# Transport Distances on Random Vectors of Measures: Recent Advances in Bayesian Nonparametrics


**Marta Catalano, Antonio Lijoi, and Igor Prünster**

**Abstract** Random vectors of measures are at the core of many recent developments in Bayesian nonparametrics. For a deep understanding of these infinite-dimensional discrete random structures and their impact on the inferential and theoretical properties of the induced models, we consider a class of transport distances based on the Wasserstein distance. The geometrical definition makes it ideal for measuring similarity between distributions with possibly different supports. Moreover, when applied to random vectors of measures with independent increments (*completely random vectors*), the interesting theoretical properties are coupled with analytical tractability. This leads to a new measure of dependence for completely random vectors and the quantification of the impact of hyperparameters in notable models for exchangeable time-to-event data.

**Keywords** Bayesian nonparametrics · Completely random measures · Completely random vectors · Compound Poisson approximation · Dependence · Lévy copula · Partial exchangeability · Wasserstein distance

## 1 Introduction

Many notable Bayesian nonparametric models allow to make inference for partially exchangeable sequences. Thanks to de Finetti's representation theorem, the law of any such sequence may be specified in terms of a random vector of probabilities

M. Catalano (✉)
ESOMAS Department, University of Torino, Italy
e-mail: marta.catalano@unito.it

A. Lijoi · I. Prünster
Department of Decision Sciences, Bocconi Institute for Data Science and Analytics (BIDSA), Bocconi University, Milano, Italy
e-mail: antonio.lijoi@unibocconi.it; igor.pruenster@unibocconi.it



D. Hernández-Hernández et al. (eds.), *Advances in Probability and Mathematical Statistics*, Progress in Probability 79, https://doi.org/10.1007/978-3-030-85325-9_4

$(\tilde{P}_1, \ldots, \tilde{P}_m)$. Let $(X_{i,j})_{j \geq 1}$, with $i = 1, \ldots, m$, be a partially exchangeable sequence on $\mathbb{X}$. Then,

$$(X_{i_1,j_1}, \ldots, X_{i_k,j_k}) | (\tilde{P}_1, \ldots, \tilde{P}_m) \sim \tilde{P}_{i_1} \times \cdots \times \tilde{P}_{i_k}; \qquad (\tilde{P}_1, \ldots, \tilde{P}_m) \sim Q;$$

for any $k \geq 1$, $i_\ell \in \{1, \ldots, m\}$, $j_\ell \in \mathbb{N} \setminus \{0\}$ such that $(i_\ell, j_\ell) \neq (i_{\ell'}, j_{\ell'})$, for $\ell \neq \ell' = 1, \ldots, k$. When $m = 1$ or $\tilde{P}_1 = \cdots = \tilde{P}_m$ almost surely (a.s.), the model degenerates to exchangeability, which can thus be seen as a special case. There have been many proposals on how to specify the law $Q$ by modeling the dependence structure between random probabilities [21]. Among the most successful specifications, many build on random vectors of measures with independent increments $(\tilde{\mu}_1, \ldots, \tilde{\mu}_m)$, which we denote as *completely random vectors* (CRVs) in analogy with the one-dimensional case of completely random measures (CRMs). Completely random vectors have appealing properties in terms of analytical tractability, typically because of the existence of a multivariate Lévy intensity that characterizes their distribution. For this reason, many random vectors of probabilities may be derived from suitable transformation of CRVs, including normalization [22], kernel mixtures for densities [9, 18] and hazards [7] and exponential transformation for survival functions [6].

The derived nonparametric models for partially exchangeable sequences [8, 10, 15–17, 23] are very flexible but often difficult to interpret, making the prior elicitation more demanding. In order to ease the interpretation and foremost the comparison between different models, we introduce a distance between CRVs, based on the Wasserstein distance. The relationship between the Wasserstein distance and optimal transport theory [25] sheds light on its intrinsically geometric definition. This makes the Wasserstein distance an ideal measure of discrepancy between distributions with possibly different support, in contrast to other common choices, such as the total variation distance, the Hellinger distance and the Kullback–Leibler divergence.

To date the transport distance between CRVs has been used in two different scenarios: to create a new measure of dependence for partially exchangeable models [3] and to measure the discrepancy between hazard rates models for exchangeable observations [2]. The dependence between random measures regulates the borrowing of information between different groups of observations with a major impact on the posterior inference. In order to elicit the prior, one needs a measure of dependence that can be expressed in terms of the hyperparameters of the model. State-of-the-art measures typically consist in linear correlation, thus capturing only a portion of the dependence structure. By leveraging the transport distance between CRVs, Catalano et al. [3] propose a new measure of dependence that goes beyond linear correlation. On the other hand, the transport distance between CRMs, i.e. one-dimensional CRVs, has been fruitfully used in the context of survival analysis. One of the most popular Bayesian nonparametric models for time-to-event data [7] represents the hazard rate function as a kernel mixture over a CRM. For a careful prior elicitation, Catalano et al. [2] find the analytical expression for the Wasserstein

distance between the hazards, as the hyperparameters of the CRMs and the kernels vary. When treating the kernel of [7], i.e. $k(t|x) = \beta(x)\mathbb{1}_{(0,t]}$, calculations are performed only for constant $\beta(x) = \beta > 0$, which is often used in applications. This assumption implies that the index of dispersion of the induced hazard $\tilde{h}(t)$ is constant in time, which is often too restrictive. We thus consider the case where $\beta(x)$ increases linearly and compare it to the constant scenario. We find informative bounds on the Wasserstein distance between these two specifications that show how the distance increases quadratically in time.

The work is structured as follows. In Sect. 2 we introduce completely random vectors and in Sect. 3 we define a class of transport distances on them. In Sect. 4 we describe how these distances may be used to define a measure of dependence, reviewing the recent results of [3], whereas in Sect. 5 we focus on its applications in survival analysis, following [2]. New results on time-varying kernels are contained in Sect. 6.

## 2 Completely Random Vectors

In this section we recall the definition of completely random vectors and their main properties. Let $\mathbb{X}$ be a Polish space with Borel $\sigma$-algebra $\mathcal{X}$. We denote by $M_{\mathbb{X}}$ the space of boundedly finite measures on $\mathbb{X}$, endowed with the weak$^\sharp$ topology [5] and the corresponding Borel $\sigma$-algebra. An $m$-dimensional random vector of measures is a measurable function $\tilde{\boldsymbol{\mu}} : \Omega \to M_{\mathbb{X}}^m$, where $(\Omega, \Sigma, \mathbb{P})$ is some probability space and $M_{\mathbb{X}}^m = \prod_{i=1}^m M_{\mathbb{X}}$ denotes the $m$-fold product space with corresponding product topology and induced Borel $\sigma$-algebra. Let $\pi_i : M_{\mathbb{X}}^m \to M_{\mathbb{X}}$ be the $i$-th projection, i.e. $\pi_i(\mu_1, \dots, \mu_m) = \mu_i$, for $i = 1, \dots, m$. We denote the marginal random measures $\tilde{\mu}_i = \pi_i \circ \tilde{\boldsymbol{\mu}} : \Omega \to M_{\mathbb{X}}$, so that $\tilde{\boldsymbol{\mu}} = (\tilde{\mu}_1, \dots, \tilde{\mu}_m)$.

**Definition 2.1** A random vector of measures $\tilde{\boldsymbol{\mu}}$ is said to be a *completely random vector* (CRV) if for every disjoint collection of bounded Borel sets $A_1, \dots, A_n$, the one-dimensional distributions $\tilde{\boldsymbol{\mu}}(A_1), \dots, \tilde{\boldsymbol{\mu}}(A_n)$ are independent.

We observe that for $i = 1, \dots, m$, the marginal random measure $\tilde{\mu}_i$ of a CRV $\tilde{\boldsymbol{\mu}}$ is a completely random measure (CRM) in the sense of [14]. Thus, we can look at CRVs $\tilde{\boldsymbol{\mu}} = (\tilde{\mu}_1, \dots, \tilde{\mu}_m)$ as vectors of dependent CRMs. This property makes them particularly appealing, since dependent CRMs offer the ground for most tractable nonparametric priors in presence of multiple populations.

We focus on CRVs without fixed atoms. Thanks to [13, Theorem 3.19], this ensures the existence of a Poisson random measure $\mathcal{N}$ on $\mathbb{R}_+^m \times \mathbb{X}$ s.t. for every $A \in \mathcal{X}$,

$$\tilde{\boldsymbol{\mu}}(A) \stackrel{\mathrm{d}}{=} \int_{\mathbb{R}_+^m \times A} \boldsymbol{s}\, \mathcal{N}(ds_1, \dots, ds_m, dx), \tag{2.1}$$

where $\stackrel{\mathrm{d}}{=}$ denotes equality in distribution and $\boldsymbol{s} = (s_1, \ldots, s_m)$. It follows that the distribution of a CRV $\tilde{\boldsymbol{\mu}}$ is characterized by a multivariate Lévy intensity $\nu(ds_1, \ldots, ds_m, dx) = \mathbb{E}(\mathcal{N}(ds_1, \ldots, ds_m, dx))$ such that (1) $\nu(\mathbb{R}_+^m \times \{x\}) = 0$ for every $x \in \mathbb{X}$; (2) for every bounded $A \in \mathcal{X}$ and every $\epsilon > 0$,

$$\int_{\mathbb{R}_+^m \times A} \min\{s_1 + \cdots + s_m, \epsilon\}\, \nu(ds_1, \ldots, ds_m, dx) < +\infty. \tag{2.2}$$

We will focus on infinitely active CRVs, i.e. such that for every Borel set $A$, the Lévy measures of the marginal CRMs satisfy

$$\int_{\mathbb{R}_+ \times A} \nu_i(ds_i, dx) = \int_{\mathbb{R}_+^m \times A} \nu(ds_1, \ldots, ds_m, dx) = +\infty. \tag{2.3}$$

In the next section we define a class of distances between laws of CRVs, whose analytical tractability heavily relies on the existence of multivariate Lévy intensities.

## 3 Transport Distances

In this section we define a class of transport distances on CRVs. These are built on the Wasserstein distance [25], whose geometric definition makes it an ideal choice for measuring the similarity between distributions.

Let $\|\cdot\|_m$ denote the Euclidean distance on $\mathbb{R}^m$ and let $\mathbb{N}^+ = \mathbb{N} \setminus \{0\}$. For any pair $\pi_1, \pi_2$ of probability measures on $(\mathbb{R}^m, \|\cdot\|_m)$, we indicate by $C(\pi_1, \pi_2)$ the Fréchet class of $\pi_1$ and $\pi_2$, i.e. the set of distributions (*couplings*) on the product space $\mathbb{R}^{2m}$ whose marginal distributions coincide with $\pi_1$ and $\pi_2$ respectively.

**Definition 3.1** The Wasserstein distance of order $p \in \mathbb{N}^+$ on $(\mathbb{R}^m, \|\cdot\|_m)$ is defined as

$$\mathcal{W}_p(\pi_1, \pi_2) = \inf_{(Z_1, Z_2) \in C(\pi_1, \pi_2)} \left\{\mathbb{E}(\|Z_1 - Z_2\|_m^p)\right\}^{\frac{1}{p}}.$$

By extension, we refer to the Wasserstein distance between two random vectors $X_1, X_2$ on $\mathbb{R}^m$ as the Wasserstein distance between their laws, i.e. $\mathcal{W}_p(X_1, X_2) = \mathcal{W}_p(\mathcal{L}(X_1), \mathcal{L}(X_2))$.

In the next proposition we show how the Wasserstein distance may be used to define a distance between CRVs in a natural way. The proof is a straightforward generalization of results in [3]. Before providing the main statement, we underline that a CRV has finite moments up to order $p \in \mathbb{N}^+$ if for every $\ell \in \{1, \ldots, p\}$,

$$\int_{\mathbb{R}_+^m \times \mathbb{X}} \boldsymbol{s}^{\ell}\, \nu(ds_1, \ldots, ds_m, dx) < \boldsymbol{+\infty}, \tag{3.1}$$

where $\boldsymbol{s}^{\ell} = (s_1^{\ell}, \ldots, s_m^{\ell})$ and $+\boldsymbol{\infty} = (+\infty, \cdots, +\infty)$. Denote by $\mathbb{P}_p(M_{\mathbb{X}}^m) = \{\mathcal{L}(\tilde{\boldsymbol{\mu}}) \text{ s.t. } \tilde{\boldsymbol{\mu}} \text{ is a CRV that satisfies (3.1)}\}$.

**Proposition 3.2** *For every $p \in \mathbb{N}^+$, the following function $d_{\mathcal{W},p} : \mathbb{P}_p(M_{\mathbb{X}}^m) \times \mathbb{P}_p(M_{\mathbb{X}}^m) \to [0, +\infty)$ defines a distance:*

$$d_{\mathcal{W},p}(\mathcal{L}(\tilde{\boldsymbol{\mu}}_1), \mathcal{L}(\tilde{\boldsymbol{\mu}}_2)) = \sup_{A \in \mathcal{X}} W_p(\tilde{\boldsymbol{\mu}}_1(A), \tilde{\boldsymbol{\mu}}_2(A)). \tag{3.2}$$

By extension, we refer to the distance $d_{\mathcal{W},p}$ between CRVs as the distance between their laws. The natural definition of $d_{\mathcal{W},p}$ is often coupled with analytical tractability, as shown in [2] and [3], which makes it particularly attractive in a number of statistical applications. In particular, in [2], $d_{\mathcal{W},1}$ was used in the one-dimensional scenario, i.e. between the laws of completely random measures, to measure the discrepancy between Bayesian nonparametric models for exchangeable time-to-event data. On the other hand, in [3], $d_{\mathcal{W},2}$ was used to measure the dependence structure of a CRV.

## 4 Measuring Dependence in Bayesian Nonparametrics

In the last 20 years Bayesian nonparametric models have gone beyond the exchangeability assumption through the introduction of dependent random measures, which provide a flexible framework for modeling the heterogeneity across multiple populations. The prior dependence between random measures regulates the borrowing of strength across different populations and thus needs a careful elicitation. The current state-of-the-art is to provide the analytical expression for the linear correlation $Corr(\tilde{\mu}_1(A), \tilde{\mu}_2(A))$, which only captures partial information about the dependence structure. In [3] the authors propose to use the distance defined in Proposition 3.2 to compare different dependence structures between CRVs with equal marginal distributions, i.e. in the same Fréchet class. In particular, one may define an overall measure of dependence of $\tilde{\boldsymbol{\mu}}$ by considering its distance from the maximally dependent CRV in the same Fréchet class, usually referred to as the comonotonic vector $\tilde{\boldsymbol{\mu}}^{\text{co}}$:

$$\text{Dep}(\tilde{\boldsymbol{\mu}}) = d_{\mathcal{W},2}(\tilde{\boldsymbol{\mu}}, \tilde{\boldsymbol{\mu}}^{\text{co}}) \tag{4.1}$$

The goal is to find tight bounds for $\text{Dep}(\tilde{\boldsymbol{\mu}})$ in terms of the hyperparameters of the model, in order to quantify their impact on the dependence structure for a principled prior elicitation. This is achieved in [3] by (1) using compound Poisson approximations; (2) finding a new upper bound on the Wasserstein distance between bivariate compound Poisson distributions; (3) finding the expression for the optimal coupling between a distribution on $\mathbb{R}^2$ and the comonotonic one in the same Fréchet class. In particular, one finds tight upper bounds for $\text{Dep}(\tilde{\boldsymbol{\mu}})$ in terms

of the underlying bivariate Lévy measures. This allows to treat many noteworthy dependence structures, such as GM-dependence [11, 16, 17], compound random measures [10, 23] and Clayton Lévy copulae [4, 8, 15, 24].

## 5 Survival Analysis in Bayesian Nonparametrics

Completely random measures play a particularly important role in Bayesian nonparametric models for time-to-event data. Here the main quantities of interest are typically the survival function, cumulative hazards and the hazard function. Consequently, the most notable Bayesian nonparametric models in survival analysis and reliability theory provide closed form estimates of these functions, in terms of underlying CRMs. Among the models for the hazard function, the one proposed by Dykstra and Laud [7] stands out for combining both flexibility and tractability. The random hazard function $\tilde{h}$ is modeled as a kernel mixture over a CRM:

$$\tilde{h}(t) = \int_{\mathbb{X}} k(t|x)\, \tilde{\mu}(dx), \tag{5.1}$$

where $t \in \mathbb{R}^+$, $k : \mathbb{X} \times \mathbb{R}^+ \to \mathbb{R}^+$ is a measurable kernel and $\tilde{\mu}$ is a CRM on $\mathbb{X}$ with Lévy intensity $\nu$. The original model in [7] was defined for $k(t|x) = \beta(x)\mathbb{1}_{(0,t]}(x)$, where $\beta : \mathbb{R}^+ \to \mathbb{R}^+$ is a measurable function, and $\tilde{\mu}$ a gamma CRM, i.e. such that the Lévy intensity satisfies

$$\nu(ds, dx) = \frac{e^{-bs}}{s}\mathbb{1}_{(0,+\infty)}(s)\, ds\, \alpha(dx), \tag{5.2}$$

for $b > 0$ and $\alpha \in M_{\mathbb{X}}$. We write $\tilde{\mu} \sim \mathrm{Ga}(b, \alpha)$. The extension to a general kernel was proposed in [19]. Later work by James [12] also extended this model to a general CRM.

The hazard model (5.1) is very flexible and incorporates a wide variety of distributional assumptions. On the other hand, it is not easy to understand how the parameters of the CRM and the kernel function impact the distribution of the random hazard. For a careful prior elicitation and sensitivity analysis, a principled quantification of the discrepancy at the level of the hazards is of fundamental need. We choose the Wasserstein distance of order 1 as a measure of discrepancy and we seek for an analytical expression of

$$\sup_{t\in[0,T]} \mathcal{W}_1(\tilde{h}_1(t), \tilde{h}_2(t)) \tag{5.3}$$

where $T > 0$ and $\tilde{h}_1, \tilde{h}_2$ are two different specifications for (5.1). We point out that $[0, T]$ may be interpreted as a time interval of interest, typically coinciding with the start and the end of the study. The analytical evaluation of a distance is a difficult task

in general, even more so since the law of the random variable in (5.1) is defined in an indirect way through the Lévy measure of the mixing CRM. Nonetheless, in [2] the authors were able to find informative bounds on the distance in (3.2) in terms of the corresponding Lévy measures by (1) leveraging compound Poisson approximations of the completely random measures; (2) bounding the Wasserstein distance between compound Poisson distributions, as first suggested by Mariucci and Reiß [20] in the context of Lévy processes. Then, the deep connections between (5.3) and (3.2) lead to the following theorem, whose proof may be found in [2]. Before providing the main statement, we stress two conditions on the kernels.

$$\lim_{t\to\infty} \int_0^t \int_{\mathbb{R}^+\times\mathbb{X}} k(u \mid x)\, s\, du\, \mathcal{N}(ds, dx) = +\infty, \tag{5.4}$$

$$\int_{\mathbb{R}^+\times\mathbb{X}} k(t|x)\, s\, \nu(ds, dx) < +\infty. \tag{5.5}$$

**Theorem 5.1** *Let $\tilde{h}_1 = \{\tilde{h}_1(t) \mid t \geq 0\}$ and $\tilde{h}_2 = \{\tilde{h}_2(t) \mid t \geq 0\}$ be random hazard rates as in (5.1) with associated infinitely active CRMs $\tilde{\mu}_i$, Lévy intensity $\nu_i$, and kernel $k_i$ that satisfy (5.4) and (5.5), for $i = 1, 2$. Then the Wasserstein distance between the marginal hazard rates is finite and for every $t \geq 0$,*

$$g_{low}(t) \leq \mathcal{W}_1(\tilde{h}_1(t), \tilde{h}_2(t)) \leq g_{up}(t),$$

*where*

$$g_{low}(t) = \left| \int_{\mathbb{R}^+\times\mathbb{X}} k_1(t|x)\, s\, \nu_1(ds, dx) - \int_{\mathbb{R}^+\times\mathbb{X}} k_2(t|x)\, s\, \nu_2(ds, dx) \right|,$$

$$g_{up}(t) = \int_0^{+\infty} \left| \int_{(u,+\infty)\times\mathbb{X}} \frac{1}{k_1(t \mid x)} \nu_1\Big(d\frac{s}{k_1(t \mid x)}, dx\Big) - \frac{1}{k_2(t \mid x)} \nu_2\Big(d\frac{s}{k_2(t \mid x)}, dx\Big) \right| du.$$

*In particular if there exists a dominating measure $\eta$ such that the Radon–Nikodym derivatives $\nu_i(s, x)$ satisfy, for $i \neq j$ in $\{1, 2\}$,*

$$\frac{1}{k_i(t|x)} \nu_i\Big(\frac{s}{k_i(t|x)}, x\Big) \leq \frac{1}{k_j(t|x)} \nu_j\Big(\frac{s}{k_j(t|x)}, x\Big) \tag{5.6}$$

*for all $(s, x) \in \mathbb{R}^+ \times \mathbb{X}$, then*

$$\mathcal{W}_1(\tilde{h}_1(t), \tilde{h}_2(t)) = \left| \int_{\mathbb{R}^+\times\mathbb{X}} k_1(t|x)\, s\, \nu_1(ds, dx) - \int_{\mathbb{R}^+\times\mathbb{X}} k_2(t|x)\, s\, \nu_2(ds, dx) \right|.$$

Theorem 5.1 was used in [2] to measure the discrepancy between hazards with kernels of the type of [7], i.e. $k(t|x) = \beta(x)\mathbb{1}_{[0,t]}(x)$, which is a popular choice when modeling increasing hazards, in the particular case where $\beta(x) = \beta$ is a constant function. This specification is very common in applications and brings to the following measurement of discrepancy, whose proof may be found in [2]. We denote by $\text{Leb}^+$ the Lebesgue measure on $(0, +\infty)$.

**Theorem 5.2** *Let $\tilde{\mu}_i \sim \text{Ga}(b_i, \text{Leb}^+)$ as defined in (5.2) and let $k_i(t|x) = \beta_i \mathbb{1}_{[0,t]}(x)$, with $b_i, \beta_i > 0$, for $i = 1, 2$. If $\tilde{h}_1$ and $\tilde{h}_2$ are the corresponding hazard rate mixtures, then*

$$\mathcal{W}_1(\tilde{h}_1(t), \tilde{h}_2(t)) = t \left| \frac{\beta_1}{b_1} - \frac{\beta_2}{b_2} \right|.$$

## 6 Time-Dependent Kernels

In this section we make some progress in the understanding of the distributional implications of the hazard rate model in (5.1) when $\tilde{\mu} \sim \text{Ga}(b, \text{Leb}^+)$ as defined in (5.2). In particular, the goal is to understand the impact of using a kernel of the type of [7] when the time influences also the functional form of the kernel and not only the support, i.e. $\beta(\cdot)$ is not constant in $(0, t]$. This scenario is of particular importance when we judge that the index of dispersion varies in time, since when $\beta(x) = \beta > 0$,

$$\frac{\text{Var}(\tilde{h}(t))}{\mathbb{E}(\tilde{h}(t))} = \frac{\beta}{b}.$$

We thus consider the scenario where $\beta(x) = \beta + \gamma x$, with $\beta, \gamma > 0$.

**Theorem 6.1** *Let $\tilde{\mu}_i \sim \text{Ga}(b_i, \text{Leb}^+)$ as defined in (5.2) and let $k_1(t|x) = \beta \mathbb{1}_{[0,t]}(x)$ and $k_2(t|x) = (\beta + \gamma x)\mathbb{1}_{[0,t]}(x)$, with $b_1, b_2, \beta, \gamma > 0$. If $\tilde{h}_1$ and $\tilde{h}_2$ are the corresponding hazard rate mixtures, then*

1. *If $b_1 \geq b_2$,*

$$\mathcal{W}_1(\tilde{h}_1(t), \tilde{h}_2(t)) = \left( \frac{1}{b_2} - \frac{1}{b_1} \right) \beta t + \frac{\gamma}{2b_2} t^2.$$

2. *If $b_1 \leq b_2$ and $t \leq \beta(b_2 - b_1)/(b_1\gamma)$*

$$\mathcal{W}_1(\tilde{h}_1(t), \tilde{h}_2(t)) = \left( \frac{1}{b_1} - \frac{1}{b_2} \right) \beta t - \frac{\gamma}{2b_2} t^2.$$

3. *Otherwise,*

$$g_{low}(t) \leq \mathcal{W}_1(\tilde{h}_1(t), \tilde{h}_2(t)) \leq g_{up}(t),$$

*where*

$$g_{low}(t) = \left(\frac{1}{b_2} - \frac{1}{b_1}\right)\beta t + \frac{\gamma}{2b_2}t^2$$

$$g_{up}(t) = \left(\frac{1}{b_2} - \frac{1}{b_1}\right)^2 \frac{\beta^2 b_2}{\gamma} + \left(\frac{1}{b_2} - \frac{1}{b_1}\right)\beta t + \frac{\gamma}{2b_2}t^2$$

***Proof*** First of all we observe that

$$\nu_{k,1}(ds, dx) = \frac{1}{k_1(t|x)}\nu_1\Big(d\frac{s}{k_1(t|x)}\, dx\Big) = \frac{e^{-\frac{s\,b_1}{\beta}}}{s}\, \mathbb{1}_{(0,+\infty)}(s)\, \mathbb{1}_{[0,t]}(x)\, ds\, dx,$$

$$\nu_{k,2}(ds, dx) = \frac{1}{k_2(t\mid x)}\nu_2\Big(d\frac{s}{k_2(t|x)}\, dx\Big) = \frac{e^{-\frac{s\,b_2}{\beta+\gamma x}}}{s}\, \mathbb{1}_{(0,+\infty)}(s)\, \mathbb{1}_{[0,t]}(x)\, ds\, dx.$$

Since $\gamma > 0$, (5.6) holds whenever $b_1 \geq b_2$. Part 1 of the statement thus holds by Theorem 5.1, by observing that

$$\int_{\mathbb{R}^+\times\mathbb{R}} k_2(t\mid x)\, s\, \nu_2(ds, dx) = \frac{\beta}{b_2}t + \frac{\gamma}{2b_2}t^2.$$

As for part 2 of the statement, it suffices to prove the expression for the upper bound. With a slight abuse of notation, indicate by $\nu_{k,i}(s, x)$ the Radon–Nikodym derivative of $\nu_{k,i}(ds, dx)$, for $i = 1, 2$. We observe that $\nu_{k,1}(s, x) \leq \nu_{k,2}(s, x)$ for every $s > 0$ and every $t \geq y \geq \beta(b_2 - b_1)/(b_1\gamma)$, so that the Radon–Nikodym derivatives are not globally ordered. We denote by $\delta = \min(\beta(b_2 - b_1)/(b_1\gamma), t)$. We then have

$$\int_0^{+\infty} \left| \int_{(u,+\infty)\times\mathbb{R}} (\nu_{k,1}(s, x) - \nu_{k,2}(s, x))\, ds\, dx \right| du$$

$$\leq \int_0^{+\infty} \int_{(u,+\infty)} \int_{\mathbb{R}} |\nu_{k,1}(s, x) - \nu_{k,2}(s, x)|\, dx\, ds\, du.$$

By interchanging the integrals thanks to Fubini's Theorem, this is equal to

$$\int_0^{+\infty} \int_{\mathbb{R}} \left(\int_0^s du\right) |\nu_{k,1}(s, x) - \nu_{k,2}(s, x)|\, dx\, ds$$

$$= \int_0^{+\infty} \int_{\mathbb{R}} |s\nu_{k,1}(s, x) - s\nu_{k,2}(s, x)|\, dx\, ds$$

$$= \int_0^{+\infty} \left( \int_0^{\delta} (e^{-\frac{s\,b_1}{\beta}} - e^{-\frac{s\,b_2}{\beta+\gamma x}})\, dx + \int_{\delta}^{t} (e^{-\frac{s\,b_2}{\beta+\gamma x}} - e^{-\frac{s\,b_1}{\beta}})\, dx \right) ds$$

$$= \int_0^{\delta} \left( \frac{\beta}{b_1} - \frac{\beta + \gamma x}{b_2} \right) dx + \int_{\delta}^{t} \left( \frac{\beta + \gamma x}{b_2} - \frac{\beta}{b_1} \right) dx.$$

The conclusion follows by simple calculations. □

Theorem 6.1 allows one to measure the impact of introducing a time dependent function in the kernel of [7]. In particular, we underline how the discrepancy grows quadratically in time, thus greatly influencing our prior opinion on the process for large $t$. Moreover, as $t \to +\infty$ we observe that the upper and lower bounds are asymptotically equivalent, providing the exact leading term for the Wasserstein distance.

## 7 Discussion and Further Work

In this paper we have discussed two different frameworks where transport distances between random vectors of measures provide deeper insights on notable Bayesian nonparametric models, favoring the elicitation of the prior. The amount of dependence in partially exchangeable models regulates the borrowing of information across groups, with a large impact on the inference. It is thus of fundamental importance to translate our prior beliefs on the dependence structure into the specification of the prior. Since exchangeability corresponds to a situation of maximal dependence, it seems natural to encode the prior beliefs on the dependence structure in terms of distance from the exchangeable scenario, as in (4.1). By choosing a subjective threshold $\tau$ for such distance, the tight upper bounds found in [3] may be set to be equal to $\tau$ by choosing appropriate values of the hyperparameters of the model. Moreover, Theorem 6.1 may be used for the prior elicitation of hazard rate models as in (5.1). The kernel $k(t|x) = \beta \mathbb{1}_{[0,t]}(x)$ with $\beta > 0$ is the most common specification in applications involving increasing hazard rates and may be treated as a reference kernel. However, if one believes that the index of dispersion varies over the time interval of interest $(0, t^*]$, it is natural to use a time varying specification as $k(t|x) = (\beta + \gamma x)\mathbb{1}_{[0,t]}(x)$, though securing a certain degree of similarity with respect to the reference kernel for every $t \in (0, t^*]$. By measuring the similarity in terms of Wasserstein distance and fixing a subjective threshold $\tau$, the exact value of the distance or its upper bound in Theorem 6.1 is maximized in $t^*$. One can then set the upper bound at time $t^*$ equal to $\tau$, so that the hyperparameter $\gamma$ may be chosen and elicited accordingly.

Completely random measures are widely used because they combine modeling flexibility with analytical tractability. In particular, there are many closed form results for the posterior distribution of the random measures given exchangeable or partially exchangeable observations. These have been used for example in [2] to evaluate approximation errors of a posterior sampling scheme in terms of the Wasserstein distance. Future research will concern the analysis of the dependence structure of the posterior distribution $\tilde{\boldsymbol{\mu}}^*$ through $\mathrm{Dep}(\tilde{\boldsymbol{\mu}}^*)$ in (4.1). The plan would

then be to use this to test whether the data supports the heterogeneity assumption across groups, along the lines of the parametric tests developed in [1].

**Acknowledgement** Antonio Lijoi and Igor Prünster are partially supported by MIUR, PRIN Project 2015SNS29B.

## References

1. Bacallado, S., Diaconis, P., Holmes, S.: de Finetti priors using Markov chain Monte Carlo computations. J. Stat. Comput. **25**, 797–808 (2015)
2. Catalano, M., Lijoi, A., Prünster, I.: Approximation of Bayesian models for time-to-event data. Electron. J. Stat. **14**, 3366–3395 (2020)
3. Catalano, M., Lijoi, A., Prünster, I.: Measuring the dependence in the Wasserstein distance for Bayesian nonparametric models. Annals of Statistics, forthcoming. DOI: https://doi.org/10.1214/21-AOS2O6S (2021)
4. Cont, R., Tankov, P.: Financial Modeling with Jump Processes. Chapman and Hall/CRC, Boca Raton (2004)
5. Daley, D.J., Vere-Jones, D.: An Introduction to the Theory of Point Processes: Volume I: Elementary Theory and Methods. Probability and Its Applications. Springer, Berlin (2002)
6. Doksum, K.: Tailfree and neutral random probabilities and their posterior distributions. Ann. Probab. **2**, 183–201 (1974)
7. Dykstra, R.L., Laud, P.: A Bayesian nonparametric approach to reliability. Ann. Stat. **9**, 356–367 (1981)
8. Epifani, I., Lijoi, A.: Nonparametric priors for vectors of survival functions. Stat. Sin. **20**, 1455–1484 (2010)
9. Ferguson, T.S.: Bayesian density estimation by mixtures of normal distributions. In: Recent Advances in Statistics, pp. 287–302. Academic Press, New York (1983)
10. Griffin, J.E., Leisen, F.: Compound random measures and their use in Bayesian non-parametrics. JRSS B **79**, 525–545 (2017)
11. Griffiths, R., Milne, K.R.: A class of bivariate Poisson processes. J. Multivar. Anal. **8**, 380–395 (1978)
12. James, L.F.: Bayesian Poisson process partition calculus with an application to Bayesian Lévy moving averages. Ann. Stat. **33**, 1771–1799 (2005)
13. Kallenberg, O.: Random Measures, Theory and Applications. Springer International Publishing, Cham (2017)
14. Kingman, J.F.C.: Completely random measures. Pacific J. Math. **21**, 59–78 (1967)
15. Leisen, F., Lijoi, A.: Vectors of two-parameter Poisson-Dirichlet processes. J. Multivar. Anal. **102**, 482–495 (2011)
16. Lijoi, A., Nipoti, B.I.: A class of hazard rate mixtures for combining survival data from different experiments. J. Am. Stat. Assoc. **20**. 802–814 (2014)
17. Lijoi, A., Nipoti, B., Prünster, I.: Bayesian inference with dependent normalized completely random measures. Bernoulli **20**, 1260–1291 (2014)
18. Lo, A.: On a class of Bayesian nonparametric estimates: I. Density estimates. Ann. Stat. **12**, 351–357 (1984)
19. Lo, A., Weng, C.: On a class of Bayesian nonparametric estimates: II. Hazard rate estimates. Ann. Inst. Stat. Math. **41**, 227–245 (1989)
20. Mariucci, E., Reiß, M.: Wasserstein and total variation distance between marginals of Lévy processes. Electron. J. Stat. **12**, 2482–2514 (2018)
21. Quintana, F.A., Müller, P., Jara, A., MacEachern, S.N.: The dependent Dirichlet process and related models. arXiv 2007.06129 (2020)

22. Regazzini, E., Lijoi, A., Prünster, I.: Distributional results for means of normalized random measures with independent increments. Ann. Stat. **31**, 560–585 (2003)
23. Riva–Palacio, A., Leisen, F.: Compound vectors of subordinators and their associated positive Lévy copulas. arXiv 1909.12112 (2019)
24. Tankov, P.: Dependence structure of spectrally positive multidimensional Lévy processes. Unpublished manuscript (2003)
25. Villani, C.: Optimal Transport: Old and New. Springer, Berlin Heidelberg (2008)