
Journal of the American Statistical Association

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

# Multivariate Temporal Point Process Regression

Xiwei Tang & Lexin Li

**To cite this article:** Xiwei Tang & Lexin Li (2021): Multivariate Temporal Point Process Regression, Journal of the American Statistical Association, DOI: 10.1080/01621459.2021.1955690

**To link to this article:** https://doi.org/10.1080/01621459.2021.1955690

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2021, VOL. 00, NO. 0, 1–16: Theory and Methods
https://doi.org/10.1080/01621459.2021.1955690



# Multivariate Temporal Point Process Regression

Xiwei Tang[a] and Lexin Li[b]

[a]Department of Statistics, University of Virginia, Charlottesville, VA; [b]Department of Biostatistics and Epidemiology, University of California, Berkeley, CA

**ABSTRACT**

Point process modeling is gaining increasing attention, as point process type data are emerging in a large variety of scientific applications. In this article, motivated by a neuronal spike trains study, we propose a novel point process regression model, where both the response and the predictor can be a high-dimensional point process. We model the predictor effects through the conditional intensities using a set of basis transferring functions in a convolutional fashion. We organize the corresponding transferring coefficients in the form of a three-way tensor, then impose the low-rank, sparsity, and subgroup structures on this coefficient tensor. These structures help reduce the dimensionality, integrate information across different individual processes, and facilitate the interpretation. We develop a highly scalable optimization algorithm for parameter estimation. We derive the large sample error bound for the recovered coefficient tensor, and establish the subgroup identification consistency, while allowing the dimension of the multivariate point process to diverge. We demonstrate the efficacy of our method through both simulations and a cross-area neuronal spike trains analysis in a sensory cortex study.

**ARTICLE HISTORY**
Received December 2020
Accepted July 2021

**KEYWORDS**
Conditional intensity function; Diverging dimension; Neuronal spike trains; Regularization; Temporal process; Tensor decomposition

## 1. Introduction

Point process modeling is drawing increasing attention, as data in the form of point process are emerging in a wide variety of scientific and business applications. Examples include forest ecology (Stoyan et al. 2000), spatial epidemiology (Diggle et al. 2010), social network modeling (Perry and Wolfe 2013), neuronal activity modeling (Brown, Kass, and Mitra 2004; Chen et al. 2019b), functional neuroimaging meta analysis (Kang et al. 2011, 2014), among others. In general, a point process is a collection of events, or points, randomly located in some domain space, for example, a spatial domain or a time domain. Our motivation is a neuronal spike trains analysis in a sensory cortex study (Okun et al. 2015). A newly developed two-photon calcium imaging technique is now greatly facilitating neuroscience studies, by enabling simultaneously recording of the dynamic activities for a population of neurons while maintaining individual neuron resolution (Ji, Freeman, and Smith 2016). In our study, there are 139 and 283 neurons imaged simultaneously from two areas of a rat's brain, the primary visual cortex area (V1) and the primary auditory cortex area (A1). In a visual activity, it is known that some locations of the primary visual cortex would respond to input from auditory and other sensory areas (Liang et al. 2013). One of the scientific goals of this study is to understand the association patterns and information transmissions between the neurons across A1 and V1, and to model potential excitation or inhibition effects of neuron firings between the two areas.

In this article, we propose a new multivariate point process regression model to address this question, where both the response and the predictor can be a high-dimensional point process. We model the predictor effects through the conditional intensities using a set of basis transferring functions in a convolutional fashion. We organize the corresponding transferring coefficients in the form of a three-way tensor, then impose the low-rank, sparsity, and subgroup structures. Both low-rank and sparsity are commonly used low-dimensional structures in high-dimensional data analysis, and are scientifically plausible in neuroscience and many other applications (Zhou, Li, and Zhu 2013a; Chen, Raskutti, and Yuan 2019a; Zhang and Han 2019; Bacry et al. 2020). Subgroup is another frequently used structure in plenty of applications, and it corresponds to ensemble neural activities in neuroscience (Okun et al. 2015). Together these structures effectively reduce the number of free parameters, and also greatly facilitate the model interpretation. We then develop a highly scalable alternating direction method of multipliers (ADMM) algorithm for parameter estimation. We establish the asymptotic properties of the penalized maximum likelihood estimator while allowing the dimensions of both the response and predictor processes to diverge. We also comment that, although motivated by a neuroscience problem, our method is equally applicable to numerous other point process applications, for example, the social infection network learning (Zhou, Zha, and Song 2013b).

There have been a large number of models targeting a spatial point process, where its local intensity is usually assumed to depend on some deterministic features or some location-relevant random variables (see, e.g., Guan 2008; Guan, Jalilian, and Waagepetersen 2015; Kang et al. 2014; Deng et al. 2017). We instead aim at a temporal point process, which is usually evolutionary in nature, in that the occurrence of

**CONTACT** Lexin Li lexinli@berkeley.edu Department of Biostatistics and Epidemiology, University of California, Berkeley, CA 94720.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

a future event depends on the historical realizations of the process. This leads to a different set of model assumptions and modeling techniques. Moreover, most classical inhomogeneous point process solutions target a univariate or bivariate process along with a limited number of predictors (Diggle et al. 2010; Waagepetersen and Guan 2009). We instead target high-dimensional response and predictor processes, and we allow the dimension of both processes to diverge. This has introduced new challenges in both modeling and theoretical analysis.

There have also been a family of models targeting a temporal point process, all of which are built on a self-exciting process called the Hawkes process (Hawkes 1971). A Hawkes process assumes that a future event is triggered by its own past events, and is widely used in neuronal spike trains analysis. In recent years, there have been a number of point process models extending the Hawkes process. Our proposal is related to but also clearly distinctive from the existing models in multiple ways. First of all, our model extends the classical Hawkes process, by simultaneously incorporating nonlinear and inhomogeneous intensities, multiple basis functions, diverging point process dimensions, and additional structures on the transferring coefficients. On the other hand, our model is more general, in that it allows a wide class of stochastic processes to be predictors, and can be applied to the scenarios where the Hawkes process is applicable, but not vice versa. We give some examples in Section 2.2. Second, Zhou, Zha, and Song (2013b) introduced low-rank and sparsity structures in a multivariate Hawkes model. Bacry et al. (2020) studied the theoretical properties of the model, while potentially allowing the point process dimension to diverge. Our proposal employs similar low-dimensional structures and explicitly studies the diverging dimension. However, there are numerous fundamental differences. Zhou, Zha, and Song (2013b), Bacry et al. (2020) both imposed the linear link function and the stationary assumption, whereas we consider a general and potentially nonlinear link, and do not require the process to be stationary. In addition, Zhou, Zha, and Song (2013b) and Bacry et al. (2020) organized the transferring coefficients in a matrix form and placed a low-rank structure on the coefficient matrix, whereas we organize the transferring coefficients in a tensor form and employ a low-rank tensor decomposition. Tensor decomposition is considerably different from matrix decomposition (Kolda and Bader 2009), and naively transforming a tensor to a matrix may lose information. More importantly, the estimation error bound obtained by Bacry et al. (2020) is to increase as the point process dimension diverges, and as such the estimator is to suffer from the increasing dimension. By contrast, in Section 4.2, we show that the diverging dimension is to benefit our penalized maximum likelihood estimator, and leads to a faster convergence rate on the asymptotic concentration. Third, Hansen et al. (2015) considered an intensity-based model for multivariate Hawkes process with a fixed dimension, and adopted the least-square estimation and $\ell_1$ regularization. Cai, Zhang, and Guan (2020) proposed a nonstationary multivariate Hawkes process model, estimated the transferring functions using B-spline approximations along with a group $\ell_1$ penalty. However, the key difference is that Hansen et al. (2015) and Cai, Zhang, and Guan (2020) modeled each individual point process *separately*, while we model multiple processes *jointly*. More specifically, even though they both targeted a multivariate point process, their loss functions were completely separable. As such, they essentially modeled each individual point process one at a time. By contrast, we model multiple response point processes in a joint fashion, in that we integrate and borrow information across different response processes. This is achieved through both the tensor latent factors that are shared across all intensity functions, as well as the subgrouping structure on the transferring coefficients that encourage information sharing among similar individual processes. Such a joint modeling strategy essentially leads to the improved estimation bound when the dimension of response processes increases. Finally, Bacry and Muzy (2016) and Chen et al. (2019b) studied some multivariate Hawkes process models using moment-based statistics, while we focus on modeling the conditional intensity function. In Section 2.2, we discuss in more detail why the intensity-based approach is more suitable than the moment-based approach in our setting.

The rest of the article is organized as follows. Section 2 introduces our proposed multivariate temporal point process regression model. Section 3 develops the estimation algorithm, and Section 4 derives the theoretical properties. Section 5 presents the simulations, and Section 6 illustrates with a neuronal spike trains data analysis. All proofs are relegated to the Supplementary Appendix.

## 2. Model

### 2.1. Background

We begin with a brief review of temporal point process, and we refer to Daley and Vere-Jones (2007) for more details. Specifically, a temporal point process is a stochastic counting process defined on the positive half of the real line $\mathbb{R}^+$, and taking nonnegative integer values. For a univariate process $X(t)$, let $t_1, t_2, \ldots \in \mathbb{R}^+$ denote the event times, under which $X(A) = \sum_{l=1} \mathbf{1}_{[t_l \in A]}$ for any $A \in \mathcal{B}(\mathbb{R}^+)$, and $\mathcal{B}(\mathbb{R}^+)$ denotes the Borel $\sigma$-algebra of $\mathbb{R}^+$. Define its mean intensity function as $\Lambda(t) = \lim_{dt \to 0} \mathbb{E}[dX(t)]/dt$, where $dX(t) = X\big([t, t+dt)\big)$, and $dt$ is an arbitrary small increment of time. A temporal point process is homogeneous if its mean intensity is a constant, and is inhomogeneous otherwise. If $\Lambda(t)$ is also a stochastic process, then it is a doubly stochastic process; for example, a Cox process. A temporal point process is usually assumed to be orderly; that is, $\Pr\{dX(t) > 1\} = o(dt)$, which implies that $\Lambda(t)dt = \Pr\{dX(t) = 1\}$. In addition, a point process $X(t)$ is *stationary* if, for arbitrary bounded Borel subsets $A_1, \ldots, A_r$ of the real line, the joint distribution of $\{X(A_1 + t), \ldots, X(A_r + t)\}$ does not depend on $t$. We also note that there are different forms of stationary definitions; see (Daley and Vere-Jones 2007, Chapter 3.2). Finally, it is straightforward to generalize the notion of a univariate point process to a multivariate point process, that is, $\boldsymbol{X}(t) = \big(X_1(t), \ldots, X_p(t)\big)^{\mathrm{T}}$.

Moment statistics are widely used in point process modeling, especially for a stationary process (Guan 2008, 2011; Chen et al. 2019b). Considering a $p$-dimensional stationary point process $\boldsymbol{X}(t)$, define its first-order moment statistic, that is, the mean

intensity, and its second-order statistic, that is, the covariance, as,

$$\Lambda_i^x = \mathbb{E}\{dX_i(t)\}/dt, \qquad i = 1, \ldots, p,$$

$$V_{ij}^x(\tau) = \mathbb{E}\{dX_i(t)dX_j(t-\tau)\}/\{dtd(t-\tau)\} - \Lambda_i^x \Lambda_j^x - \delta_{ij}(\tau)\Lambda_i^x, \quad i,j = 1, \ldots, p,$$

respectively, where $\delta_{ij}(\tau) = 0$ if $i \neq j$, and $\delta_{ij}(\tau) = \delta(\tau)$ if $i = j$, and $\delta(\cdot)$ denotes the Dirac delta function satisfying that $\delta(x) = 0$ for $x \neq 0$ and $\int_{-\infty}^{+\infty} \delta(x)dx = 1$. Write $\Lambda^x = (\Lambda_1^x, \ldots, \Lambda_p^x)^{\mathrm{T}} \in \mathbb{R}^p$, and $\boldsymbol{V}^{xx}(\cdot) = \left(V_{ij}^x(\cdot)\right) : \mathbb{R} \mapsto \mathbb{R}^{p\times p}$. Analogously, consider another $m$-dimensional stationary point process $\boldsymbol{Y}(t)$, with the mean intensity $\Lambda^y = (\Lambda_1^y, \ldots, \Lambda_m^y)^{\mathrm{T}} \in \mathbb{R}^m$, and the covariance $\boldsymbol{V}^{yy}(\cdot) = \left(V_{ij}^y(\cdot)\right) : \mathbb{R} \mapsto \mathbb{R}^{m\times m}$. The cross-covariance between $X_j(t)$ and $Y_i(t)$ is defined as,

$$C_{ji}^{xy}(\tau) = \mathbb{E}\{dX_j(t)dY_i(t-\tau)\}/\{dtd(t-\tau)\} - \Lambda_j^x \Lambda_i^y, \quad i = 1, \ldots, m, j = 1, \ldots, p.$$

Write $\boldsymbol{C}^{xy}(\cdot) = \left(C_{ji}^{xy}(\cdot)\right) : \mathbb{R} \mapsto \mathbb{R}^{p\times m}$, and $\boldsymbol{C}^{yx}(\cdot) = \left(C_{ij}^{yx}(\cdot)\right) : \mathbb{R} \mapsto \mathbb{R}^{m\times p}$.

Conditional intensity function is another extensively used tool for modeling both spatial and temporal point processes with additional covariates. For our proposed multivariate point process regression, we mainly target the conditional intensity function, as we detail in the next section.

### 2.2. Multivariate Point Process Regression

We consider a temporal regression model with a $p$-dimensional predictor process $\boldsymbol{X}(t)$ and an $m$-dimensional response process $\boldsymbol{Y}(t)$. Letting $\mathcal{H}_t$ denote the $\sigma$-algebra generated by $\{\boldsymbol{X}(t), \boldsymbol{Y}(t)\}$, then the $\mathcal{H}_t$-predictable intensity function $\lambda_i^y(t)$ of the $i$th response process $Y_i(t)$ is defined as,

$$\lambda_i^y(t)dt = \Pr\{dY_i(t) = 1|\mathcal{H}_t\}, \quad i = 1, \ldots, m.$$

We assume this conditional intensity function takes the form,

$$\lambda_i^y(t) = \phi\left\{\mu_i + \sum_{j=1}^{p}\left(\omega_{ij} * dX_j\right)(t)\right\}, \quad i = 1, \ldots, m, \quad (1)$$

where $\phi(\cdot)$ is a link function that is possibly nonlinear, for example, a rectifier function $\phi(x) = \max(0, x)$, or a sigmoid function $\phi(x) = e^x/(1+e^x)$, $\mu_i$ is the background intensity, and $\omega_{ij}(\cdot) : \mathbb{R}^+ \mapsto \mathbb{R}$ is the transferring function, $i = 1, \ldots, m, j = 1, \ldots, p$. Write $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_m)^{\mathrm{T}} \in \mathbb{R}^m$, and $\boldsymbol{\omega} = \left(\omega_{ij}(\cdot)\right) \in \mathbb{R}^{m\times p}$. To account for potential evolution over time, we further assume that the transferring function $\omega_{ij}(\cdot)$ models the historical information of the predictor process $\boldsymbol{X}(t)$ in a convolutional fashion, in that,

$$\left(\omega_{ij} * dX_j\right)(t) = \int_0^t \omega_{ij}(\Delta)dX_j(t-\Delta). \quad (2)$$

Similar formulation such as Equation (2) has been commonly used in the temporal point process literature (see, e.g., Hawkes 1971; Zhou, Zha, and Song 2013b). We also note that, Equations (1) and (2) together cover a fairly general class of models. We do *not* require a linear link function, *nor* the stationary condition. We do *not* enforce $\omega_{ij}(\cdot)$ to be nonnegative, as typically in the classical Hawkes process model, and therefore allow both "exciting" and "inhibiting" effects. In addition, we allow the predictor process $\boldsymbol{X}(t)$ to take a general form. The convolution in Equation (2) actually works for both a stochastic predictor process and a deterministic predictor process, corresponding to a stochastic integral or a Stieltjes integral, respectively.

We next outline a number of examples covered under our proposed model framework.

*Example 1.* Let $X_j(t), j = 1, \ldots, p$, be a deterministic function on $[0, \infty)$, and $\boldsymbol{Y}(t)$ in this case is a multivariate inhomogeneous Poisson process with the deterministic mean intensity function $\lambda_i^y(t), i = 1, \ldots, m$. In a neuronal activity study, $\boldsymbol{X}(t)$ may represent the designed stimulus signal.

*Example 2.* Let $X_j(t), j = 1, \ldots, p$, be a stochastic point process, and $\boldsymbol{Y}(t)$ in this case is a multivariate Cox process or a doubly-stochastic process. If the transferring function $\omega_{ij}(x)$ takes the form such as a Dirac delta function $\delta(x-t)$, then the conditional intensity $\lambda_i^y(t)$ only depends on the value of $\boldsymbol{X}(t)$ at point $t$. Consequently, our model can also be applied to nontemporal point processes, for example, a multivariate spatial point process.

*Example 3.* Let $\boldsymbol{X}(t) = \boldsymbol{Y}(t)$ be the same stochastic point process. Then our proposed model includes the Hawkes process as a special case, while the classical Hawkes process model only considers self-exciting effects, that is, $\omega_{ij}(\cdot)$ has to be nonnegative. Moreover, it usually assumes the process is stationary, which requires additional conditions on $\omega_{ij}(\cdot)$, for example, the spectral radius is smaller than one (Brémaud and Massoulié 1996). We impose neither of these constraints.

*Example 4.* For our motivating neuronal spike trains example, let $\boldsymbol{N}_{V1}(t)$, $\boldsymbol{N}_{A1}(t)$ denote the multivariate point processes for the neuron firing activities on the layers V1 and A1, respectively. If we expect that there are only directed connections between the neurons from A1 to V1 (Liang et al. 2013), then we can set $\boldsymbol{Y}(t) = \boldsymbol{N}_{V1}(t)$, and $\boldsymbol{X}(t) = \boldsymbol{N}_{A1}(t)$. Meanwhile, if we also expect potential connections between the neurons within the same layer of V1, then we can set $\boldsymbol{Y}(t) = \boldsymbol{N}_{V1}(t)$, and $\boldsymbol{X}(t) = (\boldsymbol{N}_{V1}(t), \boldsymbol{N}_{A1}(t))$, which takes the history of $\boldsymbol{N}_{V1}(t)$ into account as well. Consequently, the transferring coefficient $\boldsymbol{\omega}_{A1}$ of $\boldsymbol{N}_{A1}(t)$ in this model captures the association between $\boldsymbol{N}_{V1}(t)$ and $\boldsymbol{N}_{A1}(t)$ after controlling the cross-connections within $\boldsymbol{N}_{V1}(t)$ itself.

In our temporal point process modeling, we mainly target the conditional intensity function, instead of the moment statistics. There are several reasons of doing so. To illustrate, we consider a special case of our model (1), which takes a linear link function, that is,

$$\lambda_i^y(t) = \mu_i + \sum_{j=1}^{p}\left\{\omega_{ij} * dX_j\right\}(t), \quad i = 1, \ldots, m. \quad (3)$$

We next characterize the first- and second-order statistics under this special case.

*Proposition 1.* Consider a special case of Equation (1), such that $\boldsymbol{X}(t)$ and $\boldsymbol{Y}(t)$ satisfy the linear relation (3), and both are stationary. Then the corresponding moment statistics are of the form,

$$\begin{aligned}
&\Lambda^y = \boldsymbol{\mu} + \left\{\int_0^{+\infty} \boldsymbol{\omega}(\Delta)d\Delta\right\}\ \Lambda^x,\\
&\boldsymbol{C}^{yx}(\tau) = \boldsymbol{\omega}(\tau)\mathrm{Diag}(\Lambda^x) + \boldsymbol{\omega} * \boldsymbol{V}^{xx}(\tau), \quad \tau \geq 0\\
&\boldsymbol{V}^{yy}(\tau) = \boldsymbol{\omega} * \boldsymbol{C}^{xy}(\tau), \quad \tau > 0,\\
&\boldsymbol{V}^{yy}(0) = \boldsymbol{\omega} \star \left\{\boldsymbol{V}^{xx}(\cdot) + \mathrm{Diag}(\Lambda^x)\right\} \star \boldsymbol{\omega},
\end{aligned} \tag{4}$$

where $\boldsymbol{\omega} * \boldsymbol{C}^{xy}(\tau) = \omega_{ij}(\cdot) * C^{xy}_{ji}(\tau)$, and $f * g(t) = \int f(\Delta)g(t-\Delta)d\Delta$ denotes the convolution of two univariate functions $f$ and $g$, and $\boldsymbol{\omega} \star \left\{\boldsymbol{V}^{xx}(\cdot) + \mathrm{Diag}(\Lambda^x)\right\} \star \boldsymbol{\omega} = \int_0^{+\infty}\int_0^{+\infty} \boldsymbol{\omega}(\Delta)\left\{\mathrm{Diag}(\Lambda^x)\ \delta(\Delta' - \Delta) + \boldsymbol{V}^{xx}(\Delta' - \Delta)\right\}\boldsymbol{\omega}^{\mathrm{T}}(\Delta')d\Delta d\Delta'$.

The equations in Equation (4) belong to a class of integral equations for the Wiener-Hopf system with respect to $\boldsymbol{\omega}$. In principle, one can estimate the transferring function by solving the above equations and plugging in the estimated first- and second-order statistics. However, this strategy is not suitable for our framework, for several reasons. First, the derived integral equation system for model (3) is more complicated than the one for a regular Hawkes process, in that we require four integral equations involving not only the first- and second-order moments, but also their cross-covariance. Second, when the dimensions of $\boldsymbol{X}_t$ and $\boldsymbol{Y}_t$ are high and diverging, the terms $\boldsymbol{V}^{yy}(\tau)$, $\boldsymbol{V}^{xx}(\tau)$ and $\boldsymbol{C}^{yx}(\tau)$ would also be high-dimensional and expanding. This would make both their sample estimation, as well as carrying out certain operations like matrix inverse, difficult, sometimes even infeasible. Third, the explicit forms of equations in Equation (4) have been derived entirely based on the linear relation (3). There may be no explicit forms like Equation (4) for nonlinear associations between $\boldsymbol{X}(t)$ and $\boldsymbol{Y}(t)$. Finally, a number of moment-based estimation methods require some form of stationary properties, which can be restrictive for some applications in practice. For these reasons, we choose to adopt an intensity-based modeling approach.

### 2.3. Low-Rank Structure

In our model, the transferring function $\boldsymbol{\omega}$ in the intensity fully captures the cross-process connection pattern, and is of the primary interest. In the point process modeling literature, a common strategy is to employ basis functions to characterize the intensity (Xu, Farajtabar, and Zha 2016; Wang et al. 2016a). Adopting this strategy, we assume $\omega_{ij}(t)$ takes the form of a linear combination of a set of basis functions, $g^{(k)}(t), k = 1, \ldots, K$, in that,

$$\omega_{ij}(t) = \sum_{k=1}^{K} \beta_{ij}^k \cdot g^{(k)}(t), \quad i = 1, \ldots, m,\ j = 1, \ldots, p, \tag{5}$$

where each $g^{(k)}(t)$ is a nonnegative basis function on $[0, \infty)$, $K$ is the number of basis functions, and $\beta_{ij}^k$'s can take arbitrary real values. The choice of basis functions mostly relies on the scientific knowledge to account for specific coevolutionary effects (Hansen et al. 2015). In neuronal spike trains study, common basis functions include the exponential function, $g(t) = a\exp(-at), a > 0$ (Zhou, Zha, and Song 2013b), the logarithmic decay function, $g(t) = \log(1 + T - t)$, for the process defined on $[0, T]$ (Luo et al. 2016), and a series of piecewise constant functions, $g^{(k)}(t) = a_k \mathbf{1}(t \in \mathcal{T}_k)$ $(k = 1, \ldots, K)$, where $\{\mathcal{T}_k\}_{k=1}^K$ form a partition of $[0, +\infty]$ and $\{a_k\}_{k=1}^K$ are some nonnegative constants (Wang et al. 2016b). One may also use a mix of different types of basis functions. We later conduct a sensitivity analysis to investigate the choice of basis functions in the appendix.

Given the basis expansion in Equation (5), the conditional intensity model in Equation (1) can be rewritten as follows:

$$\lambda_i^y(t) = \phi\left[\mu_i + \sum_{j=1}^{p}\sum_{k=1}^{K} \beta_{ij}^k \cdot \left\{g^{(k)} * dX_j\right\}(t)\right], \quad i = 1, \ldots, m. \tag{6}$$

We then collect the transferring coefficients into a three-way tensor $\boldsymbol{\mathcal{B}} \in \mathbb{R}^{m\times p\times K}$, with the entry $\beta_{ij}^k$, $i = 1, \ldots, m$, $j = 1, \ldots, p, k = 1, \ldots, K$. The conditional intensity function is now fully characterized by the background intensity vector $\boldsymbol{\mu}$ and the transferring coefficient tensor $\boldsymbol{\mathcal{B}}$.

We estimate the model through a likelihood-based approach, where the joint log-likelihood function is of the form,

$$\begin{aligned}
\mathcal{L}\big(\boldsymbol{\mu}, \boldsymbol{\mathcal{B}}|\boldsymbol{Y}(t), \boldsymbol{X}(t)\big) &= \frac{1}{T}\sum_{i=1}^{m} L_i\big(\boldsymbol{\mu}, \boldsymbol{\mathcal{B}}|Y_i(t), \boldsymbol{X}(t)\big)\\
&= \frac{1}{T}\sum_{i=1}^{m}\int_0^T \left[\log\left\{\lambda_i^y\big(t; \boldsymbol{\mathcal{B}}, \boldsymbol{X}(t)\big)\right\} dY_i(t)\right.\\
&\qquad \left. - \lambda_i^y\big(t; \boldsymbol{\mathcal{B}}, \boldsymbol{X}(t)\big)dt\right].
\end{aligned} \tag{7}$$

In the following discussion, we drop $\boldsymbol{Y}(t)$ and $\boldsymbol{X}(t)$ in $\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\mathcal{B}})$ and $\lambda_i^y(t)$ for notational simplicity whenever there is no confusion.

Next, we impose that $\boldsymbol{\mathcal{B}}$ admits a low-rank CANDECOMP/PARAFAC (CP) structure, in that,

$$\boldsymbol{\mathcal{B}} = \sum_{r=1}^{R} \nu_r \boldsymbol{b}_r^y \circ \boldsymbol{b}_r^x \circ \boldsymbol{b}_r^c, \tag{8}$$

where $R$ is the tensor rank, $\boldsymbol{b}_r^y \in \mathbb{R}^m$, $\boldsymbol{b}_r^x \in \mathbb{R}^p$ and $\boldsymbol{b}_r^c \in \mathbb{R}^K$ are the normalized vectors corresponding to the modes of the response process, the predictor process, and the convolutional basis functions, respectively, $\nu_r$'s are the normalization weights, and "$\circ$" is the outer product. For notational convenience, we represent the decomposition (8) by a shorthand, $\boldsymbol{\mathcal{B}} = [[\boldsymbol{\nu}; \boldsymbol{B}^y, \boldsymbol{B}^x, \boldsymbol{B}^c]]$, where $\boldsymbol{B}^y = [\boldsymbol{b}_1^y\ \ldots \boldsymbol{b}_R^y] \in \mathbb{R}^{m\times R}$, $\boldsymbol{B}^x = [\boldsymbol{b}_1^x\ \ldots \boldsymbol{b}_R^x] \in \mathbb{R}^{p\times R}$, $\boldsymbol{B}^c = [\boldsymbol{b}_1^c\ \ldots \boldsymbol{b}_R^c] \in \mathbb{R}^{K\times R}$, and $\boldsymbol{\nu} = (\nu_1, \ldots, \nu_R)^{\mathrm{T}} \in \mathbb{R}^R$. See Kolda and Bader (2009) for a review of tensor and its decomposition. The low-rank decomposition (8) has been widely adopted in recent years in imaging-based tensor regressions (Zhou, Li, and Zhu 2013a; Sun and Li 2017; Chen, Raskutti, and Yuan 2019a). In the context of point process modeling, Zhou, Zha, and Song (2013b); Bacry et al. (2020) also adopted a low-rank structure in a linear Hawkes model, yet on a matrix form of their transferring coefficients. By contrast, we target a nonlinear, nonstationary, general temporal

point process, and consider a low-rank structure on a coefficient tensor. Even though tensor is a conceptual generalization of matrix, tensor decomposition and matrix decomposition are considerably different (Kolda and Bader 2009). Naively transforming a tensor to a matrix may lose information. Moreover, we later show in Section 4.2 that our estimator and those of Zhou, Zha, and Song (2013b); Bacry et al. (2020) have completely different asymptotic convergence properties.

Imposing the low-rank structure like Equation (8) in our point process regression has several advantages. First, it substantially reduces the number of free parameters in the transferring coefficient tensor $\mathcal{B}$, from $mpK$ to $R(m+p+K)$. In our example, if we set $\boldsymbol{Y}(t) = \boldsymbol{N}_{V1}(t)$, and $\boldsymbol{X}(t) = \boldsymbol{N}_{A1}(t)$, the dimensions of the response and predictor processes are $m = 139$ and $p = 283$, respectively. If we choose $K = 3$ basis functions, and choose the rank $R = 4$, then the number of free parameters in $\mathcal{B}$ reduces from 118,011 to 1700. Second, and perhaps more importantly, it allows us to model the multivariate point process in a *joint* fashion. Existing approaches such as Hansen et al. (2015), Cai, Zhang, and Guan (2020) modeled each response process $Y_i(t)$ separately, since their loss function is separable with respect to the individual intensity function $\lambda_i^y(t)$, each of which depends on a separate set of parameters $\boldsymbol{\beta}_i$. By contrast, our model with Equation (8) suggests that $\mathcal{B}$ relies on some underlying latent factors, $\boldsymbol{B}^y$, $\boldsymbol{B}^x$ and $\boldsymbol{B}^c$. Unlike the separate modeling strategy, the latent factors $\boldsymbol{B}^x$ and $\boldsymbol{B}^c$ are shared by all intensity functions $\lambda_i^y(t)$'s, and thus information across different response processes $Y_i(t)$'s is integrated. In the context of neuronal spike trains modeling, it implies that a particular predictor neuron in $\boldsymbol{X}(t)$ exercises similar influence on multiple response neurons in $\boldsymbol{Y}(t)$, or a particular response neuron in $\boldsymbol{Y}(t)$ enjoys similar influence from multiple predictor neurons in $\boldsymbol{X}(t)$. Such an integration leads to an improved coefficient estimator, as we show asymptotically in Section 4.2.

### 2.4. Additional Structure Pursuit: Sparsity and Subgrouping

To better accommodate scientific knowledge, facilitate the interpretation, and further reduce the number of free parameters, we consider some additional structure pursuit.

The first structure we consider is sparsity, in that each response process is affected by a subset of predictor processes. This sparsity structure simplifies the model interpretation, further reduces the number of parameters, and is scientifically plausible. In multivariate Hawkes process modeling, the sparsity on transferring functions has been widely employed (Hansen et al. 2015; Bacry et al. 2020; Cai, Zhang, and Guan 2020). Specifically, we impose a group $\ell_1$ penalty (Yuan and Lin 2006) on the coefficient tensor $\mathcal{B}$,

$$P_s(\mathcal{B}; \tau_s) = \tau_s \sum_{i=1}^{m} \sum_{j=1}^{p} \left\| \mathcal{B}[i, j, \cdot] \right\|_2, \tag{9}$$

where $\mathcal{B}[i, j, \cdot] \in \mathbb{R}^K$ is a vector of $\mathcal{B}$ with the first two indices fixed and the third index varying, which corresponds to the associations between $Y_i(t)$ and $X_j(t)$ under all basis functions, $\tau_s$ is the sparsity tuning parameter, and $\|\cdot\|_2$ denotes the $\ell_2$ norm.

The second structure we consider is subgrouping. In a neuronal spike trains study, certain subgroups of neurons are expected to share similar patterns in neuronal firing activities, and such clustering patterns are usually of great scientific interest (Kim et al. 2011). Specifically, in our example, such patterns are reflected by the underlying clustering structure in the transferring function $\boldsymbol{\omega}$. To capture this structure, we embed clustering pursuit into the proposed tensor decomposition. In principle, we can pursue clustering on the response process, or the predictor process, or both. For our motivating example, there is evidence of neuron clustering in the primary visual cortex V1, that is, the response process (Liang et al. 2013). As such, we introduce a subgrouping penalty on the decomposed factors of the response process mode $\boldsymbol{B}^y$, so to encourage clustering of the response neurons. Specifically, we impose a pairwise fusion penalty,

$$P_f(\boldsymbol{B}^y; \tau_f) = \sum_{i<i'} f_\kappa \left( \left\| \boldsymbol{B}^y[i, \cdot] - \boldsymbol{B}^y[i', \cdot] \right\|_2, \tau_f \right), \tag{10}$$

where $\boldsymbol{B}^y[i, \cdot] \in \mathbb{R}^R$ is the row vector of $\boldsymbol{B}^y$, $\tau_f$ is the fusion parameter, and the penalty function $f_\kappa(t, \tau_f) = \tau_f \int_0^t \{1 - x/(\tau_f \kappa)\}_+ dx$, with $\kappa$ being a thresholding parameter (Zhang 2010). This penalty function is to help reduce the estimation bias, as it only groups the individual predictors with similar effects on the responses through a non-convex fusion penalty (Zhu, Tang, and Qu 2019).

We also remark that, the sparsity and subgroup structures are embedded in the low-rank CP decomposition, and the three structures are related to each other. Meanwhile, they focus on different aspects of the transferring coefficients. In the neuronal spike trains example, the low-rank structure is to capture the block-wise connection pattern between groups of neurons, the sparsity is on the individual neuron effect, and the subgroup is to identify the neurons that receive signals from the same group of neurons from the other layer. These interrelated structures introduce additional difficulty to parameter estimation. Next, we develop an efficient optimization algorithm.

## 3. Estimation

### 3.1. ADMM Optimization

We develop a highly scalable ADMM type optimization algorithm (Boyd et al. 2011) to estimate the parameters in our proposed model. Consider the realizations of the predictor and response processes $\boldsymbol{X}(t)$ and $\boldsymbol{Y}(t)$ on a time interval $[0, T]$. Let $t_1^i < t_2^i < \cdots < t_{n_i}^i$ denote the time points of the $n_i$ events of the response process $Y_i(t)$ that are observed on $[0, T]$, $i = 1, \ldots, m$. Given the set of basis functions $\{g^{(k)}(\cdot)\}_{k=1}^K$, the log-likelihood function in our model can be written as follows:

$$\mathcal{L}(\boldsymbol{\mu}, \mathcal{B}) = \frac{1}{T} \sum_{i=1}^{m} \Bigg( - \int_0^T \phi \left\{ \mu_i + \langle \boldsymbol{G}(t), \mathcal{B}[i, \cdot, \cdot] \rangle \right\} dt + \sum_{l=1}^{n_i} \log \left[ \phi \left\{ \mu_i + \langle \boldsymbol{G}(t_l^i), \mathcal{B}[i, \cdot, \cdot] \rangle \right\} \right] \Bigg),$$

where $\boldsymbol{G}(t) = \left(G_{j,k}(t)\right) \in \mathbb{R}^{p \times K}$, $G_{j,k}(t) = \{g^{(k)} * dX_j\}(t)$, $\mathcal{B}[i, \cdot, \cdot] \in \mathbb{R}^{p \times K}$ is a matrix from $\mathcal{B}$ with the first index fixed

and the other two indices varying, and $\langle \cdot, \cdot \rangle$ denotes the inner product.

Incorporating the low-rank structure (8) and the two regularization structures (9) and (10), we aim at the following optimization problem,

$$\begin{aligned} &\min_{\boldsymbol{\mu},\boldsymbol{v},\boldsymbol{B}^y,\boldsymbol{B}^x,\boldsymbol{B}^c} \Big\{ -\mathcal{L}\big(\boldsymbol{\mu}, [[\boldsymbol{v}; \boldsymbol{B}^y, \boldsymbol{B}^x, \boldsymbol{B}^c]]\big) \\ &+\tau_s \textstyle\sum_{i=1}^{m} \sum_{j=1}^{p} \big\| [[\boldsymbol{v}; \boldsymbol{B}^y, \boldsymbol{B}^x, \boldsymbol{B}^c]][i, j, \cdot] \big\|_2 \\ &+ \textstyle\sum_{i<i'} f_\kappa \left( \big\| \boldsymbol{B}^y[i, \cdot] - \boldsymbol{B}^y[i', \cdot] \big\|_2, \tau_f \right) \Big\}. \end{aligned} \tag{11}$$

The optimization in Equation (11) is challenging in several ways. It involves a tensor decomposition embedded in a complicated log-likelihood function with summation of integrals and a possibly nonlinear link function $\phi$. In addition, the sparsity penalty in Equation (9) is nondifferentiable, while the fusion penalty in Equation (10) is nonconvex. Moreover, Equation (10) involves the differences of parameters, rendering those parameters inseparable in optimization. To overcome those challenges, and to achieve computational scalability, we develop an ADMM algorithm for the optimization in Equation (11).

Specifically, we introduce two sets of auxiliary variables. The first set is $\Psi \in \mathbb{R}^{m \times p \times K}$ with $\Psi[i, j, \cdot] = \boldsymbol{\psi}_{ij} \in \mathbb{R}^K$ that targets the sparsity structure (9) such that $\boldsymbol{\psi}_{ij} = \boldsymbol{\mathcal{B}}[i, j, \cdot]$, $1 \le i \le m, 1 \le j \le p$. The second set is $\Gamma \in \mathbb{R}^{m(m-1)/2 \times R}$ that stacks $\boldsymbol{\gamma}_{ii'} \in \mathbb{R}^R$ together and targets the subgroup structure (10) such that $\boldsymbol{\gamma}_{ii'} = \boldsymbol{B}^y[i, \cdot] - \boldsymbol{B}^y[i', \cdot]$, $1 \le i < i' \le m$. We then rewrite (11) in its equivalent form,

$$\begin{aligned} \min_{\boldsymbol{\mu},\boldsymbol{v},\boldsymbol{B}^y,\boldsymbol{B}^x,\boldsymbol{B}^c,\boldsymbol{\mathcal{B}},\Psi,\Gamma} \Big\{ &-\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\mathcal{B}}) + \tau_s \sum_{i=1}^{m} \sum_{j=1}^{p} \|\boldsymbol{\psi}_{ij}\|_2 \\ &+ \sum_{j<j'} f_\kappa \left( \|\boldsymbol{\gamma}_{ii'}\|_2, \tau_f \right) \Big\} \\ \text{subject to } \boldsymbol{\mathcal{B}} = [[\boldsymbol{v}; \boldsymbol{B}^y, \boldsymbol{B}^x, \boldsymbol{B}^c]], \quad &\Psi = \boldsymbol{\mathcal{B}}, \quad \Gamma = \boldsymbol{D}_m \boldsymbol{B}^y, \end{aligned} \tag{12}$$

where $\boldsymbol{D}_m \in \mathbb{R}^{m(m-1)/2 \times m}$ that stacks $\boldsymbol{d}_{ii'} \in \mathbb{R}^m$ together, with $\boldsymbol{d}_{ii'} = \boldsymbol{e}_i - \boldsymbol{e}_{i'}$, $\boldsymbol{e}_i \in \mathbb{R}^m$ has one on the $i$th position and zero elsewhere, $1 \le i < i' \le m$. To solve Equation (12), we minimize the following augmented Lagrangian objective function,

$$\begin{aligned} &-\mathcal{L}(\boldsymbol{\mathcal{B}}, \boldsymbol{\mu}) + \tau_s \sum_{i,j} \|\boldsymbol{\psi}_{ij}\|_2 + \sum_{j<j'} f_\kappa(\|\boldsymbol{\gamma}_{ii'}\|_2, \tau_f) \\ &+ \langle \boldsymbol{\mathcal{W}}_1, \boldsymbol{\mathcal{B}} - [[\boldsymbol{v}; \boldsymbol{B}^y, \boldsymbol{B}^x, \boldsymbol{B}^c]] \rangle + \langle \boldsymbol{\mathcal{W}}_2, \Psi - \boldsymbol{\mathcal{B}} \rangle + \langle \boldsymbol{W}_3, \Gamma - \boldsymbol{D}_m \boldsymbol{B}^y \rangle \\ &+ \frac{\rho}{2} \left( \big\| \boldsymbol{\mathcal{B}} - [[\boldsymbol{v}; \boldsymbol{B}^y, \boldsymbol{B}^x, \boldsymbol{B}^c]] \big\|_F^2 + \|\Psi - \boldsymbol{\mathcal{B}}\|_F^2 + \|\Gamma - \boldsymbol{D}_m \boldsymbol{B}^y\|_F^2 \right), \end{aligned}$$

where $\boldsymbol{\mathcal{W}}_1, \boldsymbol{\mathcal{W}}_2 \in \mathbb{R}^{m \times p \times K}$ and $\boldsymbol{W}_3 \in \mathbb{R}^{m(m-1)/2 \times m}$ are the corresponding Lagrangian multipliers, $\rho > 0$ is a fixed augmented parameter, and $\|\cdot\|_F$ denotes the Frobenius norm.

Next, we update the blocks of parameters, $\boldsymbol{\mu}, \boldsymbol{\mathcal{B}}, \boldsymbol{v}, \boldsymbol{B}^y, \boldsymbol{B}^x, \boldsymbol{B}^c, \Psi, \Gamma$, and the Lagrangian multipliers $\boldsymbol{\mathcal{W}}_1, \boldsymbol{\mathcal{W}}_2, \boldsymbol{W}_3$ in an alternating fashion. That is, given the estimates at the $s$th iteration, $\boldsymbol{\mathcal{B}}^{(s)}, \bar{\boldsymbol{\mathcal{B}}}^{(s)} = [[\{\boldsymbol{v}\}^{(s)}; \{\boldsymbol{B}^y\}^{(s)}, \{\boldsymbol{B}^x\}^{(s)}, \{\boldsymbol{B}^c\}^{(s)}]], \Psi^{(s)}, \Gamma^{(s)}, \boldsymbol{\mathcal{W}}_1^{(s)}, \boldsymbol{\mathcal{W}}_2^{(s)}, \boldsymbol{W}_3^{(s)}$, we update:

$$\begin{aligned} &\boldsymbol{\mu}^{(s+1)}, \boldsymbol{\mathcal{B}}^{(s+1)} = \arg\min_{\boldsymbol{\mu},\boldsymbol{\mathcal{B}}} -\mathcal{L}(\boldsymbol{\mathcal{B}}, \boldsymbol{\mu}) \\ &+ \frac{\rho}{2} \left\{ \left\| \boldsymbol{\mathcal{B}} - \bar{\boldsymbol{\mathcal{B}}}^{(s)} + \rho^{-1} \boldsymbol{\mathcal{W}}_1^{(s)} \right\|_F^2 + \left\| \boldsymbol{\mathcal{B}} - \Psi^{(s)} + \rho^{-1} \boldsymbol{\mathcal{W}}_2^{(s)} \right\|_F^2 \right\}, \end{aligned} \tag{13}$$

$$\begin{aligned} \bar{\boldsymbol{\mathcal{B}}}^{(s+1)} = \arg\min_{\boldsymbol{v},\boldsymbol{B}^y,\boldsymbol{B}^x,\boldsymbol{B}^c} &\left\| [[\boldsymbol{v}; \boldsymbol{B}^y, \boldsymbol{B}^x, \boldsymbol{B}^c]] - \boldsymbol{\mathcal{B}}^{(s+1)} \right. \\ &\left. - \rho^{-1} \boldsymbol{\mathcal{W}}_1^{(s)} \right\|_F^2 + \left\| \Gamma^{(s)} - \boldsymbol{D}_m \boldsymbol{B}^y - \rho^{-1} \boldsymbol{W}_3^{(s)} \right\|_F^2, \end{aligned} \tag{14}$$

$$\begin{aligned} \Psi^{(s+1)} = \arg\min_{\Psi} &\frac{\rho}{2} \left\| \Psi - \boldsymbol{\mathcal{B}}^{(s+1)} - \rho^{-1} \boldsymbol{\mathcal{W}}_2^{(s)} \right\|_F^2 \\ &+ \tau_s \textstyle\sum_{i,j} \|\boldsymbol{\psi}_{ij}\|_2, \end{aligned} \tag{15}$$

$$\begin{aligned} \Gamma^{(s+1)} = \arg\min_{\Gamma} &\frac{\rho}{2} \left\| \Gamma - \boldsymbol{D}_m (\boldsymbol{B}^y)^{(s+1)} - \rho^{-1} \boldsymbol{W}_3^{(s)} \right\|_F^2 \\ &+ \textstyle\sum_{j<j'} f_\kappa(\|\boldsymbol{\gamma}_{ii'}\|_2, \tau_f). \end{aligned} \tag{16}$$

$$\begin{aligned} \boldsymbol{\mathcal{W}}_1^{(s+1)} &= \boldsymbol{\mathcal{W}}_1^{(s)} + \rho \left\{ \boldsymbol{\mathcal{B}}^{(s+1)} - \bar{\boldsymbol{\mathcal{B}}}^{(s+1)} \right\}, \\ \boldsymbol{\mathcal{W}}_2^{(s+1)} &= \boldsymbol{\mathcal{W}}_2^{(s)} + \rho \left\{ \boldsymbol{\mathcal{B}}^{(s+1)} - \Psi^{(s+1)} \right\}, \\ \boldsymbol{W}_3^{(s+1)} &= \boldsymbol{W}_3^{(s)} + \rho \left\{ \boldsymbol{D}_m (\boldsymbol{B}^y)^{(s+1)} - \Gamma^{(s+1)} \right\}. \end{aligned} \tag{17}$$

We then tackle the optimization problems (13) to (16) one-by-one.

The optimization problem in Equation (13) can be split slice-by-slice for $\boldsymbol{\mathcal{B}}[i, \cdot, \cdot]$, $i = 1, \ldots, m$. That is, it can be solved with respect to each marginal response process $Y_i(t)$ in a parallel fashion. Define

$$\begin{aligned} L_i^*(\mu_i, \boldsymbol{\mathcal{B}}[i, \cdot, \cdot]) = &\textstyle\sum_{l=1}^{n_i} \log \left[ \phi \left\{ \mu_i + \langle \boldsymbol{G}(t_l^i), \boldsymbol{\mathcal{B}}[i, \cdot, \cdot] \rangle \right\} \right] \\ &- \textstyle\int_0^T \phi \left\{ \mu_i + \langle \boldsymbol{G}(t), \boldsymbol{\mathcal{B}}[i, \cdot, \cdot] \rangle \right\} dt \\ &+ \frac{\rho}{2} \left\| \boldsymbol{\mathcal{B}}[i, \cdot, \cdot] - \bar{\boldsymbol{\mathcal{B}}}^{(s)}[i, \cdot, \cdot] + \rho^{-1} \boldsymbol{\mathcal{W}}_1^{(s)}[i, \cdot, \cdot] \right\|_F^2 \\ &+ \frac{\rho}{2} \left\| \boldsymbol{\mathcal{B}}[i, \cdot, \cdot] - \Psi^{(s)}[i, \cdot, \cdot] + \rho^{-1} \boldsymbol{\mathcal{W}}_2^{(s)}[i, \cdot, \cdot] \right\|_F^2. \end{aligned}$$

The objective function $L_i^*(\mu_i, \boldsymbol{\mathcal{B}}[i, \cdot, \cdot])$ is differentiable, and with a large enough $\rho$, it is almost convex regardless of the form of the link function $\phi$. Therefore, we can minimize $L_i^*(\mu_i, \boldsymbol{\mathcal{B}}[i, \cdot, \cdot])$ efficiently using a gradient descent type algorithm. In our implementation, we employ the Newton–Raphson algorithm, as we use the linear and the logit link functions.

The optimization problem in Equation (14) turns to be a regularized CP decomposition with an $\ell_2$ penalty. It can be solved by an alternating block updating algorithm (Zhou, Li, and Zhu 2013a), which updates one block of the parameters in $\{\boldsymbol{B}^y, \boldsymbol{B}^x, \boldsymbol{B}^c\}$, while fixing the other two blocks and $\boldsymbol{v}$. For instance, $\boldsymbol{B}^y$ is updated by minimizing $\left\| \left\{ \boldsymbol{\mathcal{B}}^{(s+1)} + \rho^{-1} \boldsymbol{\mathcal{W}}_1^{(s)} \right\}_{(1)} - \boldsymbol{B}^y \left[ (\boldsymbol{B}^c)^{(s)} \odot (\boldsymbol{B}^x)^{(s)} \operatorname{diag} \{\boldsymbol{v}^{(s)}\} \right]^{\mathrm{T}} \right\|^2 + \left\| \Gamma^{(s)} - \rho^{-1} \boldsymbol{W}_3^{(s)} - \boldsymbol{D}_m \boldsymbol{B}^y \right\|^2$, with respect to $\boldsymbol{B}^y$, where $\odot$ is the Khatri-Rao product, $\boldsymbol{\mathcal{B}}_{(1)}$ denotes the mode-1 matricization of the tensor $\boldsymbol{\mathcal{B}}$, and $\operatorname{diag}(\boldsymbol{v})$ is the diagonal matrix with $\boldsymbol{v}$ as the diagonal elements. Note that this is essentially a least-square optimization problem with an $\ell_2$ penalty, which has an explicit solution. The other two blocks $\boldsymbol{B}^x$ and $\boldsymbol{B}^c$ are updated similarly. After updating each block, for instance, $\boldsymbol{B}^y$, we update $v_r$ by normalizing the

**Algorithm 1** The ADMM algorithm for parameter estimation.

[1] Initialize $\boldsymbol{\mu}^{(0)}, [[\boldsymbol{\nu}; \boldsymbol{B}^y, \boldsymbol{B}^x, \boldsymbol{B}^c]]^{(0)}, \boldsymbol{\mathcal{B}}^{(0)}, \Psi^{(0)}, \Gamma^{(0)}, \boldsymbol{\mathcal{W}}_1^{(0)}, \boldsymbol{\mathcal{W}}_2^{(0)}, \boldsymbol{W}_3^{(0)}$. Set $\rho$ and $\kappa > \rho^{-1}$.
**repeat**
[2] Update $\boldsymbol{\mu}_i^{(s+1)}, \boldsymbol{\mathcal{B}}[i,\cdot,\cdot]^{(s+1)}$ via (13) with parallel computing over $i = 1, \ldots, m$.
[3] Update $[[\boldsymbol{\nu}^{(s+1)}; \{\boldsymbol{B}^y\}^{(s+1)}, \{\boldsymbol{B}^x\}^{(s+1)}, \{\boldsymbol{B}^c\}^{(s+1)}]]$ via (14).
[4] Update $\Psi^{(s+1)} = \left\{\boldsymbol{\psi}_{ij}^{(s+1)}\right\}$ via (18) with parallel computing over $1 \le i \le m, 1 \le j \le p$.
[5] Update $\Gamma^{(s+1)} = \left\{\boldsymbol{\gamma}_{ii'}^{(s+1)}\right\}$ via (19) with parallel computing over $1 \le i < i' \le m$.
[6] Update $\boldsymbol{\mathcal{W}}_1^{(s+1)}, \boldsymbol{\mathcal{W}}_2^{(s+1)}, \boldsymbol{W}_3^{(s+1)}$ via (17).
**until** the stopping criterion is met.

corresponding vector $\boldsymbol{b}_r^y$. In addition, recognizing that the conventional cyclical alternating procedure may sometimes be unstable, we employ the maximum block improvement strategy similar to Chen et al. (2012), Tang, Bi, and Qu (2019) to ensure the algorithmic convergence of the block updating iterations; see Proposition 2. Specifically, instead of updating the blocks $\boldsymbol{B}^y$, $\boldsymbol{B}^x$ and $\boldsymbol{B}^c$ sequentially, we only update the block that yields the most improvement on the target objective function at each iteration.

The optimization problems in Equations (15) and (16) have explicit solutions, since the corresponding objective functions are convex with respect to $\boldsymbol{\psi}_{ij}$, and $\boldsymbol{\gamma}_{ii'}$ when $\kappa > \rho^{-1}$, respectively. That is,

$$\boldsymbol{\psi}_{ij}^{(s+1)} = \begin{cases} \mathbf{0} & \text{if } \|\boldsymbol{\vartheta}_{ij}^{(s+1)}\| < \sqrt{K}\tau_s/\rho, \\ \left\{1 - \frac{\sqrt{K}\tau_s/\rho}{\|\boldsymbol{\vartheta}_{ij}^{(s+1)}\|}\right\} \boldsymbol{\vartheta}_{ij}^{(s+1)} & \text{if } \|\boldsymbol{\vartheta}_{ij'}^{(s+1)}\| \ge \sqrt{K}\tau_s/\rho, \end{cases} \tag{18}$$

$$\boldsymbol{\gamma}_{ii'}^{(s+1)} = \begin{cases} \boldsymbol{\zeta}_{ii'}^{(s+1)} & \text{if } \|\boldsymbol{\zeta}_{ii'}^{(s+1)}\| \ge \kappa\tau_f, \\ \frac{\kappa\rho}{\kappa\rho - 1}\left\{1 - \frac{\tau_f/\rho}{\|\boldsymbol{\zeta}_{ii'}^{(s+1)}\|}\right\}_+ \boldsymbol{\zeta}_{ii'}^{(s+1)} & \text{if } \|\boldsymbol{\zeta}_{ii'}^{(s+1)}\| < \kappa\tau_f, \end{cases} \tag{19}$$

where $\boldsymbol{\vartheta}_{ij}^{(s+1)} = \boldsymbol{\mathcal{B}}[i,j,\cdot]^{(s+1)} + \rho^{-1}\{\boldsymbol{\mathcal{W}}_2[i,j,\cdot]\}^{(s)}$, $\boldsymbol{\zeta}_{ii'}^{(s+1)} = (\boldsymbol{B}^y[i,\cdot])^{(s+1)} - (\boldsymbol{B}^y[i',\cdot])^{(s+1)} + \rho^{-1}\boldsymbol{W}_3[l_{ii'},\cdot]^{(s)}$, and $l_{ii'} = (2m-i)(i-1)/2 + i' - i$. We note that this computation can be done in a parallel fashion over $(i,i'), i, i' = 1, \ldots, m$, and $(i,j), i = 1, \ldots, m, j = 1, \ldots, p$.

We summarize the above optimization procedures in Algorithm 1.

### 3.2. Initialization, Convergence, Tuning, and Computational Complexity

We recommend to use a warm initialization, by setting the initial values $\{\boldsymbol{\mu}^{(0)}, \boldsymbol{\mathcal{B}}^{(0)}\}$ as the unpenalized estimators without imposing any low-rank or penalty structures, while setting the other initial values at zeros. We stop the algorithm when some stopping criterion is met, for example, when the difference of the consecutive estimates is smaller than a threshold.

Algorithm 1 is guaranteed to converge to a stationary point. This can be verified by checking the conditions of Proposition 1 in Zhu, Tang, and Qu (2019).

*Proposition 2.* Suppose the log-likelihood function $\mathcal{L}$ is a Lipschitz function with respect to $\boldsymbol{\mathcal{B}}$, and the parameter space for $\boldsymbol{B}^y$, $\boldsymbol{B}^x$ and $\boldsymbol{B}^c$ is a compact set. Then the obtained estimator from Algorithm 1 converges to a stationary point of the objective function in Equation (11).

We select the tuning parameters as follows. The first is the Lagrangian augmented parameter $\rho$, which can be viewed as the learning rate of the ADMM algorithm. Our numerical results have suggested that the final estimates are not overly sensitive to the choice of $\rho$, so we simply set $\rho = 1$. The second is the thresholding parameter $\kappa$ in the fusion penalty $f_\kappa$. Again, the estimates are not sensitive to $\kappa$ as long as $\kappa > \rho^{-1}$, and we set $\kappa = 2$. The third set of tuning parameters include the rank $R$ in Equation (8), and the two regularization parameters, $\tau_s$ in Equation (9) and $\tau_f$ in Equation (10). We tune them by minimizing a Bayesian information criterion (BIC), $-2\mathcal{L}(\boldsymbol{\mu}, \boldsymbol{\mathcal{B}}) + \log(N)p_e$, where $N = \sum_{i=1}^m n_i$ is the total number of events observed on the multivariate response process $\boldsymbol{Y}(t)$, and $p_e$ is the effective number of parameters. For the tuning of $R$, $p_e = R(m + p + K - 2)$, for $\tau_s$, $p_e$ is the total number of nonzero latent parameters, and for $\tau_f$, $p_e$ is the total number of unique nonzero latent parameters. A similar BIC type criterion has been commonly adopted in low-rank tensor regressions (Zhou, Li, and Zhu 2013a; Sun and Li 2017). Moreover, to speed up tuning, we tune $R, \tau_s, \tau_f$ in a sequential manner. That is, we first tune $R$ while setting $\tau_s = \tau_f = 0$, then tune $\tau_s$ given the selected $R$ while setting $\tau_f = 0$, and finally tune $\tau_f$ given the selected $R$ and $\tau_s$. We also conduct a sensitivity analysis to investigate the choice of rank $R$ in the appendix.

Finally, we briefly discuss the computational complexity of Algorithm 1. As an example, if we use a sigmoid link function, then the overall computational complexity can be approximated by $O\left[n_{\text{iter}}\left\{mC_{\text{LR}(p)} + C_{\text{ALS}(mp)} + mp + m(m-1)/2\right\}\right]$, where $n_{\text{iter}}$ denotes the total number of ADMM iterations, $C_{\text{PLR}(p)}$ denotes the computational complexity for a logistic regression with $p$ covariates, and $C_{\text{ALS}(mpK)}$ denotes the computational complexity for a three-way tensor decomposition with the size $m \times p \times K$ using the alternating least-square (ALS) algorithm. Besides, several steps in this algorithm can be accelerated using parallel computing. In Section 5, we report the computation time of our simulated examples.

## 4. Theory

### 4.1. Regularity Conditions

We begin by introducing some notation. Let $\boldsymbol{\theta} = \left\{\boldsymbol{\mu}^{\mathrm{T}}, \text{vec}(\boldsymbol{B}^y)^{\mathrm{T}}, \text{vec}(\boldsymbol{B}^x)^{\mathrm{T}}, \text{vec}(\boldsymbol{B}^c)\right\}^{\mathrm{T}}$ collect all latent parameters in our model, including the background intensity $\boldsymbol{\mu}$, and the latent factors $\boldsymbol{B}^y, \boldsymbol{B}^x, \boldsymbol{B}^c$ from the CP decomposition (8). Without loss of generality, the normalization weight $\boldsymbol{\nu}$ is omitted to simplify the notation. Let $\boldsymbol{\beta} = \boldsymbol{\beta}(\boldsymbol{\theta}) = \left(\boldsymbol{\beta}_1^{\mathrm{T}}, \ldots, \boldsymbol{\beta}_m^{\mathrm{T}}\right)^{\mathrm{T}}$, where $\boldsymbol{\beta}_i = \left\{\mu_i, \text{vec}\left(\boldsymbol{\mathcal{B}}[i,\cdot,\cdot]\right)^{\mathrm{T}}\right\}^{\mathrm{T}}$, $i = 1, \ldots, m$. Note that the transferring coefficient $\boldsymbol{\beta}$ is a function of $\boldsymbol{\theta}$, and thus we

sometimes write it as $\boldsymbol{\beta}(\boldsymbol{\theta})$. Let $\Theta_{\boldsymbol{\theta}} \subset \mathbb{R}^{R(m+p+K)+m}$ and $\Theta_{\boldsymbol{\beta}} \subset \mathbb{R}^{mpK+m}$ denote the parameter space for $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$, respectively. For a real-valued function $f(t)$ defined on $[0,\infty)$, define the norm $\|f\|_A = \left\{\int_A f^2(t)dt\right\}^{1/2}$, where $A$ is a Borel set in $[0,\infty)$. In particular, write $\|f\|_T = \left\{\int_0^T f^2(t)dt\right\}^{1/2}$ for interval $[0,T]$. Moreover, let $\|\cdot\|_2$, $\|\cdot\|_\infty$, $\|\cdot\|_F$, and $\|\cdot\|_{\max}$ denote the $\ell_2$ norm, the $\ell_\infty$ norm, the Frobenius norm, and the maximum norm, respectively. Let $\pi_{\min}(\cdot)$ and $\pi_{\max}(\cdot)$ denote the smallest and the largest eigenvalue for a symmetric matrix.

For our theoretical analysis, we consider a general likelihood-based loss function, which encompasses our model (1) and the two penalty functions (9) and (10),

$$\begin{aligned}\mathcal{S}\{\boldsymbol{\beta}(\boldsymbol{\theta})\} &= -\mathcal{L}\{\boldsymbol{\beta}(\boldsymbol{\theta})\} + \tau\, P\{\boldsymbol{\beta}(\boldsymbol{\theta})\} \\ &= -\frac{1}{T}\sum_{i=1}^{m} L_i\{\boldsymbol{\beta}(\boldsymbol{\theta})\} + \tau\, P(\boldsymbol{\theta}), \end{aligned} \tag{20}$$

where $L_i(\cdot)$ is the log-likelihood function for the $i$th response process $Y_i(t)$, while in our model it is as specified in (7), $i = 1,\ldots,m$, $P(\cdot)$ is a nonnegative penalty function, and $\tau$ is the penalization parameter. Note that $\boldsymbol{\theta}$ is associated with the log-likelihood function $L_i(\cdot)$ only through $\boldsymbol{\beta}(\boldsymbol{\theta})$.

We next present a set of regularity conditions, where $c_1$ to $c_5$ are some finite positive constants.

1. Let $\mathcal{K}(\boldsymbol{\beta}^0) \subset \Theta_{\boldsymbol{\beta}}$ denote a neighborhood of the true value $\boldsymbol{\beta}^0$. For any $\boldsymbol{\beta}, \tilde{\boldsymbol{\beta}} \in \mathcal{K}(\boldsymbol{\beta}^0)$ and a large enough $T$, $\lambda_i^y(t;\boldsymbol{\beta}) = \lambda_i^y(t;\tilde{\boldsymbol{\beta}})$ almost surely on $[0,T]$, if and only if $\boldsymbol{\beta} = \tilde{\boldsymbol{\beta}}$.
2. For any $i = 1,\ldots,m$, $\sup_{\mathcal{V}(A)<1} \mathbb{E}\{Y_i(A)\}^2/\mathcal{V}(A) < \infty$, and for any $j = 1,\ldots,p$, $\sup_{\mathcal{V}(A)<1} \mathbb{E}\{X_j(A)\}^2/\mathcal{V}(A) < \infty$, where $A$ is a Borel-set on $[0,+\infty)$, and $\mathcal{V}$ is the Lebesgue measure.
3. For any $\boldsymbol{\beta} \in \mathcal{K}(\boldsymbol{\beta}^0)$, $i = 1,\ldots,m$, $\sup_{t\in[0,T]} \mathbb{E}\left[\left\| \frac{\partial \log\{\lambda_i^y(t)\}}{\partial \boldsymbol{\beta}_i} \lambda_i^y(t;\boldsymbol{\beta}^0)\right\|_\infty^2\right] < \infty$, $\sup_{t\in[0,T]} \mathbb{E}\left[\left\| \frac{\partial^2 \log\{\lambda_i^y(t)\}}{\partial \boldsymbol{\beta}_i \partial \boldsymbol{\beta}_i^{\mathrm{T}}} \lambda_i^y(t;\boldsymbol{\beta}^0)\right\|_{\max}^2\right] < \infty$, and $\sup_{t\in[0,T]} \mathbb{E}\left\{\lambda_i^y(t)^2\right\} < \infty$.
4. For any $\boldsymbol{\beta} \in \mathcal{K}(\boldsymbol{\beta}^0)$, $i = 1,\ldots,m$, $\pi_{\min}(\boldsymbol{J}_i) \geq c_1$ almost surely with a large $T$, where $\boldsymbol{J}_i = T^{-1}\int_0^T \boldsymbol{H}_i(t)\boldsymbol{H}_i(t)^{\mathrm{T}} \lambda_i^y(t,\boldsymbol{\beta}_i^0)dt$, and $\boldsymbol{H}_i(t) = \lambda_i^y(t)^{-1}\partial\lambda_i^y(t)/\partial\boldsymbol{\beta}_i$.
5. The link function $\phi(x)$ is a Lipschitz function satisfying that $|\phi(x_1) - \phi(x_2)| \leq c_2|x_1 - x_2|$ for any $x_1, x_2 \in \mathbb{R}$ and some $c_2$.
6. The basis function $g^{(k)}(t)$ satisfies that $\max_{1\leq k\leq K}\left\{\int_0^\infty g^{(k)}(t)^2 dt\right\}^{1/2} < c_3$ for some $c_3$.
7. The penalty function $P(\boldsymbol{\theta})$ is a nonnegative Lipschitz function in a neighborhood of the true value $\boldsymbol{\theta}^0$ satisfying that $|P(\boldsymbol{\theta}_1) - P(\boldsymbol{\theta}_2)| \leq c_4\|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2$ for some $c_4$.
8. Let $\mathcal{I}_1,\ldots,\mathcal{I}_N$ denote the true subgroup partition of the index set $\{1,\ldots,m\}$, in that $\boldsymbol{B}^y[i,\cdot] = \bar{\boldsymbol{b}}_{(s)}$ for any $i \in \mathcal{I}_s$, $s = 1,\ldots,N$, and $N$ is the number of subgroups. There is a minimum gap, such that $\min_{s\neq s'} \left\|\bar{\boldsymbol{b}}_{(s)} - \bar{\boldsymbol{b}}_{(s')}\right\|_2 > c_5$.

We make some remarks about these conditions. Condition (C1) ensures the identifiability of the intensity function with respect to $\boldsymbol{\beta}$, and is equivalent to (Ogata et al. 1978, assump. B3). Condition (C2) implies a finite mean intensity for both the response and the predictor process, and the resulting process is referred as a non-explosive point process (Daley and Vere-Jones 2007). The same condition was imposed in (Ogata et al. 1978, assump. A3) and (Hansen et al. 2015, the condition of Theorem 2). Such a condition also makes sense in neuronal activity studies, which implies that there is an upper bound for the average neuronal activity level or calcium concentration level over time. Condition (C3) is a standard regularity condition, and the same condition or its equivalent forms have been commonly adopted in the point process literature; see, for example, (Ogata et al. 1978, cond. B4, B5) for a stationary process, and (Rathbun and Cressie 1994, cond. C1, C4) for an inhomogeneous process. This condition is also easy to verify for a class of commonly used link functions, for example, a rectifier link or a sigmoid link. Condition (C4) is placed on the minimum eigenvalue of the information matrix. A similar condition was considered in (Ogata et al. 1978, Assumption B6) and (Rathbun and Cressie 1994, cond. c), and it is analogous to the usual regularity condition placed on the design matrix for regressions with random variables. Condition (C5) is usually adopted in nonlinear point process models (Brémaud and Massoulié 1996), and it holds for a range of commonly used link functions. Similarly, Condition (C6) holds for numerous basis functions, since the basis function is generally a normalized decaying kernel. Condition (C7) holds for a variety of penalty functions in a compact space, including the $\ell_1$ and $\ell_2$ penalties, the group $\ell_1$ penalty in Equation (9), and the fusion penalty in Equation (10). Finally, Condition (C8) is a standard condition to ensure the identifiability of the subgroups, and has been often assumed in subgroup analysis (Ma and Huang 2017; Zhu, Tang, and Qu 2019). This condition is not required for establishing the coefficient estimation convergence properties, but only for the subgroup identification consistency. In summary, we feel the above regularity conditions are relatively mild and reasonable. They are clearly weaker than the stationary condition. The same conditions or similar forms have been widely adopted in the asymptotic studies of temporal point process in the literature.

### 4.2. Asymptotic Convergence Properties

We next derive the asymptotic properties for the penalized likelihood estimator from Equation (20), which covers a variety of link and penalty functions, and does not assume the stationarity. We allow the point process dimension to diverge, and show that the increasing dimension is to actually benefit the estimation, leading to a faster convergence rate for the coefficient estimator and a smaller error bound for the recovered intensity function. In the interest of space, we present some supporting lemmas and all the proofs in the supplementary material (appendix).

Since the parameter spaces $\Theta_{\boldsymbol{\theta}}$ and $\Theta_{\boldsymbol{\beta}}$ grow along with the point process dimensions, we adopt the sieve idea and the large-deviation approach introduced by Shen and Wong (1994) and Shen (1998) to derive the asymptotics. Specifically, we define a restricted parameter space for $\boldsymbol{\theta}$,

$$\tilde{\Theta}_{\boldsymbol{\theta}} = \left\{\boldsymbol{\theta} \in \Theta_{\boldsymbol{\theta}} : \|\boldsymbol{\theta}\|_\infty \leq c_0, P(\boldsymbol{\theta}) \leq c_{\boldsymbol{\theta}}^2\right\},$$

where $c_0$ is a positive constant, and $c_{\boldsymbol{\theta}}$ is another constant that is allowed to increase at the rate of $O\left(\sqrt{(m+p+K)R+m}\right)$, since the dimension of $\boldsymbol{\theta}$ is $(m+p+K)R+m$ that is to diverge with $m$ and $p$. Furthermore, we define a metric based on the Kullback–Leibler (KL) pseudo-distance in $\Theta_{\boldsymbol{\theta}}$ with respect to the underlying true value $\boldsymbol{\theta}^0$ as,

$$d(\boldsymbol{\theta}, \boldsymbol{\theta}^0) = \frac{1}{\sqrt{mp}} \mathbb{E}\left[\mathcal{L}\{\boldsymbol{\beta}(\boldsymbol{\theta})\} - \mathcal{L}\{\boldsymbol{\beta}(\boldsymbol{\theta}^0)\}\right]^{1/2}.$$

Analogous to (Ogata et al. 1978, lem. 3), it is straightforward to verify that $d(\boldsymbol{\theta}, \boldsymbol{\theta}^0)$ is an appropriate distance metric for any $\boldsymbol{\theta} \in \Theta_{\boldsymbol{\theta}}$. Let $\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta} \in \tilde{\Theta}_{\boldsymbol{\theta}}} \mathcal{S}\{\boldsymbol{\beta}(\boldsymbol{\theta})\}$ denote the penalized likelihood estimator for Equation (20). The next theorem shows that $\hat{\boldsymbol{\theta}}$ converges to the true value $\boldsymbol{\theta}^0$ exponentially in probability under the KL distance.

*Theorem 1.* Suppose Conditions (C1) to (C7) hold. For some $\varepsilon_1 > 0$, there exist finite positive constants $\tilde{c}_1, \tilde{c}_2$, such that

$$\Pr\left\{d(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}^0) \geq \varepsilon_1\right\} \leq 7 \exp\left(-\tilde{c}_2 T \eta_{\boldsymbol{\theta}}^2 \varepsilon_1^2\right),$$

where

$$\eta_{\boldsymbol{\theta}} = \frac{(mpK)^{1/2}}{\{R(m+p+K)+m\}^{1/2}} \left[\log\left\{\frac{\tilde{c}_1 mpK}{\sqrt{R(m+p+K)+m}}\right\}\right]^{-1/2},$$

and the penalty parameter $\tau$ in Equation (20) satisfies that $\tau \leq O(T^{-1}\eta_{\boldsymbol{\theta}}^{-2})$.

The convergence result in Theorem 1 is established under the KL distance, which is stronger than and usually dominates some other distance measures, for example, the Hellinger metric. The next corollary establishes the convergence of the recovered transferring coefficient $\boldsymbol{\mathcal{B}}$ under the $\ell_2$ norm. Denote $\hat{\boldsymbol{\mathcal{B}}} = \boldsymbol{\mathcal{B}}(\hat{\boldsymbol{\theta}})$, $\boldsymbol{\mathcal{B}}^0 = \boldsymbol{\mathcal{B}}(\boldsymbol{\theta}^0)$, and $\tilde{d}(\hat{\boldsymbol{\mathcal{B}}}, \boldsymbol{\mathcal{B}}^0) = (mpK)^{-1/2}\|\hat{\boldsymbol{\mathcal{B}}} - \boldsymbol{\mathcal{B}}^0\|_F$.

*Corollary 1.* Suppose the conditions in Theorem 1 hold. For some $\varepsilon_2 > 0$, there exists a finite positive constant $\tilde{c}_3$, such that

$$\Pr\left\{\tilde{d}(\hat{\boldsymbol{\mathcal{B}}}, \boldsymbol{\mathcal{B}}^0) \geq \varepsilon_2\right\} \leq 7 \exp\left(-\tilde{c}_3 T \eta_{\boldsymbol{\mathcal{B}}}^2 \varepsilon_2^2\right),$$

where $\eta_{\boldsymbol{\mathcal{B}}} = [(mpK)/\{R(m+p+K)+m\}]^{1/2}$.

A few remarks are in order. First, Theorem 1 and Corollary 1 indicate that the penalized estimator and the recovered coefficient tensor achieve a convergence rate of $\sqrt{T}\eta_{\boldsymbol{\theta}}$ and $\sqrt{T}\eta_{\boldsymbol{\mathcal{B}}}$, respectively. On one hand, since the length $T$ of the observed point process plays the role of sample size as in the usual random variable based regressions, an increasing $T$ would lead to a faster convergence rate and a smaller error bound. On the other hand, the diverging point process dimensions $m$ and $p$ are to benefit the estimation as well. This is achieved *not* because of a stronger set of model and regularity conditions we impose; actually as we discuss in detail earlier that our conditions are compatible with or weaker than those in the existing literature. But it is indeed due to our proposed low-rank model structure (8). Specifically, our model substantially reduces the size of the parameter space $\Theta_{\boldsymbol{\beta}}$ through $\boldsymbol{\mathcal{B}}(\boldsymbol{\theta})$ with the latent parameter $\boldsymbol{\theta} \in \Theta_{\boldsymbol{\theta}}$. Consequently, it enables us to obtain a smaller metric entropy with bracketing on the restricted parameter space $\tilde{\Theta}_{\boldsymbol{\theta}}$, which in turn yields a tighter bound for the loss function and a faster convergence rate. Moreover, the latent factors $\boldsymbol{B}^x$ and $\boldsymbol{B}^c$ are commonly shared across the intensity functions of all response processes. This enables us to use and borrow information from the entire multivariate response process $\boldsymbol{Y}(t)$ rather than a single response $Y_i(t)$. This result of the blessing of dimensionality clearly distinguishes our method from the existing ones. For instance, Bacry et al. (2020) studied the asymptotics for a multivariate Hawkes process model and potentially allowed the process dimension to diverge, but their error bound is to increase at the rate of the logarithm of the dimension when it diverges. Second, in our current analysis, we fix the rank $R$ of the tensor decomposition (8) for simplicity. However, we can allow $R$ to diverge along with $m$ and $p$ too. By Corollary 1, as long as $R$ grows at a limited rate of $o\left(\min(m,p)\right)$, for example, $\log(m)$ or $\log(p)$, the obtained estimator still enjoys a faster convergence rate as $m$ and $p$ increase. Third, we note that the theoretical properties obtained in Theorem 1 and Corollary 1 are for the global minimizer of (20). Nevertheless, Equation (20) is a nonconvex optimization problem, and there is no guarantee that the optimization algorithm can land at the global minimizer. This is a well-known issue in almost any statistical models involving nonconvex optimization (Zhu, Shen, and Ye 2016), and it still remains an open question. In the recent years, there has been some progress to tackle this problem. For instance, Bi et al. (2018) showed that, in a tensor factorization model, the established large deviation property of a global minimizer can be generalized to an asymptotically good local optimizer. However, it was obtained with the price of imposing additional assumptions. We leave this problem as future research.

Next, we establish the subgroup structure identification consistency.

*Theorem 2.* Suppose Conditions (C1) to (C8) hold. Let $\hat{\boldsymbol{B}}^y$ denote the estimated latent factor $\boldsymbol{B}^y$ from (11). Suppose $\tau_s = o\left\{\left(T\eta_{\boldsymbol{\mathcal{B}}}^2\right)^{-1/2}\right\}$, and $\tau_f = O\left\{\left(T\eta_{\boldsymbol{\mathcal{B}}}^2\right)^{-1/2+c_f}\right\}$ for $0 < c_f < 1/2$. Then,

$$\Pr\left(\hat{\boldsymbol{B}}^y[i,\cdot] = \hat{\boldsymbol{B}}^y[i',\cdot] \mid i, i' \in \mathcal{I}_s,\ 1 \leq s \leq N\right) \to 1, \quad \text{as } T \to \infty.$$

Theorem 2 shows that, as $T \to \infty$, the true subgroup structure can be identified with the probability tending to one. We also comment that, the number of subgroups is also allowed to increase as $m$ increases, as long as Condition (C8) holds. Moreover, we may relax (C8), by allowing the minimum gap $c_5$ to decrease at a limited rate. For instance, Theorem 2 continues to holds if $c_5 \to 0$ and $c_5 T^{1/2-c_f} \to \infty$. Finally, we comment that, the imposed subgrouping structure encourages grouping of similar response processes, which helps further integrate the information across different individual processes.

Finally, we remark that, even though the CP decomposition is not unique generally, there is an easy-to-check sufficient condition to ensure the uniqueness of the decomposition in Equation (8) up to scaling and permutation (Kruskal 1988; Sidiropoulos and Bro 2000). We next give such a condition. Moreover, we note that our asymptotic convergence results do not rely on the decomposed latent parameters, but instead the entire transferring coefficient tensor $\boldsymbol{\mathcal{B}}$.

*Proposition 3.* Let $R(\boldsymbol{B}^y)$, $R(\boldsymbol{B}^x)$, and $R(\boldsymbol{B}^c)$ denote the column ranks of $\boldsymbol{B}^y$, $\boldsymbol{B}^x$, and $\boldsymbol{B}^c$, respectively. Then the rank-$R$ decomposition in Equation (8) is unique up to scaling and permutation if

$$R(\boldsymbol{B}^y) + R(\boldsymbol{B}^x) + R(\boldsymbol{B}^c) \geq 2R + 2.$$

If we have $R \geq 2$ and the three blocks $\boldsymbol{B}^y$, $\boldsymbol{B}^x$, and $\boldsymbol{B}^c$ are of full-rank, then the above inequality easily holds. In addition, in our implementation, we normalize the columns of $\boldsymbol{B}^y$, $\boldsymbol{B}^x$, and $\boldsymbol{B}^c$ to avoid possible indeterminacy in the decomposition due to scaling.

## 5. Simulations

### 5.1. Model With Low-Rank and Sparsity Structures

We study the finite-sample performance of our method under different predictor processes, link functions $\phi$, point process dimensions $m, p$ and time length $T$. We first consider a model and an implementation with only the low-rank and sparsity structures. We then consider a model with an additional subgroup structure, and an implementation with all three structures in the next section.

We generate the data following model (6). Specifically, we first generate the $p$-dimensional predictor point process $\boldsymbol{X}(t)$. We consider two predictor processes, a homogeneous Poisson process with the marginal intensity $\Lambda_j^x$, and a Hawkes process with the transferring function $\omega_{jj'}(t) = a_{jj'}e^{-\beta t}$ and the initial intensity $\Lambda_j^{(0)}$, where $\alpha_{jj'}$ is generated from a uniform distribution on $[0.2, 0.3]$, $\beta = 0.7$, and $j, j' = 1, \ldots, p$. We consider two intensity link functions $\phi$, a linear link and a logit link. For the linear link, we set the marginal intensity $\Lambda_j^x = 0.5$ for the Poisson predictor process, and set the initial intensity $\Lambda_j^{(0)} = 0.3$ for the Hawkes predictor process, $j = 1, \ldots, p$. For the logit link, we set $\Lambda_j^x = 0.2$ for the Poisson process, and set $\Lambda_j^{(0)} = 0.15$ for the Hawkes process, $j = 1, \ldots, p$. This way, the Poisson and Hawkes predictor processes are generated with similar levels of overall intensities. Next, we employ a mixture of three basis functions, $g^{(1)}(t) = \exp(-5t)$, $g^{(2)}(t) = 0.2\,\mathbf{1}(t \leq 0.1)$, and $g^{(3)}(t) = 0.05\,\mathbf{1}(t \leq 1)$. The first basis function is an exponential decaying kernel that is widely used in point process modeling. The other two basis functions are piecewise indicator functions, and they are used to capture some "short-term" effect and "long-term" effect, respectively, that are motivated by neuronal spike trains analysis. Next, we generate the transferring coefficient tensor $\boldsymbol{\mathcal{B}}$ with a rank-3 structure, $\boldsymbol{\mathcal{B}} = \sum_{r=1}^{3} \nu_r \boldsymbol{b}_r^y \circ \boldsymbol{b}_r^x \circ \boldsymbol{b}_r^c$. For the linear link, we set $\nu = (0.3, 0.2, 0.3)^{\mathrm{T}}$,

$$\boldsymbol{b}_1^y = \left((\boldsymbol{\eta}_1^y)_{m/2}^{\mathrm{T}}, \mathbf{0}_{m/2}^{\mathrm{T}}\right)^{\mathrm{T}}, \qquad \boldsymbol{b}_1^x = \left((\boldsymbol{\eta}_1^x)_{p/3}^{\mathrm{T}}, \mathbf{0}_{3p/4}^{\mathrm{T}}\right)^{\mathrm{T}},$$

$$\boldsymbol{b}_2^y = \left(\mathbf{0}_{5m/12}^{\mathrm{T}}, (\boldsymbol{\eta}_2^y)_{m/3}^{\mathrm{T}}, \mathbf{0}_{m/4}^{\mathrm{T}}\right)^{\mathrm{T}}, \quad \boldsymbol{b}_2^x = \left(\mathbf{0}_{p/6}^{\mathrm{T}}, (\boldsymbol{\eta}_2^x)_{p/3}^{\mathrm{T}}, \mathbf{0}_{p/2}^{\mathrm{T}}\right)^{\mathrm{T}},$$

$$\boldsymbol{b}_3^y = \left(\mathbf{0}_{3m/4}^{\mathrm{T}}, (\boldsymbol{\eta}_3^y)_{m/4}^{\mathrm{T}}\right)^{\mathrm{T}}, \qquad \boldsymbol{b}_3^x = \left(\mathbf{0}_{2p/3}^{\mathrm{T}}, (\boldsymbol{\eta}_3^x)_{p/4}^{\mathrm{T}}, \mathbf{0}_{p/12}^{\mathrm{T}}\right)^{\mathrm{T}},$$

and $\boldsymbol{b}_r^c$, $\boldsymbol{\eta}_r^y$ and $\boldsymbol{\eta}_r^x$, $r = 1, 2, 3$, are all generated from a normal distribution with mean one and covariance the identity matrix. Figure 1(a) shows the true association structure based on the generated coefficient $\boldsymbol{\mathcal{B}}$. For the logit link, we set $\nu = (0.2, 0.1, 0.2)^{\mathrm{T}}$, and generate $\boldsymbol{b}_r^y, \boldsymbol{b}_r^x, \boldsymbol{b}_r^c$ in the same way as for the linear link, except that we add a negative sign to each element of $\boldsymbol{\mathcal{B}}$ with probability 0.5. Finally, we set the background intensity $\boldsymbol{\mu} = \mathbf{0.01}_m$, then generate the $m$-dimensional response point process $\boldsymbol{Y}(t)$ following model (6). Given the intensity function, each individual response process is simulated following the thinning strategy (Ogata 1988). We set the dimension of the response and predictor process $m = p = \{60, 120\}$, and the observed length $T = \{800, 2000\}$. For a homogeneous process, $T$ plays the role of sample size, since it is proportional to the expected number of events. For an inhomogeneous process, this is not necessarily true, and the expected number of observed events could vary across different marginal processes.

We compare with a variation of our own method as a benchmark, plus two alternative solutions. The benchmark variation only considers the low-rank structure, but no sparsity structure. The first alternative solution is a standard baseline model that simply fits the conditional intensity functions in Equation (1) without specifying any additional structure. The second alternative solution adds a group $\ell_1$ penalty on the transferring coefficients to the baseline model for sparsity-pursuit, which is analogous to Hansen et al. (2015). We evaluate the estimation accuracy by the root mean square error (RMSE) of the estimated transferring coefficient tensor $\boldsymbol{\mathcal{B}}$.

Table 1 summarizes the results based on 50 data replications. It is seen that our method with both low-rank and sparsity structures consistently outperforms the benchmark and the two alternative solutions, by achieving the smallest RMSE across all settings. As the point process length $T$ increases, all methods improve in estimation accuracy. On the other hand, as the numbers of response and predictor processes $m$ and $p$ increase, our method continues to improve, whereas the two alternative solutions suffer. This is largely due to that our model jointly model all the processes together. Figure 2 shows the recovered

(a) True model coefficients in Section 5.1 (b) True model coefficients in Section 5.2

**Figure 1.** The slice of the true transferring coefficient tensor $\mathcal{B}$.

**Table 1.** Estimation accuracy of $\mathcal{B}$ for the model in Section 5.1.

| Link | Predictor | $m = p$ | $T$ | PP-Reg | Sp-PP-Reg | Lr-PP-Reg | LrSp-PP-Reg |
|---|---|---|---|---|---|---|---|
| Linear | Poisson | 60 | 800 | 0.281 (0.019) | 0.234 (0.015) | 0.165 (0.012) | **0.147 (0.011)** |
| | | | 2000 | 0.168 (0.010) | 0.149 (0.007) | 0.102 (0.006) | **0.094 (0.006)** |
| | | 120 | 800 | 0.319 (0.025) | 0.263 (0.021) | 0.133 (0.017) | **0.117 (0.015)** |
| | | | 2000 | 0.189 (0.011) | 0.169 (0.009) | 0.080 (0.009) | **0.066 (0.009)** |
| | Hawkes | 60 | 800 | 0.307 (0.045) | 0.279 (0.028) | 0.201 (0.027) | **0.185 (0.025)** |
| | | | 2000 | 0.226 (0.026) | 0.197 (0.021) | 0.135 (0.018) | **0.125 (0.018)** |
| | | 120 | 800 | 0.337 (0.034) | 0.289 (0.024) | 0.146 (0.018) | **0.129 (0.016)** |
| | | | 2000 | 0.245 (0.015) | 0.205 (0.010) | 0.098 (0.010) | **0.079 (0.010)** |
| Logit | Poisson | 60 | 800 | 0.548 (0.026) | 0.231 (0.015) | 0.258 (0.021) | **0.152 (0.012)** |
| | | | 2000 | 0.518 (0.015) | 0.202 (0.009) | 0.221 (0.012) | **0.121 (0.009)** |
| | | 120 | 800 | 0.844 (0.065) | 0.264 (0.025) | 0.240 (0.017) | **0.134 (0.015)** |
| | | | 2000 | 0.645 (0.017) | 0.196 (0.005) | 0.187 (0.004) | **0.101 (0.003)** |
| | Hawkes | 60 | 800 | 0.648 (0.045) | 0.258 (0.028) | 0.293 (0.027) | **0.158 (0.025)** |
| | | | 2000 | 0.583 (0.035) | 0.192 (0.018) | 0.201 (0.012) | **0.124 (0.012)** |
| | | 120 | 800 | 0.983 (0.048) | 0.289 (0.026) | 0.285 (0.016) | **0.149 (0.014)** |
| | | | 2000 | 0.725 (0.026) | 0.211 (0.017) | 0.143 (0.016) | **0.103 (0.016)** |

NOTES: Four methods are compared: the regular point process regression model (PP-Reg), the sparse point process regression with a group $\ell_1$ penalty (Sp-PP-Reg), the variation of our method with only low-rank structure (Lr-PP-Reg), and our proposed method with both low-rank and sparsity structures (LrSp-PP-Reg). Reported are the average RMSE based on 50 replications, with the standard errors in the parenthesis.

transferring structure based on the estimated $\boldsymbol{\mathcal{B}}$ with a linear link and a Poisson predictor process. It is seen that our method is capable of recovering the transferring structure successfully, while the alternative solutions cannot. Compared to the benchmark variation, our method achieves a smaller RMSE, which is more clear for the logit link. This demonstrates the advantage of incorporating the sparsity in addition to the low-rank structure.

### 5.2. Model With Additional Subgrouping Structure

We next consider a model with an additional subgrouping structure. For simplicity, we focus on the linear link $\phi$ and the Poisson predictor process. The results are similar for other combinations of link function and predictor process. We adopt the same simulation setup as in Section 5.1, except that we generate the transferring coefficient tensor $\boldsymbol{\mathcal{B}}$ in a different way. Specifically, we consider a rank-4 structure $\boldsymbol{\mathcal{B}} = \sum_{r=1}^{4} \nu_r \boldsymbol{b}_r^y \circ \boldsymbol{b}_r^x \circ \boldsymbol{b}_r^c$. We set $\boldsymbol{\nu} = (0.2, 0.2, 0.2, 0.2)^{\mathrm{T}}$,

$$\boldsymbol{b}_1^y = \left((\eta_1^y \mathbf{1})_{p/6}^{\mathrm{T}}, \mathbf{0}_{5p/6}^{\mathrm{T}}\right)^{\mathrm{T}}, \qquad \boldsymbol{b}_1^x = (\eta_1^x)_m,$$

$$\boldsymbol{b}_2^y = \left(\mathbf{0}_{p/6}^{\mathrm{T}}, (\eta_2^y \mathbf{1})_{p/6}^{\mathrm{T}}, \mathbf{0}_{2p/3}^{\mathrm{T}}\right)^{\mathrm{T}}, \quad \boldsymbol{b}_2^x = \left((\eta_2^x)_{m/2}^{\mathrm{T}}, \mathbf{0}_{m/2}^{\mathrm{T}}\right)^{\mathrm{T}},$$

$$\boldsymbol{b}_3^y = \left(\mathbf{0}_{p/3}^{\mathrm{T}}, (\eta_3^y \mathbf{1})_{p/3}^{\mathrm{T}}, \mathbf{0}_{p/3}^{\mathrm{T}}\right)^{\mathrm{T}}, \quad \boldsymbol{b}_3^x = \left(\mathbf{0}_{m/3}^{\mathrm{T}}, (\eta_3^x)_{m/3}^{\mathrm{T}}, \mathbf{0}_{m/3}^{\mathrm{T}}\right)^{\mathrm{T}},$$

$$\boldsymbol{b}_4^y = \left(\mathbf{0}_{2p/3}^{\mathrm{T}}, (\eta_4^y \mathbf{1})_{p/3}^{\mathrm{T}}\right)^{\mathrm{T}}, \qquad \boldsymbol{b}_4^x = \left(\mathbf{0}_{2m/3}^{\mathrm{T}}, (\eta_4^x)_{m/3}^{\mathrm{T}}\right)^{\mathrm{T}},$$

$\boldsymbol{b}_r^c$, $\boldsymbol{\eta}_r^x$ are all generated from a normal distribution with mean one and the identity covariance, and $\eta_r^y$, $r = 1, 2, 3, 4$, are generated from a univariate normal distribution with mean one and variance 0.1. Note that, unlike the coefficient tensor in Section 5.1, here the entries are repeated in $\boldsymbol{b}_r^y$, which in turn induces the subgrouping structure. This structure can also be seen in Figure 1(b), which shows a slice of one generated coefficient tensor $\boldsymbol{\mathcal{B}}$. We set the dimension of the response and predictor process $m = p = \{60, 120\}$, and the observed length $T = \{1200, 2400\}$.

Table 2 summarizes the results based on 50 data replications, and Figure 3 shows the recovered transferring structure. It is again seen that our proposed method consistently outperforms the two alternative solutions in terms of estimation accuracy. Moreover, Table 2 includes the rand index statistic for our proposed method, which evaluates the clustering performance. It is seen that our method achieves a high index value in all settings.

Finally, we report the computation time. For the simulation example in Section 5.1 with a linear link, a Poisson predictor process, $m = p = 60$ and $T = 2000$, the average computing time was about 1.1 minutes, and the algorithm usually converges within 30 iterations. For the example in Section 5.2 with $m = p = 60$ and $T = 2400$, the average computing time was about 1.9 minutes, and the algorithm usually converges within 50 iterations. All computations were done on a personal laptop with Intel(R) Core(TM) i7-8565U CPU@1.8GHz. We also report additional sensitivity analysis for the choice of the basis functions and the rank in the supplementary appendix.

## 6. Cross-area Neuronal Spike Trains Analysis

Ensemble neural activity analysis is of central importance in system neuroscience, which aims to understand sensory coding and associations with motor output and cognitive functions (Brown, Kass, and Mitra 2004; Kim et al. 2011). Some goals of common interest include the study of single-neuron activity with dependence on its own history, and the study of cross-neuron correlations based on spike trains similarities within the same area. Beyond those goals, it is also of key interest to understand the communication patterns in information transmission between neurons in different brain areas through neuronal spiking activities (Saalmann et al. 2012). A group of neurons could be identified within a brain area based on their similar exciting or inhibitory effects onto another group of neurons in a different brain area. This hypothesis has been suggested by several scientific studies. For instance, Liang et al. (2013) found that there might be discrete locations within the visual cortex area that respond to specific cross-modal inputs such as auditory or tactile. That is, the neurons in the V1 area are expected to be clustered in that they share similar cross-cortex-area association patterns, which needs to be inferred from the

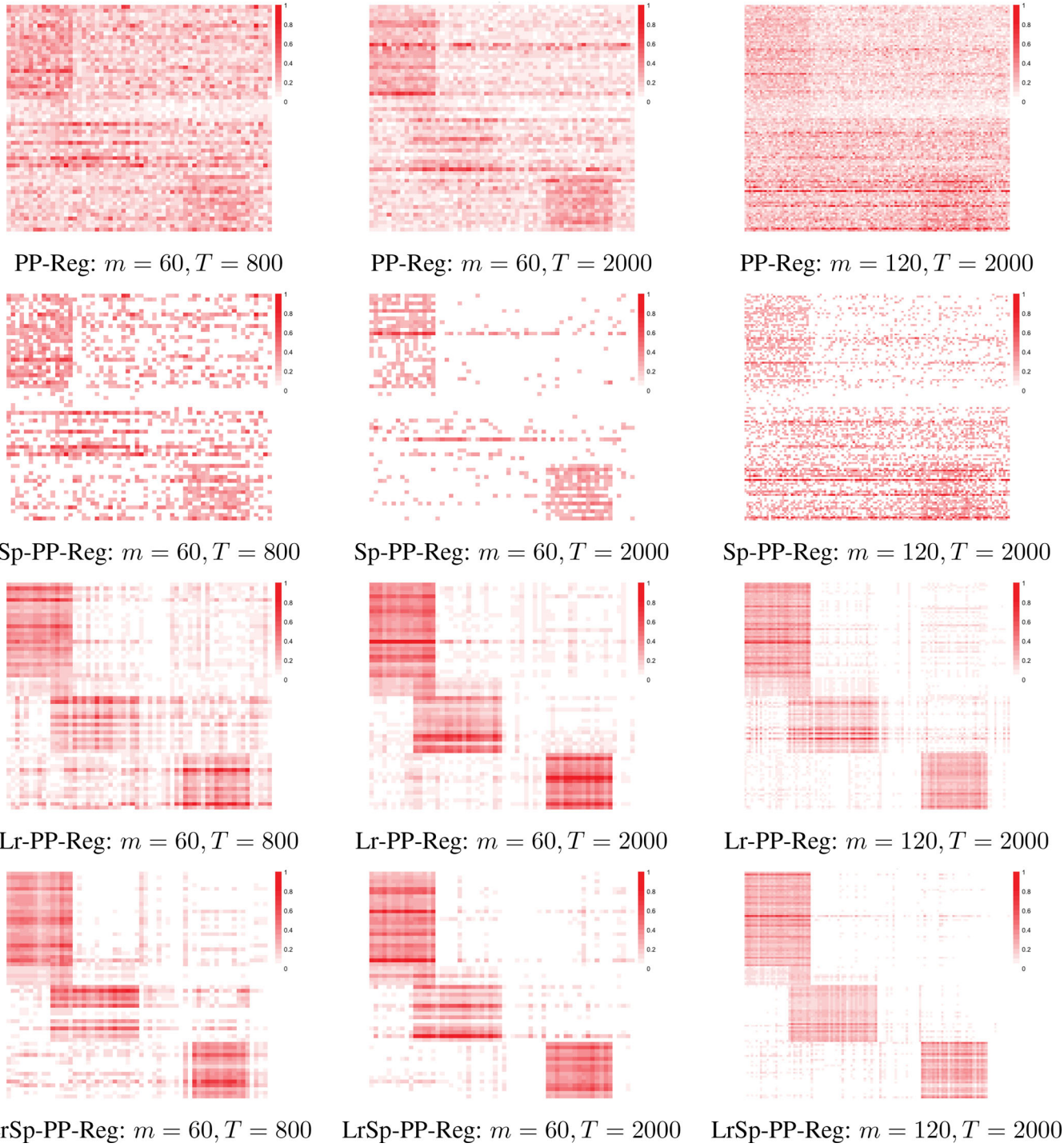



**Figure 2.** Recovered transferring coefficient tensor $\mathcal{B}$ for the model in Section 5.1. Four methods are compared: the regular point process regression model (PP-Reg), the sparse point process regression with a group $\ell_1$ penalty (Sp-PP-Reg), the variation of our method with only low-rank structure (Lr-PP-Reg), and our proposed method with both low-rank and sparsity structures (LrSp-PP-Reg).

**Table 2.** Estimation accuracy of $\mathcal{B}$ for the model in Section 5.2.

| $m = p$ | $T$ | PP-Reg | Sp-PP-Reg | Lr-PP-Reg | LrSpGr-PP-Reg {Rand Index} |
|---|---|---|---|---|---|
| 60 | 1200 | 0.205 (0.021) | 0.182 (0.015) | 0.133 (0.011) | **0.092 (0.010)** {0.847 (0.075)} |
| | 2400 | 0.161 (0.012) | 0.143 (0.008) | 0.104 (0.007) | **0.076 (0.005)** {0.893 (0.065)} |
| 120 | 1200 | 0.222 (0.024) | 0.186 (0.018) | 0.114 (0.015) | **0.073 (0.014)** {0.878 (0.099)} |
| | 2400 | 0.169 (0.008) | 0.147 (0.005) | 0.084 (0.006) | **0.059 (0.003)** {0.912 (0.071)} |

NOTES: Four methods are compared: the regular point process regression model (PP-Reg), the sparse point process regression with a group $\ell_1$ penalty (Sp-PP-Reg), the variation of our method with only low-rank structure (Lr-PP-Reg), and our proposed method with low-rank, sparsity and group structures (LrSpGr-PP-Reg). Reported are the average RMSE based on 50 replications, with the standard errors in the parenthesis.

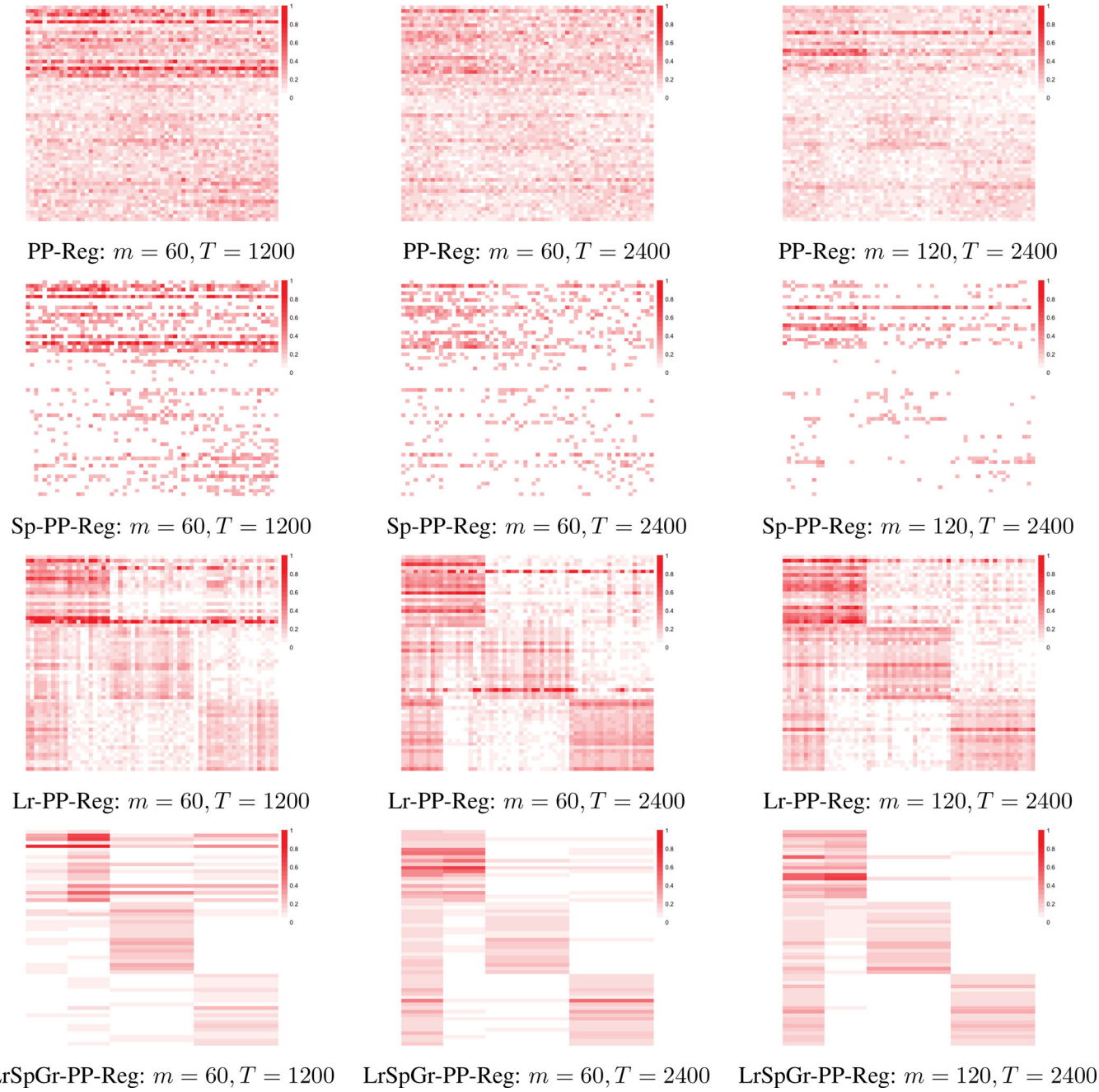
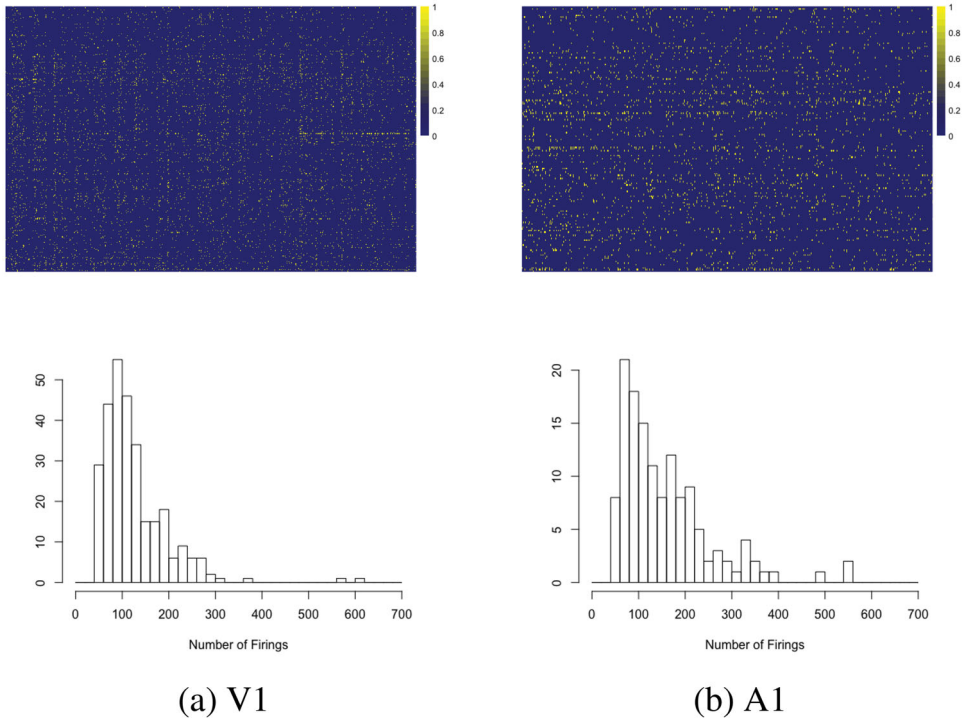



**Figure 3.** Recovered transferring coefficient tensor $\mathcal{B}$ for the model in Section 5.2. Four methods are compared: the regular point process regression model (PP-Reg), the sparse point process regression with a group $\ell_1$ penalty (Sp-PP-Reg), the variation of our method with only low-rank structure (Lr-PP-Reg), and our proposed method with low-rank, sparsity and group structures (LrSpGr-PP-Reg).

associations between the observed spike trains activities. In addition, the signal transmission takes time from one area to another, suggesting that the cross-area neuronal connection may account for a time-dependent convolutional effect rather than a simple co-firing. In recent years, benefitting from the rapid development of imaging techniques such as the calcium imaging, we are now able to monitor a large number of neurons simultaneously with a single-neuron resolution in a short time period, which produces high-dimensional point process type data of neuronal spike trains.

In our study, we simultaneously measure the neuronal spike trains activities of 139 neurons and 283 neurons from two sensory cortical areas, A1 and V1, in a rat brain, respectively. We collect the data over 192 seconds under a stable stimulus. With 50 millisecond as a unit of time, we obtain the length of time interval of $[0, 3840]$. Figure 4 shows the recorded neuron firing events over time, and the histogram summary of the numbers of observed firings for individual neurons in each of these two areas. It is seen that most neurons have their numbers of observed firing events under 200, whereas a subset of neurons have the numbers below 100.

Since a primary goal is to understand the information transmission from the A1 area to the V1 area, we fit the data using our proposed multivariate temporal point process regression, by treating the neuronal spike trains in V1 as the response point process, and the neuronal spike trains in A1 as the predictor




(a) V1 (b) A1

**Figure 4.** Neuron firings in the V1 and A1 areas. The heatmaps (upper panels) show the neuron-wise firings over time. The histograms (lower panels) show the number of firings for each neuron.

process. Since the observed firing events are sparse, we choose a logit link function. We select three basis functions, similarly as in our simulation studies: $g^{(1)}(t) = \exp(-t), g^{(2)}(t) = 0.2\,\mathbf{1}(t \leq 1)$, and $g^{(3)}(t) = 0.05\,\mathbf{1}\{t \leq 5\}$, with the time intervals in the indicator functions selected based on the existing scientific findings that the communication process between ensemble neurons across areas mostly happens within tens of milliseconds (Luo et al. 2016). In addition to our proposed model, we also fit the marginal model that takes one response process at a time. Since some neurons have very limited number of firing events, the corresponding model fittings may not converge. Actually, for our data, we have found that about one third of the individual response process fittings cannot converge. To handle this convergence issue, we add an $\ell_2$ regularization to this marginal approach, though we still refer to it as a marginal method. Moreover, we fit the marginal model with a group $\ell_1$ regularization, similarly as in our simulations.

To evaluate the model, we split the point processes into a training set, that is, the spike trains data in the time interval $[0, 2000)$, and a testing set, that is, the data in the time interval $(2000, 3800]$. We report two evaluation criteria. The main criterion is the area under the ROC curve (AUC) based on a binary prediction (Luo et al. 2016). That is, we bin the continuous point process into a sequence of binary values based on a unit of time of 50 milliseconds, with one meaning that there is a firing event in this time bin, and zero otherwise. We then produce a sequence of binary predictions based on the predicted intensity function for the testing data. The second criterion is the deviance $\|\hat{\mathcal{B}}_{\text{training}} - \hat{\mathcal{B}}_{\text{testing}}\|$. That is, we obtain the estimated coefficient tensor $\mathcal{B}$ from the training data and testing data, respectively, and evaluate the difference between the two in the Frobenius norm. Intuitively, if the firing patterns have been consistent, then this deviance measure should be small.

**Table 3.** Evaluation of model fitting for the cross-area neuronal spike trains analysis.

| | PP-Reg | SpPP-Reg | LrSpGr-PP-Reg |
|---|---|---|---|
| Deviance | 0.388 | 0.256 | **0.185** |
| AUC | 0.537 | 0.579 | **0.682** |

NOTE: Three methods are compared: the regular point process regression model (PP-Reg), the sparse point process regression with a group $\ell_1$ penalty (SpPP-Reg), and our proposed method (LrSpGr-PP-Reg).

Meanwhile, we note that, due to the non-convex nature of our optimization problem, the obtained estimator could possibly be a local optimum. To alleviate this issue, we recommend the usual strategy of trying multiple random initial values. Table 3 reports the results. It is seen that our proposed method achieves the highest AUC value and the lowest deviance value, suggesting a competitive performance of the proposed method compared to the two alternatives. We also identify five subgroups of neurons with our method, which requires future scientific validation, as we do not have relevant subgroup information for this dataset.

## Acknowledgments

The authors are grateful to the editor, the associate editor, and two referees for their constructive comments and suggestions which have improved the paper significantly. Dr. Tang's research was partially supported by NSF grants DMS 2113467. Dr. Li's research was partially supported by NIH grants R01AG061303, R01AG062542, and R01AG034570, and NSF grant CIF 2102227.

## Supplementary data

Online supplementary material provides all technical proofs, additional numerical studies and supplementary details for modeling such as the selection of basis functions and hyper parameters.

## References


Bacry, E., Bompaire, M., Gaïffas, S., and Muzy, J.-F. (2020), "Sparse and Low-Rank Multivariate Hawkes Processes," *Journal of Machine Learning Research*, 21, 1–32. [1,2,4,5,9]

Bacry, E., and Muzy, J.-F. (2016), "First-and Second-Order Statistics Characterization of Hawkes Processes and Non-Parametric Estimation," *IEEE Transactions on Information Theory*, 62, 2184–2202. [2]

Bi, X., Qu, A., and Shen, X. (2018), "Multilayer Tensor Factorization With Applications to Recommender Systems," *The Annals of Statistics*, 46, 3308–3333. [9]

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011), "Distributed Optimization and Statistical Learning Via the Alternating Direction Method of Multipliers," *Foundations and Trends® in Machine Learning*, 3, 1–122. [5]

Brémaud, P., and Massoulié, L. (1996), "Stability of Nonlinear Hawkes Processes," *The Annals of Probability*, 1563–1588. [3,8]

Brown, E. N., Kass, R. E., and Mitra, P. P. (2004), "Multiple Neural Spike Train Data Analysis: State-of-the-art and Future Challenges," *Nature Neuroscience*, 7, 456. [1,11]

Cai, B., Zhang, J., and Guan, Y. (2020), "Latent Network Structure Learning From High Dimensional Multivariate Point Processes," arXiv: 2004.03569. [2,5]

Chen, B., He, S., Li, Z., and Zhang, S. (2012), "Maximum Block Improvement and Polynomial Optimization," *SIAM Journal on Optimization*, 22, 87–107. [7]

Chen, H., Raskutti, G., and Yuan, M. (2019a), "Non-Convex Projected Gradient Descent for Generalized Low-Rank Tensor Regression," *Journal of Machine Learning Research*, 20, 1–37. [1,4]

Chen, S., Shojaie, A., Shea-Brown, E., and Witten, D. (2019b), "The Multivariate Hawkes Process in High Dimensions: Beyond Mutual Excitation," arXiv:1707.04928. [1,2]

Daley, D. J., and Vere-Jones, D. (2007), *An Introduction to the Theory of Point Processes: Volume I: Elementary Theory and Methods*, New York: Springer. [2,8]

Deng, C., Guan, Y., Waagepetersen, R. P., and Zhang, J. (2017), "Second-Order Quasi-Likelihood for Spatial Point Processes," *Biometrics*, 73, 1311–1320. [1]

Diggle, P. J., Guan, Y., Hart, A. C., Paize, F., and Stanton, M. (2010), "Estimating Individual-Level Risk in Spatial Epidemiology Using Spatially Aggregated Information on the Population at Risk," *Journal of the American Statistical Association*, 105, 1394–1402. [1,2]

Guan, Y. (2008), "On Consistent Nonparametric Intensity Estimation for Inhomogeneous Spatial Point Processes," *Journal of the American Statistical Association*, 103, 1238–1247. [1,2]

Guan, Y. (2011), "Second-Order Analysis of Semiparametric Recurrent Event Processes," *Biometrics*, 67, 730–739. [2]

Guan, Y., Jalilian, A., and Waagepetersen, R. (2015), "Quasi-Likelihood for Spatial Point Processes," *Journal of the Royal Statistical Society*, Series B, 77(3):677–697. [1]

Hansen, N. R., Reynaud-Bouret, P., and Rivoirard, V. (2015), "Lasso and Probabilistic Inequalities for Multivariate Point Processes," *Bernoulli*, 21, 83–143. [2,4,5,8,10]

Hawkes, A. G. (1971), "Spectra of Some Self-Exciting and Mutually Exciting Point Processes," *Biometrika*, 58, 83–90. [2,3]

Ji, N., Freeman, J., and Smith, S. L. (2016), "Technologies for Imaging Neural Activity in Large Volumes," *Nature Neuroscience*, 19, 1154. [1]

Kang, J., Johnson, T. D., Nichols, T. E., and Wager, T. D. (2011), "Meta Analysis of Functional Neuroimaging Data Via Bayesian Spatial Point Processes," *Journal of the American Statistical Association*, 106, 124–134. [1]

Kang, J., Nichols, T., Wager, T., and Johnson, T. (2014), "A Bayesian Hierarchical Spatial Point Process Model for Multi-Type Neuroimaging Meta-Analysis," *The Annals of Applied Statistics*, 8, 1800–1824. [1]

Kim, S., Putrino, D., Ghosh, S., and Brown, E. N. (2011), "A Granger Causality Measure for Point Process Models of Ensemble Neural Spiking Activity," *PLoS Computational Biology*, 7, e1001110. [5,11]

Kolda, T. G., and Bader, B. W. (2009), "Tensor Decompositions and Applications," *SIAM Review*, 51, 455–500. [2,4,5]

Kruskal, J.B. (1988), "Rank Decomposition and Uniqueness for 3-way and n-Way Arrays," in *Multiway Data Analysis*, eds. R. Coppi, S. Bolasco, North-Holland, Amsterdam; pp. 7–18. North-Holland. [9]

Liang, M., Mouraux, A., Hu, L., and Iannetti, G. (2013), "Primary Sensory Cortices Contain Distinguishable Spatial Patterns of Activity for Each Sense," *Nature Communications*, 4, 1979. [1,3,5,11]

Luo, X., Gee, S., Sohal, V., and Small, D. (2016), "A Point-Process Response Model for Spike Trains From Single Neurons in Neural Circuits Under Optogenetic Stimulation," *Statistics in Medicine*, 35, 455–474. [4,14]

Ma, S., and Huang, J. (2017), "A Concave Pairwise Fusion Approach to Subgroup Analysis," *Journal of the American Statistical Association*, 112, 410–423. [8]

Ogata, Y. (1978), "The Asymptotic Behaviour of Maximum Likelihood Estimators for Stationary Point Processes," *Annals of the Institute of Statistical Mathematics*, 30, 243–261. [8,9]

——— (1988), "Statistical Models for Earthquake Occurrences and Residual Analysis for Point Processes," *Journal of the American Statistical Association*, 83, 9–27. [10]

Okun, M., Steinmetz, N. A., Cossell, L., Iacaruso, M. F., Ko, H., Barthó, P., Moore, T., Hofer, S. B., Mrsic-Flogel, T. D., and Carandini, M., (2015), "Diverse Coupling of Neurons to Populations in Sensory Cortex," *Nature*, 521, 511. [1]

Perry, P. O., and Wolfe, P. J. (2013), "Point Process Modelling for Directed Interaction Networks," *Journal of the Royal Statistical Society*, Series B, 75, 821–849. [1]

Rathbun, S. L., and Cressie, N. (1994), "Asymptotic Properties of Estimators for the Parameters of Spatial Inhomogeneous Poisson Point Processes," *Advances in Applied Probability*, 26(1), 122–154. [8]

Saalmann, Y. B., Pinsk, M. A., Wang, L., Li, X., and Kastner, S. (2012), "The Pulvinar Regulates Information Transmission Between Cortical Areas Based on Attention Demands," *Science*, 337, 753–756. [11]

Shen, X. (1998), "On the Method of Penalization," *Statistica Sinica*, 8, 337–357. [8]

Shen, X., and Wong, W. H. (1994), "Convergence Rate of Sieve Estimates," *The Annals of Statistics*, 22, 580–615. [8]

Sidiropoulos, N. D., and Bro, R. (2000), "On the Uniqueness of Multilinear Decomposition of n-way Arrays," *Journal of Chemometrics*, 14, 229–239. [9]

Stoyan, D., and Penttinen, A. (2000), "Recent Applications of Point Process Methods in Forestry Statistics," *Statistical Science*, 15, 61–78. [1]

Sun, W., and Li, L. (2017), "Store: Sparse Tensor Response Regression and Neuroimaging Analysis," *Journal of Machine Learning Research*, 18, 4908–4944. [4,7]

Tang, X., Bi, X., and Qu, A. (2019), "Individualized Multilayer Tensor Learning With an Application in Imaging Analysis," *Journal of the American Statistical Association*, 115, 836–851. [7]

Waagepetersen, R., and Guan, Y. (2009), "Two-Step Estimation for Inhomogeneous Spatial Point Processes," *Journal of the Royal Statistical Society*, Series B, 71(3):685–702. [2]

Wang, Y., Du, N., Trivedi, R., and Song, L. (2016a), "Coevolutionary Latent Feature Processes for Continuous-Time User-Item Interactions," in *Advances in Neural Information Processing Systems*, eds. D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, Vol. 29, pp. 1–9. [4]

Wang, Y., Xie, B., Du, N., and Song, L. (2016b), "Isotonic Hawkes Processes," in *International Conference on Machine Learning*, eds. M. F. Balcan and K. Q. Weinberger, New York, NY, pp. 226–2234. [4]

Xu, H., Farajtabar, M., and Zha, H. (2016), "Learning Granger Causality for Hawkes Processes," in *International Conference on Machine Learning*. eds. M. F. Balcan and K. Q. Weinberger, New York, NY, PMLR; pp. 1717–1726. [4]

Yuan, M., and Lin, Y. (2006), "Model Selection and Estimation in Regression With Grouped Variables," *Journal of the Royal Statistical Society*, Series B, 68, 49–67. [5]

Zhang, A., and Han, R. (2019), "Optimal Sparse Singular Value Decomposition for High-Dimensional High-Order Data," *Journal of the American Statistical Association*, 114(528):1708–1725. [1]

Zhang, C.-H. (2010), "Nearly Unbiased Variable Selection Under Minimax Concave Penalty," *The Annals of Statistics*, 38, 894–942. [5]

Zhou, H., Li, L., and Zhu, H. (2013a), "Tensor Regression With Applications in Neuroimaging Data Analysis," *Journal of the American Statistical Association*, 108, 540–552. [1,4,6,7]

Zhou, K., Zha, H., and Song, L. (2013b), "Learning Social Infectivity in Sparse Low-Rank Networks Using Multi-Dimensional Hawkes Processes," *Artificial Intelligence and Statistics*, 31, 641–649. [1,2,3,4,5]

Zhu, Y., Shen, X., and Ye, C. (2016), "Personalized Prediction and Sparsity Pursuit in Latent Factor Models," *Journal of the American Statistical Association*, 111, 241–252. [9]

Zhu, X., Tang, X., and Qu, A. (2019), "Longitudinal Clustering for Heterogeneous Binary Data," *Statistica Sinica*. [5,7,8]