a broader class of processes with an underlying branching structure.

Acknowledgements

We are very grateful to the Editor Prof. Davy Paindaveine, an Associate Editor and three referees for their comments and suggestions that have greatly improved the paper. We are also grateful to Yacine Aït-Sahalia, Frank den Hollander, Lex Schrijver and seminar participants at Kyushu University, the University of Vienna, the University of Amsterdam, the online Stochastic Networks, Applied Probability, and Performance (SNAPP) seminar and the Conference on Stochastic Models VII at the Mathematical Research and Conference Center in Będlewo for helpful comments. This research was funded in part by the Netherlands Organization for Scientific Research under grant NWO VICI 2019/20 (Laeven) and by the Netherlands Organization for Scientific Research under the Gravitation project NETWORKS, grant 024.002.003 (Mandjes).

Supplementary Material

Supplement to "Compound multivariate Hawkes processes: Large deviations and rare event simulation"
Provides: (i) the proofs that are omitted from the main text; and (ii) all simulation experiments.

References

[1] AÏT-SAHALIA, Y., CACHO-DIAZ, J. A. & LAEVEN, R. J. A. (2015). Modeling financial contagion using mutually exciting jump processes. *J. Financ. Econ.* **117** 585–606. https://doi.org/10.1016/j.jfineco.2015.03.002

[2] AÏT-SAHALIA, Y., LAEVEN, R. J. A. & PELIZZON, L. (2014). Mutual excitation in Eurozone sovereign CDS. *J. Econometrics* **183** 151–167. https://doi.org/10.1016/j.jeconom.2014.05.006

[3] ASMUSSEN, S & GLYNN, P. (2007). *Stochastic Simulation: Algorithms and Analysis*. Springer, New York. https://doi.org/10.1007/978-0-387-69033-9

[4] BACRY, E., DELATTRE, S., HOFFMANN, M. & MUZY, J. F. (2013). Some limit theorems for Hawkes processes and application to financial statistics. *Stochastic Process. Appl.* **123** 2475–2499. https://doi.org/10.1016/j.spa.2013.04.007

[5] BALDWIN, A., GHEYAS, I., IOANNIDIS, CH., PYM, D. & WILLIAMS, J. (2017). Contagion in cyber security attacks. *J. Oper. Res. Soc.* **68** 780–791. https://doi.org/10.1057/jors.2016.37

[6] BESSY-ROLAND Y., BOUMEZOUED, A. & HILLAIRET, C. (2021). Multivariate Hawkes process for cyber risk insurance. *Ann. Actuar. Sci.* **15** 14–39. https://doi.org/10.1017/S1748499520000093

[7] BORDENAVE, C. & TORRISI, G. L. (2007). Large deviations of Poisson cluster processes. *Stoch. Models* **23** 593–625. https://doi.org/10.1080/15326340701645959

[8] BRAUNSTEINS, P. & MANDJES, M. (2023). The Cramér-Lundberg model with a fluctuating number of clients. *Insurance Math. Econom.* **112** 1–22. https://doi.org/10.1016/j.insmatheco.2023.05.007

[9] CHAUMONT, L. & LIU, R. (2014). Coding multitype forests: Application to the law of the total population of branching forests. *Trans. Amer. Math. Soc.* **368** 2723–2747. https://doi.org/10.1090/tran/6421

[10] COLLAMORE, J. F. (1996). Hitting probabilities and large deviations. *Ann. Probab.* **24** 2065–2078. https://doi.org/10.1214/aop/1041903218

[11] CRANE, R. & SORNETTE, D. (2008). Robust dynamic classes revealed by measuring the response function of a social system. *Proc. Natl. Acad. Sci. USA* **105** 15649–15653. https://doi.org/10.1073/pnas.0803685105

[12] DALEY, D. J. & VERE-JONES, D. (2003). *An Introduction to the Theory of Point Processes. Volume 1: Elementary Theory and Methods. Volume 2: General Theory and Structure.* Second Edition. Springer, New York. https://doi.org/10.1007/b97277, https://doi.org/10.1007/978-0-387-49835-5

[13] DASSIOS, A., JANG, J. & ZHAO, H. (2015). A risk model with renewal shot-noise Cox process. *Insurance Math. Econom.* **65** 55–65. https://doi.org/10.1016/j.insmatheco.2015.08.009

[14] DEMBO, A. & ZEITOUNI, O. (1998). *Large Deviations Techniques and Applications*. Second Edition. Springer, New York. https://doi.org/10.1007/978-3-642-03311-7

[15] DUFFIELD, N. & O'CONNELL, N. (1995). Large deviations and overflow probabilities for the general single-server queue, with applications. *Math. Proc. Cambridge Philos. Soc.* **118** 363–374. https://doi.org/10.1017/S0305004100073709

[16] FENG, M., CAI, S.-M., TANG, M., & LAI, Y.-C. (2019). Equivalence and its invalidation between non-Markovian and Markovian spreading dynamics on complex networks. *Nat. Commun.* **10** 3748. https://doi.org/10.1038/s41467-019-11763-z

[17] GAO, X. & ZHU, L. (2018). Large deviations and applications for Markovian Hawkes processes with a large initial intensity. *Bernoulli* **24** 2875–2905. https://doi.org/10.3150/17-BEJ948

[18] GAO, X. & ZHU, L. (2018). Limit theorems for Markovian Hawkes processes with a large initial intensity. *Stochastic Process. Appl.* **128** 3807–3839. https://doi.org/10.1016/j.spa.2017.12.001

[19] GAO, X. & ZHU, L. (2019). Affine point processes: Refinements to large-time asymptotics. Preprint. Available at https://arxiv.org/pdf/1903.06371.pdf.

[20] GAO, F. & ZHU, L. (2021). Precise deviations for Hawkes processes. *Bernoulli* **27** 221–248. https://doi.org/10.3150/20-BEJ1235

[21] GUO, Q., REMILLARD, B. & SWISHCHUK, A. (2020). Multivariate general compound point processes in limit order books. *Risks* **8** 98. https://doi.org/10.3390/risks8030098

[22] HAWKES, A. G. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika* **58** 83–90. https://doi.org/10.1093/biomet/58.1.83

[23] HAWKES, A. G. & OAKES, D. (1974). A cluster representation of a self-exciting process. *J. Appl. Probab.* **11** 493–503. https://doi.org/10.2307/3212693

[24] HAWKES, A. G. (2018). Hawkes processes and their applications to finance: A review. *Quant. Finance* **18** 193–198. https://doi.org/10.1080/14697688.2017.1403131

[25] IKEFUJI, M., LAEVEN, R. J. A., MAGNUS, J. R. & YUE, Y. (2022). Earthquake risk embedded in property prices: Evidence from five Japanese cities. *J. Amer. Statist. Assoc.* **117** 82–93. https://doi.org/10.1080/01621459.2021.1928512

[26] JACOD, J. & MÉMIN, J (1976). Caractéristiques locales et conditions de continuité absolue pour les semi-martingales. *Probab. Theory Related Fields* **35** 1–37. https://doi.org/10.1007/BF00532597

[27] JAGERS, P. (1975). *Branching Processes with Biological Applications.* First edition. Wiley, London.

[28] KARABASH, D. & ZHU, L. (2015). Limit theorems for marked Hawkes processes with application to a risk model. *Stoch. Models* **31** 433–451. https://doi.org/10.1080/15326349.2015.1024868

[29] KARIM, R. S., LAEVEN, R. J. A. & MANDJES, M. (2021). Exact and asymptotic analysis of general multivariate Hawkes processes and induced population processes. Preprint. Available at https://arxiv.org/pdf/2106.03560.pdf.

[30] KARIM, R. S., LAEVEN, R. J. A. & MANDJES, M. (2024). Supplement to "Compound multivariate Hawkes processes: Large deviations and rare event simulation". *Bernoulli*.

[31] LIN, Z.-H., FENG, M., TANG, M., LIU, Z., XU, C., HUI, P. M. & LAI, Y.-C. (2020). Non-Markovian recovery makes complex networks more resilient against large-scale failures. *Nat. Commun.* **11** 2490. https://doi.org/10.1038/s41467-020-15860-2

[32] LINIGER, T. J. (2009). *Multivariate Hawkes Processes.* PhD Thesis, ETH Zürich.

[33] MANDJES, M. & BLANCHET, J. (2009). Rare-event simulation for queues. In: G. Rubino and B. Tuffin (eds.): *Rare Event Simulation Using Monte Carlo Methods*, Wiley, Chichester, pp. 87–124. https://doi.org/10.1002/9780470745403

[34] NYRHINEN, H. (1994). Rough limit results for level crossing probabilities. *J. Appl. Probab.* **31** 373–382. https://doi.org/10.2307/3215030

[35] OGATA, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes. *J. Amer. Statist. Assoc.* **83** 9–27. https://doi.org/10.1080/01621459.1988.10478560

[36] Pojer, S. & Thonhauser, S. (2023). Ruin probabilities in a Markovian shot-noise environment. *J. Appl. Probab.* **60** 542–556. https://doi.org/10.1017/jpr.2022.63

[37] Reynaud-Bouret, P. & Schbath, S. (2010). Adaptive estimation for Hawkes processes; application to genome analysis. *Ann. Statist.* **38** 2781–2822. https://doi.org/10.1214/10-AOS806

[38] Siegmund, D. (1976). Importance sampling in the Monte Carlo study of sequential tests. *Ann. Statist.* **4** 673–684. https://doi.org/10.1214/aos/1176343541

[39] Stabile, G. & Torrisi, G. L. (2010). Risk processes with non-stationary Hawkes claims arrivals. *Methodol. Comput. Appl. Probab.* **12** 415–429. https://doi.org/10.1007/s11009-008-9110-6

[40] Tu, L. W. (2011). *An Introduction to Manifolds.* Second edition. Springer, New York. https://doi.org/10.1007/978-1-4419-7400-6

[41] Yao, N. (2018). Moderate deviations for multivariate Hawkes processes. *Statist. Probab. Lett.* **140** 71–76. https://doi.org/10.1016/j.spl.2018.04.023

[42] Zhang, X., Blanchet, J., Giesecke, K. & Glynn, P. (2015). Affine point processes: approximation and efficient simulation. *Math. Oper. Res.* **40** 797–819. https://doi.org/10.1287/moor.2014.0696

[43] Zhu, L. (2013). Ruin probabilities for risk processes with non-stationary arrivals and subexponential claims. *Insurance Math. Econom.* **53** 544–550. https://doi.org/10.1016/j.insmatheco.2013.08.008

[44] Zhu, L. (2014). Process-level large deviations for nonlinear Hawkes point processes. *Ann. Inst. Henri Poincaré Probab. Stat.* **50** 845–871. https://doi.org/10.1214/12-AIHP532

[45] Zhu, L. (2015). Large deviations for Markovian nonlinear Hawkes processes. *Ann. Appl. Probab.* **25** 548–581. https://doi.org/10.1214/14-AAP1003