The Annals of Statistics
2016, Vol. 44, No. 2, 771–812
DOI: 10.1214/15-AOS1387

AMPLITUDE AND PHASE VARIATION OF POINT PROCESSES[1]

BY VICTOR M. PANARETOS AND YOAV ZEMEL

Ecole Polytechnique Fédérale de Lausanne

We develop a canonical framework for the study of the problem of registration of multiple point processes subjected to warping, known as the problem of separation of amplitude and phase variation. The amplitude variation of a real random function $\{Y(x) : x \in [0, 1]\}$ corresponds to its random oscillations in the y-axis, typically encapsulated by its (co)variation around a mean level. In contrast, its phase variation refers to fluctuations in the x-axis, often caused by random time changes. We formalise similar notions for a point process, and nonparametrically separate them based on realisations of i.i.d. copies $\{\Pi_i\}$ of the phase-varying point process. A key element in our approach is to demonstrate that when the classical phase variation assumptions of Functional Data Analysis (FDA) are applied to the point process case, they become equivalent to conditions interpretable through the prism of the theory of optimal transportation of measure. We demonstrate that these induce a natural Wasserstein geometry tailored to the warping problem, including a formal notion of bias expressing over-registration. Within this framework, we construct nonparametric estimators that tend to avoid over-registration in finite samples. We show that they consistently estimate the warp maps, consistently estimate the structural mean, and consistently register the warped point processes, even in a sparse sampling regime. We also establish convergence rates, and derive $\sqrt{n}$-consistency and a central limit theorem in the Cox process case under dense sampling, showing rate optimality of our structural mean estimator in that case.

1. Introduction. When analysing the (co)variation of a real random function $\{Y(x) : x \in K\}$ over a continuous compact domain K, it can be broadly said that one may distinguish two layers of variation. The first is *amplitude variation*. This is the "classical" variation that one would also encounter in multivariate analysis, and refers to the stochastic fluctuations around a mean level, usually encoded in its covariance kernel, at least up to second order. In short, this is variation "in the y-axis."

The second layer of variation is a non-linear variation peculiar to continuous domain stochastic processes, and is rarely—if ever—encountered in multivariate analysis. It arises as the result of random changes (or deformations) in the time scale (or the spatial domain) of definition of the process. It can be conceptualised as

Received September 2014; revised September 2015.

[1]Supported by a European Research Council Starting Grant Award.

MSC2010 subject classifications. Primary 62M; secondary 60G55, 62G.

Key words and phrases. Doubly stochastic Poisson process, Fréchet mean, geodesic variation, Monge problem, optimal transportation, length space, registration, warping, Wasserstein metric.

a composition of the stochastic process with a random transformation acting on its domain, or as variation "in the x-axis," typically referred to as a *warp function*. The terminology on amplitude/phase variation is adapted from trigonometric functions, which may vary in amplitude or phase.

Phase variation arises quite naturally in the study of random phenomena where there is no absolute notion of time or space, but every realisation of the phenomenon evolves according to a time-scale that is intrinsic to the phenomenon itself, and (unfortunately) unobservable. Processes related to physiological measurements (such as growth curves, neuronal signals, or brain images), are usual suspects, where phase variability arises at the level of individual (see the extensive discussion in Ramsay and Silverman [30, 31]); but examples abound in diverse fields of application of stochastic processes, perhaps quite prominently in environmental sciences (e.g., Sampson and Guttorp [33], and references therein) and pattern recognition (for instance, handwriting analysis, e.g., Ramsay [28], or speech analysis, e.g., Hadjipantelis, Aston and Evans [19]).

Natural as the confluence of these two types of variation may be, failing to recognise and correct for their entanglement can obscure or even entirely distort the findings of a statistical analysis of the random function (see Section 2). Consequently, it is an important problem to be able to separate the two, thus correctly accounting for the distinct contribution of each. If one is able to only observe a single realisation of the random function $\{Y(x)\}$ in question, the separation problem is not well-defined unless further modelling assumptions are introduced. For example, one could assume that a process should be stationary or otherwise have some invariance property in the x-domain that is measurably perturbed by the phase variation; and attempt to unwarp it on the basis of this assumption. Such models can be found in the analysis of random fields (see, e.g., Sampson and Guttorp [33], Anderes and Stein [3], Anderes and Chatterjee [2]), and of points processes alike (see, e.g., Schoenberg [34], Senoussi, Chadoef and Allard [35]).

In the field of functional data analysis, however, one has the good fortune of being able to observe multiple i.i.d. realisations $\{Y_1(x), \ldots, Y_n(x)\}$ of the random function in question. When this is the case, one may attempt to separate phase and amplitude variation under less stringent assumptions—in fact in a nonparametric fashion. Indeed, there is a substantial amount of work on this topic in the field of functional data, as the problem is in some sense one of the distinguishing characteristics of FDA as compared to multivariate statistics (see Section 2).

The purpose of this paper is to investigate the problem of separation of amplitude and phase variation in the case where one observes multiple realisations $\{\Pi_1, \ldots, \Pi_n\}$ of *random point processes* rather than *random functions*. Though the study of multiple realisations of point processes has been considered prior to the emergence of FDA (see, e.g., Karr [22]), treating realisations of point processes as individual data objects within a functional data analysis context is a more recent development offering important advantages; a key paper is that of Wu, Müller and Zhang [42] (also see Chiou and Müller [10] and Chiang, Wang and Huang

[9]). Such data may be an object of interest in themselves (see, e.g., Wu, Müller and Zhang [42], Arribas-Gil and Müller [4], Wu and Srivastava [43]) but may also arise as landmark data in an otherwise classical functional data analysis (see, e.g., Gasser and Kneip [16], Arribas-Gil and Müller [4]). The recent surge of interest is exemplified in an upcoming discussion paper by Wu and Srivastava [44], whose discussion documents early progress and challenges in the field. One of the main complications arising in the point process case is that a point processes, when viewed as a single *datum*, is a discrete random measure. The nature of such a datum gives rise to different sets of challenges as compared to FDA. Their ambient space is not a vector space, so point process variation—whether due to amplitude or phase—is intrinsically non-linear, calling for an analysis either via a suitable transformation, or via consideration of an alternative space where their covariation structure can be suitably analysed. Nevertheless, this special nature can be seen as a blessing, rather than a curse, as the case of point processes enjoys important advantages that considerably simplify the analysis relative to more general functions.

Specifically, we argue that the problem of amplitude and phase variation in point process data admits a *canonical* framework through the theory of optimal transportation of measure. Indeed, we show that this formulation follows unequivocally when employing the classical phase variation assumptions of functional data analysis to the point process case (Section 3.2, Assumptions 1). These are proven to be *equivalent* to a geometrical characterisation of the problem by means of geodesic variation around a Fréchet mean with respect to the Wasserstein metric (Section 3.3, Proposition 1). We show that the special nature of the problem in the case of point processes renders it identifiable (Section 3.3, Proposition 2) and also allows for the elucidation of what "over" and "under" registering means, through a notion of *unbiased registration* (Section 5). We construct easily implementable *nonparametric* estimators that separate amplitude and phase (Section 4) and develop their asymptotic theory, establishing consistency in a genuinely nonparametric framework (Section 6, Theorem 1) even under sparse sampling (Remark 1). In the special case of Cox processes (randomly warped Poisson processes, see Section 3.5), we derive rates of convergence (Theorem 2), and provide conditions for $\sqrt{n}$-consistency. We also obtain a central limit theorem for the estimator of the structural mean (Theorem 3), which shows our estimator attains the optimal rate under dense sampling and allows for uncertainty quantification (Remark 5). The finite sample performance methodology is illustrated by means of examples in Section 8, and a simulation study in the supplementary material [27].

2. Amplitude and phase variation of functional data. In order to motivate our framework for modelling amplitude and phase variation in point processes, we first revisit the case of functional data, that is, n independent realisations of a random element of $L^2[0,1]$, say $\{Y_i(x) : x \in [0,1]; i = 1,\dots,n\}$. One typically understands *amplitude variation* as corresponding to linear stochastic variability

in the observations. That is, assuming that the mean function is $\mu(x) \in L^2[0,1]$, amplitude variation enters the model through

$$Y_i(x) = \mu(x) + Z_i(x), \qquad i = 1, \dots, n,$$

where the $Z_i(x)$ are mean zero i.i.d. stochastic processes with covariance kernel $\kappa(s,t)$, typically assumed to be continuous (equivalently, Z_i are assumed continuous in mean square). In this setup, the covariation structure of Y can be probed by means of the Karhunen–Loève expansion,

$$Y(x) = \mu(x) + \sum_{n=1}^{\infty} \xi_n \varphi_n(x), \tag{2.1}$$

the optimal Fourier representation of Y in the ortho-normal system of eigenfunctions of κ. The equality is understood in $\mathbb{P}-$mean square, uniformly in x. This expansion explains the term *amplitude variation*: Y varies about μ by random amplitude oscillations of the functions $\{\varphi_n\}$. A key feature of this expansion is the separation of the stochastic component (in the countable collection $\{\xi_n\}$) and the functional component (in the deterministic collection $\{\varphi_n\}$).

On the other hand, *phase variation* is understood as the presence of non-linear variation. Heuristically, this means that there is an initial random change of time scale, followed by amplitude variation, yielding *time-warped* curves $\widetilde{Y}_i$,

$$\begin{aligned} \widetilde{Y}_i(x) = Y_i\big(T_i^{-1}(x)\big) &= \mu\big(T_i^{-1}(x)\big) + Z_i\big(T_i^{-1}(x)\big) \\ &= \mu\big(T_i^{-1}(x)\big) + \sum_{n=1}^{\infty} \xi_n \varphi_n\big(T_i^{-1}(x)\big). \end{aligned} \tag{2.2}$$

The warp functions $T_i : [0,1] \to [0,1]$ are typically assumed to be random *increasing functions* independent of the Z_i and with $\mathbb{E}[T_i(x)] = x$. Consequently, one has

$$\begin{aligned} \mathbb{E}\big[\widetilde{Y}(x)|T\big] &= \mu\big(T^{-1}(x)\big) = \widetilde{\mu}(x); \\ \operatorname{cov}\big\{\widetilde{Y}(x), \widetilde{Y}(y)\big\} &= \mathbb{E}\big[\kappa\big(T^{-1}(x), T^{-1}(y)\big)\big] + \operatorname{cov}\big\{\widetilde{\mu}(x), \widetilde{\mu}(y)\big\}, \end{aligned}$$

and thus notices that the right-hand side of equation (2.2) is no longer interpretable as the Karhunen–Loève expansion of $\widetilde{Y}_i$ [the $\varphi_n(T^{-1}(x))$ are *not* eigenfunctions of the covariance kernel $\operatorname{cov}\{\widetilde{Y}(x), \widetilde{Y}(y)\}$]. Indeed, if one ignores phase variation, and proceeds to analyse the $\widetilde{Y}_i$'s by their own Karhunen–Loève expansion, the analysis will be seriously distorted: the eigenfunctions will be more diffuse and less interpretable (owing to the effect of attempting to capture horizontal variation via vertical variation, i.e., local features by global expansions) and the spectral decay of the covariance operator will be far slower (requiring the retention of a larger number of components in an eventual principal component analysis).

The data will then usually come in the form of discrete measurements on a grid $\{t_j\}_{j=1}^m \subset [0,1]$ subject to additive white noise of variance $\sigma^2 > 0$,

$$\widetilde{y}_{ij} = \widetilde{Y}_i(t_j) + \varepsilon_{ij}, \qquad i = 1, \ldots, n;\ j = 1, \ldots, m, \tag{2.3}$$

assuming of course that the Y_i are continuous. The problem of separation of amplitude and phase variation can now be seen as that of recovering the T_i and Y_i from the data $\{\widetilde{y}_{ij}\}_{i=1}^n$, and therefore separating phase variation (fluctuations of T_i) and amplitude variation (fluctuations of Y_i). Doing so successfully depends on the nature of T (e.g., to guarantee identifiability), the crystallisation of which is a matter of assumption. Specifically, more assumptions are needed further to monotonicity and the expected value being the identity. Indeed, there does not appear to be a single universally accepted formulation. In landmark registration, for example, the T are estimated by assuming that clearly defined landmarks (such as local maxima of the curves or their derivatives) be optimally aligned across curves (Gasser and Kneip [16]; see also Gervini and Gasser [17] for a more flexible setup). Template methods iteratively register curves to a template, minimising an overall discrepancy; the template is then updated, for example, starting from the overall mean (Wang and Gasser [40]; Ramsay and Li [29]). Moment-based registration proceeds by an alignment of the moments of inertia of the curves (James [20]). Pairwise separation proceeds by iteratively registering pairs of observations by means of a penalised sums of square criterion, and takes advantage of a moment assumption on T being the identity on average to derive a global alignment (Tang and Müller [37]). Approaches of a semi-parametric flavour assume a functional form for T that is known, except for a finite dimensional parameter, and proceed by likelihood methods in a random-effects type setup (Rønn [32]; Gervini and Gasser [18]). Principal components based registration registers the data so that the resulting curves have a parsimonious representation by means of a principal components analysis (the "least second eigenvalue" principle; Kneip and Ramsay [24]). Elastic registration defines a metric between curves that is invariant under joint elastic deformation of two curves by the same warp function, and registers by means of computing averages with respect to this metric (Tucker, Wu and Srivastava [38]). Multiresolution methods have also been proposed, leading to the notion of "warplets" (Claeskens, Silverman and Slaets [11]). In recent work, Marron et al. [26] consider comparisons between different registration techniques.

The literature is very rich, and a more in-depth review would be beyond the scope of the present paper. However, we note that a key conceptual aspect that recurs in several different estimation approaches in the literature is the postulate that a registration procedure should attempt to *minimise phase variability (a fit criterion)* subject to the constraint that the *registration maps ought to be smooth and as close to the identity map as possible (a regularity/parsimony criterion)*. With these key assumptions and principles in mind, we now turn to consider the case of point process data, and see how these ideas might be adapted.

3. Amplitude and phase variation of point processes.

3.1. *Amplitude variation*. Let Π be a point process on $[0,1]$, viewed as a random discrete measure, with the property that $\mathbb{E}\{(\int_0^1 d\Pi)^2\} < \infty$. Defining its mean measure as

$$\lambda(A) = \mathbb{E}\big\{\Pi(A)\big\}, \qquad A \in \mathscr{B}$$

on the collection of Borel sets $\mathscr{B}$ of $[0,1]$, we may understand amplitude variation as being encoded in the *covariance measure*,

$$\kappa(A \times B) = \operatorname{cov}\big\{\Pi(A), \Pi(B)\big\} = \mathbb{E}\big[\Pi(A)\Pi(B)\big] - \lambda(A)\lambda(B), \tag{3.1}$$

a signed Radon measure over Borel subsets of $[0,1]^2$. The covariance measure captures the second order fluctuations of $\Pi(A)$ around its mean value $\lambda(A)$, as well as their dependence on the corresponding fluctuations of $\Pi(B)$ around $\lambda(B)$. It naturally generalises the notion of a covariance operator for functional data to the case of point process data. Without loss of generality, we may assume that $\lambda(A)$ is renormalised to be a probability measure. In the absence of phase variation, estimation of the covariation structure of Π on the basis of n i.i.d. realisations $\Pi_1, \ldots, \Pi_n$ can be carried out by means of the empirical versions of λ and κ,

$$\widehat{\lambda}_n(A) = \frac{1}{n}\sum_{i=1}^{n} \Pi_i(A); \qquad \widehat{\kappa}_n(A \times B) = \frac{1}{n}\sum_{i=1}^{n} \Pi_i(A)\Pi_i(B) - \widehat{\lambda}_n(A)\widehat{\lambda}_n(B).$$

These are both strongly consistent (in the sense of weak convergence of measures with probability 1) as $n \to \infty$, and in fact one has the usual central limit theorem in that $\sqrt{n}(\widehat{\lambda}_n - \lambda)$ converges in law to a centred Gaussian random measure on $[0,1]$ with covariance measure κ (see, e.g., Karr [22], Proposition 4.8).

3.2. *Phase variation*: *First principles*. Phase variation may be introduced by direct analogy to the functional case. Assuming that $T_i : [0,1] \to [0,1]$ are i.i.d. random homeomorphisms, warped versions of the $\Pi_1, \ldots, \Pi_n$ can be defined as

$$\widetilde{\Pi}_i = T_{i\#}\Pi_i, \qquad i = 1, \ldots, n,$$

with $T_{i\#}\Pi_i(A) = \Pi_i(T_i^{-1}(A))$ the push-forward of Π_i through T_i. It is natural to assume that the collection $\{T_i\}$ is independent of the collection $\{\Pi_i\}$. Defining the random measures $\Lambda_i(A) = \lambda(T_i^{-1}(A)) = T_{i\#}\lambda(A)$, one also observes that the conditional mean and covariance measures of Π_i given T_i are

$$\begin{aligned} \mathbb{E}\{\widetilde{\Pi}|T\} &= \Lambda; \\ \operatorname{cov}\big\{\widetilde{\Pi}(A), \widetilde{\Pi}(B)\big\} &= \mathbb{E}\big\{\kappa\big(T^{-1}(A), T^{-1}(B)\big)\big\} + \operatorname{cov}\big\{\Lambda(A), \Lambda(B)\big\}, \end{aligned}$$

in analogy to the functional case. Furthermore, if $\Pi_i([0,t)) - \lambda([0,t))$ is mean-square continuous (equivalently, if $\mathrm{var}[\Pi(0,t)]$ is continuous), we have an expansion similar to that of equation (2.1) for the compensated process, and the warped compensated process

$$\Pi_i\big([0,t)\big) - \lambda\big([0,t)\big) = \sum_{n=1}^{\infty} \zeta_n \psi_n(t);$$

$$\widetilde{\Pi}_i\big([0,t)\big) - (T_{\#}\lambda)\big([0,t)\big) = \sum_{n=1}^{\infty} \zeta_n \psi_n\big(T^{-1}(t)\big),$$

where $\{\psi_n\}$ are the eigenfunctions of $\kappa(s,t) = \kappa\{[0,s],[0,t]\}$, in analogy with equation (2.2). The task of separation of amplitude and phase variation amounts to constructing estimators $\{\widehat{T}_i\}$ and $\{\widehat{\Pi}_i\}$ of the random maps T_i and of the unwarped (registered) point processes $\{\Pi_i\}$, respectively, on the basis of $\widetilde{\Pi}_1, \ldots, \widetilde{\Pi}_n$. Phase variation is then attributed to the $\{\widehat{T}_i\}$ and amplitude variation to the $\{\widehat{\Pi}_i\}$. As with the case of random curves, if consistent separation is to be achievable, we will need to impose some basic assumptions on the precise stochastic and analytic nature of the $\{T_i\}$. These will come in the form of *unbiasedness* and *regularity*.

ASSUMPTIONS 1. The maps $T_i : [0,1] \to [0,1]$ are i.i.d. random homeomorphisms distributed as T, independently of the point processes $\{\Pi_i\}$. The random map T satisfies the following two conditions:

(A1) *Unbiasedness*: $\mathbb{E}[T(x)] = x$ almost everywhere on $[0,1]$.
(A2) *Regularity*: T is monotone increasing almost surely.

Assumption (A1) asks that the average time change $\mathbb{E}[T(x)]$ be the identity: on average, the "objective" time-scale should be maintained, so that time is not overall sped up or slowed down. Now, since T is already a homeomorphism, it is bound to be monotone, either increasing or decreasing. The regularity assumption (A2) asks that T represent a proper warping of time (time change): if (A2) were to fail, we would have a time reversal, which is rather problematic in most applied settings. Indeed, these assumptions are arguably *sine qua non* in the classical FDA phase variation literature, perhaps supplemented with further conditions as discussed earlier. We will now see that now such further conditions are unnecessary in the point process case, as they *derive* from the basic assumptions (A1) and (A2).

3.3. *Phase variation*: *Geometry.* Though our unbiasedness and regularity assumptions stem from first principles related to warping, they in fact are fully compatible with an elegant geometrical interpretation of phase variation—indeed one that opens the way for its consistent separation.

One may consider the space of all diffuse probability measures on $[0,1]$ as a metric space, endowed with the so-called L^2-Wasserstein distance (also known as Mallows' distance, or earth-mover's distance),

$$d(\mu,\nu) = \inf_{Q\in\Gamma(\mu,\nu)} \sqrt{\int_0^1 |Q(x)-x|^2 \mu(dx)}, \tag{3.2}$$

where $\Gamma(\mu,\nu)$ is the collection of mappings $Q:[0,1]\to[0,1]$ such that $Q_{\#}\mu=\nu$. The metric d is related to the so-called Monge problem of optimally transferring the mass of μ onto ν, with the cost of transferring a unit of mass from x to y being equal to their squared distance, $|x-y|^2$. In the case of diffuse measures (μ,ν), the infimum in equation (3.2) is attained at a unique map $T\in\Gamma(\mu,\nu)$ that is explicitly given by

$$T = F_\nu^{-1}\circ F_\mu,$$

where $F_\mu(t)=\int_0^t \mu(dx)$, $F_\nu(t)=\int_0^t \nu(dx)$ are the cumulative distribution functions corresponding to the two measures, and F_ν^{-1} is the quantile function $F_\nu^{-1}(p)=\inf\{y\in[0,1]: F_\nu(y)\geq p\}$ (see Villani [39], Chapter 7; Bickel and Freedman [5]). Consequently, the optimal map T inherits the regularity properties of the measures μ and ν, and does not require any further regularising assumptions. For example, if both measures admit continuous densities strictly positive on $[0,1]$, then T is a homeomorphism, but further smoothness assumptions on the densities will carry over to smoothness properties of the optimal maps.

When equipped with the metric d, the space of all diffuse probability measures on $[0,1]$ is a *length space* (also known as *inner metric space*), and the optimal Monge maps T, known as optimal transport maps, generate the geodesic structure of this space. Specifically, given any diffuse pair (μ,ν), there is a unique geodesic curve $\{\gamma(t): t\in[0,1]\}$ with endpoints μ and ν that is explicitly given by

$$\gamma(t) = \big[tT+(1-t)I\big]_{\#}\mu, \qquad t\in[0,1],$$

where T is the optimal coupling map of μ and ν, and I is the identity mapping [39], equation (5.11). The following proposition demonstrates how this optimal transportation geometry is inextricably linked with the first principles of phase variation, as encapsulated in assumptions (A1) and (A2).

PROPOSITION 1. *Let λ have strictly positive density with respect to Lebesgue measure on $[0,1]$. A random map $T:[0,1]\to[0,1]$ satisfies assumptions* (A1) *and* (A2), *if and only if it satisfies assumptions* (B1) *and* (B2) *as stated below*:

(B1) *Unbiasedness*: *Given any diffuse probability measure γ on $[0,1]$, we have*

$$\mathbb{E}\big\{d^2(T_{\#}\lambda,\lambda)\big\} \leq \mathbb{E}\big\{d^2(T_{\#}\lambda,\gamma)\big\}.$$

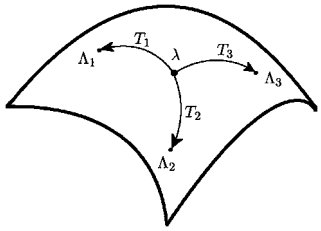

FIG. 1. *Schematic representation of the geometry of phase variation implied by our assumptions.*

(B2) *Regularity*: *Whenever* $T_{\#}\lambda = Q_{\#}\lambda$, *for some homeomorphism* $Q : [0, 1] \to [0, 1]$, *it must be that*

$$\int_0^1 \big|T(x) - x\big|^2 \lambda(dx) \leq \int_0^1 \big|Q(x) - x\big|^2 \lambda(dx) \qquad \textit{almost surely.}$$

In the optimal transportation geometry, the equivalent assumptions (B1) and (B2) have a clear-cut interpretation. Assumption (B2) implies that the conditional means $\Lambda_i = T_{i\#}\lambda$ of the warped processes correspond to perturbations of the structural mean measure λ along geodesics (see Figure 1). Furthermore, in the presence of (B2), assumption (B1) stipulates that these geodesic perturbations are "zero mean" in that the structural mean measure λ is a Fréchet mean of the Λ_i,

$$\mathbb{E}\big\{d^2(\Lambda, \lambda)\big\} \leq \mathbb{E}\big\{d^2(\Lambda, \gamma)\big\} \qquad \text{for any probability measure } \gamma.$$

Notice how these assumptions also mimic the additional estimation principles encountered in the phase variation of functional data (as discussed in the end of Section 2): we ask that the warp maps be such that *phase variability around the structural mean be minimised* [*our unbiasedness assumption* (B1)] subject to the constraint that the *registration maps deviate as least as possible from the identity map* [*our regularity assumption* (B2)]. In this case, however, these principles are *equivalent* to the basic assumptions, and do not have to be added as supplementary.

Furthermore, the following proposition establishes that if λ is a Fréchet mean of each Λ_i, then it is the unique such Fréchet mean. Our assumptions, therefore, suffice to guarantee identifiability of the structural mean (and hence, of the warping maps). We note that the cumulative distribution function of $\Lambda = T_{\#}\lambda$ is strictly increasing almost surely, as a composition of two such functions.

PROPOSITION 2 (Identifiability). *Let* Λ *be a diffuse random probability measure on* $[0, 1]$ *with a strictly increasing CDF almost surely. Then the minimiser of the functional*

$$\gamma \mapsto \mathbb{E}\big\{d^2(\Lambda, \gamma)\big\},$$

defined over probability measures γ *on* $[0, 1]$, *exists and is unique*.

3.4. *Phase variation*: *Measures vs. densities*. One should note that postulating that $\widetilde{\Pi} = T_{\#}\Pi$ induces phase variation of the conditional mean measure relative to the structural mean measure, $\Lambda = T_{\#}\lambda$. This is *not* equivalent to phase variation at the level of the *conditional mean density*, say f_Λ, relative to the *structural mean density*, say f_λ. Indeed, if $\Lambda = T_{\#}\lambda$ then

$$f_\Lambda(x) = \left[\frac{d}{dx}\big(T^{-1}(x)\big)\right] f_\lambda\big(T^{-1}(x)\big), \qquad x \in [0,1].$$

Thus, our framework cannot be equivalent to a model that directly models phase variation at the level of densities, by postulating (say) that $f_\Lambda(x) = f_\lambda(T(x))$. In such a model, phase variation immediately induces further amplitude variation, as the lack of a correcting factor $\frac{d}{dx}(T^{-1}(x))$ means that the new density is no longer a probability density, and thus the total measure of $[0,1]$ varies as a result of the variation of T (an overall amplitude variation effect).

An example of phase variation at the level of densities is the model of Wu and Srivastava [44], where the smoothed point processes are viewed as random density functions. These are then registered by employing the (extended) Fisher–Rao metric, using the algorithm of Srivastava et al. [36]. The authors of [36] argue that the Fisher–Rao approach consistently recovers phase variation for models of the type $f(x) = U \times g(T(x))$, where g is a deterministic function, U is a real random variable, and T is the phase map. In the particular case where phase variation is of densities, the model for the densities becomes

$$f_\Lambda(x) = U \times f_\lambda\big(T(x)\big).$$

Comparing the last two displayed equations, we see that the two setups are compatible when the T are assumed to be linear maps. In this case, unless $T(x) = x$ almost surely, our two conditions (A1) and (A2) cannot be consolidated: if we require $\mathbb{E}[T(x)] = x$, for a non-trivial random map (i.e., $\mathbb{P}[\|T - \mathrm{id}\|_{L^2} > 0] > 0$), then T cannot be an almost surely strictly increasing homeomorphism on the finite interval $[0,1]$.

Whether phase variation is formalised at the level of measure or density is to some extent a modelling decision. However, it is worth pointing out that if we wish to understand phase variation as the result of a *non-linear deformation of the underlying space* (e.g., a smooth deformation of the coordinate system), then the model postulating $\Lambda = T_{\#}\lambda$ appears to be the natural choice.

3.5. *Phase variation*: *The* (*warped*) *Poisson process case*. Just as Gaussian processes are the archetypal ones in the analysis of functional data, Poisson processes are so when it comes to point processes. It is hence worth to briefly consider the effect of phase variation as encoded in (A1) and (A2) [and their equivalent versions (B1) and (B2)] on a Poisson process.

Assume that Π is a Poisson point process with mean measure λ, and let $\widetilde{\Pi} = T_{\#}\Pi$ be the warped process, as before. Then, for any disjoint Borel sets

$\{A_1, \ldots, A_k\} \subset \mathscr{B}$, the random variables $\{\widetilde{\Pi}(A_j)\}_{j=1}^k$ are independent conditional on the random warp map T. This is because $\{T^{-1}(A_j)\}_{j=1}^k$ must also be disjoint Borel sets, combined with the fact that $\{\widetilde{\Pi}(A_1), \ldots, \widetilde{\Pi}(A_k)\} = \{\Pi(T^{-1}(A_1)), \ldots, \Pi(T^{-1}(A_k))\}$, with Π being Poisson. Furthermore, for any $A \in \mathscr{B}$,

$$\mathbb{P}\big[\widetilde{\Pi}(A) = k|T\big] = \mathbb{P}\big[\Pi\big(T^{-1}(A)\big) = k|T\big] = e^{-\lambda(T^{-1}(A))}\frac{\lambda^k(T^{-1}(A))}{k!}.$$

In other words, conditional on T, the process $\widetilde{\Pi}(A)$ is Poisson with mean measure $T_\#\lambda$. This establishes that $\widetilde{\Pi} = T_\#\Pi$ is distributionally equivalent to a *Cox process* with directing random measure $T_\#\lambda = \Lambda$. Consequently, our model for phase variation reduces to asking that the law of the warped point process is that of a Cox process, where the random directing measure Λ is non-linearly varying with a Fréchet mean (with respect to the Wasserstein distance) equal to the structural mean. Thus, in the Poissonian case, the compounding of phase and amplitude variation can be viewed as *double stochasticity*: the phase variation is attributed to the random directing measure, and the amplitude variation is attributed to the Poisson fluctuations conditional on the directing measure. It is worth comparing this with the framework introduced by Wu, Müller and Zhang [42], where point processes are modelled as Cox processes whose driving log-densities are linearly varying functional data.

4. Estimation.

4.1. *Overview of the estimation and registration procedure.* Armed with the intuition furnished by the geometrical interpretation of our assumptions, we may now formulate an estimation strategy. Since the structural mean measure λ is the Fréchet mean of the random measures $\Lambda_i = T_{i\#}\lambda$ in the Wasserstein metric, the natural estimator of λ would be the *empirical Fréchet–Wasserstein mean* of $\{\Lambda_1, \ldots, \Lambda_n\}$. Of course, the true $\{\Lambda_i\}$ are unobservable, and instead we observe the point processes $\{\widetilde{\Pi}_i\}$. However, since

$$T_{i\#}\lambda = \Lambda_i = \mathbb{E}\{\widetilde{\Pi}_i|T_i\},$$

a sensible strategy is to use proxies (estimates) of the $\{\Lambda_1, \ldots, \Lambda_n\}$ constructed on the basis of $\{\widetilde{\Pi}_1, \ldots, \widetilde{\Pi}_n\}$, and attempt to use these to approximate the empirical Fréchet–Wasserstein mean. Our procedure will follow the steps:

1. Estimate the random measures Λ_i. This may be done, for example, by carrying out classical density estimation on each $\widetilde{\Pi}_i$, viewed as a point process with mean measure Λ_i. Call these estimators $\widehat{\Lambda}_i$, with corresponding cumulative distribution functions $\widehat{F}_i(t) = \int_0^t \widehat{\Lambda}_i(dx)$.

2. Estimate λ by the empirical Fréchet mean of $\widehat{\Lambda}_1, \ldots, \widehat{\Lambda}_n$ (with respect to the Wasserstein metric d). We call this estimator the *regularised Fréchet–Wasserstein* mean, and denote it by $\widehat{\lambda}$, with corresponding cumulative distribution function $\widehat{F}(t) = \int_0^t \widehat{\lambda}(dx)$.

3. Estimate each T_i by the corresponding optimal transportation map of $\widehat{\lambda}$ onto $\widehat{\Lambda}_i$. In light of the discussion in the previous section, this is given by $\widehat{T}_i = \widehat{F}_i^{-1} \circ \widehat{F}$. Equivalently, one may estimate the registration maps by $\widehat{T}_i^{-1} = \widehat{T}_i^{-1} = \widehat{F}^{-1} \circ \widehat{F}_i$.

4. Register the point processes by pushing them forward through the registration maps,

$$\widehat{\Pi}_i = \widehat{T}_i^{-1}{}_{\#}\widetilde{\Pi}_i, \qquad i = 1, \ldots, n. \tag{4.1}$$

Of these steps, the last poses no difficulty once the first three have been carried out. We consider these in more detail in the following three subsections.

Before doing so, we comment on how these estimators are modified in the case where the true mean measure is not a probability measure. In this case, the true measure, say μ, can always be written as $\mu = c\lambda$, where $c = \mu([0, 1])$ and λ is a probability measure. The parameter c can be easily estimated (consistently) by $\widehat{c}_n = \frac{1}{n}\sum_{i=1}^n \widetilde{\Pi}_i([0, 1])$ and the remaining estimators can be constructed by normalising the $\widehat{\Lambda}_i$ to be probability measures (see, e.g., Section 4.2).

4.2. *Estimation of the conditional mean measures.* The probability measures Λ_i can be estimated by various means; here we will employ kernel density estimation. For $\sigma > 0$, let $\psi_\sigma(x) = \sigma^{-1}\psi(x/\sigma)$, with ψ a smooth symmetric probability density function strictly positive throughout the real line and such that $\int x^2\psi(x)\,dx = 1$. Let Ψ be the corresponding distribution function, $\Psi(t) = \int_{-\infty}^t \psi(x)\,dx$.

We consider the following smoothing procedure on a set of points $x_1, \ldots, x_m$. For $y \in [0, 1]$, construct a diffuse probability measure μ_y on $[0, 1]$ with the strictly positive density

$$\psi_\sigma(x - y) + 2b_2\psi_\sigma(x - y)\mathbf{1}\{x > y\} + 2b_1\psi_\sigma(x - y)\mathbf{1}\{x < y\} + 4b_1b_2, \qquad x \in [0, 1],$$

where $b_1 = 1 - \Psi((1 - y)/\sigma)$ and $b_2 = \Psi(-y/\sigma)$. Indeed, integration gives

$$\int_0^1 \psi_\sigma(x - y)\,dx = 1 - b_1 - b_2; \qquad \int_y^1 \psi_\sigma(x - y)\,dx = \frac{1}{2} - b_1;$$

$$\int_0^y \psi_\sigma(x - y)\,dx = \frac{1}{2} - b_2.$$

The intuition behind this construction is the following. First, we smooth the Dirac measure δ_y by the kernel ψ around y, and restrict it to $[0, 1]$; this yields a measure with total mass $1 - b_1 - b_2$. Then we construct the two one-sided versions of

ψ around y with total masses b_1 and b_2, respectively, and again restrict them to $[0, 1]$. The remaining mass, $4b_1b_2$, is distributed uniformly across $[0, 1]$—it does not really matter what we do with this mass, and we could have re-distributed it in any diffuse way. Finally, we construct the estimator

$$\widehat{\Lambda}_i = \frac{1}{m_i}\sum_{j=1}^{m_i} \mu_{x_j}, \qquad m_i = \widetilde{\Pi}_i([0,1]), \tag{4.2}$$
$$(\widehat{\Lambda}_i = \text{Lebesgue measure if } m_i = 0),$$

where the $\{x_j\}_{j=1}^{m_i}$ are the points corresponding to $\widetilde{\Pi}_i$.

Our construction was slightly more complicated than usual in order to: (1) ensure that $\widehat{\Lambda}_i$ is everywhere positive on $[0, 1]$; and, (2) allow us to suitably bound the Wasserstein distance between the smoothed measure and the discrete measure $\widetilde{\Pi}_i/\widetilde{\Pi}_i([0, 1])$. Both these properties will be instrumental in our theoretical results. Indeed, regarding (2), we have the following.

LEMMA 1. *In the notation of the current section, when* $\widetilde{\Pi}_i([0, 1]) > 0$ *and* $\sigma \le 1/4$, *we have the bound*

$$\begin{aligned} &d^2\big(\widehat{\Lambda}_i, \widetilde{\Pi}_i/\widetilde{\Pi}_i([0,1])\big) \\ &\quad\le 3\sigma^2 + 4\max\big(\Psi(-1/\sqrt{\sigma}), 1 - \Psi(1/\sqrt{\sigma})\big). \end{aligned} \tag{4.3}$$

4.3. *Estimation of the structural mean measure.* Given our discussion in Section 3.3, it makes sense to use an M-estimation approach in order to construct an estimator for λ. Since λ arises as a minimum of the population functional $M(\gamma) = \mathbb{E}[d^2(\Lambda, \gamma)]$, with $\Lambda = T_{\#}\lambda$, we would like to define an estimator by minimising the sample functional

$$M_n(\gamma) = \frac{1}{n}\sum_{i=1}^{n} d^2(\Lambda_i, \gamma).$$

Unfortunately, the $\{\Lambda_i\}$ are unobservable, so that they need to be replaced by their estimators (4.2), leading to the proxy functional

$$\widehat{M}_n(\gamma) = \frac{1}{n}\sum_{i=1}^{n} d^2(\widehat{\Lambda}_i, \gamma).$$

If this functional has a unique minimum, then this is the sample Fréchet mean of the $\{\widehat{\Lambda}_i\}$. This type of optimisation problem rarely admits a closed-form solution. Gangbo and Święch [15] have considered this in the form of a multi-coupling problem, and Agueh and Carlier [1] in the barycentric formulation given above. They provide general results on existence and uniqueness (not restricted to the 1-dimensional case), and characterising equations. Remarkably, in the 1-dimensional

case, these yield an explicit solution. This can also be determined directly, using elementary arguments: by our assumption on $\{T_i\}$ being homeomorphisms and λ being diffuse, we know that the measures $\{\Lambda_i\}$ are diffuse measures supported on $[0, 1]$ with probability 1. It follows that (see, e.g., Villani [39], Theorem 2.18)

$$\widehat{M}_n(\gamma) = \frac{1}{n}\sum_{i=1}^{n} d^2(\widehat{\Lambda}_i, \gamma) = \frac{1}{n}\sum_{i=1}^{n}\int_0^1 |\widehat{F}_i^{-1}(x) - F_\gamma^{-1}(x)|^2\,dx$$
$$= \frac{1}{n}\sum_{i=1}^{n}\|\widehat{F}_i^{-1} - F_\gamma^{-1}\|_{L^2}^2,$$

with $\|\cdot\|_{L^2}$ the usual norm on $L^2[0, 1]$. Therefore, if there exists an optimum of

$$\widehat{L}_n(Q) = \frac{1}{n}\sum_{i=1}^{n}\|\widehat{F}_i^{-1} - Q\|_{L^2}^2$$

and this optimum is a valid quantile function, it must be that the probability measure corresponding to this quantile function is an optimum of $\widehat{M}_n(\gamma)$. Indeed, $\widehat{L}_n$ does admit a unique minimum $\bar{Q}$ given by the empirical mean of the $\{\widehat{F}_i^{-1}\}$,

$$\bar{Q}(x) = \frac{1}{n}\sum_{i=1}^{n}\widehat{F}_i^{-1}(x).$$

Furthermore, $\bar{Q}$ is non-decreasing and continuous, since each of the $\widehat{F}_i^{-1}$ is so. It is therefore a valid quantile function [clearly $\bar{Q}(0) = 0$ and $\bar{Q}(1) = 1$]. We conclude that $\widehat{M}_n(\gamma)$ attains a unique minimum at the measure

$$\widehat{\lambda}(A) = \int_A \frac{d}{dx}\left(\frac{1}{n}\sum_{i=1}^{n}\widehat{F}_i^{-1}\right)^{-1}(x)\,dx,$$

that is, the probability measure with cumulative distribution function $\widehat{F} = (\frac{1}{n}\sum_{i=1}^{n}\widehat{F}_i^{-1})^{-1}$.

4.4. *Estimation of the registration maps*. Once the conditional mean measures $\{\Lambda_i\}$ and the structural mean measure λ have been estimated, we automatically get the estimators for the warp and registration maps, respectively,

$$\widehat{T}_i^{-1} = \left(\frac{1}{n}\sum_{j=1}^{n}\widehat{F}_j^{-1}\right)\circ\widehat{F}_i \quad \text{and} \quad \widehat{T}_i = (\widehat{T}_i^{-1})^{-1}. \tag{4.4}$$

Note here that if T is the optimal transportation map of μ onto ν, the change of variables formula immediately implies that T^{-1} is the optimal transportation map of ν onto μ.

4.4.1. *Regularity of the optimal maps.* As was foretold in the end of Section 3.2, the estimation of the warp/registration maps did not require additional smoothness constraints (and by means of tuned penalties) on T. Since $\widehat{T}_i^{-1} = (\frac{1}{n}\sum_{j=1}^n \widehat{F}_j^{-1}) \circ \widehat{F}_i$, we immediately note that the estimated maps will be as regular as the estimators of λ and Λ_i are, or equivalently, as smooth as the $\widehat{F}_j$. It follows that the smoothness of the estimated maps will be directly inherited from any smoothness constraints we place on the estimated mean and conditional mean measures, and will not require the addition of any further smoothness penalties.

5. Bias and over-registering. Note that our geometrical framework essentially induces a loss function in the estimation problem for the structural mean,

$$\mathscr{L}(\lambda, \delta) = d^2(\lambda, \delta),$$

where $\delta = \delta(\Lambda_1, \ldots, \Lambda_n)$ is a candidate estimator of λ. Under this loss function, one can consider the class of *unbiased estimators of the structural mean* (in the general sense of Lehmann [25]), that is, estimators $\delta = \delta(\Lambda_1, \ldots, \Lambda_n)$ satisfying

$$\mathbb{E}_\lambda d^2(\lambda, \delta) = \mathbb{E}_\lambda \mathscr{L}(\lambda, \delta) \leq \mathbb{E}_\lambda \mathscr{L}(\gamma, \delta) = \mathbb{E}_\lambda d^2(\gamma, \delta)$$

for all diffuse measures λ and γ on $[0, 1]$. A *biased* estimator $\psi = \psi(\Lambda_1, \ldots, \Lambda_n)$ would be such that for some measure γ,

$$\mathbb{E}_\lambda d^2(\lambda, \psi) > \mathbb{E}_\lambda d^2(\gamma, \psi).$$

Thus, using a biased estimator in order to estimate the warp functions, may (on average) occasionally produce registrations that appear to be "successful" in the sense that the residual phase variation is small; but on the other hand, they would be registering to the wrong reference measure (a bias-variance tradeoff). It would thus appear that *unbiasedness* is a reasonable requirement in this setup, protecting us against overfitting (or "over-registering," to be more precise).

Interestingly, unbiased estimators can be characterised in terms of their quantile functions; in particular, the empirical Fréchet mean of $\{\Lambda_1, \ldots, \Lambda_n\}$ is unbiased.

PROPOSITION 3 (Unbiased estimators). *Let $\Lambda_1, \ldots, \Lambda_n$ be i.i.d. random probability measures on $[0, 1]$ with positive density with respect to Lebesgue measure. Let λ be their (unique) Fréchet mean in the Wasserstein metric. A random measure δ is unbiased for λ if and only if its expected quantile function is the quantile function of λ, that is,*

$$\mathbb{E}F_\delta^{-1}(x) = F_\lambda^{-1}(x) \tag{5.1}$$

for almost any x. In particular, the (unique) empirical Fréchet–Wasserstein mean of $\Lambda_1, \ldots, \Lambda_n$ is an unbiased estimator of λ.

We can thus interpret our regularised Fréchet–Wasserstein estimator $\hat{\lambda}$ as *approximately unbiased*, since it is a proxy for the unobservable empirical Fréchet–Wasserstein mean.

6. Asymptotic theory. We now turn to establishing the consistency of the estimators constructed in the previous section, and the rate of convergence of the estimator of the structural mean. In the functional case, as encapsulated in equation (2.3), one would need to assume that the number of observed curves, n, as well as the number of sampled observations per curve, m, diverge. Similarly, we will need to construct a framework for asymptotics where the number of point processes n, and the number of points per observed (warped) point process, $\int_0^1 \widetilde{\Pi}(dx)$, diverge. To allow for this, we shall assume that the processes $\{\Pi_i\}$ are infinitely divisible.

THEOREM 1 (Consistency). *Let λ be a diffuse probability measure whose support is $[0,1]$, and let $\{\Pi_1^{(n)}, \dots, \Pi_n^{(n)}\}_{n=1}^{\infty}$ be a triangular array of row independent and identically distributed infinitely divisible point processes with mean measure $\tau_n\lambda$, with $\tau_n > 0$ a scalar. Let $\{T_1, \dots, T_n\}$ be independent and identically distributed random homeomorphisms on $[0,1]$, stochastically independent of $\{\Pi_i^{(n)}\}$, and satisfying assumptions* (B1) *and* (B2) *relative to λ. Let $\widetilde{\Pi}_i^{(n)} = T_{i\#}\Pi_i^{(n)}$, and $\Lambda_i = T_{i\#}\lambda = \tau_n^{-1}\mathbb{E}\{\widetilde{\Pi}_i^{(n)}|T_i\}$. (We shall suppress the dependency on n, but we notice that, by construction, Λ_i does not depend on n.) If $\sigma_n \to 0$ and $\tau_n/\log n \to \infty$ as $n \uparrow \infty$, then:*

1. *The conditional mean measure estimators of Section 4.2 (constructed with bandwidth $\sigma = \sigma_n$) are Wasserstein-consistent,*

$$d(\widehat{\Lambda}_i, \Lambda_i) \xrightarrow{p} 0 \qquad \text{as } n \uparrow \infty, \forall i.$$

2. *The regularised Fréchet–Wasserstein estimator of the structural mean measure (as described in Section 4.3) is strongly Wasserstein-consistent,*

$$d(\widehat{\lambda}, \lambda) \xrightarrow{a.s.} 0 \qquad \text{as } n \uparrow \infty.$$

3. *The warp functions and registration maps estimators of Section 4.4 are uniformly consistent,*

$$\sup_{x\in[0,1]} \left|\widehat{T}_i(x) - T_i(x)\right| \xrightarrow{p} 0 \quad \text{and} \quad \sup_{x\in[0,1]} \left|\widehat{T}_i^{-1}(x) - T_i^{-1}(x)\right| \xrightarrow{p} 0$$

$$\text{as } n \uparrow \infty, \forall i.$$

4. *The registration procedure in equation (4.1) is Wasserstein-consistent,*

$$d\left(\frac{\widehat{\Pi}_i}{\widehat{\Pi}_i([0,1])}, \frac{\Pi_i}{\Pi_i([0,1])}\right) \xrightarrow{p} 0 \qquad \text{as } n \uparrow \infty, \forall i.$$

Under the additional conditions that $\sum_{n=1}^{\infty} \tau_n^{-2} < \infty$ and $\mathbb{E}[\Pi_1^{(1)}([0,1])]^4 < \infty$, the convergence in (1), (3) *and* (4) *holds almost surely.*

REMARK 1. The assumption that $\tau_n / \log n \to \infty$ is only needed in order to avoid empty point processes. It requires that the number of observed processes should not grow too rapidly relative to the mean number of points observed per process. This condition can be compared to similar conditions relating the number of discrete observations per curve in classical FDA. In a sense, it separates the so-called sparse from the dense sampling regime (see also Wu, Müller and Zhang [42]) and shows that even sparse designs lead to consistency. Notice that no assumption on the precise rate of convergence of σ_n to 0 is required, and in particular its decay is independent of τ_n. Indeed, σ_n can even be random (e.g., sample dependent), provided it converges to zero in probability (see also Remark 6).

REMARK 2. Any (cluster) Poisson process is infinitely divisible, so that this assumption is not overly restrictive, and allows for the phase varying point process to be of Cox type, as discussed in Section 3.5 (as a matter of fact, a point process is infinitely divisible if and only if its finite dimensional distributions are infinitely divisible; see Daley and Vere-Jones [13], Section 10.2, for a detailed discussion). It allows us to mathematically translate the increasing expected number of points per process, to a sort of "i.i.d." sampling framework more similar to the classical FDA one.

REMARK 3. In conclusion (4), the random quantity $\widehat{\Pi}_i([0,1]) = \Pi_i([0,1])$ is the number of points observed for the ith process. Normalisation by this factor is a technicality ensuring that the quantities involved are probability measures (or else the Wasserstein distance would not be well-defined). The actual distance $d(\frac{\widehat{\Pi}_i}{\widehat{\Pi}_i([0,1])}, \frac{\Pi_i}{\Pi_i([0,1])})$ only depends on the point patterns themselves, and not on the normalisation.

In the case of Cox processes, when the processes are Poisson prior to warping, if we impose a mild constraint on the decay rate of σ_n, we can also establish rates of convergence of the estimator $\widehat{\lambda}_n$ of the structural mean measure λ.

THEOREM 2 (Rate of convergence). *Assume the conditions of Theorem 1, and suppose in addition that the processes* $\{\Pi_1^{(n)}, \ldots, \Pi_n^{(n)}\}_{n=1}^{\infty}$ *are Poisson. If the kernel* Ψ *used for the smoothing has a finite fourth moment* $\int_{-\infty}^{\infty} x^4 \, d\Psi(x) < \infty$, *then* $\widehat{\lambda}_n$ *satisfies*

$$d(\widehat{\lambda}_n, \lambda) \leq O_{\mathbb{P}}\left(\frac{1}{\sqrt{n}}\right) + O_{\mathbb{P}}\left(\frac{1}{\sqrt[4]{\tau_n}}\right) + O_{\mathbb{P}}\left(\frac{1}{n}\sum_{i=1}^{n} \sigma_i^{(n)}\right).$$

Here, $\sigma_i^{(n)}$ *is the bandwidth used for constructing* $\widehat{\Lambda}_i$, *and it is assumed that* $\sigma_n = \max_{1 \leq i \leq n} \sigma_i^{(n)} \to 0$ *in probability.*

REMARK 4. The first term corresponds to the phase variation, the standard $\sqrt{n}$ rate resulting from the approximation of a theoretical expectation by a sample mean. The second term corresponds to the amplitude variation. The third term corresponds to the bias incurred by the smoothing.

Theorem 2 allows us to conclude that for $\tau_n \geq O(n^2)$ and $\max_{1\leq i\leq n} \sigma_i^{(n)} \leq O_{\mathbb{P}}(n^{-1/2})$ we have $\sqrt{n}$-consistency when dealing with Cox processes, attaining the optimal rate under dense sampling. Indeed, even more can be said in the dense sampling regime:

THEOREM 3 (Asymptotic normality). *In addition to the conditions of Theorem 2, assume that* $\tau_n/n^2 \to \infty$, $\max_{1\leq i\leq n} \sigma_i^{(n)} = o_{\mathbb{P}}(n^{-1/2})$ *and that the density of* λ *is bounded below by a strictly positive constant. Then* $\widehat{\lambda}_n$ *is asymptotically Gaussian, in the sense that*

$$\sqrt{n}(S_n - \mathrm{id}) \xrightarrow{d} Z \qquad \text{in } L^2([0,1]),$$

where S_n *is the optimal transport map from* λ *to* $\widehat{\lambda}_n$, $\mathrm{id} : [0,1] \to [0,1]$ *is the identity map and* Z *is a mean-square continuous Gaussian process with covariance kernel*

$$\kappa(x,y) = \mathrm{cov}\{T(x), T(y)\},$$

for T *a random warp map distributed as the* $\{T_1, \ldots, T_n\}$.

REMARK 5 (Uncertainty quantification). Since we have uniformly consistent estimators of the maps $\{T_1, \ldots, T_n\}$, we can construct an empirical estimate of $\mathrm{cov}\{T(x), T(y)\}$, which would allow us to carry out uncertainty quantification on our structural mean estimate (for example in the form of pointwise confidence intervals of its CDF).

REMARK 6. The statements allow the bandwidth $\sigma_i^{(n)}$ to be random. It follows from Lemma 3 that the (minimal) number of points is of the order $O(\tau_n)$. Consequently, if one chooses the bandwidth by $\sigma_i^{(n)} = \Pi_i^{(n)}([0,1])^{-\alpha}$ for some $\alpha > 0$, then with probability one, $\sigma_n = \max_{1\leq i\leq n} \sigma_i^{(n)} \leq O(\tau_n^{-\alpha})$. The condition $\sigma_n = o_{\mathbb{P}}(n^{-1/2})$ then translates to $\tau_n/n^{1/2\alpha} \to \infty$, which automatically holds for $\alpha \geq 1/4$ due to the independent assumption that $\tau_n/n^2 \to \infty$. Under Rosenblatt's rule $\alpha = 1/5$, one needs the stronger requirement $\tau_n/n^{5/2} \to \infty$ for asymptotic normality to hold.

7. Proofs of formal statements.

PROOF OF PROPOSITION 1. We begin by showing that conditions (A2) and (B2) are equivalent in their own right. Then we will show that subject to (B2) being true, conditions (A1) and (B1) are equivalent. In the language of optimal transportation, condition (B2) requires that T should be the optimal transport map between the diffuse measure λ and $T_{\#}\lambda$. By Brenier's theorem ([39], Theorem 2.12), it must be that T is monotone increasing (as the gradient of a convex function on [0, 1]), and thus (A2) is implied. Conversely, assume that (A2) holds true. We know that there is a unique optimal map between λ and $T_{\#}\lambda$ by λ being diffuse. By Brenier's theorem, this map must be monotone increasing, and hence it must be T itself. This implies (B2).

Consider now condition (B1), which stipulates that given γ a diffuse measure with everywhere positive density [0, 1], we have

$$\mathbb{E}\{d^2(T_{\#}\lambda, \lambda)\} \leq \mathbb{E}\{d^2(T_{\#}\lambda, \gamma)\}.$$

In the presence of (B2), we know that T is an optimal map. It follows that the left-hand side is

$$d^2(T_{\#}\lambda, \lambda) = \int |T(x) - x|^2 \, d\lambda.$$

Keeping this in mind, we focus on the right-hand side. Since γ is absolutely continuous, it can be written as $Q_{\#}\lambda$, for some monotone increasing function Q, and in fact Q is the optimal plan between λ and γ (since any two diffuse measures have a unique optimal map, which must be monotone increasing). It follows that

$$d^2(T_{\#}\lambda, \gamma) = d^2(T_{\#}\lambda, Q_{\#}\lambda) = \int |F^{-1}_{T_{\#}\lambda}(x) - F^{-1}_{Q_{\#}\lambda}(x)|^2 \, dx.$$

Now we note that $F_{T_{\#}\lambda}(x) = F_\lambda(T^{-1}(x))$, since T is increasing, and thus $F^{-1}_{T_{\#}\lambda}(x) = T(F^{-1}_\lambda(x))$; similarly, Q is increasing too, so $F^{-1}_{Q_{\#}\lambda}(x) = Q(F^{-1}_\lambda(x))$. Consequently,

$$\begin{aligned} d^2(T_{\#}\lambda, Q_{\#}\lambda) &= \int |F^{-1}_{T_{\#}\lambda}(x) - F^{-1}_{Q_{\#}\lambda}(x)|^2 \, dx = \int |T(F^{-1}_\lambda(x)) - Q(F^{-1}_\lambda(x))|^2 \, dx \\ &= \int |T(F^{-1}_\lambda(x)) - Q(F^{-1}_\lambda(x))|^2 \frac{f_\lambda(F^{-1}_\lambda(x))}{f_\lambda(F^{-1}_\lambda(x))} \, dx, \end{aligned}$$

where f_λ is the density of λ, which we assumed earlier to be positive everywhere on [0, 1]. Now we change variables, setting $y = F^{-1}_\lambda(x)$, and observing that $dx = f_\lambda(y)\, dy$, we have

$$d^2(T_{\#}\lambda, Q_{\#}\lambda) = \int |T(y) - Q(y)|^2 f_\lambda(y)\, dy = \int |T(y) - Q(y)|^2 \lambda(dy).$$

As a result of our calculations, we see that, in the presence of (B2), condition (B1) is equivalent to

$$\mathbb{E}\int |T(x)-x|^2\lambda(dx) \le \mathbb{E}\int |T(x)-Q(x)|^2\lambda(dx) = \int \mathbb{E}|T(x)-Q(x)|^2\lambda(dx),$$

for all monotone increasing functions Q, where the last equality follows from Tonelli's theorem. The last condition is satisfied if and only if $\mathbb{E}[T(x)] = x$, λ-almost everywhere. Thus, when λ has positive density with respect to Lebesgue measure everywhere on $[0,1]$, we have established that, if (B2) holds, then (A1) is equivalent to (B1). This completes the proof. $\square$

PROOF OF PROPOSITION 2. Since Λ is diffuse and strictly positive, we may re-express the functional of interest as

$$M(\gamma) = \mathbb{E}\big[d^2(\Lambda,\gamma)\big] = \mathbb{E}\bigg[\int_0^1 |F_\Lambda^{-1}(x) - F_\gamma^{-1}(x)|^2\,dx\bigg] = \mathbb{E}\|F_\Lambda^{-1} - F_\gamma^{-1}\|_{L^2}^2,$$

with $\|\cdot\|_{L^2}$ the usual L^2 norm. Therefore, if there exists an optimum of

$$L(Q) = \mathbb{E}\|F_\Lambda^{-1} - Q\|_{L^2}^2, \qquad Q \in L_2\big([0,1]\big)$$

and this optimum is a valid quantile function, it must be that the probability measure corresponding to this quantile function is an optimum of $M(\gamma)$. Indeed, L does admit a unique minimum given by $\Gamma(x) = \mathbb{E}[F_\Lambda^{-1}(x)]$, $x \in [0,1]$, which we claim is a valid quantile function. Note first that F_Λ^{-1} is, in fact, a proper inverse of the continuous, strictly increasing mapping $F_\Lambda(x) = \Lambda([0,x])$.

1. Since $F_\Lambda^{-1}(0) = 0$ and $F_\Lambda^{-1}(1) = 1$ almost surely, we have $\Gamma(0) = 0$ and $\Gamma(1) = 1$.

2. If $x \le y$, then $F_\Lambda^{-1}(x) \le F_\Lambda^{-1}(y)$ almost surely. Consequently, $\mathbb{E}[F_\Lambda^{-1}(x)] \le \mathbb{E}[F_\Lambda^{-1}(y)]$ also, proving that Γ is non-decreasing.

3. If $x_k \to x$ in $[0,1]$, then $X_k = F_\Lambda^{-1}(x_k) \to F_\Lambda^{-1}(x) = X$ almost surely. Since $|X_k|$ is bounded by 1, the bounded convergence theorem implies that $\mathbb{E}[X_k] \to \mathbb{E}[X]$, proving that $\Gamma(x)$ is continuous at x (and hence everywhere in $[0,1]$ by arbitrary choice of x). $\square$

PROOF OF PROPOSITION 3. Requiring an estimator ψ to be unbiased translates to

$$\mathbb{E}_\lambda\|F_\lambda^{-1} - F_\psi^{-1}\|_{L^2}^2 \le \mathbb{E}_\lambda\|F_\gamma^{-1} - F_\psi^{-1}\|_{L^2}^2.$$

Since L^2 is a linear space, and using Tonelli's theorem to exchange expectation and integration, the unbiasedness condition is equivalent to requiring that

$$\mathbb{E}_\lambda\big[F_\psi^{-1}(x)\big] = F_\lambda^{-1}(x) \qquad \text{almost everywhere.}$$

To show that this is indeed the case for the empirical Wasserstein mean δ, we note that

$$F_{\Lambda_i}^{-1} = F_{(T_i)\#\lambda}^{-1} = \big(F_\lambda \circ T_i^{-1}\big)^{-1} = T_i \circ F_\lambda^{-1},$$

and so, by Proposition 1, it follows that

$$\mathbb{E}_\lambda\big[F_{\Lambda_i}^{-1}(x)\big] = \mathbb{E}_\lambda\big[T_i\big(F_\lambda^{-1}(x)\big)\big] = F_\lambda^{-1}(x), \qquad i = 1, \ldots, n$$

almost everywhere on $[0, 1]$. Since $F_\delta^{-1}(x) = n^{-1}\sum F_{\Lambda_i}^{-1}(x)$ (see Section 4.3), $\mathbb{E}_\lambda[F_\delta^{-1}(x)] = F_\lambda^{-1}(x)$ also holds a.e., and the unbiasedness of δ has been established. $\square$

PROOF OF LEMMA 1. The squared Wasserstein distance is bounded by the cost of sending all the mass in μ_{x_i} to x_i. The squared distance between μ_y and δ_y is

$$\begin{aligned}
&\int_0^1 (x-y)^2\psi_\sigma(x-y)\,dx + 2b_1\int_y^1 (x-y)^2\psi_\sigma(x-y)\,dx \\
&\quad + 2b_2\int_0^y (x-y)^2\psi_\sigma(x-y)\,dx + 4b_1b_2\int_0^1 (x-y)^2\,dx \\
&\quad \le (1+2b_1+2b_2)\int_0^1 (x-y)^2\psi_\sigma(x-y)\,dx + 4b_1b_2 \\
&\quad \le (1+2b_1+2b_2)\int_{\mathbb{R}} (x-y)^2\psi_\sigma(x-y)\,dx + 4b_1b_2 \\
&\quad \le 3\int_{\mathbb{R}} x^2\psi_\sigma(x)\,dx + 4b_1b_2 = 3\sigma^2 + 4b_1b_2 \qquad (\text{since } b_1+b_2 \le 1).
\end{aligned}$$

The reason we needed the one-sided kernels in addition to the standard two-sided one is that either b_1 or b_2 can be large (e.g., if $y = 0$, then $b_2 = 1/2$), but they cannot both be large simultaneously. Indeed, when $y \ge \sqrt{\sigma}$, we have $b_2 \le \Psi(-1/\sqrt{\sigma})$ and when $1 - y \ge \sqrt{\sigma}$, $b_1 \le 1 - \Psi(1/\sqrt{\sigma})$. When $\sigma \le 1/4$, at least one of these possibilities holds, and since $0 \le b_i \le 1$, this implies that

$$b_1b_2 \le \max\big(\Psi(-1/\sqrt{\sigma}), 1 - \Psi(1/\sqrt{\sigma})\big).$$

This bound holds for any $y \in [0, 1]$, and the conclusion follows. $\square$

In order to prove Theorem 1, we first need to eliminate the possibility of having empty point processes (this is the only reason we assume $\tau_n/\log n \to \infty$). To this aim, we will use a seemingly unrelated technical result for binomial distributions.

LEMMA 2 (Chernoff bound for binomial distributions). *Let $N \sim B(\tau, q)$, then*

$$\mathbb{P}(N \le \tau q/2) \le \beta^\tau, \qquad \beta = \beta(q) = 2\big((1-q)/(2-q)\big)^{1-q/2} < 1.$$

PROOF. For any $t \geq 0$, we have

$$\mathbb{P}\left(N \leq \frac{\tau q}{2}\right) = \mathbb{P}\left(\exp(-Nt) \geq \exp\left(-t\frac{\tau q}{2}\right)\right) \leq \mathbb{E}\exp(-Nt)\exp\left(t\frac{\tau q}{2}\right)$$
$$= \left[s^{q/2}\left(1 - q + \frac{q}{s}\right)\right]^{\tau},$$

where $s = e^t \geq 1$. A straightforward calculation shows that this is minimised when $s = (2-q)/(1-q) > 1$. The objective value at this point, β, must be smaller than the objective value at $s = 1$, which is 1. $\square$

LEMMA 3 [Number of points per process is $O(\tau_n)$]. *If $\tau_n / \log n \to \infty$, then there exists a constant $C_\Pi > 0$, depending only on the distribution of the Π's, such that*

$$\liminf_{n\to\infty} \frac{\min_{1\leq i\leq n} \Pi_i^{(n)}([0,1])}{\tau_n} \geq C_\Pi \qquad a.s.$$

In particular, there are no empty point processes, so the normalisation is well-defined.

PROOF. Let us denote for simplicity by Π_τ ($\tau > 0$) a point process that follows the same infinitely divisible distribution as $\Pi_i^{(n)}$, but with mean measure $\tau\lambda$. Let p be the probability that Π_1 has no points (clearly, $p < 1$, since Π_1 has one point in average). It follows from the infinite divisibility that for any rational τ, the probability that Π_τ has no points is p^τ. By a continuity argument, this can be extended to any real value of τ: indeed, the Laplace functional of Π_1 takes the form (Kallenberg [21], Chapter 6)

$$L_1(f) = \mathbb{E}\left[e^{-\Pi_1 f}\right] = \exp\left(-\int \left(1 - e^{-\rho f}\right) d\mu(\rho)\right), \qquad f \in F([0,1]),$$

where $F[0,1]$ is the set of Borel measurable functions $f : [0,1] \to \mathbb{R}_+$, and μ is a Radon measure on the set $P([0,1])$. It follows that $L_\tau(f)$, the Laplace functional of Π_τ, is $(L_1(f))^\tau$ when τ is rational, which simply corresponds to multiplying μ by the scalar τ. By considering the measure $\tau\mu$ for any real τ, we obtain $L_\tau(f) = (L_1(f))^\tau$ for any value of τ. The Laplace functional completely determines the distribution of the process; in particular, the probability of Π_τ having no points is obtained as the limit

$$\lim_{m\to\infty} L_\tau(m) = \lim_{m\to\infty} \left(L_1(m)\right)^\tau = p^\tau,$$

by the bounded convergence theorem, where $L_\tau(m) = L_\tau(f)$ for the constant function $f \equiv m$.

Denote the total number of points by $N_i^{(n)} = \Pi_i^{(n)}([0,1])$, and assume momentarily that the τ_n's are integers. Then $N_i^{(n)}$ is the sum of τ_n i.i.d. integer valued random variables X_i, each having a probability of $p < 1$ to equal zero. (In the Poisson

case, $p = e^{-1}$.) Each X_i is larger than $\mathbf{1}\{X_i \geq 1\}$, which follows a Bernoulli distribution with parameter $q = 1 - p$, and $N_i^{(n)} = \sum X_i \geq \sum \mathbf{1}\{X_i \geq 1\}$. It follows that for any m,

$$\mathbb{P}\big(N_i^{(n)} \leq m\big) \leq \mathbb{P}\big(B(\tau_n, q) \leq m\big).$$

Since $N_i^{(n)}$ are i.i.d. across i, specifying $m = \tau_n q/2$ and applying Lemma 2 yields

$$\mathbb{P}\Big(\min_{1\leq i\leq n} N_i^{(n)} \leq \frac{\tau_n q}{2}\Big) = 1 - \Big[1 - \mathbb{P}\Big(N_1^{(n)} \leq \frac{\tau_n q}{2}\Big)\Big]^n \leq 1 - \big(1 - \beta^{\tau_n}\big)^n$$
$$\leq 1 - \big(1 - n\beta^{\tau_n}\big),$$

by the Bernoulli inequality $(1-x)^n \geq 1 - nx$ (valid for $x \leq 1$ and n integer; easily proved by induction on n). The right-hand side is n^{a+1} for $a = (\log\beta)\tau_n/\log n$. Since $\tau_n/\log n \to \infty$ and $\beta < 1$, we have $a \to -\infty$ as $n \to \infty$ so this is smaller than n^{-2} for sufficiently large n. By the Borel–Cantelli lemma, the result holds for $C_\Pi = q/2$.

If τ_n is not an integer, then $N_i^{(n)}$ is the sum of $\lfloor\tau_n\rfloor$ (the largest integer $\leq \tau_n$) i.i.d. random variables X_i with probability $p' = p^{\tau_n/\lfloor\tau_n\rfloor} \leq p$ to equal zero. Letting $q' = 1 - p' \geq q$ and observing that $\mathbb{P}(B(k, q') \leq m) \leq \mathbb{P}(B(k, q) \leq m)$ for any k and any m [or that $\beta(q') \leq \beta(q)$], we obtain

$$\mathbb{P}\Big(\min_{1\leq i\leq n} N_i^{(n)} \leq \frac{\lfloor\tau_n\rfloor q}{2}\Big) \leq \mathbb{P}\Big(\min_{1\leq i\leq n} N_i^{(n)} \leq \frac{\lfloor\tau_n\rfloor q'}{2}\Big) \leq n\beta^{\lfloor\tau_n\rfloor} = n^{a+1},$$
$$a = \log\beta\,\frac{\lfloor\tau_n\rfloor}{\log n}.$$

We still have $a \to -\infty$ and since $\tau_n/[\tau_n] \to 1$, any $C_\Pi < q/2$ will qualify. Thus, the lemma holds with $C_\Pi = q/2$. □

REMARK 7. As the proof shows, the condition $\tau_n/\log n \to \infty$ can be slightly weakened to

$$\liminf_{n\to\infty}(\tau_n/\log n) > 2/-\log\beta$$

and the lower bound equals 9.75 in the Poisson case.

PROOF OF THEOREM 1. Maintaining the notation $N_i = N_i^{(n)} = \Pi_i([0,1]) = \widetilde{\Pi}_i([0,1])$, we begin by proving (1). Without loss of generality, assume that τ_n takes integer values [otherwise, work with t_n, the greatest integer smaller than τ_n, that is, replace τ_n by t_n and Λ_i by $(\tau_n/t_n)\Lambda_i$]. Let i be a fixed integer. Since the processes $\{\Pi_i\}$ are infinitely divisible, it is clear that the $\{\widetilde{\Pi}_i\}$ must be so too. Consequently, we note that a single realisation of a point process with mean measure $\tau_n\Lambda_i$ is equivalent in law to a superposition of τ_n independent and identically distributed processes $\{P_j^{(n)}\}_{j=1}^{\tau_n}$, each with mean Λ_i. We can assume that $P_j^{(n)}$ are

constructed as the push-forward through T_i of independent and identically distributed point processes $Q_j^{(n)}$ with mean measure λ, that are independent of T_i. It follows that as $n \to \infty$, (e.g., Karr [22], Chapter 4)

$$\frac{1}{\tau_n}\widetilde{\Pi}_i \stackrel{d}{=} \frac{1}{\tau_n}\sum_{j=1}^{\tau_n} P_j^{(n)} \stackrel{w}{\to} \Lambda_i \qquad \text{in probability,}$$

with "$\stackrel{w}{\to}$" denoting weak convergence of measures. Since $N_i/\tau_n \stackrel{p}{\to} 1$, it follows by Slutsky's theorem that

$$\widetilde{\Pi}_i/N_i \stackrel{w}{\to} \Lambda_i \qquad \text{in probability.} \tag{7.1}$$

As $[0,1]$ is compact, we conclude that this last convergence also holds in Wasserstein distance [39], Theorem 7.12, in probability. Noting that by (4.3) and since $\sigma_n \to 0$ as $n \to \infty$,

$$\sup_{\Omega} d(\widehat{\Lambda}_i, \widetilde{\Pi}_i/N_i) \to 0, \qquad n \to \infty,$$

an application of the triangle inequality shows that $d(\widehat{\Lambda}_i, \Lambda_i) \stackrel{p}{\to} 0$, establishing claim (1). For convergence almost surely, we fix $a \in [0,1]$ and set

$$S_n = \sum_{j=1}^{\tau_n} X_{nj}, \qquad X_{nj} = P_j^{(n)}([0,a]) - \Lambda_i([0,a]), \qquad j = 1, \ldots, \tau_n.$$

One sees that $S_n^4 = \varphi(Q_1^{(n)}, \ldots, Q_k^{(n)}, T_i)$, where $k = \tau_n$ and

$$\varphi(q_1, \ldots, q_k, f) = \left[\sum_{j=1}^{k} f_{\#}q_j([0,a]) - f_{\#}\lambda([0,a])\right]^4,$$

$$f \in \mathrm{Hom}[0,1]; q_j \in M_R,$$

(where M_R is the collection of Radon measures on $[0,1]$ endowed with the topology of weak convergence, and $\mathrm{Hom}[0,1]$ is the space of homeomorphisms of $[0,1]$ endowed with the supremum norm) is a measurable function (since it is continuous). It is also integrable because $0 \le f_{\#}\lambda([0,a]) \le 1$ and $\mathbb{E}[T_{i\#}Q_j^{(n)}([0,a])]^4 \le \mathbb{E}[Q_j^{(n)}([0,1])]^4 < \infty$ by the hypothesis.

Since the arguments of φ are independent, the proof of [14], Lemma 6.2.1, can be adapted to show that $\mathbb{E}[S_n^4|\sigma(T_i)] = g(T_i)$, where (with a slight abuse of notation)

$$g(f) = \mathbb{E}_Q[\varphi(Q_1^{(n)}, \ldots, Q_k^{(n)}, f)] = \int dq_1 \int dq_2 \cdots \int dq_k \varphi(q_1, \ldots, q_k, f),$$

$$f \in \mathrm{Hom}[0,1].$$

The same idea shows that for each j,

$$\begin{aligned}\mathbb{E}\big[X_{nj}|\sigma(T_i)\big] &= \int dq_j T_{i\#} q_j\big([0,a]\big) - T_{i\#}\lambda\big([0,a]\big)\\ &= \lambda\big(T_i^{-1}\big([0,a]\big)\big) - \lambda\big(T_i^{-1}\big([0,a]\big)\big) = 0.\end{aligned}$$

In words, conditional on $\sigma(T_i)$, $\{X_{nj}\}_{j=1}^{\tau_n}$ are mean zero independent and identically distributed random variables. One readily verifies that (see the proof of [14], Theorem 2.3.5, for the details)

$$\begin{aligned}\mathbb{E}\big[S_n^4|\sigma(T_i)\big] &= \sum_{j=1}^{\tau_n}\mathbb{E}\big[X_{nj}^4|\sigma(T_i)\big] + \sum_{j<l}\mathbb{E}\big[X_{nj}^2X_{nl}^2|\sigma(T_i)\big]\\ &= \tau_n\mathbb{E}\big[X_{11}^4|\sigma(T_i)\big] + 3\tau_n(\tau_n-1)\mathbb{E}\big[X_{11}^2X_{12}^2|\sigma(T_i)\big].\end{aligned}$$

Taking again expected values and applying Markov's inequality,

$$\mathbb{P}\bigg[\bigg(\frac{S_n}{\tau_n}\bigg)^4 > \varepsilon\bigg] \le \frac{\mathbb{E}[S_n^4]}{\varepsilon^4\tau_n^4} = \frac{\tau_n\mathbb{E}[X_{11}^4] + 3\tau_n(\tau_n-1)\mathbb{E}[X_{11}^2X_{12}^2]}{\varepsilon^4\tau_n^4}.$$

The numerator is finite, and the sum over n of the right-hand side converges when $\sum_n \tau_n^{-2} < \infty$. As ε is arbitrary, $S_n/\tau_n \overset{\text{a.s.}}{\to} 0$ by the Borel–Cantelli lemma.

Repeating this argument countably many times, we have

$$\mathbb{P}\bigg(\frac{\widetilde{\Pi}_i([0,a])}{\tau_n} - \Lambda_i\big([0,a]\big) \to 0 \text{ for any rational } a\bigg) = 1.$$

If a is irrational, choose $a_k \nearrow a \swarrow b_k$ rational. We have the inequalities

$$\begin{aligned}\frac{\widetilde{\Pi}_i([0,a])}{\tau_n} - \Lambda_i\big([0,a]\big) &\le \frac{\widetilde{\Pi}_i([0,b_k])}{\tau_n} - \Lambda_i\big([0,b_k]\big) + \Lambda_i\big([0,b_k]\big) - \Lambda_i\big([0,a]\big);\\ \frac{\widetilde{\Pi}_i([0,a])}{\tau_n} - \Lambda_i\big([0,a]\big) &\ge \frac{\widetilde{\Pi}_i([0,a_k])}{\tau_n} - \Lambda_i\big([0,a_k]\big) + \Lambda_i\big([0,a_k]\big) - \Lambda_i\big([0,a]\big),\end{aligned}$$

from which one concludes that almost surely, for any k,

$$\begin{aligned}-\Lambda_i\big((a_k,a]\big) &\le \liminf_{n\to\infty}\frac{\widetilde{\Pi}_i([0,a])}{\tau_n} - \Lambda_i\big([0,a]\big) \le \limsup_{n\to\infty}\frac{\widetilde{\Pi}_i([0,a])}{\tau_n} - \Lambda_i\big([0,a]\big)\\ &\le \Lambda_i\big((a,b_k]\big).\end{aligned}$$

Letting $k\to\infty$, we see that convergence holds for any continuity point a of Λ_i. But Λ_i is a continuous measure by construction. One then easily shows the almost sure analogue of (7.1) (take $a=1$) and concludes (1) as above.

In order to prove (2), we note that λ being a minimiser of the functional $M(\gamma) = \mathbb{E}[d^2(\Lambda,\gamma)]$ implies that it must be the unique such minimiser (this follows by Proposition 2), since $\Lambda = T_\#\lambda$ is diffuse and everywhere positive on $[0,1]$, and T is

a homeomorphism. To establish the purported convergence, we therefore study the convergence of $\widehat{M}_n(\gamma) = \frac{1}{n}\sum_{i=1}^n d^2(\widehat{\Lambda}_i, \gamma)$ to M, both viewed as being defined over $P([0,1])$, the space of probability measures supported on $[0,1]$. Using the triangle inequality, we may interject the functionals

$$M_n(\gamma) = \frac{1}{n}\sum_{i=1}^n d^2(\Lambda_i, \gamma) \tag{7.2}$$

that is, the empirical functional assuming that the Λ_i could be observed; and

$$M_n^*(\gamma) = \frac{1}{n}\sum_{i=1}^n d^2\Big(\frac{\widetilde{\Pi}_i}{N_i}, \gamma\Big) \tag{7.3}$$

(which is well-defined for n sufficiently large by Lemma 3), and write

$$|\widehat{M}_n(\gamma) - M(\gamma)| \le |\widehat{M}_n(\gamma) - M_n^*(\gamma)| + |M_n^*(\gamma) - M_n(\gamma)| + |M_n(\gamma) - M(\gamma)|.$$

We shall show that each of the three terms in the right-hand side converges to 0 uniformly.

For any three probability measures μ, ν, ρ on $[0,1]$, one has

$$\begin{aligned} d^2(\mu,\nu) &\le \sup_{\theta\in P([0,1]^2)} \int_{[0,1]}\int_{[0,1]} |x-y|^2\theta(dx\times dy) \\ &\le \sup_{x,y\in[0,1]} |x-y|^2 = 1; \end{aligned} \tag{7.4}$$

$$\begin{aligned} |d^2(\mu,\rho) - d^2(\nu,\rho)| &= |d(\mu,\rho)+d(\nu,\rho)||d(\mu,\rho)-d(\nu,\rho)| \\ &\le 2d(\nu,\mu), \end{aligned} \tag{7.5}$$

and consequently

$$|\widehat{M}_n(\gamma) - M_n^*(\gamma)| \le \frac{1}{n}\sum_{i=1}^n \Big| d^2(\widehat{\Lambda}_i,\gamma) - d^2\Big(\frac{\widetilde{\Pi}_i}{N_i},\gamma\Big)\Big| \le \frac{2}{n}\sum_{i=1}^n d\Big(\widehat{\Lambda}_i, \frac{\widetilde{\Pi}_i}{N_i}\Big).$$

The right-hand side is independent of γ and converges to 0 by application of (4.3).

Similarly,

$$\sup_{\gamma\in P([0,1])} |M_n(\gamma) - M_n^*(\gamma)| \le \frac{2}{n}\sum_{i=1}^n d\Big(\Lambda_i, \frac{\widetilde{\Pi}_i}{N_i}\Big) = \frac{2}{n}\sum_{i=1}^n X_{ni} = 2\overline{X}_n.$$

Now X_{ni} is a function of T_i and $\Pi_i^{(n)}$, so by construction they are i.i.d. across i. Setting $Y_{ni} = X_{ni} - \mathbb{E}X_{ni}$, we obtain mean zero random variables that are i.i.d. across i and $|Y_{ni}| \le 1$ because $0 \le X_{ni} \le 1$ by (7.4). Applying the argument in [14], Theorem 2.3.5, again, one obtains

$$\mathbb{P}\big((\overline{X}_n - \mathbb{E}\overline{X}_n)^4 > \varepsilon\big) = \mathbb{P}\big(\overline{Y}_n^4 > \varepsilon\big) \le \frac{n\mathbb{E}[Y_{ni}^4] + 3n(n-1)\mathbb{E}[Y_{ni}^2]}{\varepsilon^4 n^4} \le \frac{3}{\varepsilon^4 n^2}.$$

By the Borel–Cantelli lemma and arbitrariness of $\varepsilon > 0$, we have $|\overline{X}_n - \mathbb{E}\overline{X}_n| \xrightarrow{\text{a.s.}} 0$. But $X_{n1} \xrightarrow{p} 0$ as $n \to \infty$ by (7.1), and the bounded convergence theorem yields $\mathbb{E}[\overline{X}_n] = \mathbb{E}[X_{n1}] \to 0$.

Turning to the term $|M_n(\gamma) - M(\gamma)|$, we remark that the strong law of large numbers yields

$$M_n(\gamma) \xrightarrow{\text{a.s.}} M(\gamma),$$

for all γ. To upgrade to uniform convergence over γ, observe that by (7.5), both M_n and M are 2-Lipschitz. By compactness of $P([0,1])$, given $\varepsilon > 0$, we can choose an ε-cover $\gamma_1, \ldots, \gamma_k$. For any γ, we have $d(\gamma, \gamma_j) < \varepsilon$ for some j, so

$$\begin{aligned}\big|M_n(\gamma) - M(\gamma)\big| &\le \big|M_n(\gamma_j) - M_n(\gamma)\big| + \big|M_n(\gamma_j) - M(\gamma_j)\big| + \big|M(\gamma_j) - M(\gamma)\big| \\ &\le 4d(\gamma, \gamma_j) + \big|M_n(\gamma_j) - M(\gamma_j)\big| \\ &\le 4\varepsilon + \big|M_n(\gamma_j) - M(\gamma_j)\big|.\end{aligned}$$

Taking $n \to \infty$, then $\varepsilon \to 0$, we conclude

$$\sup_{\gamma} \big|M_n(\gamma) - M(\gamma)\big| \xrightarrow{\text{a.s.}} 0, \qquad n \to \infty.$$

Summarising, we have established that $\sup_\gamma |\widehat{M}_n(\gamma) - M(\gamma)| \xrightarrow{\text{a.s.}} 0$. Let λ_n be a minimiser of $\widehat{M}_n$. By compactness of $P([0,1])$, $\lambda_{n_k} \to \mu$, for some subsequence and some μ. Then $\widehat{M}_{n_k}(\widehat{\lambda}_{n_k}) \to M(\mu)$ by the uniform convergence and continuity of $\widehat{M}_n$ and M. Since $\widehat{M}_{n_k}(\widehat{\lambda}_{n_k}) \le \widehat{M}_{n_k}(\lambda) \to M(\lambda)$, we get $M(\mu) \le M(\lambda)$, which, by uniqueness of λ as a minimiser of M, implies that $\mu = \lambda$. This establishes $\widehat{\lambda}_n \xrightarrow{\text{a.s.}} \lambda$ with respect to the Wasserstein distance.

To prove part (3), let F, G, F_n and G_n denote the distribution functions of λ, Λ_i, $\widehat{\lambda}_n$ and $\widehat{\Lambda}_i$, respectively, restricted to $[0,1]$. Since F and G are continuous functions, we have $F_n \to F$ and $G_n \to G$ pointwise on $[0,1]$ (either in probability or almost surely, depending on the assumptions). Furthermore, all these functions are strictly increasing and continuous, thus invertible. Our goal is to show

$$G_n^{-1} \circ F_n = \widehat{T}_i \to T_i = G^{-1} \circ F \qquad \text{uniformly on } [0,1].$$

Lemma 4 below shows that it will suffice to establish pointwise convergence, as uniform convergence will immediately follow in our current setup. To this aim, we remark that since G is continuous on a compact set, it maps closed sets to closed sets. Being a bijection, this implies that G^{-1} is continuous as well.

We proceed by showing that $G_n^{-1}(t) \to G^{-1}(t)$ for $0 < t < 1$ (this is obvious when $t \in \{0,1\}$). Let x be the unique number such that $G(x) = t$ and let $\varepsilon > 0$. Then $G_n(x + \varepsilon) \to G(x + \varepsilon) > t$ so that $x + \varepsilon \ge G_n^{-1}(t)$, at least for n large. Similarly, $x - \varepsilon \le G_n^{-1}(t)$ for n large and, ε being arbitrary, we conclude that $G_n^{-1}(t) \to x = G^{-1}(t)$.

By Lemma 4, G_n^{-1} converges uniformly to G^{-1} on $[0,1]$, where the latter is (uniformly) continuous. Given $\varepsilon > 0$, let δ such that $|t-s| \le \delta \Rightarrow |G^{-1}(t) - G^{-1}(s)| \le \varepsilon$. When n is large, $\|F_n - F\|_\infty \le \delta$ and $\|G_n^{-1} - G^{-1}\|_\infty \le \varepsilon$. Then, for any $x \in [0,1]$, $|F_n(x) - F(x)| < \delta$, whence

$$G_n^{-1}\big(F_n(x)\big) \le G_n^{-1}\big(F(x)+\delta\big) \le G^{-1}\big(F(x)+\delta\big) + \varepsilon \le G^{-1}\big(F(x)\big) + 2\varepsilon;$$

$$G_n^{-1}\big(F_n(x)\big) \ge G_n^{-1}\big(F(x)-\delta\big) \ge G^{-1}\big(F(x)-\delta\big) - \varepsilon \ge G^{-1}\big(F(x)\big) - 2\varepsilon.$$

In other words, $\|\widehat{T}_i - T_i\|_\infty \le 2\varepsilon$ for any large enough n, and (3) is proven. Since the functions $\widehat{T}_i$ and T_i are again strictly increasing, it also follows that $\widehat{T}_i^{-1}$ converges to T_i^{-1} uniformly.

Now, we turn to part (4). Recall that

$$\widehat{\Pi}_i = \widehat{T}_i^{-1}{}_{\#}\widetilde{\Pi}_i = \big(\widehat{T}_i^{-1} \circ T_i\big)_{\#}\Pi_i, \qquad i = 1, \dots, n.$$

It follows that $\widehat{T}_i^{-1} \circ T_i$ is a transport plan of Π_i onto $\widehat{\Pi}_i$. Consequently,

$$d^2\left(\frac{\widehat{\Pi}_i}{N_i}, \frac{\Pi_i}{N_i}\right) \le \int_0^1 \big|\widehat{T}_i^{-1}\big(T_i(x)\big) - x\big|^2 \frac{\Pi_i(dx)}{N_i} \le \sup_{x\in[0,1]} \big|\widehat{T}_i^{-1}\big(T_i(x)\big) - x\big|^2.$$

Note, however, that since $T_i \in \mathrm{Hom}[0,1]$,

$$\begin{aligned}\sup_{x\in[0,1]} \big|\widehat{T}_i^{-1}\big(T_i(x)\big) - x\big| &= \sup_{x\in[0,1]} \big|\widehat{T}_i^{-1}\big(T_i\big(T_i^{-1}(x)\big)\big) - T_i^{-1}(x)\big| \\ &= \sup_{x\in[0,1]} \big|\widehat{T}_i^{-1}(x) - T_i^{-1}(x)\big|,\end{aligned}$$

and the latter converges to zero in probability (or almost surely, depending on the assumptions) as $n \to \infty$ from part (3). □

The following elementary result is stated without proof.

LEMMA 4. *Let $F_n : [a,b] \to \mathbb{R}$ be non-decreasing and converge pointwise to a continuous limit function F. Then the convergence is uniform.*

PROOF OF THEOREM 2. Let λ_n be the minimiser of the empirical functional $M_n(\gamma) = \frac{1}{n}\sum_{i=1}^n d^2(\Lambda_i, \gamma)$. For a probability measure $\theta \in P([0,1])$, denote its quantile function $F_\theta^{-1} \in L^2([0,1])$ by $g(\theta)$. Then [39], Theorem 2.18, says that g is an isometry: $d(\theta,\gamma) = \|g(\theta) - g(\gamma)\|$. Now

$$\sqrt{n}\big(g(\lambda_n) - g(\lambda)\big) = \sqrt{n}\left(\frac{1}{n}\sum_{i=1}^n F_{\Lambda_i}^{-1} - F_\lambda^{-1}\right).$$

These are i.i.d. mean zero random elements in L^2, whose norm is bounded by 1. Therefore, the above expression converges in distribution to a Gaussian limit GP with $\mathbb{E}\|GP\|^2 < \infty$ as $n \to \infty$. In particular,

$$d(\lambda_n, \lambda) = \big\|g(\lambda_n) - g(\lambda)\big\| = O_{\mathbb{P}}\big(n^{-1/2}\big).$$

The error resulting from approximating λ_n by $\widehat{\lambda}_n$, the minimiser of $\widehat{M}_n$, is

$$\begin{aligned}\|g(\lambda_n) - g(\widehat{\lambda}_n)\| &= \left\| \frac{1}{n}\sum_{i=1}^{n} F_{\Lambda_i}^{-1} - \frac{1}{n}\sum_{i=1}^{n} F_{\widehat{\Lambda}_i}^{-1} \right\| \le \frac{1}{n}\sum_{i=1}^{n} \| F_{\Lambda_i}^{-1} - F_{\widehat{\Lambda}_i}^{-1} \| \\ &= \frac{1}{n}\sum_{i=1}^{n} d(\Lambda_i, \widehat{\Lambda}_i),\end{aligned}$$

which, by the triangle inequality, is bounded by

$$\frac{1}{n}\sum_{i=1}^{n} d(\Lambda_i, \widehat{\Lambda}_i) \le \frac{1}{n}\sum_{i=1}^{n} d\Big(\Lambda_i, \frac{\widetilde{\Pi}_i^{(n)}}{N_i^{(n)}}\Big) S_{ni} + \frac{1}{n}\sum_{i=1}^{n} d\Big(\frac{\widetilde{\Pi}_i^{(n)}}{N_i^{(n)}}, \widehat{\Lambda}_i\Big) S_{ni} + \frac{1}{n}\sum_{i=1}^{n} V_{ni},$$

where $S_{ni} = 1 - V_{ni} = \mathbf{1}\{N_i^{(n)} > 0\}$. The first term on the right-hand side corresponds to the amplitude variation, while the second corresponds to the smoothing bias. The third term was introduced to accommodate empty processes. The inequality follows from the convention that $\widehat{\Lambda}_i$ is Lebesgue measure when $N_i^{(n)} = 0$ and the distance between any two measures is no larger than one. This term is negligible by Lemma 3: $\mathbb{P}(\sum V_{ni} = 0) \to 1$ so this term "converges" to 0 at any rate.

Denote the distances of the amplitude variation by $X_{ni} \in [0, 1]$. For fixed n, X_{ni} are i.i.d. across i. Since

$$\mathbb{P}\left(a_n \frac{1}{n}\sum_{i=1}^{n} X_{ni} > \varepsilon \right) \le \frac{a_n \mathbb{E}\sum_{i=1}^{n} X_{ni}}{n\varepsilon} = \frac{a_n \mathbb{E} X_{n1}}{\varepsilon},$$

we seek to find the rate at which $\mathbb{E}X_{n1}$ vanishes. Let W_1 denote the 1-Wasserstein distance. Then equations (7.4) and (2.48) in Villani [39] and Fubini's theorem imply that

$$\begin{aligned}\mathbb{E}X_{n1}^2 \le \mathbb{E}S_{n1} W_1\Big(\Lambda_1, \frac{\widetilde{\Pi}_1^{(n)}}{\widetilde{\Pi}_1^{(n)}([0,1])}\Big) &= \int_0^1 \mathbb{E}\Big| \Lambda_1([0,t]) - \frac{\widetilde{\Pi}_1^{(n)}([0,t])}{N_1^{(n)}} \Big| S_{n1}\, dt \\ &= \int_0^1 \mathbb{E}|B_t|\, dt,\end{aligned}$$

where B_t is defined by the above equation. Let $t \in [0, 1]$ be fixed. Since $\widetilde{\Pi}_1^{(n)}$ is a Cox process with random mean measure Λ_1, conditional on Λ_1 and on $N_1^{(n)} = k \ge 1$, B_t follows a centred renormalised binomial distribution; $B_t = B(k, q)/k - q$ with $q = \Lambda_1([0, t])$. Since B_t is centred, the conditional expectation of B_t^2 equals its conditional variance, $q(1 - q)/k \le 1/(4k)$ (or 0 if $k = 0$). This bound is independent of Λ_1, so we conclude that $\mathbb{E}B_t^2 | N_1^{(n)} \le \mathbf{1}\{N_1^{(n)} > 0\}/(4N_1^{(n)})$.

Now $N_1^{(n)}$ follows a Poisson distribution with parameter τ_n. Note that if $X \sim \text{Poisson}(\theta)$ then $\mathbb{E}X^{-1}\mathbf{1}\{X > 0\} \leq 2/\theta$, which can be seen by applying the inequality $1/k \leq 2/(k+1)$ for $k \geq 1$:

$$\sum_{k=1}^{\infty} \frac{1}{k} e^{-\theta} \frac{\theta^k}{k!} \leq \sum_{k=1}^{\infty} 2e^{-\theta} \frac{\theta^k}{(k+1)!} = 2\theta^{-1} \sum_{k=1}^{\infty} e^{-\theta} \frac{\theta^{k+1}}{(k+1)!} = \frac{2}{\theta}\big(1 - e^{-\theta} - \theta e^{-\theta}\big).$$

Thus, taking expected values again, we conclude that $\mathbb{E}B_t^2 \leq (2\tau_n)^{-1}$ so that the integrand above is $\mathbb{E}|B_t| \leq (2\tau_n)^{-1/2}$. It follows that $\mathbb{E}X_{n1}^2 \leq (2\tau_n)^{-1/2}$ and so $\mathbb{E}X_{n1} \leq (2\tau_n)^{-1/4}$. Summarising, the amplitude variation is of order at most $O_{\mathbb{P}}(\tau_n^{-1/4})$.

As for the smoothing bias, it has been shown in the proof of Theorem 1 that each of the summands is bounded by $G(\sigma_i^{(n)})$, where

$$G(\sigma) = \sqrt{3\sigma^2 + 4\max\bigg(\Psi\bigg(\frac{-1}{\sqrt{\sigma}}\bigg), 1 - \Psi\bigg(\frac{1}{\sqrt{\sigma}}\bigg)\bigg)}.$$

If (the distribution corresponding to) Ψ has tails of order $O(t^{-4})$, then the first summand above dominates, so that $G(\sigma) \leq R_\Psi \sigma$ for some finite constant R_Ψ and all $\sigma \geq 0$, and

$$\frac{1}{n}\sum_{i=1}^{n} d\bigg(\frac{\widetilde{\Pi}_i^{(n)}}{N_i^{(n)}}, \widehat{\Lambda}_i\bigg) S_{ni} \leq \frac{1}{n}\sum_{i=1}^{n} G\big(\sigma_i^{(n)}\big) \leq \frac{1}{n}\sum_{i=1}^{n} R_\Psi \sigma_i^{(n)} = R_\Psi \frac{1}{n}\sum_{i=1}^{n} \sigma_i^{(n)}.$$

The result now follows from $d(\widehat{\lambda}_n, \lambda) \leq d(\widehat{\lambda}_n, \lambda_n) + d(\lambda_n, \lambda)$. □

PROOF OF THEOREM 3. The conditions of the theorem imply that $\sqrt{n}(g(\widehat{\lambda}_n) - g(\lambda_n))$ converges weakly to 0, so that

$$\sqrt{n}\big(F_{\widehat{\lambda}_n}^{-1} - F_\lambda^{-1}\big) = \sqrt{n}\big(g(\widehat{\lambda}_n) - g(\lambda)\big) \xrightarrow{\mathrm{D}} GP,$$

where GP is the Gaussian process defined above. So the first statement follows from Slutsky's theorem. The assumption that the density of λ is positively bounded below implies that $u = F_\lambda$ satisfies the hypothesis of Lemma 5 stated after the end of the proof, so that right composition is continuous on $L^2[0,1]$. By the continuous mapping theorem

$$\sqrt{n}(S_n - \mathrm{id}) = \sqrt{n}\big(F_{\widehat{\lambda}_n}^{-1} \circ F_\lambda - F_\lambda^{-1} \circ F_\lambda\big) = \big[\sqrt{n}\big(F_{\widehat{\lambda}_n}^{-1} - F_\lambda^{-1}\big)\big] \circ F_\lambda \xrightarrow{\mathrm{D}} GP \circ F_\lambda,$$

where S_n is the optimal map from λ to $\widehat{\lambda}_n$.

Now $Z = GP \circ F_\lambda$ is also the weak limit of the process

$$\sqrt{n}\bigg(\frac{1}{n}\sum_{i=1}^{n} F_{\Lambda_i}^{-1} \circ F_\lambda - F_\lambda^{-1} \circ F_\lambda\bigg) = \sqrt{n}\bigg(\frac{1}{n}\sum_{i=1}^{n} T_i - \mathrm{id}\bigg),$$

where T_i is the random warp function from λ to Λ_i. Since these are i.i.d. elements in L^2, we see that the covariance of Z is $\mathbb{E}(T - \mathrm{id}) \otimes (T - \mathrm{id})$, that is, the kernel is

$$\kappa(s,t) = \mathbb{E}\big[\big(T(s) - s\big)\big(T(t) - t\big)\big] = \mathrm{cov}\big(T(s), T(t)\big), \qquad s,t \in [0,1].$$

It easily follows from $Z(t) = GP(F_\lambda(t))$ that Z is a Gaussian process. □

LEMMA 5 (Composition and continuity). *Let $u : [0,1] \to [0,1]$ be strictly increasing piecewise continuously differentiable. Suppose that the derivative of u is bounded below by $\delta > 0$. Then the composition from the right $f \mapsto f \circ u$ from $L^p[0,1]$ takes values in $L^p[0,1]$ and it is $\delta^{-1/p}$-Lipschitz.*

PROOF. Since composition from the right is linear, it is sufficient to prove continuity around zero. This follows from the change of variables formula

$$\begin{aligned}\|f \circ u\|^p &= \int_0^1 \big|f^p\big(u(s)\big)\big|\,ds = \int_{u(0)}^{u(1)} \big|f^p(t)\big|\frac{1}{u'(u^{-1}(t))}\,dt \le \frac{1}{\delta}\int_{u(0)}^{u(1)} \big|f^p(t)\big|\,dt \\ &\le \frac{1}{\delta}\|f\|^p,\end{aligned}$$

since $0 \le u(0) \le u(1) \le 1$. The statement for $p = \infty$ holds trivially without any assumptions on $u : [0,1] \to [0,1]$. □

8. Illustrative examples. In order to illustrate the estimation framework put forth in the previous sections, we consider two scenarios involving warped Poisson processes (equivalently, Cox processes, see Section 3.5). More detailed simulations, including comparisons with the Fisher–Rao approach [36], may be found in the supplementary material [27].

8.1. *Explicit classes of warp maps.* We first introduce a flexible mixture class of warp maps that provably satisfies assumptions (A1) and (A2). This can be seen as an extension of the class considered by Wang and Gasser in [40, 41]. Let k be an integer and define $\zeta_k : [0,1] \to [0,1]$ by

$$\zeta_0(x) = x, \qquad \zeta_k(x) = x - \frac{\sin(\pi k x)}{|k|\pi}, \qquad k \in \mathbb{Z} \setminus \{0\}. \tag{8.1}$$

These are strictly increasing smooth functions satisfying $\zeta_k(0) = 0$ and $\zeta_k(1) = 1$ for any k. Plots of ζ_k for $|k| \le 3$ are presented in Figure 2(a). These maps can be made random by replacing k by an integer-valued random variable K. If the distribution of K is symmetric (around 0), then it is straightforward to see that

$$\mathbb{E}\big[\zeta_K(x)\big] = x \qquad \forall x \in [0,1].$$

This discrete family of random maps can be made continuous by means of mixtures: for $J > 1$ let $\{K_j\}_{j=1}^J$ be i.i.d. integer-valued symmetric random variables,

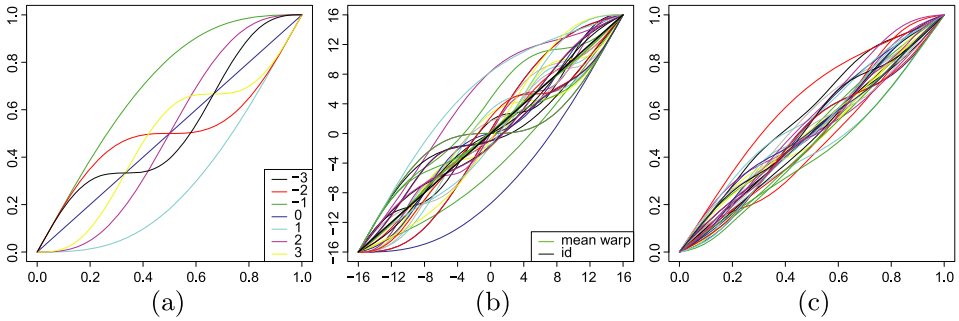

FIG. 2. (a) *The functions* $\{\zeta_k\}$ *for* $|k| \le 3$; (b) *Realisations of* T *defined as in equation* (8.2) *with* $J = 2$ *and* $K_j \overset{d}{=} V_1 V_2$ *where* V_1 *is Poisson with mean* 3, *and* $\mathbb{P}[V_2 = +1] = \mathbb{P}[V_2 = -1] = 1/2$, *independently of* V_1; (c) *Realisations of* T *defined as in equation* (8.2) *with* $J = 10$ *and* K_j *as in* (b).

and $\{U_{(j)}\}_{j=1}^{J-1}$ be the order statistics of $J-1$ i.i.d. uniform random variables on $[0,1]$, independent of $\{K_j\}_{j=1}^{J}$. The random map

$$T(x) = U_{(1)}\zeta_{K_1}(x) + \sum_{j=2}^{J-1}(U_{(j)} - U_{(j-1)})\zeta_{K_j}(x) + (1 - U_{(J-1)})\zeta_{K_J}(x), \tag{8.2}$$

$$x \in [0,1],$$

satisfies assumptions (A1) and (A2). The parameter J can be seen as controlling the *variance* of T: the larger J is, the more variables are being averaged, and so a law of large numbers effect yields maps that deviate only slightly from the identity [see Figure 2(b) and 2(c)].

8.2. *Bimodal Cox processes.* We first focus on a scenario where assumptions (B1) and (B2) hold true. We consider a structural mean measure that is a mixture of three independent components: two Gaussian distributions (of unit variance), restricted to the interval $[-16, 16]$, and a beta background with parameters $(1.5, 1.5)$, restricted on the interval $[-12, 12]$. We wish to discern the two clear modes (located at ± 8), but these may be smeared by phase variation. The structural mean density is

$$f(x) = \frac{1-\varepsilon}{2}\big[\varphi(x-8) + \varphi(x+8)\big] + \frac{\varepsilon}{24}\beta_{1.5,1.5}\left(\frac{x+12}{24}\right),$$

where φ denotes a standard Gaussian density, $\beta_{\alpha,\beta}$ is the Beta(α, β) density, and $\varepsilon = 0.1$ is the strength of the background. We generated 30 independent Poisson processes with this structural mean measure and $\tau = 93$, and warped them by means of 30 independent warp maps $\{T_i\}$, obtaining 30 warped point processes [Figure 3(c)]. The warp maps $\{T_k\}$ are affinely transformed versions of the maps

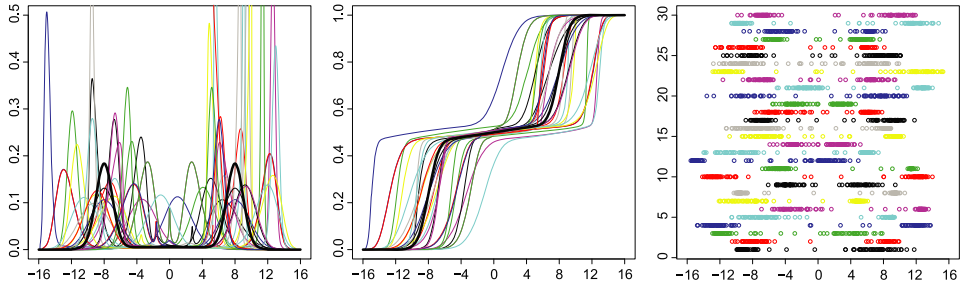

FIG. 3. (a) *Thirty warped bimodal densities, with the structural mean bimodal density* $93 \times f$ *in solid black*; (b) *Their corresponding distribution functions, with the structural mean distribution function in solid black*; (c) *Thirty Cox processes, constructed as follows: first we generate* Π_i *as i.i.d. Poisson processes with mean density* f*, then we warp them by forming* $T_{\#}\Pi_i$*, where* T *are the maps appearing in Figure* 2(b).

shown in Figure 2(b) according to the mapping

$$g(x) \mapsto 32g\left(\frac{x+16}{32}\right) - 16$$

in order to re-scale their support to $[-16, 16]$. Recall that the warp maps in Figure 2(b) were generated using the definition in equation (8.2), taking $J = 2$ and K_j are i.i.d., distributed as V_1V_2, where V_1 is Poisson with mean 3, and $\mathbb{P}[V_2 = +1] = \mathbb{P}[V_2 = -1] = 1/2$, independently of V_1. These correspond to rather violent phase variation, as can be seen by the plots of the conditional density/distribution of the warped processes given the corresponding T_i in Figure 3(a) and 3(b).

Using the 30 warped spike trains depicted in Figure 3(c), we construct the "regularised Fréchet–Wasserstein" estimator as described in Section 4. A slight deviation is that we use a Gaussian kernel with bandwidth chosen by unbiased cross validation, rather than the special kernels developed for the asymptotic theory (with no essential effect on finite sample performance). We thus obtain estimates of the warp maps $\{\widehat{T}_i\}_{i=1}^{30}$ (using the definitions in Section 4.4), depicted in Figure 4(b), which can be used to register the point processes (Figure 5). The final estimate of the structural mean distribution function (the regularised Fréchet–Wasserstein estimator) is depicted in Figure 4(a), and contrasted with the true structural CDF, as well as with the naive estimate produced by ignoring warping and averaging the empirical distributions across trains. We notice that the regularised Fréchet–Wasserstein estimator performs quite well at discerning the two modes of the structural mean measure, in contrast with the naive estimator which seems to fail to resolve them. This effect is more clearly portrayed in Figure 4(c), which plots kernel estimators of structural mean density constructed using the original (warped) point processes, and the registered point processes. It is important to remark that the minor fluctuations in the density estimate observed are *not* related to our method

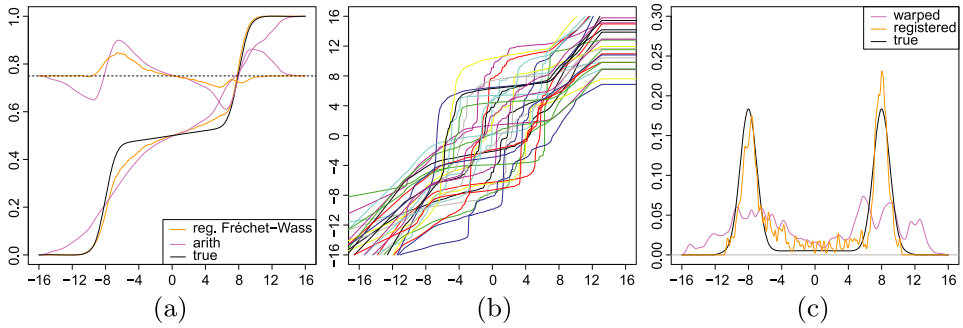

FIG. 4. (a) *The empirical arithmetic mean, our estimated regularised Fréchet–Wasserstein mean, and the true mean CDF (the curves oscillating about the horizontal line* $y = 3/4$ *are residual curves, centred at* $3/4$*);* (b) *The estimated warp functions*; (c) *Kernel estimates of the density function of the true structural mean, based on the original spike trains, and on the registered spike trains.*

of estimation, but are due to the sampling variation of the spike trains (i.e., they are not intrinsic to our registration procedure, but to the kernel density estimation procedure), and could be reduced by more careful choice of bandwidth. Figure 6 presents the sampling variation of the regularised Fréchet–Wasserstein estimator, and contrasts it with the sampling variation of the naive arithmetic estimator for 20 independent replications of the same experiment. We notice that the naive estimator is clearly biased in the neighbourhoods around the two peaks, and appears to fluctuate around a straight line. In contrast, the smoothed Fréchet mean—though presenting fluctuations around the two peaks—appears approximately unbiased. Indeed, its variation is very clearly not fluctuation around a line—to the contrary it suggests two clear elbows in the CDF, which correspond to the two peaks.

It is also interesting to note that the empirical Fréchet mean was observed to be insensitive to the choice of the bandwidth parameter used in the construction of the estimated conditional mean measures $\widehat{\Lambda}_i$. Of course, the warp functions $\widehat{T}_i$

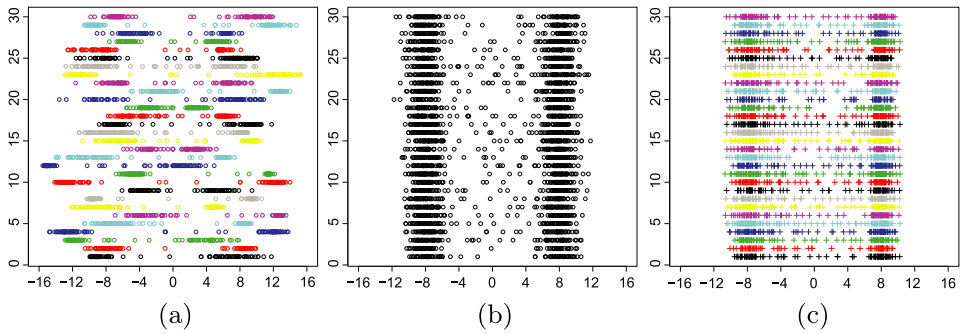

FIG. 5. *Bimodal Cox processes*: (a) *The warped point processes*; (b) *The original point processes*; (c) *The registered point processes.*

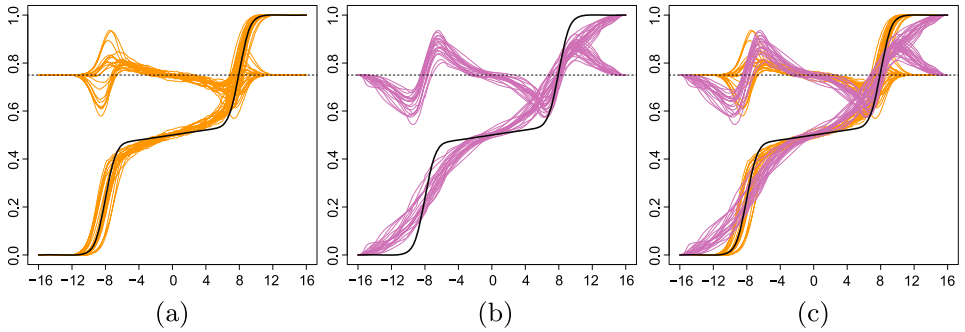

FIG. 6. (a) *Comparison of our estimated regularised Fréchet–Wasserstein mean, and the true mean CDF, for* 20 *independent replications of the experiment*; (b) *Comparison of the arithmetic mean, and the true mean CDF, for the same* 20 *replications*; (c) *Superposition of* (a) *and* (b). *In all three cases, the curves oscillating about the horizontal line* $y = 3/4$ *are residual curves, centred at* $3/4$.

themselves (and hence the registered processes) would depend on this parameter, since these couple $\widehat{\Lambda}_i$ and $\widehat{\lambda}$—and while the latter is insensitive to the choice of bandwidth parameter, the former is clearly not.

Further simulations carried out in the supplementary material [27] reaffirm these findings for different "sample sizes" τ and choices of smoothing parameter. Furthermore, numerical comparisons also carried out in the supplement suggest that Fréchet–Wasserstein registration outperforms Fisher–Rao registration (carried out as in [36] at the level of CDFs), in terms of how close the registered processes are to the original point processes (prior to warping), where "closeness" is measured by means of the ℓ_2 distance of the ordered points. This is not surprising given our unbiasedness considerations (Proposition 3), since the Fisher–Rao estimator is generally not d-unbiased.

8.3. *Triangular Cox processes.* We now treat a second scenario that somewhat deviates from our model assumptions, because it involves linear warp functions. Consequently, phase variation can also be seen at the level of densities (see Section 3.4). Consider the family of triangular densities of support length $2h$ and height $1/h$, and their corresponding distribution functions (see Figure 7)

$$f_h(t) = \frac{1}{h}\left(1 - \frac{1}{h}|t|\right), \qquad |t| \le h, h > 0,$$

$$F_h(t) = \begin{cases} \dfrac{1}{2h^2}(t+h)^2, & -h \le t \le 0, \\ 1 - \dfrac{1}{2h^2}(h-t)^2, & 0 \le t \le h. \end{cases}$$

Our example will consist in phase varying Poisson processes, with structural mean distribution equal to F_1 (i.e., the triangular distribution function with $h = 1$).

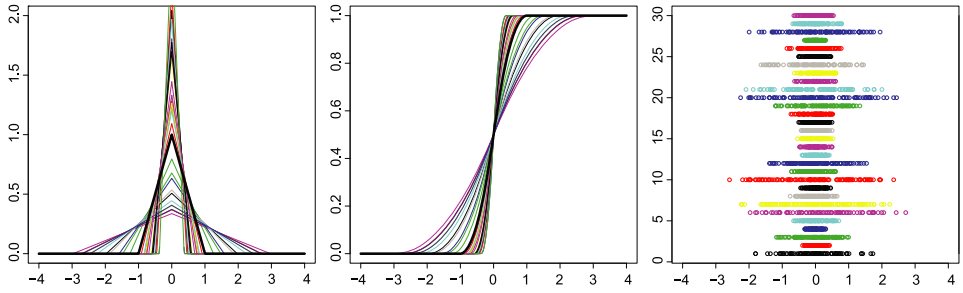

FIG. 7. (a) *Thirty triangular densities* $f_{h_i}(t)$, *with* f_1 *in solid black*; (b) *Their corresponding distribution functions* $F_{h_i}(t)$, *with* F_1 *in solid black*; (c) *Thirty Cox processes, constructed as follows: first, we generate* Π_i *as i.i.d. Poisson processes with mean density* f_1, *then we warp them by forming* $T_i \# \Pi_i$.

To this aim, let h be a random variable valued in $(0, C]$, so that the random measures have a common support $I = [-C, C]$, but they are not strictly positive there. Following the same steps as in the proof of Proposition 2, it can be seen that the random measure with distribution function F_h has a unique theoretical Fréchet mean with distribution function $F_{\mathbb{E}[h]}$, in the sense that for all distribution functions $G \neq F_{\mathbb{E}[h]}$, we have (allowing for a slight abuse of notation) $\mathbb{E}[d^2(F_{\mathbb{E}[h]}, F_h)] < \mathbb{E}[d^2(G, F_h)]$ (note that Proposition 2 and its proof remain valid as long as the measures have no atoms; they do not need to be strictly increasing). The warp map corresponding to an h is $W_h(x) = hx$, and it is not a homeomorphism of I (unless $h = 1$), thus violating our assumptions (see Section 3.4). To construct our phase-varying point processes, we generate 30 i.i.d. copies $\{h_j\}_{j=1}^{30}$ of a random variable h following the mixture of uniform distributions $\alpha\mathcal{U}[0.35, 1] + (1 - \alpha)\mathcal{U}[0.35, 3]$, where $\alpha = 0.675$ is chosen so that $\mathbb{E}[h] = 1$. Then we generate 30 Poisson processes, with cumulative mean measure $\tau \times F_1$ [i.e., $h = 1$, see Figure 7(c)], $\tau = 93$, and warp them by the maps $\{T_i = W_{h_i}\}_{i=1}^{30}$. This yields 30 Cox processes, each with a realised directing measure $93 \times \Lambda_1, \ldots, 93 \times \Lambda_{30}$, respectively, where the $\Lambda_1, \ldots, \Lambda_{30}$ have distribution functions $F_{h_1}, \ldots, F_{h_{30}}$ [depicted in Figure 7(b)]. The resulting warped spike trains are displayed in Figure 7(c).

Assuming that the parametric form of the model is unknown to us, we carry out the separation of amplitude and phase variation nonparametrically, as described in Section 4. We smooth each spike train using a Gaussian kernel with bandwidth chosen by unbiased cross validation to obtain the estimators $\{\widehat{\Lambda}_i\}_{i=1}^{30}$ (strictly speaking, not in line with our discussion in Section 4.2, but this has no practical effect), estimate the warp functions $\{\widehat{T}_i\}_{i=1}^{30}$, as described in Section 4.4, and produce a registration of the point processes using these (Figure 8). We see that these warp functions [Figure 9(b)] are indeed nearly linear (besides numerical instabilities at the boundary of the domain). The regularised Fréchet–Wasserstein mean of

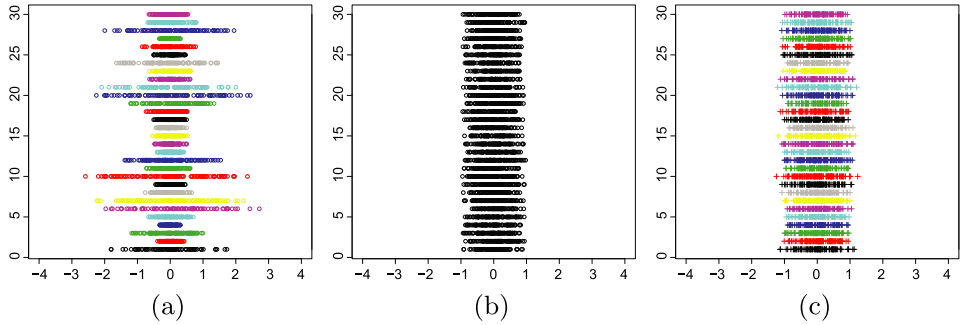

FIG. 8. *Triangular Cox processes*: (a) *The warped point processes*; (b) *The original point processes*; (c) *The registered point processes*.

$\{F_{\widehat{\Lambda}_i}\}_{i=1}^{30}$ is depicted in Figure 9(a), contrasted with the arithmetic mean and the true structural mean. Note that the regularised Fréchet–Wasserstein mean is supported on a subset of the domain, as is the true structural mean; by contrast, the arithmetic mean is supported almost on the entire domain, which is visible in Figure 9(a), where it has left-and-right tails that persist. Though both the regularised Fréchet–Wasserstein and the arithmetic mean perform well near the point of symmetry of the structural mean (which is to be expected, at least for the arithmetic mean, since the location of the structural measure is invariant to the warp action), the regularised Fréchet–Wasserstein mean estimates the support and tails of the structural measure visibly better. These observations are more clearly depicted in the residual plots contained in Figure 10, where the residual curves of the deviation between the arithmetic/Fréchet means and the estimand are considered, for 20 independent repetitions of the same simulation experiment. It is seen in that diagram that the arithmetic mean is clearly biased, especially near the boundaries of the support of the true structural mean.

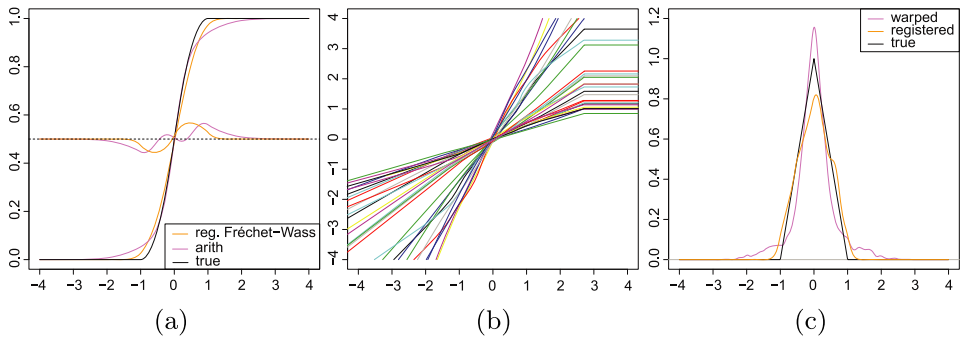

FIG. 9. (a) *The empirical arithmetic mean, our estimated regularised Fréchet–Wasserstein mean, and the true mean CDF*; (b) *The estimated warp functions*; (c) *Kernel estimates of the density function of the true structural mean, based on the original spike trains, and on the registered spike trains*.

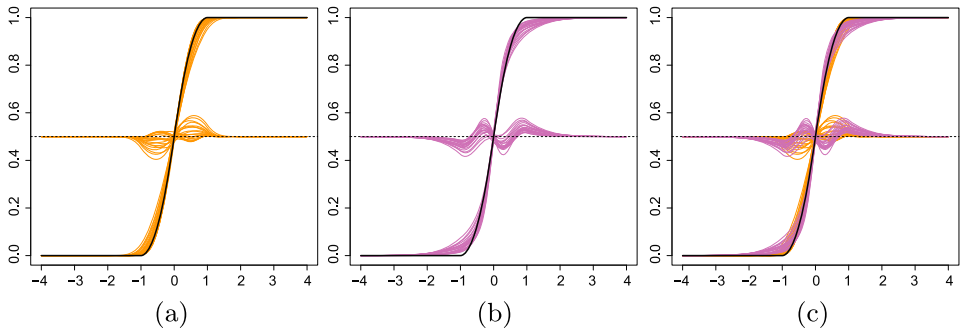

FIG. 10. (a) *Comparison of our estimated regularised Fréchet–Wasserstein mean, and the true mean CDF, for* 20 *independent replications of the experiment*; (b) *Comparison of the arithmetic mean, and the true mean CDF, for the same* 20 *replications*; (c) *Superposition of* (a) *and* (b). *In all three cases, the curves oscillating about the horizontal line* $y = 1/2$ *are residual curves, centred at* $1/2$.

To gauge the effectiveness of the registration carried out, we also constructed kernel estimators of the density of the structural mean, based on the original (warped) point processes, and on the registered (aligned) point processes. These are shown in Figure 9(c). They illustrate that the density estimate based on the raw data overestimates the mode as well as the tails of the true density, whereas the density estimate based on the registered data fits both the bulk and the tails of the density quite nicely. As in the previous example, the minor fluctuations of these density estimates are not intrinsic to our registration procedure, but to the kernel density estimation procedure.

Stability of the estimated structural mean CDF with respect to the smoothing parameter was also observed in this example, and persisted in additional simulations (presented in the supplementary material [27]), where different sample sizes were also considered. Simulation comparisons showed that also in this scenario our approach performs at least as well as the Fisher–Rao approach in terms of registration of the point processes.

9. Discussion. We have introduced a framework formalising the confounding of amplitude and phase variation in point process data, and demonstrated how this can be used for their consistent nonparametric separation on the basis of independent realisations thereof. The key ingredient of our approach was the observation that for the point process warping problem, the classical functional data assumptions on warp functions are equivalent to the geometry of the Monge problem of optimal transportation.

A particularly attractive aspect of the present framework is that it yields an identifiable setup, with a clear notion of over/under registration through the concept of bias. Indeed, we prove that consistent estimation of the warp functions *is* possible

in our framework for point process data, circumventing the so-called "pinching effect" (see, e.g., Kneip and Ramsay [24], Section 2.4) even under very sparse sampling regimes (Remark 1). Furthermore, our consistency results present some appealing features: there is no finite-dimensional parameterisation, and the unknown warp functions and measures are allowed to be genuinely functional, that is, infinite dimensional (contrary to, say Tang and Müller [37]; Gervini and Gasser [18]; Rønn [32]); though the consistency of the warp functions is in the uniform metric, there is no need for the introduction of additional smoothness penalties on the warp functions, and no tuning parameter needs be selected to impose this (the regularity is inherited directly from the underlying regularity of the structural and conditional mean point process measures themselves; in the functional case, this corresponds to the regularity of the curves themselves); consistency is established with reference to a population, that is, the number of "individuals" (processes) is allowed to grow along with the "density of their sampling" (with a clearly identified relationship between the two), instead of establishing consistency conditional on the sample (i.e., with a fixed number of curves, assuming only that the density of sampling for each curve increasing, with no reference to a more general "curve population," as in, e.g., Kneip and Engel [23], Wang and Gasser [41], and Gervini and Gasser [17]). In our experience, when consistency results are given in the functional warping literature, they typically feature at least one of these restrictions. We do not mention these characteristics as a claim to superiority, but rather point them out as a special feature of the problem in the point process case, afforded by the optimal transportation geometry (since the very warping process is inextricably linked with the metric structure of the space). Nevertheless, it is interesting to note that the functional form of the warp function estimator (4.4) is strikingly similar with the pairwise synchronisation estimator of Tang and Müller [37], equation (7).

Further to consistency, we are able to obtain detailed rates of convergence. These show $\sqrt{n}$-consistency and a central limit theorem in the special case of warped Poisson processes (Cox processes) under dense sampling. These can serve as a basis for uncertainty quantification, but also indicate that our estimator can attain the optimal rate of convergence under dense sampling.

Though we have demonstrated that the optimal transportation geometry is canonical if warping occurs at the level of the spike train observations (at the level of measures), it is possible to introduce warping at the level of the density of the underlying mean measure (see Section 3.4). In such a framework, there are options other than the optimal transportation geometry that be may better suited for the formalisation of the warping problem. For example, in the case of functional data, Tucker, Wu and Srivastava [38] attack the warping problem by imbedding the data in a quotient space modulo warp functions. This is done by employing a Fisher–Rao-type metric, which is invariant with respect to the action of a warping group. Recent work by Wu and Srivastava [44] extends their approach to the

case of spike trains, by smoothing the spike trains and considering them as densities in the Fisher–Rao space. This geometry may be more natural than the optimal transportation one to model phase variation at the level of densities.

A natural question for further work is that of *multivariate phase variation*. For example, is the "canonicity" of the optimal transportation framework preserved, and can one fruitfully proceed in a similar manner? The key challenge in this case is that, in the case of measures on subsets of $\mathbb{R}^d$, $d > 1$, evaluation of the empirical Fréchet mean in closed form is impossible (see, e.g., Agueh and Carlier [1]). Approximations can be sought, for example, via Gaussian assumptions (Cuturi and Doucet [12]) or via reduction to several 1D problems (Bonneel et al. [8]). Indeed, during the final preparation of this manuscript, we became aware of interesting independent work in parallel by Boissard, LeGuic and Loubes [7], who consider the problem of estimating Wasserstein barycentres for measures on $\mathbb{R}^d$, and define "admissible" groups of deformations that mimic the 1D case, thus allowing for consistent estimation and evaluation of the sample barycentre by calculating successive means between pairs (i.e., by an iterated barycentre).

Finally, it should be mentioned that once phase and amplitude variation have been separated, they could each be subjected to a further analysis of their own. The amplitude variation clearly would be analysed by means of *linear PCA* tools, along the lines described in Section 3.1. On the other hand, the phase variation can be analysed by making further use of the geometrical properties described in Section 3.3: for instance, via tangent space PCA (see, e.g., Boissard, LeGuic and Loubes [7]) or via geodesic PCA (see, e.g., Bigot et al. [6]). Indeed, the form of the limiting covariance function in our central limit theorem (Theorem 3) suggests that strong connections can be established between Wasserstein PCA methodology and the separation of amplitude and phase variation.

Acknowledgments. This paper grew out of work presented at the Mathematical Biosciences Institute (Ohio State University), during the "Statistics of Time Warping and Phase Variation" Workshop, November 2012. We wish to acknowledge the stimulating environment offered by the Institute. We are grateful to an Associate Editor and three referees for their insightful and constructive comments. The paper has genuinely improved as a result of the review process.

SUPPLEMENTARY MATERIAL

"Amplitude and phase variation of point processes" (DOI: 10.1214/15-AOS1387SUPP; .pdf). The online supplement contains more detailed simulation experiments.

REFERENCES

[1] AGUEH, M. and CARLIER, G. (2011). Barycenters in the Wasserstein space. *SIAM J. Math. Anal.* **43** 904–924. MR2801182

[2] ANDERES, E. and CHATTERJEE, S. (2009). Consistent estimates of deformed isotropic Gaussian random fields on the plane. *Ann. Statist.* **37** 2324–2350. MR2543694
[3] ANDERES, E. B. and STEIN, M. L. (2008). Estimating deformations of isotropic Gaussian random fields on the plane. *Ann. Statist.* **36** 719–741. MR2396813
[4] ARRIBAS-GIL, A. and MÜLLER, H.-G. (2014). Pairwise dynamic time warping for event data. *Comput. Statist. Data Anal.* **69** 255–268. MR3146893
[5] BICKEL, P. J. and FREEDMAN, D. A. (1981). Some asymptotic theory for the bootstrap. *Ann. Statist.* **9** 1196–1217. MR0630103
[6] BIGOT, J., GOUET, R., KLEIN, T. and LOPEZ, A. (2013). Geodesic PCA in the Wasserstein space. Available at arXiv:1307.7721.
[7] BOISSARD, E., LE GOUIC, T. and LOUBES, J.-M. (2015). Distribution's template estimate with Wasserstein metrics. *Bernoulli* **21** 740–759. MR3338645
[8] BONNEEL, N., RABIN, J., PEYRÉ, G. and PFISTER, H. (2015). Sliced and Radon Wasserstein barycenters of measures. *J. Math. Imaging Vision* **51** 22–45. MR3300482
[9] CHIANG, C.-T., WANG, M.-C. and HUANG, C.-Y. (2005). Kernel estimation of rate function for recurrent event data. *Scand. J. Stat.* **32** 77–91. MR2136803
[10] CHIOU, J.-M. and MÜLLER, H.-G. (2009). Modeling hazard rates as functional data for the analysis of cohort lifetables and mortality forecasting. *J. Amer. Statist. Assoc.* **104** 572–585. MR2751439
[11] CLAESKENS, G., SILVERMAN, B. W. and SLAETS, L. (2010). A multiresolution approach to time warping achieved by a Bayesian prior-posterior transfer fitting strategy. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **72** 673–694. MR2758241
[12] CUTURI, M. and DOUCET, A. (2013). Fast computation of Wasserstein barycenters. Available at arXiv:1310.4375.
[13] DALEY, D. J. and VERE-JONES, D. (2008). *An Introduction to the Theory of Point Processes. Vol. II*: *General Theory and Structure*, 2nd ed. Springer, New York. MR2371524
[14] DURRETT, R. (2010). *Probability*: *Theory and Examples*, 4th ed. Cambridge Univ. Press, Cambridge. MR2722836
[15] GANGBO, W. and ŚWIĘCH, A. (1998). Optimal maps for the multidimensional Monge–Kantorovich problem. *Comm. Pure Appl. Math.* **51** 23–45. MR1486630
[16] GASSER, T. and KNEIP, A. (1995). Searching for structure in curve samples. *J. Amer. Statist. Assoc.* **90** 1179–1188.
[17] GERVINI, D. and GASSER, T. (2004). Self-modelling warping functions. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **66** 959–971. MR2102475
[18] GERVINI, D. and GASSER, T. (2005). Nonparametric maximum likelihood estimation of the structural mean of a sample of curves. *Biometrika* **92** 801–820. MR2234187
[19] HADJIPANTELIS, P. Z., ASTON, J. A. D. and EVANS, J. P. (2012). Characterizing fundamental frequency in Mandarin: A functional principal component approach utilizing mixed effect models. *Journal of the Acoustical Society of America* **131** 4651–4664.
[20] JAMES, G. M. (2007). Curve alignment by moments. *Ann. Appl. Stat.* **1** 480–501. MR2415744
[21] KALLENBERG, O. (1986). *Random Measures*, 4th ed. Academic Press, London. MR0854102
[22] KARR, A. F. (1991). *Point Processes and Their Statistical Inference*, 2nd ed. *Probability*: *Pure and Applied* **7**. Dekker, New York. MR1113698
[23] KNEIP, A. and ENGEL, J. (1995). Model estimation in nonlinear regression under shape invariance. *Ann. Statist.* **23** 551–570. MR1332581
[24] KNEIP, A. and RAMSAY, J. O. (2008). Combining registration and fitting for functional models. *J. Amer. Statist. Assoc.* **103** 1155–1165. MR2528838
[25] LEHMANN, E. L. (1951). A general concept of unbiasedness. *Ann. Math. Statistics* **22** 587–592. MR0047296
[26] MARRON, J. S., RAMSAY, J. O., SANGALLI, L. M. and SRIVASTAVA, A. (2014). Statistics of time warpings and phase variations. *Electron. J. Stat.* **8** 1697–1702. MR3273584

[27] PANARETOS, V. M. and ZEMEL, Y. (2015). Supplement to "Amplitude and phase variation of point processes." DOI:10.1214/15-AOS1387SUPP.
[28] RAMSAY, J. O. (2000). Functional components of variation in handwriting. *J. Amer. Statist. Assoc.* **95** 9–15.
[29] RAMSAY, J. O. and LI, X. (1998). Curve registration. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **60** 351–363. MR1616045
[30] RAMSAY, J. O. and SILVERMAN, B. W. (2002). *Applied Functional Data Analysis*: *Methods and Case Studies*. Springer, New York. MR1910407
[31] RAMSAY, J. O. and SILVERMAN, B. W. (2005). *Functional Data Analysis*, 2nd ed. Springer, New York. MR2168993
[32] RØNN, B. B. (2001). Nonparametric maximum likelihood estimation for shifted curves. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **63** 243–259. MR1841413
[33] SAMPSON, P. D. and GUTTORP, P. (1992). Nonparametric estimation of non stationary spatial covariance structure. *J. Amer. Statist. Assoc.* **87** 108–119.
[34] SCHOENBERG, F. (1999). Transforming spatial point processes into Poisson processes. *Stochastic Process. Appl.* **81** 155–164. MR1694573
[35] SENOUSSI, R., CHADŒUF, J. and ALLARD, D. (2000). Weak homogenization of point processes by space deformations. *Adv. in Appl. Probab.* **32** 948–959. MR1808906
[36] SRIVASTAVA, A., WU, W., KURTEK, S., KLASSEN, R. and MARRON, J. S. (2011). Registration of functional data using the Fisher–Rao metric. Available at arXiv:1103.3817v2.
[37] TANG, R. and MÜLLER, H.-G. (2008). Pairwise curve synchronization for functional data. *Biometrika* **95** 875–889. MR2461217
[38] TUCKER, J. D., WU, W. and SRIVASTAVA, A. (2013). Generative models for functional data using phase and amplitude separation. *Comput. Statist. Data Anal.* **61** 50–66. MR3063000
[39] VILLANI, C. (2003). *Topics in Optimal Transportation*. *Graduate Studies in Mathematics* **58**. Amer. Math. Soc., Providence, RI. MR1964483
[40] WANG, K. and GASSER, T. (1997). Alignment of curves by dynamic time warping. *Ann. Statist.* **25** 1251–1276. MR1447750
[41] WANG, K. and GASSER, T. (1999). Synchronizing sample curves nonparametrically. *Ann. Statist.* **27** 439–460. MR1714722
[42] WU, S., MÜLLER, H.-G. and ZHANG, Z. (2013). Functional data analysis for point processes with rare events. *Statist. Sinica* **23** 1–23. MR3076156
[43] WU, W. and SRIVASTAVA, A. (2013). Estimating summary statistics in the spike-train space. *J. Comput. Neurosci.* **34** 391–410. MR3061973
[44] WU, W. and SRIVASTAVA, A. (2014). Analysis of spike train data: Alignment and comparisons using the extended Fisher–Rao metric. *Electron. J. Stat.* **8** 1776–1785. MR3273594

SECTION DE MATHÉMATIQUES
ECOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE
1015 LAUSANNE
SWITZERLAND
E-MAIL: victor.panaretos@epfl.ch
yoav.zemel@epfl.ch