Chapter 4
Phase Variation and Fréchet Means

Why is it relevant to construct the Fréchet mean of a collection of measures with respect to the Wasserstein metric? A simple answer is that this kind of average will often express a more natural notion of "typical" realisation of a random probability distribution than an arithmetic average.[1] Much more can be said, however, in that the Wasserstein–Fréchet mean and the closely related notion of an optimal multicoupling arise canonically as the appropriate framework for the formulation and solution to the problem of *separation of amplitude and phase variation of a point process*. It would almost seem that Wasserstein–Fréchet means were "made" for precisely this problem.

When analysing the (co)variation of a real-valued stochastic process $\{Y(x) : x \in K\}$ over a convex compact domain K, it can be broadly said that one may distinguish two layers of variation:

- *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 the covariance kernel, at least up to second order.

  In short, this is variation "in the y-axis" (ordinate).

Electronic Supplementary Material The online version of this chapter (https://doi.org/10.1007/978-3-030-38438-8_4) contains supplementary material.

[1] For instance, the arithmetic average of two scalar Gaussians $N(\mu_1, 1)$ and $N(\mu_2, 1)$ will be their mixture with equal weights, but their Fréchet–Wasserstein average will be the Gaussian $N(\frac{1}{2}\mu_1 + \frac{1}{2}\mu_2, 1)$ (see Lemma 4.2.1), which is arguably more representative from an intuitive point of view. In much the same way, the Fréchet–Wasserstein average of probability measures representing some type of object (e.g., normalised greyscale images of faces) will also be an object of the same type. This sort of phenomenon is well-known in manifold statistics, more generally, and is arguably one of the key motivations to account for the non-linear geometry of the sample space, rather than imbed it into a larger linear space and use the addition operation.

V. M. Panaretos, Y. Zemel, *An Invitation to Statistics in Wasserstein Space*,
SpringerBriefs in Probability and Mathematical Statistics,
https://doi.org/10.1007/978-3-030-38438-8_4

- *Phase variation*. This is a second layer of 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 a composition of the stochastic process with a random transformation (warp map) acting on its domain.

  This is variation "in the x-axis" (abscissa).

The terminology on amplitude/phase variation is adapted from random trigonometric functions, which may vary in amplitude (oscillations in the range of the function) or phase (oscillations in the domain of the function). Failing to properly account for the superposition of these two forms of variation may entirely distort the findings of a statistical analysis of the random function (see Sect. 4.1.1). Consequently, it is an important problem to be able to separate the two, thus correctly accounting for the distinct contribution of each. The problem of separation is also known as that of *registration* (Ramsay and Li [108]), *synchronisation* (Wang and Gasser [129]), or *multireference alignment* (Bandeira et al. [16]), though in some cases these terms refer to a simpler problem where there is no amplitude variation at all.

Phase variation naturally arises 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* and *neuronal signals*, are usual suspects. Growth curves can be modelled as *continuous random functions (functional data)*, whereas neuronal signals are better modelled as *discrete random measures (point processes)*. We first describe amplitude/phase variation in the former[2] case, as that is easier to appreciate, before moving on to the latter case, which is the main subject of this chapter.

4.1 Amplitude and Phase Variation

4.1.1 The Functional Case

Let K denote the unit cube $[0,1]^d \subset \mathbb{R}^d$. A real random function $Y = (Y(x) : x \in K)$ can, broadly speaking, have two types of variation. The first, *amplitude variation*, results from $Y(x)$ being a random variable for every x and describes its fluctuations around the mean level $m(x) = \mathbb{E}Y(x)$, usually encoded by the variance $\mathrm{var}Y(x)$. For this reason, it can be referred to as "variation in the y-axis". More

[2] As the functional case will only serve as a motivation, our treatment of this case will mostly be heuristic and superficial. Rigorous proofs and more precise details can be found in the books by Ferraty and Vieu [51], Horváth and Kokoszka [70], or Hsing and Eubank [71]. The notion of amplitude and phase variation is discussed in the books by Ramsay and Silverman [109, 110] that are of a more applied flavour. One can also consult the review by Wang et al. [127], where amplitude and phase variation are discussed in Sect. 5.2.

generally, for any finite set $x_1,\dots,x_n$, the $n\times n$ covariance matrix with entries $\kappa(x_i,x_j)=\operatorname{cov}[Y(x_i),Y(x_j)]$ encapsulates (up to second order) the stochastic deviations of the random vector $(Y(x_1),\dots,Y(x_n))$ from its mean, in analogy with the multivariate case. Heuristically, one then views amplitude variation as the collection $\kappa(x,y)$ for $x,y\in K$ in a sense we discuss next.

One typically views Y as a random element in the separable Hilbert space $L_2(K)$, assumed to have $\mathbb{E}\|Y\|^2<\infty$ and continuous sample paths, so that in particular $Y(x)$ is a random variable for all $x\in K$. Then the *mean function*

$$m(x)=\mathbb{E}Y(x),\quad x\in K$$

and the *covariance kernel*

$$\kappa(x,y)=\operatorname{cov}[Y(x),Y(y)],\quad x,y\in K$$

are well-defined and finite; we shall assume that they are continuous, which is equivalent to Y being *mean-square continuous*:

$$\mathbb{E}[Y(y)-Y(x)]^2\to 0,\qquad y\to x.$$

The covariance kernel κ gives rise to the *covariance operator* $\mathscr{R}:L_2(K)\to L_2(K)$, defined by

$$(\mathscr{R}f)(y)=\int_K \kappa(x,y)f(x)\,\mathrm{d}x,$$

a self-adjoint positive semidefinite Hilbert–Schmidt operator on $L_2(K)$. The justification to this terminology is the observation that when $m=0$, for all bounded $f,g\in L_2(K)$,

$$\mathbb{E}\,\langle Y,f\rangle\,\langle Y,g\rangle=\mathbb{E}\left[\int_{K^2} Y(x)f(x)Y(y)g(y)\,\mathrm{d}(x,y)\right]=\int_K g(y)(\mathscr{R}f)(y)\,\mathrm{d}y,$$

and so, without the restriction to $m=0$,

$$\operatorname{cov}\left[\langle Y,f\rangle,\langle Y,g\rangle\right]=\int_K g(y)(\mathscr{R}f)(y)\,\mathrm{d}y=\langle g,\mathscr{R}f\rangle.$$

The covariance operator admits an eigendecomposition $(r_k,\phi_k)_{k=1}^{\infty}$ such that $r_k\searrow 0$, $\mathscr{R}\phi_k=r_k\phi_k$ and (ϕ_k) is an orthonormal basis of $L_2(K)$. One then has the celebrated *Karhunen–Loève expansion*

$$Y(x)=m(x)+\sum_{k=1}^{\infty}\langle Y-m,\phi_k\rangle\,\phi_k(x)=m(x)+\sum_{k=1}^{\infty}\xi_k\phi_k(x).$$

A major feature in this expansion is the separation of the functional part from the stochastic part: the functions $\phi_k(x)$ are deterministic; the random variables ξ_k are scalars. This separation actually holds for any orthonormal basis; the role of choosing the eigenbasis of $\mathscr{R}$ is making ξ_k *uncorrelated*:

$$\operatorname{cov}(\xi_k, \xi_l) = \operatorname{cov}\left[\langle Y, \phi_k\rangle, \langle Y, \phi_l\rangle\right] = \langle \phi_l, \mathscr{R}\phi_k\rangle$$

vanishes when $k \neq l$ and equals r_k otherwise. For this reason, it is not surprising that using as ϕ_k the eigenfunctions yields the optimal representation of Y. Here, optimality is with respect to truncations: for any other basis (ψ_k) and any M,

$$\mathbb{E}\left\| Y - m - \sum_{k=1}^{M} \langle Y - m, \psi_k\rangle \psi_k \right\|^2 \geq \mathbb{E}\left\| Y - m - \sum_{k=1}^{M} \langle Y - m, \phi_k\rangle \phi_k \right\|^2$$

so that (ϕ_k) provides the best finite-dimensional approximation to Y. The approximation error on the right-hand side equals

$$\mathbb{E}\left\| \sum_{k=M+1}^{\infty} \xi_k \phi_k \right\|^2 = \sum_{k=M+1}^{\infty} r_k$$

and depends on how quickly the eigenvalues of $\mathscr{R}$ decay.

One carries out inference for m and κ on the basis of a sample $Y_1, \ldots, Y_n$ by

$$\widehat{m}(x) = \frac{1}{n}\sum_{i=1}^{n} Y_i(x), \qquad x \in K$$

and

$$\widehat{\kappa}(x,y) = \frac{1}{n}\sum_{i=1}^{n} Y_i(x)Y_i(y) - \widehat{m}(x)\widehat{m}(y),$$

from which one proceeds to estimate $\mathscr{R}$ and its eigendecomposition.

We have seen that amplitude variation in the sense described above is linear and dealt with using linear operations. There is another, qualitatively different type of variation, *phase variation*, that is non-linear and does not have an obvious finite-dimensional analogue. It arises when in addition to the randomness in the values $Y(x)$ itself, an extra layer of stochasticity is present in its domain of definition. In mathematical terms, there is a random invertible *warp function* (sometimes called *deformation* or *warping*) $T : K \to K$ and instead of $Y(x)$, one observes realisations from

$$\tilde{Y}(x) = Y(T^{-1}(x)), \qquad x \in K.$$

For this reason, phase variation can be viewed as "variation in the x-axis". When $d = 1$, the set K is usually interpreted as a time interval, and then the model stipulates that each individual has its own time scale. Typically, the warp function is assumed to be a homeomorphism of K independent of Y and often some additional smoothness is imposed, say $T \in C^2$. One of the classical examples is growth curves of children, of which a dataset from the Berkeley growth study (Jones and Bayley [73]) is shown in Fig. 4.1. The curves are the derivatives of the height of a sample of ten girls as a function of time, from birth until age 18. One clearly notices the presence

of the two types of variation in the figure. The initial velocity for all children is the highest immediately or shortly after birth, and in most cases decreases sharply during the first 2 years. Then follows a period of acceleration for another year or so, and so on. Despite presenting qualitatively similar behaviour, the curves differ substantially not only in the magnitude of the peaks but also in their location. For instance, one red curve has a local minimum at the age of three, while a green one has a local maximum at almost that same time point. It is apparent that if one tries to estimate the mean function by averaging the curves at each time x, the shape of the resulting estimate would look very different from each of the curves. Thus, this pointwise averaging (known as the *cross-sectional mean*) fails to represent the typical behaviour. This phenomenon is seen more explicitly in the next example. The terminology of amplitude and phase comes from trigonometric functions, from

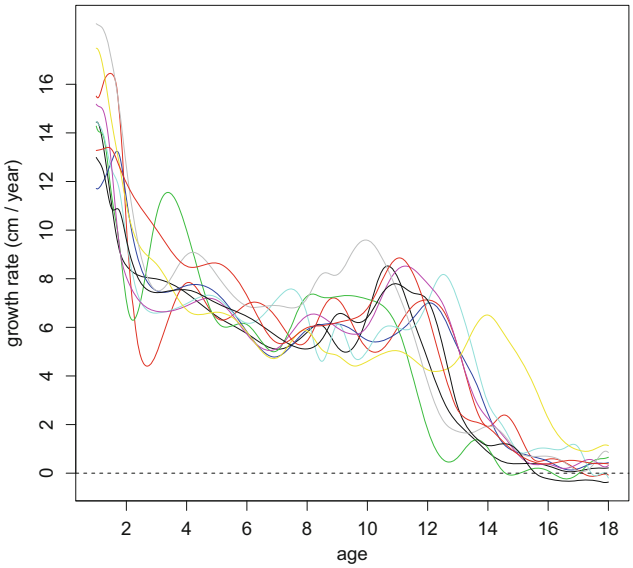


Fig. 4.1: Derivatives of growth curves of ten girls from the Berkeley dataset. The data and the code for the figure are from the R package fda (Ramsay et al. [111])

which we derive an artificial example that illustrates the difficulties of estimation in the presence of phase variation. Let A and B be symmetric random variables and consider the random function

$$\tilde{Y}(x) = A\sin[8\pi(x+B)]. \tag{4.1}$$

(Strictly speaking, $x \mapsto x+B$ is not from $[0,1]$ to itself; for illustration purposes, we assume in this example that $K = \mathbb{R}$.) The random variable A generates the amplitude

variation, while B represents the phase variation. In Fig. 4.2, we plot four realisations and the resulting empirical means for the two extreme scenarios where $B = 0$ (no phase variation) or $A = 1$ (no amplitude variation). In the left panel of the figure, we see that the sample mean (in thick blue) lies between the observations and has a similar form, so can be viewed as the curve representing the typical realisation of the random curve. This is in contrast to the right panel, where the mean is qualitatively different from all curves in the sample: though periodicity is still present, the peaks and troughs have been flattened, and the sample mean is much more diffuse than any of the observations.

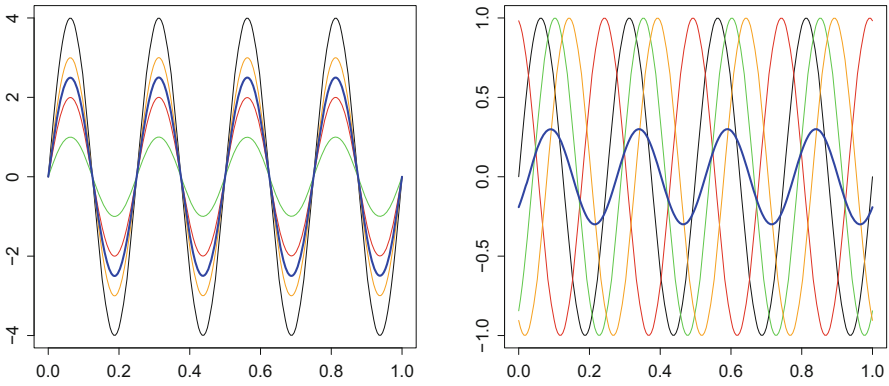


Fig. 4.2: Four realisations of (4.1) with means in thick blue. Left: amplitude variation ($B = 0$); right: phase variation ($A = 1$)

The phenomenon illustrated in Fig. 4.2 is hardly surprising, since as mentioned earlier amplitude variation is linear while phase variation is not, and taking sample means is a linear operation. Let us see in formulae how this phenomenon occurs. When $A = 1$ we have

$$\mathbb{E}\tilde{Y}(x) = \sin(8\pi x)\mathbb{E}[\cos(8\pi B)] + \cos(8\pi x)\mathbb{E}[\sin(8\pi B)].$$

Since B is symmetric the second term vanishes, and unless B is trivial the expectation of the cosine is smaller than one in absolute value. Consequently, the expectation of $\tilde{Y}(x)$ is the original function $\sin 8\pi x$ multiplied by a constant of magnitude strictly less than one, resulting in peaks of smaller magnitude.

In the general case, where $\tilde{Y}(x) = Y(T^{-1}(x))$ and Y and T are independent, we have

$$\mathbb{E}\tilde{Y}(x) = \mathbb{E}[m(T^{-1}(x))]$$

and

$$\operatorname{cov}[\tilde{Y}(x), \tilde{Y}(y)] = \mathbb{E}[\kappa(T^{-1}(x), T^{-1}(y))] + \operatorname{cov}[m(T^{-1}(x), m(T^{-1}(y))].$$

From this, several conclusions can be drawn. Let $\tilde{\mu} = \mu(T^{-1}(x))$ be the conditional mean function given T. Then the value of the mean function itself, $\mathbb{E}\tilde{\mu}$, at x_0 is determined not by a single point, say x, but rather by all the values of m at the possible outcomes of $T^{-1}(x)$. In particular, if x_0 was a local maximum for m, then $\mathbb{E}[\tilde{\mu}(x_0)]$ will typically be strictly smaller than $m(x_0)$; the phase variation results in smearing m.

At this point an important remark should be made. Whether or not phase variation is problematic depends on the specific application. If one is interested indeed in the mean and covariance functions of $\tilde{Y}$, then the standard empirical estimators will be consistent, since $\tilde{Y}$ itself is a random function. But if it is rather m, the mean of Y, that is of interest, then the confounding of the amplitude and phase variation will lead to inconsistency. This can also be seen from the formula

$$\tilde{Y}(x) = m(T^{-1}(x)) + \sum_{k=1}^{\infty} \xi_k \phi_k(T^{-1}(x)).$$

The above series is *not* the Karhunen–Loève expansion of $\tilde{Y}$; the simplest way to notice this is the observation that $\phi_k(T^{-1}(x))$ includes both the functional component ϕ_k and the random component $T^{-1}(x)$. The true Karhunen–Loève expansion of $\tilde{Y}$ will in general be qualitatively very different from that of Y, not only in terms of the mean function but also in terms of the covariance operator and, consequently, its eigenfunctions and eigenvalues. As illustrated in the trigonometric example, the typical situation is that the mean $\mathbb{E}\tilde{Y}$ is more diffuse than m, and the decay of the eigenvalues $\tilde{r}_k$ of the covariance operator is slower than that of r_k; as a result, one needs to truncate the sum at high threshold in order to capture a substantial enough part of the variability. In the toy example (4.1), the Karhunen–Loève expansion has a single term besides the mean if $B = 0$, while having two terms if $A = 1$.

When one is indeed interested in the mean m and the covariance κ, the random function T pertaining to the phase variation is a nuisance parameter. Given a sample $\tilde{Y}_i = Y_i \circ T_i^{-1}$, $i = 1, \ldots, n$, there is no point in taking pointwise means of $\tilde{Y}_i$, because the curves are *misaligned*; $\tilde{Y}_1(x) = Y_1(T_1^{-1}(x))$ should not be compared with $\tilde{Y}_2(x)$, but rather with $Y_2(T_1^{-1}(x)) = \tilde{Y}_2(T_1^{-1}(T_2(x))$. To overcome this difficulty, one seeks estimators $\widehat{T}_i$ such that

$$\widehat{Y}_i(x) = \tilde{Y}_i(\widehat{T}_i(x)) = Y_i(T_i^{-1}(\widehat{T}_i(x)))$$

is approximately $Y_i(x)$. In other words, one tries to align the curves in the sample to have a common time scale. Such a procedure is called *curve registration*. Once registration has been carried out, one proceeds the analysis on $\widehat{Y}_i(x)$ assuming only amplitude variation is now present: estimate the mean m by

$$\widehat{m}(x) = \frac{1}{n} \sum_{i=1}^{n} \widehat{Y}_i(x)$$

and the covariance κ by its analogous counterpart. Put differently, registering the curves amounts to *separating the two types of variation*. This step is crucial regardless of whether the warp functions are considered as nuisance or an analysis of the warp functions is of interest in the particular application.

There is an obvious identifiability problem in the model $\tilde{Y} = Y \circ T^{-1}$. If S is any (deterministic) invertible function, then the model with (Y, T) is statistically indistinguishable from the model with $(Y \circ S, T \circ S)$. It is therefore often assumed that $\mathbb{E}T = \mathbf{i}$ is the identity and in addition, in nearly all application, that T is monotonically increasing (if $d = 1$).

Discretely observed data. One cannot measure the height of person at every single instant of her life. In other words, it is rare in practice that one has access to the entire curve. A far more common situation is that one observes the curves *discretely*, i.e., at a finite number of points. The conceptually simplest setting is that one has access to a grid $x_1, \ldots, x_J \in K$, and the data come in the form

$$\tilde{y}_{ij} = \tilde{Y}_i(t_j),$$

possibly with measurement error. The problem is to find, given $\tilde{y}_{ij}$, consistent estimators of T_i and of the original, aligned functions Y_i.

In the bibliographic notes, we review some methods for carrying out this separation of amplitude and phase variation. It is fair to say that no single registration method arises as the canonical solution to the functional registration problem. Indeed, most need to make additional structural and/or smoothness assumptions on the warp maps, further to the basic identifiability conditions requiring that T be increasing and that $\mathbb{E}T$ equal the identity. We will eventually see that, in contrast, the case of point processes (viewed as discretely observed random measures) admits a canonical framework, without needing additional assumptions.

4.1.2 The Point Process Case

A point process is the mathematical object that represents the intuitive notion of a random collection of points in a space $\mathscr{X}$. It is formally defined as a measurable map Π from a generic probability space into the space of (possibly infinite) Borel integer-valued measures of $\mathscr{X}$ in such a way that $\Pi(B)$ is a measurable real-valued random variable for all Borel subsets B of $\mathscr{X}$. The quantity $\Pi(B)$ represents the random number of points observed in the set B. Among the plethora of books on point processes, let us mention Daley and Vere-Jones [41] and Karr [79]. Kallenberg [75] treats more general objects, *random measures*, of which point processes are a peculiar special case. We will assume for convenience that Π is a measure on a compact subset $K \subset \mathbb{R}^d$.

Amplitude variation of Π can be understood in analogy with the functional case. One defines the mean measure

$$\lambda(A) = \mathbb{E}[\Pi(A)], \qquad A \subset K \text{ Borel}$$

and, provided that $\mathbb{E}[\Pi(K)]^2 < \infty$, the covariance measure

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

the latter being a finite signed Borel measure on K. Just like in the functional case, these two objects encapsulate the (second-order) amplitude variation[3] properties of the law of Π.

Given a sample $\Pi_1, \ldots, \Pi_n$ of independent point processes distributed as Π, the natural estimators

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

are consistent and the former asymptotically normal [79, Proposition 4.8].

Phase variation then pertains to a random warp function $T : K \to K$ (independent of Π) that deforms Π: if we denote the points of Π by $x_1, \ldots, x_K$ (with K random), then instead of (x_i), one observes $T(x_1), \ldots, T(x_K)$. In symbols, this means that the data arise as $\widetilde{\Pi} = T\#\Pi$. We refer to Π as the *original point processes*, and $\widetilde{\Pi}$ as the *warped point processes*. An example of 30 warped and unwarped point processes is shown in Fig. 4.3. The point patterns in both panels present a qualitatively similar structure: there are two peaks of high concentration of points, while few points appear between these peaks. The difference between the two panels is in the position and concentration of those peaks. In the left panel, only amplitude variation is present, and the location/concentration of the peaks is the same across all observations. In contrast, phase variation results in shifting the peaks to different places for each of the observations, while also smearing or sharpening them. Clearly, estimation of the mean measure of a subset A by averaging the number of observed points in A would not be satisfactory as an estimator of λ when carried out with the warped data. As in the functional case, it will only be consistent for the measure $\tilde{\lambda}$ defined by

$$\tilde{\lambda}(A) = \mathbb{E}[\lambda(T^{-1}(A))], \qquad A \subseteq \mathscr{X},$$

and $\tilde{\lambda} = \mathbb{E}[T\#\lambda]$ misses most (or at least a significant part) of the bimodal structure of λ and is far more diffuse.

[3] If the cumulative count process $\Gamma(t) = \Pi[0,t)$ is mean-square continuous, then the use of the term "amplitude variation" can be seen to remain natural, as $\Gamma(t)$ will admit a Karhunen–Loève expansion, with all stochasticity being attributable to the random amplitudes in the expansion.

Since Π and T are independent, the conditional expectation of $\widetilde{\Pi}$ given T is

$$\mathbb{E}[\widetilde{\Pi}(A)|T] = \mathbb{E}[\Pi(T^{-1}(A))|T] = \lambda(T^{-1}(A)) = [T\#\lambda](A).$$

Consequently, we refer to $\Lambda = T\#\lambda$ as the *conditional mean measure*. The problem of separation of amplitude and phase variation can now be stated as follows. On the basis of a sample $\widetilde{\Pi}_1, \ldots, \widetilde{\Pi}_n$, find estimators of (T_i) and (Π_i). Registering the point processes amounts to constructing estimators, *registration maps* $\widehat{T_i^{-1}}$, such that the aligned points

$$\widehat{\Pi_i} = \widehat{T_i^{-1}}\#\widetilde{\Pi}_i = [\widehat{T_i^{-1}} \circ T_i]\#\Pi_i$$

are close to the original points Π_i.

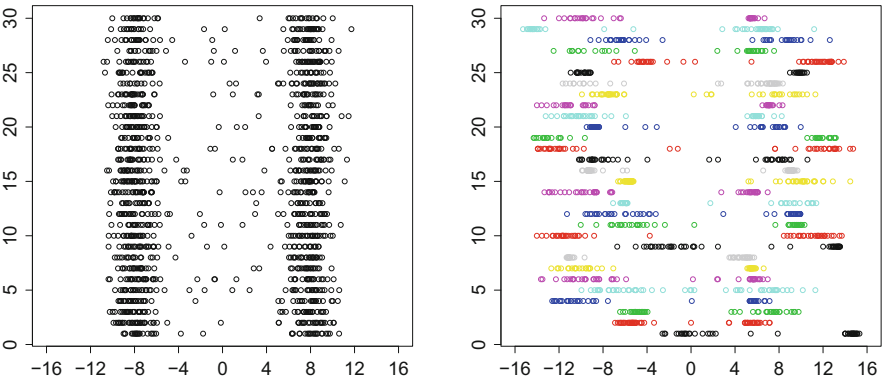


Fig. 4.3: Unwarped (left) and warped Poisson point processes

Remark 4.1.1 (Poisson Processes) *A special but important case is that of a Poisson process. Gaussian processes probably yield the most elegant and rich theory in functional data analysis, and so do Poisson processes when it comes to point processes. We say that Π is a* Poisson process *when the following two conditions hold. (1) For any disjoint collection $(A_1, \ldots, A_n)$ of Borel sets, the random variables $\Pi(A_1), \ldots, \Pi(A_n)$ are independent; and (2) for every Borel $A \subseteq K$, $\Pi(A)$ follows a Poisson distribution with mean $\lambda(A)$:*

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

Conditional upon T, the random variables $\widetilde{\Pi}(A_k) = \Pi(T^{-1}(A_k))$, $k = 1, \ldots, n$ are independent as the sets $(T^{-1}(A_k))$ are disjoint, and $\widetilde{\Pi}(A)$ follows a Poisson distribution with mean $\lambda(T^{-1}(A)) = \Lambda(A)$. This is precisely the definition of a Cox process*: conditional upon the* driving measure Λ*, $\widetilde{\Pi}$ is a Poisson process with mean measure λ. For this reason, it is also called a* doubly stochastic process*; in our context, the*

phase variation is associated with the stochasticity of Λ while the amplitude one is associated with the Poisson variation conditional upon Λ.

As in the functional case there are problems with identifiability: the model (Π, T) cannot be distinguished from the model $(S\#\Pi, T \circ S^{-1})$ for any invertible $S : K \to K$. It is thus natural to assume that $\mathbb{E}T$ is the identity map[4] (otherwise set $S = \mathbb{E}T$, i.e., replace Π by $[\mathbb{E}T]\#\Pi$ and T by $T \circ [\mathbb{E}T]^{-1}$).

Constraining T to have mean identity is nevertheless not sufficient for the model $\widetilde{\Pi} = T\#\Pi$ to be identifiable. The reason is that given the two point sets $\widetilde{\Pi}$ and Π, there are many functions that push forward the latter to the former. This ambiguity can be dealt with by assuming some sort of *regularity* or *parsimony* for T. For example, when $K = [a,b]$ is a subset of the real line, imposing T to be monotonically increasing guarantees its uniqueness. In multiple dimensions, there is no obvious analogue for increasing functions. One possible definition is the monotonicity described in Sect. 1.7.2:

$$\langle T(y) - T(x), y - x \rangle \geq 0, \qquad x, y \in K.$$

This property is rather weak in a sense we describe now. Let $K \subseteq \mathbb{R}^2$ and write $y \geq x$ if and only if $y_i \geq x_i$ for $i = 1,2$. It is natural to expect the deformations to maintain the *lexicographic order* in $\mathbb{R}^2$:

$$y \geq x \quad \Longrightarrow \quad T(y) \geq T(x).$$

If we require in addition that the ordering must be preserved for all quadrants: for $z = T(x)$ and $w = T(y)$

$$\{y_1 \geq x_1, y_2 \leq x_2\} \qquad \Longrightarrow \{w_1 \geq z_1, w_2 \leq z_2\},$$

then monotonicity is automatically satisfied. In that sense, it is arguably not very restrictive.

Monotonicity is weaker than cyclical monotonicity (see (1.10) with $y_i = T(x_i)$), which is itself equivalent to the property of being the subgradient of a convex function. But if extra smoothness is present and T is a gradient of some function $\phi : K \to \mathbb{R}$, then ϕ must be convex and T is then cyclically monotone. Consequently, we will make the following assumptions:

- the expected value of T is the identity;
- T is a gradient of a convex function.

In the functional case, at least on the real line, these two conditions are imposed on the warp functions in virtually all applications, often accompanied with additional assumptions about smoothness of T, its structural properties, or its distance from the identity. In the next section, we show how these two conditions alone lead to the Wasserstein geometry and open the door to consistent, fully nonparametric separation of the amplitude and phase variation.

[4] This can be defined as Bochner integral in the space of measurable bounded $T : K \to K$.

4.2 Wasserstein Geometry and Phase Variation

4.2.1 Equivariance Properties of the Wasserstein Distance

A first hint to the relevance of Wasserstein metrics in $\mathscr{W}_p(\mathscr{X})$ for deformations of the space $\mathscr{X}$ is that for all $p \geq 1$ and all $x, y \in \mathscr{X}$,

$$W_p(\delta_x, \delta_y) = \|x - y\|,$$

where δ_x is as usual the Dirac measure at $x \in \mathscr{X}$. This is in contrast to metrics such as the bounded Lipschitz distance (that metrises weak convergence) or the total variation distance on $P(\mathscr{X})$. Recall that these are defined by

$$\|\mu - \nu\|_{\mathrm{BL}} = \sup_{\|\varphi\|_{\mathrm{BL}} \leq 1} \left| \int_{\mathscr{X}} \varphi \, \mathrm{d}\mu - \int_{\mathscr{X}} \varphi \, \mathrm{d}\nu \right|; \qquad \|\mu - \nu\|_{\mathrm{TV}} = \sup_A |\mu(A) - \nu(A)|,$$

so that

$$\|\delta_x - \delta_y\|_{\mathrm{BL}} = \min(1, \|x - y\|); \qquad \|\delta_x - \delta_y\|_{\mathrm{TV}} = \begin{cases} 1 & x \neq y \\ 0 & x = y. \end{cases}$$

In words, the total variation metric "does not see the geometry" of the space $\mathscr{X}$. This is less so for the bounded Lipschitz distance that does take small distances into account but not large ones.

Another property (shared by BL and TV) is equivariance with respect to translations. It is more convenient to state it using the probabilistic formalism of Sect. 1.2. Let $X \sim \mu$ and $Y \sim \nu$ be random elements in $\mathscr{X}$, a be a fixed point in $\mathscr{X}$, $X' = X + a$ and $Y' = Y + a$. Joint couplings $Z' = (X', Y')$ are precisely those that take the form $(a, a) + Z$ for a joint coupling $Z = (X, Y)$. Thus

$$W_p(\mu * \delta_a, \nu * \delta_a) = W_p(X' + a, Y' + a) = W_p(X, Y) = W_p(\mu, \nu),$$

where δ_a is a Dirac measure at a and $*$ denotes convolution.

This carries over to Fréchet means in an obvious way.

Lemma 4.2.1 (Fréchet Means and Translations) *Let Λ be a random measure in $\mathscr{W}_2(\mathscr{X})$ with finite Fréchet functional and $a \in \mathscr{X}$. Then γ is a Fréchet mean of Λ if and only if $\gamma * \delta_a$ is a Fréchet mean of $\Lambda * \delta_a$.*

The result holds for other values of p, in the formulation sketched in the bibliographic notes of Chap. 2. In the quadratic case, one has a simple extension to the case where only one measure is translated. Denote the first moment (mean) of $\mu \in \mathscr{W}_1(\mathscr{X})$ by

$$m : \mathscr{W}_1(\mathscr{X}) \to \mathscr{X} \qquad m(\mu) = \int_{\mathscr{X}} x \, \mathrm{d}\mu(x).$$

(When $\mathscr{X}$ is infinite-dimensional, this can be defined as the unique element $m \in \mathscr{X}$ satisfying

$$\langle m, y \rangle = \int_{\mathscr{X}} \langle x, y \rangle \, \mathrm{d}\mu(x), \qquad y \in \mathscr{X}.)$$

By an equivalence of couplings similar to above, we obtain

$$W_2^2(\mu * \delta_a, \nu) = W_2^2(\mu, \nu) + (a - [m(\mu) - m(\nu)])^2 - [m(\mu) - m(\nu)]^2,$$

which is minimised at $a = m(\mu) - m(\nu)$. This leads to the following conclusion:

Proposition 4.2.2 (First Moment of Fréchet Mean) *Let Λ be a random measure in $\mathscr{W}_2(\mathscr{X})$ with finite Fréchet functional and Fréchet mean γ. Then*

$$\int_{\mathscr{X}} x \, \mathrm{d}\gamma(x) = \mathbb{E}\left[\int_{\mathscr{X}} x \, \mathrm{d}\Lambda(x) \right].$$

4.2.2 Canonicity of Wasserstein Distance in Measuring Phase Variation

The purpose of this subsection is to show that the standard functional data analysis assumptions on the warp function T, having mean identity and being increasing, are equivalent to purely geometric conditions on T and the conditional mean measure $\Lambda = T\#\lambda$. Put differently, if one is willing to assume that $\mathbb{E}T = \mathbf{i}$ and that T is increasing, then one is led *unequivocally* to the problem of estimation of Fréchet means in the Wasserstein space $\mathscr{W}_2(\mathscr{X})$. When $\mathscr{X} \neq \mathbb{R}$, "increasing" is interpreted as being the gradient of a convex function, as explained at the end of Sect. 4.1.2.

The total mass $\lambda(\mathscr{X})$ is invariant under the push-forward operation, and when it is finite, we may assume without loss of generality that it is equal to one, because all the relevant quantities scale with the total mass. Indeed, if $\lambda = \tau\mu$ with μ probability measure and $\tau > 0$, then $T\#\lambda = \tau \times T\#\mu$, and the Wasserstein distance (defined as the infimum-over-coupling integrated cost) between $\tau\mu$ and $\tau\nu$ is $\tau W_p(\mu, \nu)$ for μ, ν probabilities.

We begin with the one-dimensional case, where the explicit formulae allow for a more transparent argument, and for simplicity we will assume some regularity.

Assumptions 2 *The domain $K \subset \mathbb{R}$ is a nonempty compact convex set (an interval), and the continuous and injective random map $T : K \to \mathbb{R}$ (a random element in $C_b(K)$) satisfies the following two conditions:*

(A1) Unbiasedness: $\mathbb{E}[T(x)] = x$ *for all* $x \in K$.
(A2) Regularity: T *is monotone increasing.*

The relevance of the Wasserstein geometry to phase variation becomes clear in the following proposition that shows that Assumptions 2 are equivalent to geometric assumptions on the Wasserstein space $\mathscr{W}_2(\mathbb{R})$.

Proposition 4.2.3 (Mean Identity Warp Functions and Fréchet Means in $\mathscr{W}_2(\mathbb{R})$) *Let $\phi \subset K \subset \mathbb{R}$ compact and convex and $T : K \to \mathbb{R}$ continuous. Then Assumptions 2 hold if and only if, for any $\lambda \in \mathscr{W}_2(K)$ supported on K such that $\mathbb{E}[W_2^2(T\#\lambda, \lambda)] < \infty$, the following two conditions are satisfied:*

(B1) Unbiasedness: *for any $\theta \in \mathscr{W}_2(\mathbb{R})$*

$$\mathbb{E}[W_2^2(T\#\lambda, \lambda)] \leq \mathbb{E}[W_2^2(T\#\lambda, \theta)].$$

(B2) Regularity: *if $Q : K \to \mathbb{R}$ is such that $T\#\lambda = Q\#\lambda$, then with probability one*

$$\int_K \left|T(x) - x\right|^2 \mathrm{d}\lambda(x) \leq \int_K \left|Q(x) - x\right|^2 \mathrm{d}\lambda(x), \qquad \textit{almost surely}.$$

These assumptions have a clear interpretation: (B1) stipulates that λ is a Fréchet mean of the random measure $\Lambda = T\#\lambda$, while (B2) states that T must be the optimal map from λ to Λ, that is, $T = \mathbf{t}_\lambda^\Lambda$.

Proof. If T satisfies (B2) then, as an optimal map, it must be nondecreasing λ-almost surely. Since λ is arbitrary, T must be nondecreasing on the entire domain K. Conversely, if T is nondecreasing, then it is optimal for any λ. Hence (A2) and (B2) are equivalent.

Assuming (A2), we now show that (A1) and (B1) are equivalent. Condition (B1) is equivalent to the assertion that for all $\theta \in \mathscr{W}_2(\mathbb{R})$,

$$\mathbb{E}\|F_{T\#\lambda}^{-1} - F_\lambda^{-1}\|_{L_2(0,1)}^2 = \mathbb{E}[W_2^2(T\#\lambda, \lambda)] \leq \mathbb{E}[W_2^2(T\#\lambda, \theta)] = \mathbb{E}\|F_{T\#\lambda}^{-1} - F_\theta^{-1}\|_{L_2(0,1)}^2,$$

which is in turn equivalent to $\mathbb{E}[F_{T\#\lambda}]^{-1}] = \mathbb{E}[F_\Lambda^{-1}] = F_\lambda^{-1}$ (see Sect. 3.1.4). Condition (A2) and the assumptions on T imply that $F_\Lambda(x) = F_\lambda(T^{-1}(x))$. Suppose that F_λ is invertible (i.e., continuous and strictly increasing on K). Then $F_\Lambda^{-1}(u) = T(F_\lambda^{-1}(u))$. Thus (B1) is equivalent to $\mathbb{E}T(x) = x$ for all x in the range of F_λ^{-1}, which is K. The assertion that (A1) implies (B1), even if F_λ is not invertible, is proven in the next theorem (Theorem 4.2.4) in a more general context.

The situation in more than one dimension is similar but the proof is less transparent. To avoid compactness assumptions, we introduce the following power growth condition (taken from Agueh and Carlier [2]) of continuous functions that grow like $\|\cdot\|^q$ ($q \geq 0$):

$$G_q(\mathscr{X}) = (1 + \|\cdot\|^q)C_b(\mathscr{X}) = \left\{ f : \mathscr{X} \to \mathbb{R} \text{ continuous} : \sup_{x \in \mathscr{X}} \frac{|f(x)|}{1 + \|x\|^q} < \infty \right\}$$

with the norm $\|f\|_{G_q} = \sup |f(x)|/(1 + \|x\|^q) = \|f/(1 + \|\cdot\|^q)\|_\infty$. The space $G_q(\mathscr{X}, \mathscr{X})$ is defined similarly, with f taking values in $\mathscr{X}$ instead of $\mathbb{R}$, and the norm will be denoted in the same way. These are nonseparable Banach spaces.

Theorem 4.2.4 (Mean Identity Warp Functions and Fréchet Means) *Fix $\lambda \in P(\mathscr{X})$ and let $\mathbf{t} \in G_1(\mathscr{X}, \mathscr{X})$ be a (Bochner measurable) random optimal map*

with (Bochner) mean identity and such that $\mathbb{E}\|\mathbf{t}\|_{G_1} < \infty$. *Then* $\Lambda = \mathbf{t}\#\lambda$ *has Fréchet mean* λ*:*

$$\mathbb{E}[W_2^2(\lambda,\Lambda)] \le \mathbb{E}[W_2^2(\theta,\Lambda)] \qquad \forall \theta \in \mathscr{W}_2(\mathscr{X}).$$

The generalisation with respect to the one-dimensional result is threefold. Firstly, since our main interest is the implication (A1–A2) $\Rightarrow$ (B1–B2), we need not assume T to be injective. Secondly, the support of λ is not required to be compact. Lastly, the result holds in arbitrary dimension, including infinite-dimensional separable Hilbert spaces $\mathscr{X}$. In particular, if $\mathbf{t}$ is a linear map, then $\|\mathbf{t}\|_{G_1}$ coincides with the operator norm of $\mathbf{t}$, so the assumption is that $\mathbf{t}$ be a bounded self-adjoint nonnegative operator with mean identity and finite expected operator norm.

Proof. Optimality of $\mathbf{t}$ ensures that it has a convex potential ϕ, and strong and weak duality give

$$W_2^2(\lambda,\Lambda) = \int_{\mathscr{X}} \left(\frac{1}{2}\|x\|^2 - \phi(x)\right) \mathrm{d}\lambda(x) + \int_{\mathscr{X}} \left(\frac{1}{2}\|y\|^2 - \phi^*(y)\right) \mathrm{d}\Lambda(y);$$
$$W_2^2(\theta,\Lambda) \ge \int_{\mathscr{X}} \left(\frac{1}{2}\|x\|^2 - \phi(x)\right) \mathrm{d}\theta(x) + \int_{\mathscr{X}} \left(\frac{1}{2}\|y\|^2 - \phi^*(y)\right) \mathrm{d}\Lambda(y).$$

Formally taking expectations, using Fubini's theorem and that $\mathbb{E}\phi = \|\cdot\|^2/2$ (since $\mathbb{E}\mathbf{t}$ is the identity) yields

$$\mathbb{E}[W_2^2(\theta,\Lambda)] \ge \int_{\mathscr{X}} \left(\frac{1}{2}\|x\|^2 - \mathbb{E}\phi(x)\right) \mathrm{d}\theta(x) + \mathbb{E}\left[\int_{\mathscr{X}} \left(\frac{1}{2}\|y\|^2 - \phi^*(y)\right) \mathrm{d}\Lambda(y)\right] = \mathbb{E}[W_2^2(\lambda,\Lambda)]$$

as required. The rigorous mathematical justification for this is given on page 88 in the supplement.

Remark 4.2.5 *The "natural" space for* $\mathbf{t}$ *would be* $\mathscr{L}_2(\lambda)$*, but without the continuity assumption, the result may fail (Álvarez-Esteban et al. [9, Example 3.1]). A simple argument shows that the growth condition imposed by the* G_1 *assumption is minimal; see page 89 in the supplement or Galasso et al. [58].*

Remark 4.2.6 *The same statement holds if* $\mathscr{X}$ *is replaced by a (Borel) convex subset* K *thereof. The integrals will then be taken on* K*, showing that* λ *minimises the Fréchet functional among measures supported on* K*, and, by continuity, on* $\overline{K}$*. By Proposition 3.2.4,* λ *is a Fréchet mean.*

4.3 Estimation of Fréchet Means

4.3.1 Oracle Case

In view of the canonicity of the Wasserstein geometry in Sect. 4.2.2, separation of amplitude and phase variation of the point processes $\widetilde{\Pi}_i$ essentially requires comput-

ing Fréchet means in the 2-Wasserstein space. It is both conceptually important and technically convenient to introduce the case where an oracle reveals the conditional mean measures $\Lambda = T\#\lambda$ entirely. Thus, assuming that $\lambda \in \mathscr{W}_2(\mathscr{X})$ is the unique Fréchet mean of a random measure Λ, the goal is to estimate the structural mean λ on the basis of independent and identically distributed realisations $\Lambda_1, \ldots, \Lambda_n$ of λ.

Given that λ is defined as the minimiser of the Fréchet functional

$$F(\gamma) = \frac{1}{2}\mathbb{E}W_2^2(\Lambda, \gamma), \qquad \gamma \in \mathscr{W}_2(\mathscr{X}),$$

it is natural to estimate λ by a minimiser, say λ_n, of the empirical Fréchet functional

$$F_n(\gamma) = \frac{1}{2n}\sum_{i=1}^{n} W_2^2(\Lambda_i, \gamma), \qquad \gamma \in \mathscr{W}_2(\mathscr{X}).$$

A minimiser λ_n exists by Corollary 3.1.3. When $\mathscr{X} = \mathbb{R}$, λ_n can be seen to be an *unbiased* estimator of λ in a generalised sense of Lehmann [88] (see Sect. 4.3.5).

The warp maps (and their inverses) can then be estimated as the optimal maps from λ_n to each Λ_i (see Sect. 4.3.4).

4.3.2 Discretely Observed Measures

In practice, one does not have the fortune of fully observing the inherently infinite-dimensional objects $\Lambda_1, \ldots, \Lambda_n$. A far more realistic scenario is that one only has access to a discrete version of Λ_i, say $\widetilde{\Lambda}_i$. The simplest situation is when $\widetilde{\Lambda}_i$ arises as an empirical measure of the form $\tau^{-1}\sum_{i=1}^{\tau}\delta\{Y_j\}$, where Y_j are independent with distribution Λ_i. More generally, $\widetilde{\Lambda}_i$ can be a normalised point process $\widetilde{\Pi}_i$ with mean measure $\tau\Lambda_i$, i.e.

$$\widetilde{\Lambda}_i = \frac{1}{\widetilde{\Pi}_i(\mathscr{X})}\widetilde{\Pi}_i \quad \text{with} \quad \mathbb{E}[\widetilde{\Pi}_i(A)|\Lambda_i] = \tau\Lambda_i(A), \qquad A \subseteq \mathscr{X} \text{ Borel}.$$

This encapsulates the case of empirical measure when τ is an integer and $\widetilde{\Pi}_i$ is a *binomial point process*. The parameter τ is the expected number of observed points over the entire space $\mathscr{X}$; the larger τ is, the more information $\widetilde{\Pi}_i$ gives on Λ_i.

Except if $\widetilde{\Lambda}_i$ is an empirical measure, there is one difficulty in the above setting that needs to be addressed. Unless $\widetilde{\Pi}_i$ is binomial, there is a positive probability that $\widetilde{\Pi}_i(\mathscr{X}) = 0$ and no points pertaining to Λ_i are observed. In the asymptotic setup below, conditions will be imposed to ensure that this probability becomes negligible as $n \to \infty$. For concreteness we define $\widetilde{\Lambda}_i = \lambda^{(0)}$ for some fixed measure $\lambda^{(0)}$ that will be of minor importance. This can be a Dirac measure at 0, a certain fixed Gaussian measure, or (normalised) Lebesgue measure on some bounded set in case $\mathscr{X} = \mathbb{R}^d$. We can now replace the estimator λ_n by $\widetilde{\lambda}_n$, defined as any minimiser of

$$\widetilde{F}_n(\gamma) = \frac{1}{2n}\sum_{i=1}^{n} W_2^2(\widetilde{\Lambda}_i, \gamma), \qquad \gamma \in \mathscr{W}_2(\mathscr{X}),$$

which exists by Corollary 3.1.3.

As a generalisation of the discrete case discussed in Sect. 1.3, the Fréchet mean of discrete measures can be computed exactly. Suppose that $N_i = \widetilde{\Pi}_i(\mathscr{X})$ is nonzero for all i. Then each $\widetilde{\Lambda}_i$ is a discrete measure supported on N_i points. One can then recast the multimarginal formulation (see Sect. 3.1.2) as a finite linear program, solve it, and "average" the solution as in Proposition 3.1.2 in order to obtain $\widetilde{\lambda}_n$ (an alternative linear programming formulation for finding a Fréchet mean is given by Anderes et al. [14]). Thus, $\widetilde{\lambda}_n$ can be computed in finite time, even when $\mathscr{X}$ is infinite-dimensional.

Finally, a remark about measurability is in order. Point processes can be viewed as random elements in $M_+(\mathscr{X})$ endowed with the *vague topology* induced from convergence of integrals of continuous functions with compact support. If μ_n converge to μ vaguely, and a_n are numbers that converge to a, then $a_n\mu_n \to a\mu$ vaguely. Thus, $\widetilde{\Lambda}_i$ is a continuous function of the pair $(\widetilde{\Pi}_i, \widetilde{\Pi}_i(\mathscr{X}))$ and can be viewed as a random measure with respect to the vague topology. The restriction of the vague topology to probability measures is equivalent to the weak topology,[5] and therefore vague, weak, and Wasserstein measurability are all equivalent.

4.3.3 Smoothing

Even when the computational complexity involved in calculating $\widetilde{\lambda}_n$ is tractable, there is another reason not to use it as an estimator for λ. If one has a-priori knowledge that λ is smooth, it is often desirable to estimate it by a smooth measure. One way to achieve this would be to apply some smoothing technique to $\widetilde{\lambda}_n$ using, e.g., kernel density estimation. However, unless the number of observed points from each measure is the same $N_1 = \cdots = N_n = N$, $\widetilde{\lambda}_n$ will usually be concentrated on many points, essentially $N_1 + \cdots + N_n$ of them. In other words, the Fréchet mean is concentrated on many more points than each of the measures $\widetilde{\Lambda}_i$, thus potentially hindering its usefulness as a mean because it will not be a representative of the sample.

This is most easily seen when $\mathscr{X} = \mathbb{R}$, in which case each $\widetilde{\Lambda}_i$ is a discrete uniform measure on points $x_1^i < x_2^i < \cdots < x_{N_i}^i$, where we assume for simplicity that the points are not repeated (this will happen with probability one if Λ_i is diffuse). If we now set G_i to be the distribution function of $\widetilde{\Lambda}_i$, then the quantile function G_i^{-1} is piecewise constant on each interval $(k, k+1]/N_i$ with jumps at

$$G_i^{-1}(k/N_i) = x_k^i, \qquad k = 1, 2, \ldots, N_i.$$

[5] In finite dimensional (or more generally, locally compact metric) spaces. If $\mathscr{X}$ is an infinite-dimensional Hilbert space, the vague topology is trivial. This is stated and proved as Lemma 5 on page 27 in the supplement.

The Fréchet mean has quantile function $G^{-1}(u) = n^{-1}\sum G_i^{-1}(u)$ and will have jumps at every point of the form k/N_i for $k \leq N_i$ and $i = 1,\ldots,n$. In the worst-case scenario, when no pair from N_i has a common divisor, there will be

$$\left(\sum_{i=1}^{n} N_i - 1\right) + 1 = \left(\sum_{i=1}^{n} N_i\right) - n + 1$$

jumps for G^{-1}, which is the number of points on which the Fréchet mean will be supported. (All the G_i^{-1}'s have a jump at one which thus needs to be counted once rather than n times.)

By counting the number of redundancies in the constraints matrix of the linear program, one can show that this is in general an upper bound on the number of support points of the Fréchet mean.

An alternative approach is to first smooth each observation $\widetilde{\lambda}_n$ and then calculate the Fréchet mean. Since it is easy to bound the Wasserstein distances when dealing with convolutions, we will employ kernel density estimation, although other smoothing approaches could be used as well.

To simplify the exposition, we provide the technical details only when $\mathscr{X} = \mathbb{R}^d$, but a similar construction will work when the dimension of $\mathscr{X}$ is infinite. Let $\psi : \mathbb{R}^d \to (0,\infty)$ be a continuous, bounded, strictly positive isotropic density function with unit variance: $\psi(x) = \psi_1(\|x\|)$ with ψ_1 nonincreasing and

$$\int_{\mathbb{R}^d} \|x\|^2 \psi(x)\,\mathrm{d}x = 1 = \int_{\mathbb{R}^d} \psi(x)\,\mathrm{d}x.$$

(Besides the boundedness all these properties can be relaxed, and if $\mathscr{X} = \mathbb{R}$ even boundedness is not necessary.) A classical example for ψ is the standard Gaussian density in $\mathbb{R}^d$. Define the rescaled version $\psi_\sigma(x) = \sigma^{-d}\psi(x/\sigma)$ for all $\sigma > 0$. We can then replace $\widetilde{\Lambda}_i$ by a smooth proxy $\widetilde{\Lambda}_i * \psi_\sigma$. If $\widetilde{\Lambda}_i$ is a sum of Dirac masses at $x_1,\ldots,x_{N_i}$, then

$$\widetilde{\Lambda}_i * \psi_\sigma \quad \text{has density} \quad g(x) = \frac{1}{N_i}\sum_{j=1}^{N_i} \psi_\sigma(x - x_i).$$

If $N_i = 0$, one can either use $\lambda^{(0)}$ or $\lambda^{(0)} * \psi_\sigma$; this event will have negligible probability anyway.

For the purpose of approximating $\widetilde{\Lambda}_i$, this convolution is an acceptable estimator, because as was seen in the proof of Theorem 2.2.7,

$$W_2^2(\widetilde{\Lambda}_i, \widetilde{\Lambda}_i * \psi_\sigma) \leq \sigma^2.$$

But the measure $\widetilde{\Lambda}_i$ has a strictly positive density throughout $\mathbb{R}^d$. If we know that Λ is supported on a convex compact $K \subset \mathbb{R}^d$, it is desirable to construct an estimator that has the same support K. The first idea that comes to mind is to project $\widetilde{\Lambda}_i * \psi_\sigma$ to K (see Proposition 3.2.4), as this will further decrease the Wasserstein distance, but

the resulting measure will then have positive mass on the boundary of K, and will not be absolutely continuous. We will therefore use a different strategy: eliminate all the mass outside K and redistribute it on K. The simplest way to do this is to restrict $\widetilde{\Lambda}_i * \psi_\sigma$ to K and renormalise the restriction to be a probability measure. For technical reasons, it will be more convenient to bound the Wasserstein distance when the restriction and renormalisation is done separately on each point of $\widetilde{\Lambda}_i$. This yields the measure

$$\widehat{\Lambda}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} \frac{\delta\{x_j\} * \psi_\sigma}{[\delta\{x_j\} * \psi_\sigma](K)}\bigg|_K, \tag{4.2}$$

Lemma 4.4.2 below shows that $W_2^2(\widetilde{\Lambda}_i, \widehat{\Lambda}_i) \leq C\sigma^2$ for some finite constant C. It is apparent that $\widehat{\Lambda}_i$ is a continuous function of $\widetilde{\Lambda}_i$ and σ, so $\widehat{\Lambda}_i$ is measurable; in any case this is not particularly important because σ will vanish, so $\widehat{\Lambda}_i = \widetilde{\Lambda}_i$ asymptotically and the latter is measurable.

Our final estimator $\widehat{\lambda}_n$ for λ is defined as the minimiser of

$$\widehat{F}_n(\gamma) = \frac{1}{2n} \sum_{i=1}^{n} W_2^2(\widehat{\Lambda}_i, \gamma), \qquad \gamma \in \mathscr{W}_2(\mathscr{X}).$$

Since the measures $\widehat{\Lambda}_i$ are absolutely continuous, $\widehat{\lambda}_n$ is unique. We refer to $\widehat{\lambda}_n$ as the *regularised Fréchet–Wasserstein estimator*, where the regularisation comes from the smoothing and the possible restriction to K.

In the case $\mathscr{X} = \mathbb{R}$, $\widehat{\lambda}_n$ can be constructed via averaging of quantile functions. Let $\widehat{G}_i$ be the distribution function of $\widehat{\Lambda}_i$. Then $\widehat{\lambda}_n$ is the measure with quantile function

$$F^{-1}_{\widehat{\lambda}_n}(u) = \frac{1}{n} \sum_{i=1}^{n} \widehat{G}_i^{-1}(u), \qquad u \in (0,1),$$

and distribution function

$$F_{\widehat{\lambda}_n}(x) = [F^{-1}_{\widehat{\lambda}_n}]^{-1}(x).$$

By construction, the $\widehat{G}_i$ are continuous and strictly increasing, so the inverses are proper inverses and one does not to use the right-continuous inverse as in Sect. 3.1.4.

If $\mathscr{X} = \mathbb{R}^d$ and $d \geq 2$, then there is no explicit expression for $\widehat{\lambda}_n$, although it exists and is unique. In the next chapter, we present a steepest descent algorithm that approximately constructs $\widehat{\lambda}_n$ by taking advantage of the differentiability properties of the Fréchet functional $\widehat{F}_n$ in Sect. 3.1.6.

4.3.4 Estimation of Warpings and Registration Maps

Once estimators $\widehat{\Lambda}_i$, $i = 1, \ldots, n$ and $\widehat{\lambda}_n$ are constructed, it is natural to estimate the map $T_i = \mathbf{t}_\lambda^{\Lambda_i}$ and its inverse $T_i^{-1} = \mathbf{t}_{\Lambda_i}^{\lambda}$ (when Λ_i are absolutely continuous; see the discussion after Assumptions 3 below) by the plug-in estimators

$$\widehat{T}_i = \mathbf{t}_{\widehat{\lambda}_n}^{\widehat{\Lambda}_i}, \qquad \widehat{T_i^{-1}} = (\widehat{T}_i)^{-1} = \mathbf{t}_{\widehat{\Lambda}_i}^{\widehat{\lambda}_n}.$$

The latter, the registration maps, can then be used in order to register the points Π_i via

$$\widehat{\Pi_i^{(n)}} = \widehat{T_i^{-1}} \# \widetilde{\Pi}_i^{(n)} = \left[\widehat{T_i^{-1}} \circ T_i\right] \# \Pi_i^{(n)}.$$

It is thus reasonable to expect that if $\widehat{T_i^{-1}}$ is a good estimator, then its composition with T_i should be close to the identity and $\widehat{\Pi}_i$ should be close to Π_i.

4.3.5 Unbiased Estimation When $\mathscr{X} = \mathbb{R}$

In the same way, Fréchet means extend the notion of mean to non-Hilbertian spaces, they also extend the definition of unbiased estimators. Let H be a separable Hilbert space (or a convex subset thereof) and suppose that $\widehat{\theta}$ is a random element in H whose distribution μ_θ depends on a parameter $\theta \in H$. Then $\widehat{\theta}$ is *unbiased* for θ if for all $\theta \in H$

$$\mathbb{E}_\theta \widehat{\theta} = \int_H x \, \mathrm{d}\mu_\theta(x) = \theta.$$

(We use the standard notation $\mathbb{E}_\theta g(\widehat{\theta}) = \int g(x) \, \mathrm{d}\mu_\theta(x)$ in the sequel.) This is equivalent to

$$\mathbb{E}_\theta \|\theta - \widehat{\theta}\|^2 \leq \mathbb{E}_\theta \|\gamma - \widehat{\theta}\|^2, \qquad \forall \theta, \gamma \in H.$$

In view of that, one can define unbiased estimators of $\lambda \in \mathscr{W}_2$ as measurable functions $\delta = \delta(\Lambda_1, \ldots, \Lambda_n)$ for which

$$\mathbb{E}_\lambda W_2^2(\lambda, \delta) \leq \mathbb{E}_\lambda W_2^2(\gamma, \delta), \qquad \forall \gamma, \theta \in \mathscr{W}_2.$$

This definition was introduced by Lehmann [88].

Unbiased estimators allow us to avoid the problem of over-registering (the so-called pinching effect; Kneip and Ramsay [82, Section 2.4]; Marron et al. [90, p. 476]). An extreme example of over-registration is if one "aligns" all the observed patterns into a single fixed point x_0. The registration will then seem "successful" in the sense of having no residual phase variation, but the estimation is clearly biased because the points are not registered to the correct reference measure. Thus, requiring the estimator to be unbiased is an alternative to penalising the registration maps.

Due to the Hilbert space embedding of $\mathscr{W}_2(\mathbb{R})$, it is possible to characterise unbiased estimators in terms of a simple condition on their quantile functions. As a corollary, λ_n, the Fréchet mean of $\{\Lambda_1, \ldots, \Lambda_n\}$, is unbiased. Our regularised Fréchet–Wasserstein estimator $\hat{\lambda}_n$ can then be interpreted as *approximately unbiased*, since it approximates the unobservable λ_n.

Proposition 4.3.1 (Unbiased Estimators in $\mathscr{W}_2(\mathbb{R})$) *Let Λ be a random measure in $\mathscr{W}_2(\mathbb{R})$ with finite Fréchet functional and let λ be the unique Fréchet mean of Λ (Theorem 3.2.11). An estimator δ constructed as a function of a sample $(\Lambda_1, \dots, \Lambda_n)$ is unbiased for λ if and only if the left-continuous representatives (in $L_2(0,1)$) satisfy $\mathbb{E}[F_\delta^{-1}(x)] = F_\lambda^{-1}(x)$ for all $x \in (0,1)$.*

Proof. The proof is straightforward from the definition: δ is unbiased if and only if for all λ and all γ,

$$\mathbb{E}_\lambda \|F_\lambda^{-1} - F_\delta^{-1}\|_{L_2}^2 \leq \mathbb{E}_\lambda \|F_\gamma^{-1} - F_\delta^{-1}\|_{L_2}^2,$$

which is equivalent to $\mathbb{E}_\lambda[F_\delta^{-1}] = F_\lambda^{-1}$. In other words, these two functions must equal almost everywhere on $(0,1)$, and their left-continuous representatives must equal everywhere (the fact that $\mathbb{E}_\lambda[F_\delta^{-1}]$ has such a representative was established in Sect. 3.1.4).

To show that $\delta = \lambda_n$ is unbiased, we simply invoke Theorem 3.2.11 twice to see that

$$\mathbb{E}[F_\delta^{-1}] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n F_{\Lambda_i}^{-1}\right] = \mathbb{E}[F_\Lambda^{-1}] = F_\lambda^{-1},$$

which proves unbiasedness of δ.

4.4 Consistency

In functional data analysis, one often assumes that the number of curves n and the number of observed points per curve m both diverge to infinity. An analogous framework for point processes would similarly require the number of point processes n as well as the expected number of points τ per processes to diverge. A technical complication arises, however, because the mean measures do not suffice to characterise the distribution of the processes. Indeed, if one is given a point processes Π with mean measure λ (not necessarily a probability measure), and τ is an integer, there is no unique way to define a process $\Pi^{(\tau)}$ with mean measure $\tau\lambda$. One can define $\Pi^{(\tau)} = \tau\Pi$, so that every point in Π will be counted τ times. Such a construction, however, can never yield a consistent estimator of λ, even when $\tau \to \infty$.

Another way to generate a point process with mean measure $\tau\lambda$ is to take a superposition of τ independent copies of Π. In symbols, this means

$$\Pi^{(\tau)} = \Pi_1 + \cdots + \Pi_\tau,$$

with (Π_i) independent, each having the same distribution as Π. This superposition scheme gives the possibility to use the law of large numbers. If τ is not an integer, then this construction is not well-defined but can be made so by assuming that the distribution of Π is *infinitely divisible*. The reader willing to assume that τ is always

an integer can safely skip to Sect. 4.4.1; all the main ideas are developed first for integer values of τ and then extended to the general case.

A point process Π is infinitely divisible if for every integer m there exists a collection of m independent and identically distributed $\Pi_i^{(1/m)}$ such that

$$\Pi = \Pi_1^{(1/m)} + \cdots + \Pi_m^{(1/m)} \qquad \text{in distribution.}$$

If Π is infinitely divisible and $\tau = k/m$ is rational, then can define $\pi^{(\tau)}$ using km independent copies of $\Pi^{(1/m)}$:

$$\Pi^{(\tau)} = \sum_{i=1}^{km} \Pi_i^{(1/m)}.$$

One then deals with irrational τ via duality and continuity arguments, as follows. Define the *Laplace functional* of Π by

$$L_\Pi(f) = \mathbb{E}\left[e^{-\Pi f}\right] = \mathbb{E}\left[\exp\left(-\int_{\mathscr{X}} f \,\mathrm{d}\Pi\right)\right] \in [0,1], \qquad f : \mathscr{X} \to \mathbb{R}_+ \quad \text{Borel.}$$

The Laplace functional characterises the distribution of the point process, generalising the notion of Laplace transform of a random variable or vector (Karr [79, Theorem 1.12]). By definition, it translates convolutions into products. When $\Pi = \Pi^{(1)}$ is infinitely divisible, the Laplace functional L_1 of Π takes the form (Kallenberg [75, Chapter 6]; Karr [79, Theorem 1.43])

$$L_1(f) = \mathbb{E}\left[e^{-\Pi^{(1)} f}\right] = \exp\left[-\int_{M_+(\mathscr{X})} (1 - e^{-\mu f}) \,\mathrm{d}\rho(\mu)\right] \quad \text{for some } \rho \in M_+(M_+(\mathscr{X})).$$

The Laplace functional of $\Pi^{(\tau)}$ is $L_\tau(f) = [L_1(f)]^\tau$ for any rational τ, which simply amounts to multiplying the measure ρ by the scalar τ. One can then do the same for an irrational τ, and the resulting Laplace functional determines the distribution of $\Pi^{(\tau)}$ for all $\tau > 0$.

4.4.1 Consistent Estimation of Fréchet Means

We are now ready to define our asymptotic setup. The following assumptions will be made. Notice that the Wasserstein geometry does not appear explicitly in these assumptions, but is rather *derived* from them in view of Theorem 4.2.4. The compactness requirement can be relaxed under further moment conditions on λ and the point process Π; we focus on the compact case for the simplicity and because in practice the point patterns will be observed on a bounded observation window.

Assumptions 3 *Let $K \subset \mathbb{R}^d$ be a compact convex nonempty set, λ an absolutely continuous probability measure on K, and τ_n a sequence of positive numbers. Let Π be a point processes on K with mean measure λ. Finally, define $U = \text{int}K$.*

- *For every n, let $\{\Pi_1^{(n)}, \ldots, \Pi_n^{(n)}\}$ be independent point processes, each having the same distribution as a superposition of τ_n copies of Π.*
- *Let T be a random injective function on K (viewed as a random element in $C_b(K,K)$ endowed with the supremum norm) such that $T(x) \in U$ for $x \in U$ (that is, $T \in C_b(U,U)$) with nonsingular derivative $\nabla T(x) \in \mathbb{R}^{d\times d}$ for almost all $x \in U$, that is a gradient of a convex function. Let $\{T_1, \ldots, T_n\}$ be independent and identically distributed as T.*
- *For every $x \in U$, assume that $\mathbb{E}[T(x)] = x$.*
- *Assume that the collections $\{T_n\}_{n=1}^{\infty}$ and $\{\Pi_i^{(n)}\}_{i\le n, n=1,2,\ldots}$ are independent.*
- *Let $\widetilde{\Pi}_i^{(n)} = T_i \# \Pi_i^{(n)}$ be the warped point processes, having* conditional mean measures $\Lambda_i = T_i \# \lambda = \tau_n^{-1} \mathbb{E}\left\{ \widetilde{\Pi}_i^{(n)} \middle| T_i \right\}$.
- *Define $\widehat{\Lambda}_i$ by the smoothing procedure (4.2), using bandwidth $\sigma_i^{(n)} \in [0,1]$ (possibly random).*

The dependence of the estimators on n will sometimes be tacit. But Λ_i does not depend on n.

By virtue of Theorem 4.2.4, λ is a Fréchet mean of the random measure $\Lambda = T\#\lambda$. Uniqueness of this Fréchet mean will follow from Proposition 3.2.7 if we show that Λ is absolutely continuous with positive probability. This is indeed the case, since T is injective and has a nonsingular Jacobian matrix; see Ambrosio et al. [12, Lemma 5.5.3]. The Jacobian assumption can be removed when $\mathscr{X} = \mathbb{R}$, because Fréchet means are always unique by Theorem 3.2.11.

Notice that there is no assumption about the dependence between rows. Assumptions 3 thus cover, in particular, two different scenarios:

- *Full independence*: here the point processes are independent across rows, that is, $\Pi_i^{(n)}$ and $\Pi_i^{(n+1)}$ are also independent.
- *Nested observations*: here $\Pi_i^{(n+1)}$ includes the same points as $\Pi_i^{(n)}$ and additional points, that is, $\Pi_i^{(n+1)}$ is a superposition of $\Pi_i^{(n)}$ and another point process distributed as $(\tau_{n+1} - \tau_n)\Pi$.

Needless to say, Assumptions 3 also encompass binomial processes when τ_n are integers, as well as Poisson processes or, more generally, Poisson cluster processes.

We now state and prove the consistency result for the estimators of the conditional mean measures Λ_i and the structural mean measure λ.

Theorem 4.4.1 (Consistency) *If Assumptions 3 hold, $\sigma_n = n^{-1}\sum_{i=1}^n \sigma_i^{(n)} \to 0$ almost surely and $\tau_n \to \infty$ as $n \to \infty$, then:*

1. *The estimators $\widehat{\Lambda}_i$ defined by (4.2), constructed with bandwidth $\sigma = \sigma_i^{(n)}$, are Wasserstein-consistent for the conditional mean measures: for all i such that $\sigma_i^{(n)} \xrightarrow{p} 0$*

$$W_2\left(\widehat{\Lambda}_i, \Lambda_i\right) \xrightarrow{p} 0, \qquad \text{as } n \to \infty;$$

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

$$W_2(\widehat{\lambda}_n, \lambda) \xrightarrow{a.s.} 0, \qquad \text{as } n \to \infty.$$

Convergence in 1. holds almost surely under the additional conditions that $\sum_{n=1}^{\infty} \tau_n^{-2} < \infty$ *and* $\mathbb{E}\left[\Pi(\mathbb{R}^d)\right]^4 < \infty$. *If* $\sigma_n \to 0$ *only in probability, then convergence in 2. still holds in probability.*

Theorem 4.4.1 still holds without smoothing ($\sigma_n = 0$). In that case, $\widehat{\lambda}_n = \tilde{\lambda}_n$ is possibly not unique, and the theorem should be interpreted in a set-valued sense (as in Proposition 1.7.8): almost surely, *any* choice of minimisers $\tilde{\lambda}_n$ converges to λ as $n \to \infty$.

The preceding paragraph notwithstanding, we will usually assume that some smoothing *is* present, in which case $\widehat{\lambda}_n$ is unique and absolutely continuous by Proposition 3.1.8. The uniform Lipschitz bounds for the objective function show that if we restrict the relevant measures to be absolutely continuous, then $\widehat{\lambda}_n$ is a continuous function of $(\widehat{\Lambda_1}, \ldots, \widehat{\Lambda_n})$ and hence $\widehat{\lambda}_n : (\Omega, \mathscr{F}, \mathbb{P}) \to \mathscr{W}_2(K)$ is measurable; this is again a minor issue because many arguments in the proof hold for each $\omega \in \Omega$ separately. Thus, even if $\widehat{\lambda}_n$ is not measurable, the proof shows that the convergence holds outer almost surely or in outer probability.

The first step in proving consistency is to show that the Wasserstein distance between the unsmoothed and the smoothed estimators of Λ_i vanishes with the smoothing parameter. The exact rate of decay will be important to later establish the rate of convergence of $\widehat{\lambda}_n$ to λ, and is determined next.

Lemma 4.4.2 (Smoothing Error) *There exists a finite constant* $C_{\psi,K}$, *depending only on* ψ *and on* K, *such that*

$$W_2^2\left(\widehat{\Lambda}_i, \widetilde{\Lambda}_i\right) \leq C_{\psi,K}\sigma^2 \quad \text{if } \sigma \leq 1. \tag{4.3}$$

Since the smoothing parameter will anyway vanish, this restriction to small values of σ is not binding. The constant $C_{\psi,K}$ is explicit. When $\mathscr{X} = \mathbb{R}$, a more refined construction allows to improve this constant in some situations, see Panaretos and Zemel [100, Lemma 1].

Proof. The idea is that (4.2) is a sum of measures with mass $1/N_i$ that can be all sent to the relevant point x_j, and we refer to page 98 in the supplement for the precise details.

Proof (Proof of Theorem 4.4.1). The proof, detailed on page 97 of the supplement, follows the following steps: firstly, one shows the convergence in probability of $\widehat{\Lambda}_i$ to Λ_i. This is basically a corollary of Karr [79, Proposition 4.8] and the smoothing bound from Lemma 4.4.2.

To prove claim (2) one considers the functionals, defined on $\mathscr{W}_2(K)$:

$$
\begin{aligned}
F(\gamma) &= \frac{1}{2}\mathbb{E}W_2^2(\Lambda,\gamma);\\
F_n(\gamma) &= \frac{1}{2n}\sum_{i=1}^{n} W_2^2(\Lambda_i,\gamma);\\
\tilde{F}_n(\gamma) &= \frac{1}{2n}\sum_{i=1}^{n} W_2^2(\widetilde{\Lambda}_i,\gamma), \qquad \widetilde{\Lambda}_i = \frac{\widetilde{\Pi}_i^{(n)}}{N_i^{(n)}} \qquad \text{or } \lambda^{(0)} \text{ if } N_i^{(n)} = 0;\\
\widehat{F}_n(\gamma) &= \frac{1}{2n}\sum_{i=1}^{n} W_2^2(\widehat{\Lambda}_i,\gamma), \qquad \widehat{\Lambda}_i = \lambda^{(0)} \text{ if } N_i^{(n)} = 0.
\end{aligned}
$$

Since K is compact, they are all locally Lipschitz, so their differences can be controlled by the distances between Λ_i, $\widetilde{\Lambda}_i$, and $\widehat{\Lambda}_i$. The first distance vanishes since the intensity $\tau \to \infty$, and the second by the smoothing bound. Another compactness argument yields that $\widehat{F}_n \to F$ uniformly on $\mathscr{W}_2(K)$, and so the minimisers converge.

The almost sure convergence in (1) is proven as follows. Under the stronger conditions at the end of the theorem's statement, for any fixed $a = (a_1,\ldots,a_d) \in \mathbb{R}^d$,

$$
\mathbb{P}\left(\frac{\widetilde{\Pi}_i^{(n)}((-\infty,a])}{\tau_n} - \Lambda_i((-\infty,a]) \to 0\right) = 1
$$

by the law of large numbers. This extends to all rational a's, then to all a by approximation. The smoothing error is again controlled by Lemma 4.4.2.

4.4.2 Consistency of Warp Functions and Inverses

We next discuss the consistency of the warp and registration function estimators. These are key elements in order to align the observed point patterns $\widetilde{\Pi}_i$. Recall that we have consistent estimators $\widehat{\Lambda}_i$ for Λ_i and $\widehat{\lambda}_n$ for λ. Then $T_i = \mathbf{t}_{\lambda}^{\Lambda_i}$ is estimated by $\mathbf{t}_{\widehat{\lambda}_n}^{\widehat{\Lambda}_i}$ and T_i^{-1} is estimated by $\mathbf{t}_{\widehat{\Lambda}_i}^{\widehat{\lambda}_n}$. We will make the following extra assumptions that lead to more transparent statements (otherwise one needs to replace K with the set of Lebesgue points of the supports of λ and Λ_i).

Assumptions 4 (Strictly Positive Measures) *In addition to Assumptions 3 suppose that:*

1. *λ has a positive density on K (in particular,* $\operatorname{supp}\lambda = K$*);*
2. *T is almost surely surjective on $U = \operatorname{int}K$ (thus a homeomorphism of U).*

As a consequence $\operatorname{supp}\Lambda = \operatorname{supp}(T\#\lambda) = \overline{T(\operatorname{supp}\lambda)} = K$ almost surely.

Theorem 4.4.3 (Consistency of Optimal Maps) *Let Assumptions 4 be satisfied in addition to the hypotheses of Theorem 4.4.1. Then for any i such that $\sigma_i^{(n)} \xrightarrow{p} 0$ and any compact set $S \subseteq \mathrm{int}K$,*

$$\sup_{x\in S} \|\widehat{T_i^{-1}}(x) - T_i^{-1}(x)\| \xrightarrow{\mathrm{p}} 0, \qquad \sup_{x\in S} \|\widehat{T}_i(x) - T_i(x)\| \xrightarrow{\mathrm{p}} 0.$$

Almost sure convergence can be obtained under the same provisions made at the end of the statement of Theorem 4.4.1.

A few technical remarks are in order. First and foremost, it is not clear that the two suprema are measurable. Even though T_i and T_i^{-1} are random elements in $C_b(U, \mathbb{R}^d)$, their estimators are only defined in an L_2 sense. The proof of Theorem 4.4.3 is done ω-wise. That is, for any ω in the probability space such that Theorem 4.4.1 holds, the two suprema vanish as $n \to \infty$. In other words, the convergence holds in outer probability or outer almost surely.

Secondly, assuming positive smoothing, the random measures $\widehat{\Lambda}_i$ are smooth with densities bounded below on K, so $\widehat{T_i^{-1}}$ are defined on the whole of U (possibly as set-valued functions on a Λ_i-null set). But the only known regularity result for $\widehat{\lambda}_n$ is an upper bound on its density (Proposition 3.1.8), so it is unclear what is its support and consequently what is the domain of definition of $\widehat{T}_i$.

Lastly, when the smoothing parameter σ is zero, $\widehat{T}_i$ and $\widehat{T_i^{-1}}$ are not defined. Nevertheless, Theorem 4.4.3 still holds in the set-valued formulation of Proposition 1.7.11, of which it is a rather simple corollary:

Proof (Proof of Theorem 4.4.3). The proof amounts to setting the scene in order to apply Proposition 1.7.11 of stability of optimal maps. We define

$$\mu_n = \widehat{\Lambda}_i; \qquad \nu_n = \widehat{\lambda}_n; \qquad \mu = \Lambda_i; \qquad \nu = \lambda; \qquad u_n = \widehat{T}_i^{-1}; \qquad u = T_i^{-1},$$

and verify the conditions of the proposition. The weak convergence of μ_n to μ and ν_n to ν is the conclusion of Theorem 4.4.1; the finiteness is apparent because K is compact and the uniqueness follows from the assumed absolute continuity of Λ_i. Since in addition T_i^{-1} is uniquely defined on $U = \mathrm{int}K$ which is an open convex set, the restrictions on Ω in Proposition 1.7.11 are redundant. Uniform convergence of $\widehat{T}_i$ to T_i is proven in the same way.

Corollary 4.4.4 (Consistency of Point Pattern Registration) *For any i such that $\sigma_i^{(n)} \xrightarrow{p} 0$,*

$$W_2\left(\frac{\widehat{\Pi_i^{(n)}}}{N_i^{(n)}}, \frac{\Pi_i^{(n)}}{N_i^{(n)}}\right) \xrightarrow{\mathrm{p}} 0.$$

The division by the number of observed points ensures that the resulting measures are probability measures; the relevant information is contained in the point patterns themselves, and is invariant under this normalisation.

Proof. The law of large numbers entails that $N_i^{(n)}/\tau_n \to 1$, so in particular $N_i^{(n)}$ is almost surely not zero when n is large. Since $\widehat{\Pi_i^{(n)}} = (\widehat{T_i^{-1}} \circ T_i)\#\Pi_i^{(n)}$, we have the upper bound

$$W_2^2\left(\frac{\widehat{\Pi_i^{(n)}}}{N_i^{(n)}}, \frac{\Pi_i^{(n)}}{N_i^{(n)}}\right) \le \int_K \|\widehat{T_i^{-1}}(T_i(x)) - x\|^2 \,\mathrm{d}\frac{\Pi_i^{(n)}}{N_i^{(n)}}.$$

Fix a compact $\Omega \subseteq \mathrm{int}K$ and split the integral to Ω and its complement. Then

$$\int_{K\setminus\Omega} \|\widehat{T_i^{-1}}(T_i(x)) - x\|^2 \,\mathrm{d}\frac{\Pi_i^{(n)}}{N_i^{(n)}} \le d_K^2 \frac{\Pi_i^{(n)}(K\setminus\Omega)}{\tau_n}\frac{\tau_n}{N_i^{(n)}} \xrightarrow{\text{as}} d_K^2 \lambda(K\setminus\Omega),$$

by the law of large numbers, where d_K is the diameter of K. By writing int$K$ as a countable union of compact sets (and since λ is absolutely continuous), this can be made arbitrarily small by choice of Ω.

We can easily bound the integral on Ω by

$$\int_\Omega \|\widehat{T_i^{-1}}(T_i(x)) - x\|^2 \,\mathrm{d}\frac{\Pi_i^{(n)}}{N_i^{(n)}} \le \sup_{x\in\Omega} \|\widehat{T_i^{-1}}(T_i(x)) - x\|^2 = \sup_{y\in T_i(\Omega)} \|\widehat{T_i^{-1}}(y) - T_i^{-1}(y)\|^2.$$

But $T_i(\Omega)$ is a compact subset of $U = \mathrm{int}K$, because $T_i \in C_b(U,U)$. The right-hand side therefore vanishes as $n \to \infty$ by Theorem 4.4.3, and this completes the proof.

Possible extensions pertaining to the boundary of K are discussed on page 33 of the supplement.

4.5 Illustrative Examples

In this section, we illustrate the estimation framework put forth in this chapter by considering an example of a structural mean λ with a bimodal density on the real line. The unwarped point patterns Π originate from Poisson processes with mean measure λ and, consequently, the warped points $\widetilde{\Pi}$ are Cox processes (see Sect. 4.1.2). Another scenario involving triangular densities can be found in Panaretos and Zemel [100].

4.5.1 Explicit Classes of Warp Maps

As a first step, we introduce a class of random warp maps satisfying Assumptions 2, that is, increasing maps that have as mean the identity function. The construction is a mixture version of similar maps considered by Wang and Gasser [128, 129].

For any integer k define $\zeta_k : [0,1] \to [0,1]$ by

$$\zeta_0(x) = x, \qquad \zeta_k(x) = x - \frac{\sin(\pi k x)}{|k|\pi}, \quad k \in \mathbb{Z} \setminus \{0\}. \tag{4.4}$$

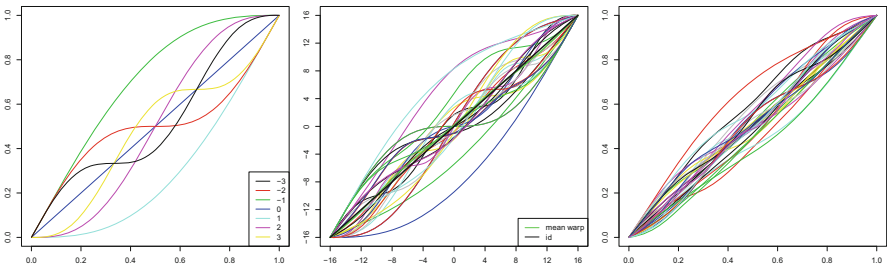
Clearly $\zeta_k(0) = 0$, $\zeta_k(1) = 1$ and ζ_k is smooth and strictly increasing for all k. Figure 4.4a plots ζ_k for $k = -3, \ldots, 3$. To make ζ_k a random function, we let k be an integer-valued random variable. If the latter is symmetric, then we have

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

By means of mixtures, we replace this discrete family by a continuous one: let $J > 1$ be an integer and $V = (V_1, \ldots, V_J)$ be a random vector following the flat Dirichlet distribution (uniform on the set of nonnegative vectors with $v_1 + \cdots + v_J = 1$). Take independently k_j following the same distribution as k and define

$$T(x) = \sum_{j=1}^{J} V_j \zeta_{k_j}(x). \tag{4.5}$$

Since V_j is positive, T is increasing and as (V_j) sums up to unity T has mean identity. Realisations of these warp functions are given in Fig. 4.4b and c for $J = 2$ and $J = 10$, respectively. The parameters (k_j) were chosen as symmetrised Poisson random variables: each k_j has the law of XY with X Poisson with mean 3 and $\mathbb{P}(Y = 1) = \mathbb{P}(Y = -1) = 1/2$ for Y and X independent. When $J = 10$ is large, the function T deviates only mildly from the identity, since a law of large numbers begins to take effect. In contrast, $J = 2$ yields functions that are quite different from the identity. Thus, it can be said that the parameter J controls the variance of the random warp function T.



(a) (b) (c)

Fig. 4.4: (**a**) The functions $\{\zeta_{-3}, \ldots, \zeta_3\}$; (**b**) realisations of T defined by (4.5) with $J = 2$ and k_j symmetrisations of Poisson random variables with mean 3; (**c**) realisations of T defined by (4.5) with $J = 10$ and k_j as in (**b**)

4.5.2 Bimodal Cox Processes

Let the structural mean measure λ be a mixture of a bimodal Gaussian distribution (restricted to $K = [-16, 16]$) and a beta background on the interval $[-12, 12]$, so that mass is added at the centre of K but not near the boundary. In symbols this is given as follows. Let φ be the standard Gaussian density and let $\beta_{\alpha,\beta}$ denote the density of a the beta distribution with parameters α and β. Then λ is chosen as the measure with density

$$f(x) = \frac{1-\varepsilon}{2}[\varphi(x-8) + \varphi(x+8)] + \frac{\varepsilon}{24}\beta_{1.5,1.5}\left(\frac{x+12}{24}\right), \qquad x \in [-16, 16], \tag{4.6}$$

where $\varepsilon \in [0, 1]$ is the weight of the beta background. (We ignore the loss of a negligible amount of mass due to the restriction of the Gaussians to $[-16, 16]$.) Plots of the density and distribution functions are given in Fig. 4.5.

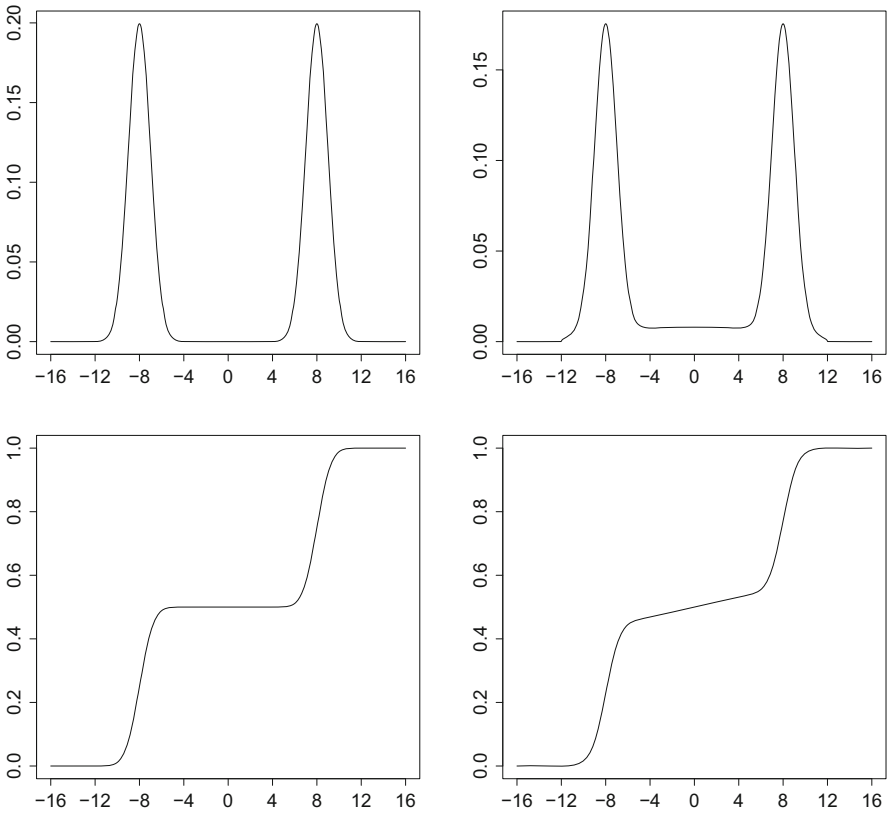


Fig. 4.5: Density and distribution functions corresponding to (4.6) with $\varepsilon = 0$ and $\varepsilon = 0.15$

The main criterion for the quality of our regularised Fréchet–Wasserstein estimator will be its success in discerning the two modes at ± 8; these will be smeared by the phase variation arising from the warp functions.

We next simulated 30 independent Poisson processes with mean measure λ, $\varepsilon = 0.1$, and total intensity (expected number of points) $\tau = 93$. In addition, we generated warp functions as in (4.5) but rescaled to $[-16, 16]$; that is, having the same law as the functions

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

from K to K. These cause rather violent phase variation, as can be seen by the plots of the densities and distribution functions of the conditional measures $\Lambda = T\#\lambda$ presented in Fig. 4.6a and b; the warped points themselves are displayed in Fig. 4.6c.

Using these warped point patterns, we construct the regularised Fréchet–Wasserstein estimator employing the procedure described in Sect. 4.3. Each $\widetilde{\Pi}_i$ was smoothed with a Gaussian kernel and bandwidth chosen by unbiased cross validation. We deviate slightly from the recipe presented in Sect. 4.3 by not restricting the resulting estimates to the interval $[-16, 16]$, but this has no essential effect on the finite sample performance. The regularised Fréchet–Wasserstein estimator $\widehat{\lambda}_n$ serves as the estimator of the structural mean λ and is shown in Fig. 4.7a. It is contrasted with λ at the level of distribution functions, as well as with the empirical arithmetic mean; the latter, the *naive estimator*, is calculated by ignoring the warping and simply averaging linearly the (smoothed) empirical distribution functions across the observations. We notice that $\widehat{\lambda}_n$ is rather successful at locating the two modes of λ, in contrast with the naive estimator that is more diffuse. In fact, its distribution function increases approximately linearly, suggesting a nearly constant density instead of the correct bimodal one.

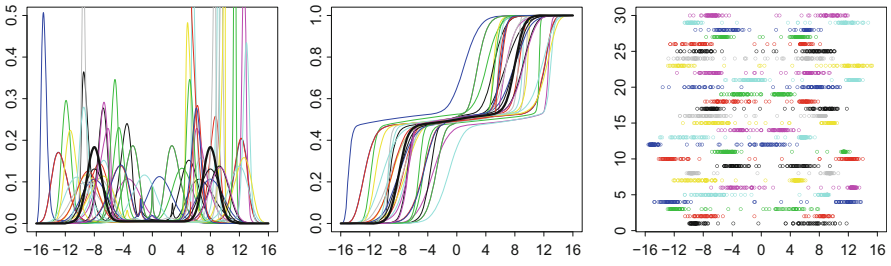
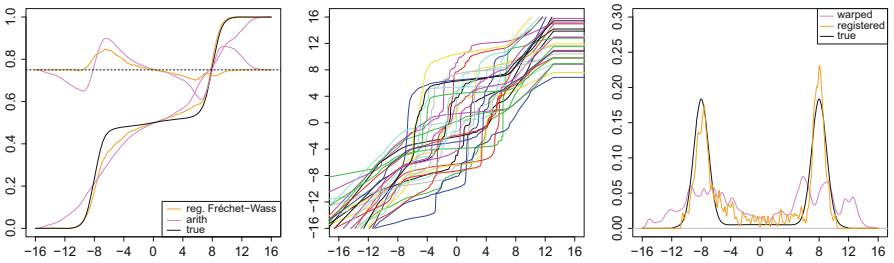


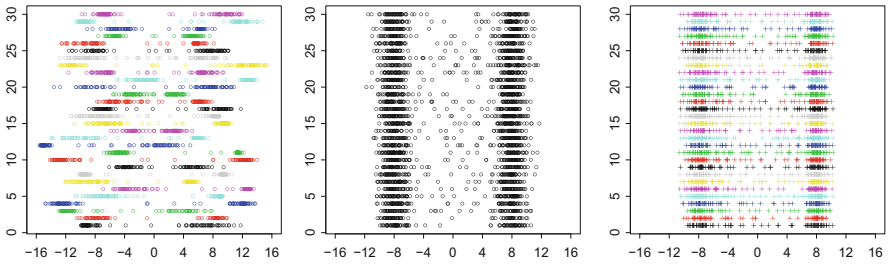
Fig. 4.6: (**a**) 30 warped bimodal densities, with density of λ given by (4.6) in solid black; (**b**) their corresponding distribution functions, with that of λ in solid black; (**c**) 30 Cox processes, constructed as warped versions of Poisson processes with mean intensity $93f$ using as warp functions the rescaling to $[-16.16]$ of (4.5)



(a) (b) (c)

Fig. 4.7: (**a**) Comparison between the regularised Fréchet–Wasserstein estimator, the empirical arithmetic mean, and the true distribution function, including residual curves centred at $y = 3/4$; (**b**) The estimated warp functions; (**c**) Kernel estimates of the density function f of the structural mean, based on the warped and registered point patterns

Estimators of the warp maps $\widehat{T}_i$, depicted in Fig. 4.7b, and their inverses, are defined as the optimal maps between $\widehat{\lambda}_n$ and the estimated conditional mean measures, as explained in Sect. 4.3.4. Then we register the point patterns by applying to them the inverse estimators $\widehat{T_i^{-1}}$ (Fig. 4.8). Figure 4.7c gives two kernel estimators of the density of λ constructed from a superposition of all the warped points and all the registered ones. Notice that the estimator that uses the registered points is much more successful than the one using the warped ones in discerning the two density peaks. This is not surprising after a brief look at Fig. 4.8, where the unwarped, warped, and registered points are displayed. Indeed, there is very high concentration of registered points around the true location of the peaks, ± 8. This is not the case for the warped points because of the phase variation that translates the centres of concentration for each individual observation. It is important to remark that the fluctuations in the density estimator in Fig. 4.7c are not related to the registration procedure, and could be reduced by a better choice of bandwidth. Indeed, our procedure does not attempt to estimate the density, but rather the distribution function.



(a) (b) (c)

Fig. 4.8: Bimodal Cox processes: (**a**) the observed warped point processes; (**b**) the unobserved original point processes; (**c**) the registered point processes

Figure 4.9 presents a superposition of the regularised Fréchet–Wasserstein estimators for 20 independent replications of the experiment, contrasted with a similar superposition for the naive estimator. The latter is clearly seen to be biased around the two peaks, while the regularised Fréchet–Wasserstein seems approximately unbiased, despite presenting fluctuations. It always captures the bimodal nature of the density, as is seen from the two clear elbows in each realisation.

To illustrate the consistency of the regularised Fréchet–Wasserstein estimator $\widehat{\lambda}_n$ for λ as shown in Theorem 4.4.1, we let the number of processes n as well as the expected number of observed point per process τ to vary. Figures 4.10 and 4.11 show the sampling variation of $\widehat{\lambda}_n$ for different values of n and τ. We observe that as either of these increases, the realisations $\widehat{\lambda}_n$ indeed approach λ. The figures suggest that, in this scenario, the amplitude variation plays a stronger role than the phase variation, as the effect of τ is more substantial.

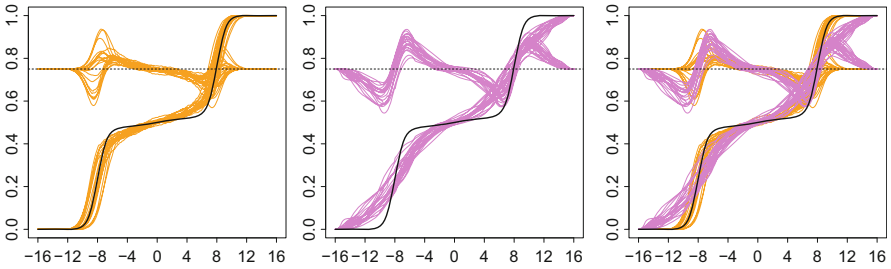


Fig. 4.9: (**a**) Sampling variation of the regularised Fréchet–Wasserstein mean $\widehat{\lambda}_n$ and the true mean measure λ for 20 independent replications of the experiment; (**b**) sampling variation of the arithmetic mean, and the true mean measure λ for the same 20 replications; (**c**) superposition of (**a**) and (**b**). For ease of comparison, all three panels include residual curves centred at $y = 3/4$

4.5.3 Effect of the Smoothing Parameter

In order to work with measures of strictly positive density, the observed point patterns have been smoothed using a kernel function. This necessarily incurs an additional bias that depends on the bandwidth σ_i. The asymptotics (Theorem 4.4.1) guarantee the consistency of the estimators, in particular the regularised Fréchet–Wasserstein estimator $\widehat{\lambda}_n$, provided that $\max_{i=1}^{n} \sigma_i \to 0$. In our simulations, we choose σ_i in a data-driven way by employing unbiased cross validation. To gauge for the effect of the smoothing, we carry out the same estimation procedure but with σ_i multiplied by a parameter s. Figure 4.12 presents the distribution function of $\widehat{\lambda}_n$ as a function of s. Interestingly, the curves are nearly identical as long as $s \leq 1$, whereas when $s > 1$, the bias becomes more substantial.

These findings are reaffirmed in Fig. 4.13 that show the registered point processes again as a function of s. We see that only minor differences are present as s varies from 0.1 to 1, for example, in the grey (8), black (17), and green (19) processes. When $s = 3$, the distortion becomes quite more substantial. This phenomenon repeats itself across all combinations of n, τ, and s tested.

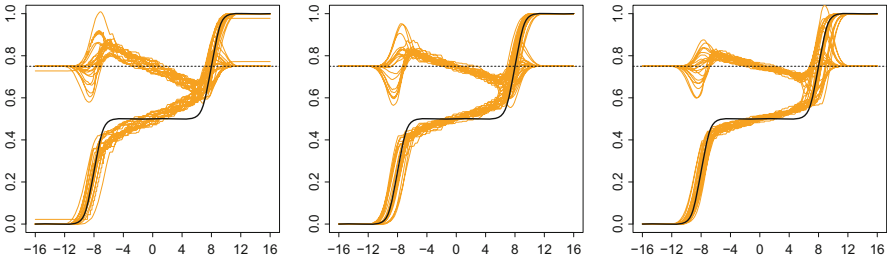


Fig. 4.10: Sampling variation of the regularised Fréchet–Wasserstein mean $\widehat{\lambda}_n$ and the true mean measure λ for 20 independent replications of the experiment, with $\varepsilon = 0$ and $n = 30$. Left: $\tau = 43$; middle: $\tau = 93$; right: $\tau = 143$. For ease of comparison, all three panels include residual curves centred at $y = 3/4$

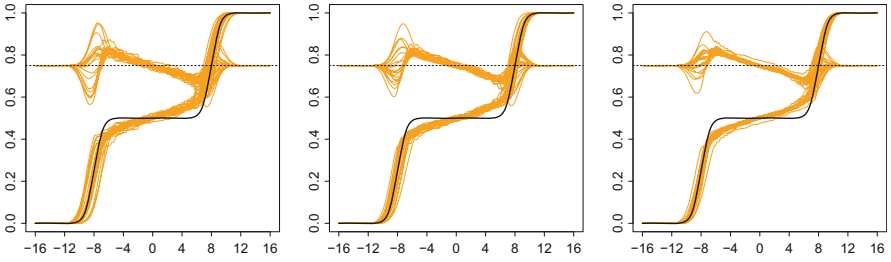


Fig. 4.11: Sampling variation of the regularised Fréchet–Wasserstein mean $\widehat{\lambda}_n$ and the true mean measure λ for 20 independent replications of the experiment, with $\varepsilon = 0$ and $\tau = 93$. Left: $n = 30$; middle: $n = 50$; right: $n = 70$. For ease of comparison, all three panels include residual curves centred at $y = 3/4$.

4.6 Convergence Rates and a Central Limit Theorem on the Real Line

Since the conditional mean measures Λ_i are discretely observed, the rate of convergence of our estimators will be affected by the rate at which the number of observed points per process $N_i^{(n)}$ increases to infinity. The latter is controlled by the next lemma, which is valid for any complete separable metric space $\mathscr{X}$.

Lemma 4.6.1 (Number of Points Grows Linearly) *Let $N_i^{(n)} = \Pi_i^{(n)}(\mathscr{X})$ denote the total number of observed points. If $\tau_n/\log n \to \infty$, then there exists a constant $C_\Pi > 0$, depending only on the distribution of Π, such that almost surely*

$$\liminf_{n\to\infty} \frac{\min_{1\le i\le n} N_i^{(n)}}{\tau_n} \ge C_\Pi.$$

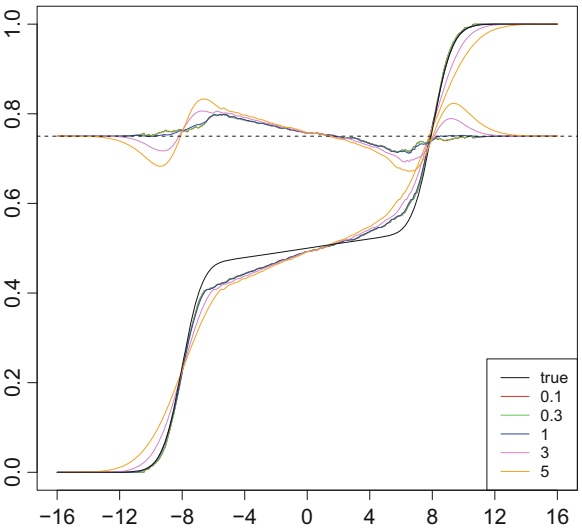


Fig. 4.12: Regularised Fréchet–Wasserstein mean as a function of the smoothing parameter multiplier s, including residual curves. Here, $n = 30$ and $\tau = 143$

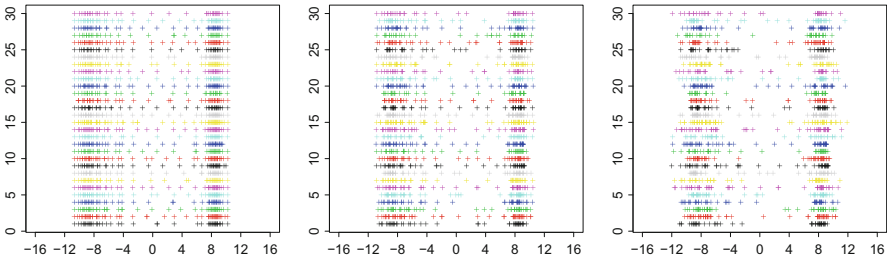


Fig. 4.13: Registered point processes as a function of the smoothing parameter multiplier s. Left: $s = 0.1$; middle: $s = 1$; right: $s = 3$. Here, $n = 30$ and $\tau = 43$

In particular, there are no empty point processes, so the normalisation is well-defined. If Π is a Poisson process, then we have the more precise result

$$\lim_{n\to\infty} \frac{\min_{1\le i\le n} N_i^{(n)}}{\tau_n} = 1 \qquad \textit{almost surely}.$$

Remark 4.6.2 *One can also show that the limit superior of the same quantity is bounded by a constant C'_Π. If $\tau_n/\log n$ is bounded below, then the same result holds but with worse constants. If only $\tau_n \to \infty$, then the result holds for each i separately but in probability.*

The proof is a simple application of Chernoff bounds; see page 108 in the supplement.

With Lemma 4.6.1 under our belt, we can replace terms of the order $\min_i N_i^{(n)}$ by the more transparent order τ_n. As in the consistency proof, the idea is to write

$$F - \widehat{F}_n = (F - F_n) + (F_n - \tilde{F}_n) + (\tilde{F}_n - \widehat{F}_n)$$

and control each term separately. The first term corresponds to the phase variation, and comes from the approximation of the theoretical expectation F by a sample mean F_n. The second term is associated with the amplitude variation resulting from observing Λ_i discretely. The third term can be viewed as the bias incurred by the smoothing procedure. Accordingly, the rate at which $\widehat{\lambda}_n$ converges to λ is a sum of three separate terms. We recall the standard $O_{\mathbb{P}}$ terminology: if X_n and Y_n are random variables, then $X_n = O_{\mathbb{P}}(Y_n)$ means that the sequence (X_n/Y_n) is *bounded in probability*, which by definition is the condition

$$\forall \varepsilon > 0 \; \exists M : \quad \sup_n \mathbb{P}\left(\left| \frac{X_n}{Y_n} \right| > M \right) < \varepsilon.$$

Instead of $X_n = O_{\mathbb{P}}(Y_n)$, we will sometimes write $Y_n \geq O_{\mathbb{P}}(X_n)$. The former notation emphasises the condition that X_n grows no faster than Y_n, while the latter stresses that Y_n grows at least as fast as X_n (which is of course the same assertion). Finally, $X_n = o_{\mathbb{P}}(Y_n)$ means that $X_n/Y_n \to 0$ in probability.

Theorem 4.6.3 (Convergence Rates on $\mathbb{R}$) *Suppose in addition to Assumptions 3 that $d = 1$, $\tau_n/\log n \to \infty$ and that Π is either a Poisson process or a binomial process. Then*

$$W_2(\widehat{\lambda}_n, \lambda) \leq O_{\mathbb{P}}\left(\frac{1}{\sqrt{n}}\right) + O_{\mathbb{P}}\left(\frac{1}{\sqrt[4]{\tau_n}}\right) + O_{\mathbb{P}}(\sigma_n), \qquad \sigma_n = \frac{1}{n}\sum_{i=1}^{n} \sigma_i^{(n)},$$

where all the constants in the $O_{\mathbb{P}}$ terms are explicit.

Remark 4.6.4 *Unlike classical density estimation, no assumptions on the rate of decay of σ_n are required, because we only need to estimate the distribution function and not the derivative. If the smoothing parameter is chosen to be $\sigma_i^{(n)} = [N_i^{(n)}]^{-\alpha}$ for some $\alpha > 0$ and $\tau_n/\log n \to \infty$, then by Lemma 4.6.1 $\sigma_n \leq \max_{1 \leq i \leq n} \sigma_i^{(n)} = O_{\mathbb{P}}(\tau_n^{-\alpha})$. For example, if* Rosenblatt's rule *$\alpha = 1/5$ is employed, then the $O_{\mathbb{P}}(\sigma_n)$ term can be replaced by $O_{\mathbb{P}}(1/\sqrt[5]{\tau_n})$.*

One can think about the parameter τ as separating the *sparse* and *dense* regimes as in classical functional data analysis (see also Wu et al. [132]). If τ is bounded, then the

setting is *ultra sparse* and consistency cannot be achieved. A sparse regime can be defined as the case where $\tau_n \to \infty$ but slower than $\log n$. In that case, consistency is guaranteed, but some point patterns will be empty. The *dense* regime can be defined as $\tau_n \gg n^2$, in which case the amplitude variation is negligible asymptotically when compared with the phase variation.

The exponent $-1/4$ of τ_n can be shown to be optimal without further assumptions, but it can be improved to $-1/2$ if $\mathbb{P}(f_\Lambda \geq \varepsilon \text{ on } K) = 1$ for some $\varepsilon > 0$, where f_Λ is the density of Λ (see Sect. 4.7). In terms of T, the condition is that $\mathbb{P}(T' \geq \varepsilon) = 1$ for some ε and λ has a density bounded below. When this is the case, τ_n needs to compared with n rather than n^2 in the next paragraph and the next theorem.

Theorem 4.6.3 provides conditions for the optimal parametric rate $\sqrt{n}$ to be achieved: this happens if we set σ_n to be of the order $O_{\mathbb{P}}(n^{-1/2})$ or less and if τ_n is of the order n^2 or more. But if the last two terms in Theorem 4.6.3 are *negligible* with respect to $n^{-1/2}$, then a sort of *central limit theorem* holds for $\widehat{\lambda}_n$:

Theorem 4.6.5 (Asymptotic Normality) *In addition to the conditions of Theorem 4.6.3, assume that $\tau_n/n^2 \to \infty$, $\sigma_n = o_{\mathbb{P}}(n^{-1/2})$ and λ possesses an invertible distribution function F_λ on K. Then*

$$\sqrt{n}\left(\mathbf{t}_\lambda^{\widehat{\lambda}_n} - \mathbf{i}\right) \longrightarrow Z \quad \textit{weakly in } L_2(\lambda),$$

for a zero-mean Gaussian process Z with the same covariance operator of T (the latter viewed as a random element in $L_2(\lambda)$), namely with covariance kernel

$$\kappa(x,y) = \operatorname{cov}\Big\{T(x), T(y)\Big\}.$$

If the density f_λ exists and is (piecewise) continuous and bounded below on K, then the weak convergence also holds in $L_2(K)$.

In view of Sect. 2.3, Theorem 4.6.5 can be interpreted as asymptotic normality of $\widehat{\lambda}_n$ in the *tangential* sense: $\sqrt{n}\log_\lambda(\widehat{\lambda}_n)$ converges to a Gaussian random element in the tangent space Tan_λ, which is a subset of $L_2(\lambda)$. The additional smoothness conditions allow to switch to the space $L_2(K)$, which is independent of the unknown template measure λ.

See pages 109 and 110 in the supplement for detailed proofs of these theorems. Below we sketch the main ideas only.

Proof (Proof of Theorem 4.6.3). The quantile formula $W_2(\gamma,\theta) = \|F_\theta^{-1} - F_\gamma^{-1}\|_{L^2(0,1)}$ from Sect. 1.5 and the average quantile formula for the Fréchet mean (Sect. 3.1.4) show that the oracle empirical mean $F_{\lambda_n}^{-1}$ follows a central limit theorem in $L_2(0,1)$. Since we work in the Hilbert space $L_2(0,1)$, Fréchet means are simple averages, so the errors in the Fréchet mean have the same rate as the errors in the Fréchet functionals. The smoothing term is easily controlled by Lemma 4.4.2.

Controlling the amplitude term is more difficult. Bounds can be given using the machinery sketched in Sect. 4.7, but we give a more elementary proof by reducing

to the 1-Wasserstein case (using (2.2)), which can be more easily handled in terms of distributions functions (Corollary 1.5.3).

Proof (Proof of Theorem 4.6.5). The hypotheses guarantee that the amplitude and smoothing errors are negligible and

$$\sqrt{n}\left(F^{-1}_{\widehat{\lambda}_n} - F^{-1}_{\lambda}\right) \to GP \quad \text{weakly in } L_2(0,1),$$

where GP is the Gaussian process defined in the proof of Theorem 4.6.3. One then employs a composition with F_λ.

4.7 Convergence of the Empirical Measure and Optimality

One may find the term $O_{\mathbb{P}}(1/\sqrt[4]{\tau_n})$ in Theorem 4.6.3 to be somewhat surprising, and expect that it ought to be $O_{\mathbb{P}}(1/\sqrt{\tau_n})$. The goal of this section is to show why the rate $1/\sqrt[4]{\tau_n}$ is optimal without further assumptions and discuss conditions under which it can be improved to the optimal rate $1/\sqrt{\tau_n}$. For simplicity, we concentrate on the case $\tau_n = n$ and assume that the point process Π is binomial; the Poisson case being easily obtained from the simplified one (using Lemma 4.6.1). We are thus led to study rates of convergence of empirical measures in the Wasserstein space. That is to say, for a fixed exponent $p \geq 1$ and a fixed measure $\mu \in \mathscr{W}_p(\mathscr{X})$, we consider independent random variables $X_1, \ldots$ with law μ and the *empirical measure* $\mu_n = n^{-1}\sum_{i=1}^n \delta\{X_i\}$. The first observation is that $\mathbb{E}W_p(\mu,\mu_n) \to 0$:

Lemma 4.7.1 *Let $\mu \in P(\mathscr{X})$ be any measure. Then*

$$\mathbb{E}W_p(\mu,\mu_n)\begin{cases} = \infty & \mu \notin \mathscr{W}_p(\mathscr{X}) \\ \to 0 & \mu \in \mathscr{W}_p(\mathscr{X}). \end{cases}$$

Proof. This result has been established in an almost sure sense in Proposition 2.2.6. To extend to convergence in expectation observe that

$$W_p^p(\mu,\mu_n) \leq \int_{\mathscr{X}^2} \|x-y\|^p \,\mathrm{d}\mu \otimes \mu_n(x,y) = \frac{1}{n}\sum_{i=1}^n \int_{\mathscr{X}} \|x - X_i\|^p \,\mathrm{d}\mu(x).$$

Thus, the random variable $0 \leq Y_n = W_p^p(\mu,\mu_n)$ is bounded by the sample average Z_n of a random variable $V = \int_{\mathscr{X}} \|x - X_1\|^p \,\mathrm{d}\mu(x)$ that has a finite expectation. A version of the dominated converge theorem (given on page 111 in the supplement) implies that $\mathbb{E}Y_n \to 0$. Now invoke Jensen's inequality.

Remark 4.7.2 *The sequence $\mathbb{E}W_p(\mu,\mu_n)$ is not monotone, as the simple example $\mu = (\delta_0 + \delta_1)/2$ shows (see page 111 in the supplement).*

The next question is how quickly $\mathbb{E}W_p(\mu,\mu_n)$ vanishes when $\mu \in \mathscr{W}_p(\mathscr{X})$. We shall begin with two simple general lower bounds, then discuss upper bounds in the one-

dimensional case, put them in the context of Theorem 4.6.3, and finally briefly touch the d-dimensional case.

Lemma 4.7.3 ($\sqrt{n}$ Lower Bound) *Let $\mu \in P(\mathscr{X})$ be nondegenerate. Then there exists a constant $c(\mu) > 0$ such that for all $p \geq 1$ and all n*

$$\mathbb{E}W_p(\mu_n, \mu) \geq \frac{c(\mu)}{\sqrt{n}}.$$

Proof. Let $X \sim \mu$ and let $a \neq b$ be two points in the support μ. Consider $f(x) = \min(1, \|x-a\|)$, a bounded 1-Lipschitz function such that $f(a) = 0 < f(b)$. Then

$$\sqrt{n}\mathbb{E}W_p(\mu_n, \mu) \geq \sqrt{n}\mathbb{E}W_1(\mu_n, \mu) \geq \mathbb{E}\left| n^{-1/2} \sum_{i=1}^{n} f(X_i) - \mathbb{E}f(X) \right| \to \sqrt{\frac{2\mathrm{var}f(X)}{\pi}} > 0$$

by the central limit theorem and the Kantorovich–Rubinstein theorem (1.11).

For discrete measures, the rates scale badly with p. More generally:

Lemma 4.7.4 (Separated Support) *Suppose that there exist Borel sets $A, B \subset \mathscr{X}$ such that $\mu(A \cup B) = 1$,*

$$\mu(A)\mu(B) > 0 \qquad \textit{and} \qquad d_{\min} = \inf_{x \in A, y \in B} \|x - y\| > 0.$$

Then for any $p \geq 1$ there exists $c_p(\mu) > 0$ such that $\mathbb{E}W_p(\mu_n, \mu) \geq c_p(\mu) n^{-1/(2p)}$.

Any nondegenerate finitely discrete measure μ satisfies this condition, and so do "non-pathological" countably discrete ones. (An example of a "pathological" measure is one assigning positive mass to any rational number.)

Proof. Let $k \sim B(n, q = \mu(A))$ denote the number of points from the sample $(X_1, \ldots, X_n)$ that fall in A. Then a mass of $|k/n - q|$ must travel between A and B, a distance of at least $d_{\min}$. Thus, $W_p^p(\mu_n, \mu) \geq d_{\min}^p |k/n - q|$, and the result follows from the central limit theorem for k; see page 112 in the supplement for the full details.

These lower bounds are valid on any separable metric space. On the real line, it is easy to obtain a sufficient condition for the optimal rate $n^{-1/2}$ to be attained for W_1: since $F_n(t) \sim B(n, F(t))$ has variance $F(t)(1-F(t))/n$, we have (by Fubini's theorem and Jensen's inequality)

$$\mathbb{E}W_1(\mu_n, \mu) = \int_{\mathbb{R}} \mathbb{E}|F_n(t) - F(t)| \, \mathrm{d}t \leq n^{-1/2} \int_{\mathbb{R}} \sqrt{F(t)(1-F(t))} \, \mathrm{d}t,$$

so that $W_1(\mu_n, \mu)$ is of the optimal order $n^{-1/2}$ if

$$J_1(\mu) := \int_{\mathbb{R}} \sqrt{F(t)(1-F(t))} \, \mathrm{d}t < \infty.$$

Since the integrand is bounded by $1/2$, this is certainly satisfied if μ is compactly supported. The J_1 condition is essentially a moment condition, since for any $\delta > 0$, we have for $X \sim \mu$ that $\mathbb{E}|X|^{2+\delta} < \infty \Longrightarrow J_1(\mu) < \infty \Longrightarrow \mathbb{E}|X|^2 < \infty$. It turns out that this condition is necessary, and has a more subtle counterpart for any $p \geq 1$. Let f denote the density of the absolutely continuous part of μ (so $f \equiv 0$ if μ is discrete).

Theorem 4.7.5 (Rate of Convergence of Empirical Measures) *Let $p \geq 1$ and $\mu \in \mathscr{W}_p(\mathbb{R})$. The condition*

$$J_p(\mu) = \int_{\mathbb{R}} \frac{[F(t)(1-F(t))]^{p/2}}{[f(t)]^{p-1}}\,\mathrm{d}t < \infty, \qquad (0^0 = 1)$$

is necessary and sufficient for $\mathbb{E}W_p(\mu_n, \mu) = O(n^{-1/2})$.

See Bobkov and Ledoux [25, Theorem 5.10] for a proof for the J_p condition, and Theorems 5.1 and 5.3 for the values of the constants and a stronger result.

When $p > 1$, for $J_p(\mu)$ to be finite, the support of μ must be connected; this is not needed when $p = 1$. Moreover, the J_p condition is satisfied when f is bounded below (in which case the support of μ must be compact). However, smoothness alone does not suffice, even for measures with positive density on a compact support. More precisely, we have:

Proposition 4.7.6 *For any rate $\varepsilon_n \to 0$ there exists a measure μ on $[-1,1]$ with positive C^∞ density there, and such that for all n*

$$\mathbb{E}W_p(\mu_n, \mu) \geq C(p,\mu) n^{-1/(2p)} \varepsilon_n.$$

The rate $n^{-1/(2p)}$ from Lemma 4.7.4 is the worst among compactly supported measures on $\mathbb{R}$. Indeed, by Jensen's inequality and (2.2), for any $\mu \in P([0,1])$,

$$\mathbb{E}W_p(\mu_n, \mu) \leq \left[\mathbb{E}W_p^p(\mu_n, \mu)\right]^{1/p} \leq \left[\mathbb{E}W_1(\mu_n, \mu)\right]^{1/p} \leq n^{-1/(2p)}.$$

The proof of Proposition 4.7.6 is done by "smoothing" the construction in Lemma 4.7.4, and is given on page 113 in the supplement.

Let us now put this in the context of Theorem 4.6.3. In the binomial case, since each $\Pi_i^{(n)}$ and each Λ_i are independent, we have

$$\mathbb{E}W_2(\Lambda_i, \widetilde{\Lambda}_i)|\Lambda_i \leq \sqrt{2J_2(\Lambda_i)}\,\frac{1}{\sqrt{\tau_n}}.$$

(In the Poisson case, we need to condition on $N_i^{(n)}$ and then estimate its inverse square root as is done in the proof of Theorem 4.6.3.) Therefore, a sufficient condition for the rate $1/\sqrt{\tau_n}$ to hold is that $\mathbb{E}\sqrt{J_2(\Lambda)} < \infty$ and a necessary condition is that $\mathbb{P}(\sqrt{J_2(\Lambda)} < \infty) = 1$. These hold if there exists $\delta > 0$ such that with probability one Λ has a density bounded below by δ. Since $\Lambda = T\#\lambda$, this will happen

provided that λ itself has a bounded below density and T has a bounded below derivative. Bigot et al. [23] show that the rate $\sqrt{\tau_n}$ cannot be improved.

We conclude by proving a lower bound for absolutely continuous measures and stating, without proof, an upper bound.

Proposition 4.7.7 *Let $\mu \in \mathscr{W}_1(\mathbb{R}^d)$ have an absolutely continuous part with respect to Lebesgue measure, and let ν_n be any discrete measure supported on n points (or less). Then there exists a constant $C(\mu) > 0$ such that*

$$W_p(\mu, \nu_n) \geq W_1(\mu, \nu_n) \geq C(\mu) n^{-1/d}.$$

Proof. Let f be the density of the absolutely continuous part μ_c, and observe that for some finite number M,

$$2\delta = \mu_c(\{x : f(x) \leq M\}) > 0.$$

Let $x_1, \ldots, x_n$ be the support points of ν_n and $\varepsilon > 0$. Let $\mu_{c,M}$ be the restriction of μ_c to the set where the density is smaller than M. The union of balls $B_\varepsilon(x_i)$ has $\mu_{c,M}$-measure of at most

$$M \sum_{i=1}^{n} \mathrm{Leb}(B_\varepsilon(x_i)) = Mn\varepsilon^d \mathrm{Leb}_d(B_1(0)) = Mn\varepsilon^d C_d = \delta,$$

if $\varepsilon^d = \delta(nMC_d)^{-1}$. Thus, a mass $2\delta - \delta = \delta$ must travel more than ε from ν_n to μ in order to cover $\mu_{c,M}$. Hence

$$W_1(\nu_n, \mu) \geq \delta\varepsilon = \delta(\delta/MC_d)^{1/d}\, n^{-1/d}.$$

The lower bound holds because we need ε^{-d} balls of radius ε in order to cover a sufficiently large fraction of the mass of μ. The determining quantity for *upper* bounds on the empirical Wasserstein distance is the *covering numbers*

$$N(\mu, \varepsilon, \tau) = \text{ minimal number of balls whose union has } \mu \text{ mass } \geq 1 - \tau.$$

Since μ is tight, these are finite for all $\varepsilon, \tau > 0$, and they increase as ε and τ approach zero. To put the following bound in context, notice that if μ is compactly supported on $\mathbb{R}^d$, then $N(\mu, \varepsilon, 0) \leq K\varepsilon^{-d}$.

Theorem 4.7.8 *If for some $d > 2p$, $N(\mu, \varepsilon, \varepsilon^{dp/(d-2p)}) \leq K\varepsilon^{-d}$, then $\mathbb{E}W_p \leq C_p n^{-1/d}$.*

Comparing this with the lower bound in Lemma 4.7.4, we see that in the high-dimensional regime $d > 2p$, absolutely continuous measures have a worse rate than discrete ones. In the low-dimensional regime $d < 2p$, the situation is opposite. We also obtain that for $d > 2$ and a compactly supported absolutely continuous $\mu \in \mathscr{W}_1(\mathbb{R}^d)$, $\mathbb{E}W_1(\mu_n, \mu) \sim n^{-1/d}$.

4.8 Bibliographical Notes

Our exposition in this chapter closely follows the papers Panaretos and Zemel [100] and Zemel and Panaretos [134].

Books on functional data analysis include Ramsay and Silverman [109, 110], Ferraty and Vieu [51], Horváth and Kokoszka [70], and Hsing and Eubank [71], and a recent review is also available (Wang et al. [127]). The specific topic of amplitude and phase variation is discussed in [110, Chapter 7] and [127, Section 5.2]. The next paragraph gives some selective references.

One of the first functional registration techniques employed dynamic programming (Wang and Gasser [128]) and dates back to Sakoe and Chiba [118]. Landmark registration consists of identifying salient features for each curve, called *landmarks*, and aligning them (Gasser and Kneip [61]; Gervini and Gasser [63]). In pairwise synchronisation (Tang and Müller [122]) one aligns each pair of curves and then derives an estimator of the warp functions by linear averaging of the pairwise registration maps. Another class of methods involves a template curve, to which each observation is registered, minimising a discrepancy criterion; the template is then iteratively updated (Wang and Gasser [129]; Ramsay and Li [108]). James [72] defines a "feature function" for each curve and uses the moments of the feature function to guarantee identifiability. Elastic registration employs the Fisher–Rao metric that is invariant to warpings and calculates averages in the resulting quotient space (Tucker et al. [123]). Other techniques include semiparametric modelling (Rønn [115]; Gervini and Gasser [64]) and principal components registration (Kneip and Ramsay [82]). More details can be found in the review article by Marron et al. [90]. Wrobel et al. [131] have recently developed a registration method for functional data with a discrete flavour. It is also noteworthy that a version of the Wasserstein metric can also be used in the functional case (Chakraborty and Panaretos [34]).

The literature on the point processes case is more scarce; see the review by Wu and Srivastava [133].

A parametric version of Theorem 4.2.4 was first established by Bigot and Klein [22, Theorem 5.1] in $\mathbb{R}^d$, extended to a compact nonparametric formulation in Zemel and Panaretos [134]. There is an infinite-dimensional linear version in Masarotto et al. [91]. The current level of generality appears to be new.

Theorem 4.4.1 is a stronger version of Panaretos and Zemel [100, Theorem 1] where it was assumed that τ_n must diverge to infinity faster than $\log n$. An analogous construction under the Bayesian paradigm can be found in Galasso et al. [58]. Optimality of the rates of convergence in Theorem 4.6.3 is discussed in detail by Bigot et al. [23], where finiteness of the functional J_2 (see Sect. 4.7) is assumed and consequently $O_{\mathbb{P}}(\tau_n^{-1/4})$ is improved to $O_{\mathbb{P}}(\tau_n^{-1/2})$.

As far as we know, Theorem 4.6.5 (taken from [100]) is the first central limit theorem for Fréchet means in Wasserstein space. When the measures Λ_i are observed exactly (no amplitude variation: $\tau_n = \infty$ and $\sigma = 0$) Kroshnin et al. [84] have recently proven a central limit theorem for random Gaussian measures in arbitrary dimension, extending a previous result of Agueh and Carlier [3]. It seems likely that

in a fully nonparametric setting, the rates of convergence (compare Theorem 4.6.3) might be slower than $\sqrt{n}$; see Ahidar-Coutrix et al. [4].

The magnitude of the amplitude variation in Theorem 4.6.3 pertains to the rates of convergence of $\mathbb{E}W_p(\mu_n, \mu)$ to zero (Sect. 4.7). This is a topic of intense research, dating back to the seminal paper by Dudley [46], where a version of Theorem 4.7.8 with $p = 1$ is shown for the bounded Lipschitz metric. The lower bounds proven in this section were adapted from [46], Fournier and Guillin [54], and Weed and Bach [130].

The version of Theorem 4.7.8 given here can be found in [130] and extends Boissard and Le Gouic [27]. Both papers [27, 130] work in a general setting of complete separable metric spaces. An additional $\log n$ term appears in the limiting case $d = 2p$, as already noted (for $p = 1$) by [46], and the classical work of Ajtai et al. [5] for μ uniform on $[0, 1]^2$. More general results are available in [54]. A longer (but far from being complete) bibliography is given in the recent review by Panaretos and Zemel [101, Subsection 3.3.1], including works by Barthe, Dobrić, Talagrand, and coauthors on almost sure results and deviation bounds for the empirical Wasserstein distance.

The J_1 condition is due to del Barrio et al. [43], who showed it to be necessary and sufficient for the empirical process $\sqrt{n}(F_n - F)$ to converge in distribution to $\mathbb{B} \circ F$, with $\mathbb{B}$ Brownian bridge. The extension to $1 \leq p \leq \infty$ (and a lot more) can be found in Bobkov and Ledoux [25], employing order statistics and beta distributions to reduce to the uniform case. Alternatively, one may consult Mason [92], who uses weighted approximations to Brownian bridges.

An important aspect that was not covered here is that of statistical inference of the Wasserstein distance on the basis of the empirical measure. This is a challenging question and results by del Barrio, Munk, and coauthors are available for one-dimensional, elliptical, or discrete measures, as explained in [101, Section 3].