ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistical Methodology
Series B

J. R. Statist. Soc. B (2018)
80, *Part* 1, *pp.* 219–238

Inference for empirical Wasserstein distances on finite spaces

Max Sommerfeld

University of Göttingen, Germany

and Axel Munk

University of Göttingen and Max Planck Institute for Biophysical Chemistry, Göttingen, Germany

[Received February 2016. Final revision April 2017]

Summary. The Wasserstein distance is an attractive tool for data analysis but statistical inference is hindered by the lack of distributional limits. To overcome this obstacle, for probability measures supported on finitely many points, we derive the asymptotic distribution of empirical Wasserstein distances as the optimal value of a linear programme with random objective function. This facilitates statistical inference (e.g. confidence intervals for sample-based Wasserstein distances) in large generality. Our proof is based on directional Hadamard differentiability. Failure of the classical bootstrap and alternatives are discussed. The utility of the distributional results is illustrated on two data sets.

Keywords: Bootstrap; Central limit theorem; Directional Hadamard derivative; Hypothesis testing; Optimal transport; Wasserstein distance

1. Introduction

The *Wasserstein distance* (Vasershtein, 1969), which is also known as the Mallows distance (Mallows, 1972), the Monge–Kantorovich–Rubinstein distance in the physical sciences (Kantorovich and Rubinstein, 1958; Rachev, 1985; Jordan *et al.*, 1998), the earth mover's distance in computer science (Rubner *et al.*, 2000) or the optimal transport distance in optimization (Ambrosio, 2003), is one of the most fundamental metrics on the space of probability measures. Besides its prominence in probability (e.g. Dobrushin (1970) and Gray (1988)) and finance (e.g. Rachev and Ruschendorf (1998)) it has deep connections to the asymptotic theory of partial differential equations of diffusion type (Otto (2001), Villani (2003, 2008) and references therein). In a statistical setting it has mainly been used as a tool to prove weak convergence in the context of limit laws (e.g. Bickel and Freedman (1981), Shorack and Wellner (1986), Johnson and Samworth (2005), Dümbgen *et al.* (2011) and Dorea and Ferreira (2012)) as it metrizes weak convergence together with convergence of moments. However, recently the empirical (i.e. estimated from data) Wasserstein distance has also been recognized as a central quantity itself in many applications, among them clinical trials (Munk and Czado, 1998; Freitag *et al.*, 2007), metagenomics (Evans and Matsen, 2012), medical imaging (Ruttenberg *et al.*, 2013),

Address for correspondence: Axel Munk, Institute for Mathematical Stochastics, University of Göttingen, Goldschmidtstraße 7, 37077 Göttingen, Germany.
E-mail: munk@math.uni-goettingen.de

 1369–7412/18/80219

goodness-of-fit testing (Freitag and Munk, 2005; Del Barrio *et al.*, 1999), biomedical engineering (Oudre *et al.*, 2012), computer vision (Gangbo and McCann, 2000; Ni *et al.*, 2009), cell biology (Orlova *et al.*, 2016) and model validation (Halder and Bhattacharya, 2011). The barycentre with respect to the Wasserstein metric (Agueh and Carlier, 2011) has been shown to elicit important structure from complex data and to be a promising tool, e.g. in deformable models (Boissard *et al.*, 2015; Agulló-Antolín *et al.*, 2015). It has also been used in large-scale Bayesian inference to combine posterior distributions from subsets of the data (Srivastava *et al.*, 2015).

Generally speaking three characteristics of the Wasserstein distance make it particularly attractive for various applications. First, it incorporates a ground distance on the space in question. This often makes it more adequate than competing metrics such as total variation or χ^2-metrics which are oblivious to any metric or similarity structure on the ground space. As an example, the success of the Wasserstein distance in metagenomics applications can largely be attributed to this fact (see Evans and Matsen (2012) and also our application in Section 3.3).

Second, it has a clear and intuitive interpretation as the amount of 'work' required to transform one probability distribution into another and the resulting transport can be visualized (see Section 3.2). This is also interesting in applications where probability distributions are used to represent actual physical mass and spatiotemporal changes must be tracked.

Third, it is well established (Rubner *et al.*, 2000) that the Wasserstein distance performs exceptionally well at capturing human perception of similarity. This motivates its popularity in computer vision and related fields.

Despite these advantages, the use of the empirical Wasserstein distance in a statistically rigorous way is severely hampered by a lack of inferential tools. We argue that this issue stems from considering too large classes of candidate distributions (e.g. those which are absolutely continuous with respect to the Lebesgue measure if the ground space has dimension 2 or more). In this paper, we therefore discuss the Wasserstein distance on finite spaces, which enables us to solve this issue. We argue that the restriction to finite spaces is not merely an approximation to the truth, but rather that this setting is sufficient for many practical situations as measures often already come naturally discretized (e.g. two- or three-dimensional images—see also our applications in Section 3).

We remark that from our methodology further inferential procedures can be derived, e.g. a (multivariate) analysis-of-variance type of analysis and multiple comparisons of Wasserstein distances based on their p-values (see for example Benjamini and Hochberg (1995)). Our techniques also extend immediately to dependent samples (X_i, Y_i) with marginals $\mathbf{r}$ and $\mathbf{s}$.

1.1. Wasserstein distance

Let $(\mathcal{X}, d)$ be a complete metric space with metric $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_{\geqslant 0}$. The *Wasserstein distance of order* p ($p \geqslant 1$) between two Borel probability measures μ_1 and μ_2 on $\mathcal{X}$ is defined as

$$W_p(\mu_1, \mu_2) = \left\{ \inf_{\nu \in \Pi(\mu_1, \mu_2)} \int_{\mathcal{X} \times \mathcal{X}} d^p(x, x') \, \nu(\mathrm{d}x, \mathrm{d}x') \right\}^{1/p},$$

where $\Pi(\mu_1, \mu_2)$ is the set of all Borel probability measures on $\mathcal{X} \times \mathcal{X}$ with marginals μ_1 and μ_2.

1.2. Wasserstein distance on finite spaces

If we restrict in the above definition $\mathcal{X} = \{x_1, \ldots, x_N\}$ to be a finite space, every probability measure on $\mathcal{X}$ is given by a vector $\mathbf{r}$ in $\mathcal{P}_{\mathcal{X}} = \{\mathbf{r} = (r_x)_{x \in \mathcal{X}} \in \mathbb{R}^{\mathcal{X}}_{>0} : \Sigma_{x \in \mathcal{X}} r_x = 1\}$, via $P_{\mathbf{r}}(\{x\}) = r_x$.

We shall not distinguish between the vector $\mathbf{r}$ and the measure that it defines. The *Wasserstein distance of order* p between two finitely supported probability measures $\mathbf{r}, \mathbf{s} \in \mathcal{P}_{\mathcal{X}}$ then becomes

$$W_p(\mathbf{r}, \mathbf{s}) = \left\{ \min_{\mathbf{w} \in \Pi(\mathbf{r}, \mathbf{s})} \sum_{x, x' \in \mathcal{X}} d^p(x, x') w_{x,x'} \right\}^{1/p}, \tag{1}$$

where $\Pi(\mathbf{r}, \mathbf{s})$ is the set of all probability measures on $\mathcal{X} \times \mathcal{X}$ with marginal distributions $\mathbf{r}$ and $\mathbf{s}$ respectively. All our methods and results concern this Wasserstein distance on finite spaces.

1.3. Overview of main results

1.3.1. Distributional limits

The basis for inferential procedures for the Wasserstein distance on finite spaces is a limit theorem for its empirical version $W_p(\hat{\mathbf{r}}_n, \hat{\mathbf{s}}_m)$. Here, the empirical measure that is generated by independent random variables $X_1, \ldots, X_n \sim \mathbf{r}$ is given by $\hat{\mathbf{r}}_n = (\hat{r}_{n,x})_{x \in \mathcal{X}}$, where $\hat{r}_{n,x} = (1/n)\#\{k : X_k = x\}$. Let $\hat{\mathbf{s}}_m$ be generated from independently and identically distributed (IID) $Y_1, \ldots, Y_m \sim \mathbf{s}$ in the same fashion. Under the null hypothesis $\mathbf{r} = \mathbf{s}$ we prove that

$$\left(\frac{nm}{n+m} \right)^{1/(2p)} W_p(\hat{\mathbf{r}}_n, \hat{\mathbf{s}}_m) \Rightarrow \left(\max_{\mathbf{u} \in \Phi_p^*} \langle \mathbf{G}, \mathbf{u} \rangle \right)^{1/p}, \qquad n, m \to \infty. \tag{2}$$

Here, '$\Rightarrow$' means convergence in distribution, $\mathbf{G}$ is a mean 0 Gaussian random vector with covariance depending on $\mathbf{r} = \mathbf{s}$ and Φ_p^* is the convex set of dual solutions to the Wasserstein problem depending on the metric d only (see theorem 1). In Section 3.2 we use this result to assess the statistical significance of the differences between real and synthetically generated fingerprints in the fingerprint verification competition (Maio *et al.*, 2002).

We give analogous results under the alternative $\mathbf{r} \neq \mathbf{s}$. This extends the scope of our results beyond the classical two-sample (or goodness-of-fit) test as it enables confidence statements on $W_p(\mathbf{r}, \mathbf{s})$ when the null hypothesis of equality is likely or even *known to be false*. An example for this is given by our application to metagenomics (Section 3.3) where samples from the same person taken at different times are typically statistically different but our asymptotic results enable us to assert with statistical significance that interpersonal are larger than intrapersonal distances.

1.3.2. Proof strategy

We prove these results by showing that the Wasserstein distance is *directionally Hadamard differentiable* (Shapiro, 1990) and the right-hand side of expression (2) is its derivative evaluated at the Gaussian limit of the empirical multinomial process (see theorem 4 in Section 2.3). This notion generalizes Hadamard differentiability by allowing *non-linear* derivatives but still enables a refined delta method (Römisch (2004) and theorem 3 in Section 2.2). Notably, the Wasserstein distance is not Hadamard differentiable in the usual sense.

1.3.3. Explicit limiting distribution for tree metrics

When the space $\mathcal{X}$ is the vertices of a tree and the metric d is given by path length we give an explicit expression for the limiting distribution in expression (2) (see theorem 5 in Section 2.5). In contrast with the general case, this explicit formula enables fast and direct simulation of the limiting distribution. This extends a previous result of Samworth and Johnson (2004) who considered a finite number of point masses on the real line. The Wasserstein distance on trees has, to the best of our knowledge, only been considered in two papers: Kloeckner (2013) studied

the geometric properties of the Wasserstein space of measures on a tree and Evans and Matsen (2012) used the Wasserstein distance on phylogenetic trees to compare microbial communities.

1.3.4. The bootstrap

Directional Hadamard differentiability is not enough to guarantee the consistency of the naive (n out of n) bootstrap (Dümbgen, 1993; Fang and Santos, 2014)—in contrast with the usual notion of Hadamard differentiability. This implies that the bootstrap is *not* consistent for the Wasserstein distance (1) (see theorem 1 in Section 2.1). In contrast, the m out of n bootstrap for $m/n \to 0$ is known to be consistent in this setting (Dümbgen, 1993) and can be applied to the Wasserstein distance. Under the null hypothesis $\mathbf{r} = \mathbf{s}$, however, there is a more direct way of obtaining an approximation of the limiting distribution. In the on-line appendix, we discuss this alternative resampling scheme based on ideas of Fang and Santos (2014), which essentially consists of plugging in a bootstrap version of the underlying empirical process in the derivative. We show that this scheme, which we shall call the *directional bootstrap*, is consistent for the Wasserstein distance (see theorem 1, part (b)).

1.4. Related work

1.4.1. Empirical Wasserstein distances

In very general terms, we study a particular case (finite spaces) of the following question and its two-sample analogue: given the empirical measure μ_n based on n IID random variables taking variables in a metric space with law μ, what can be inferred about $W_p(\mu_n, \mu_0)$ for a reference measure μ_0 which may be equal to μ?

It is a well-known and straightforward consequence of the strong law of large numbers that if the pth moments are finite for μ and μ_0 then $W_p(\mu_n, \mu_0)$ converges to $W_p(\mu, \mu_0)$, almost surely, as the sample size n approaches ∞ (Villani (2008), corollary 6.11). Determining the exact rate of this convergence is the subject of an impressive body of literature that has been developed over the last decades starting with the seminal work of Ajtai *et al.* (1984) considering for μ_0 the uniform distribution on the unit square, followed by Talagrand (1992, 1994) for the uniform distribution in higher dimensions and Horowitz and Karandikar (1994) giving bounds on mean rates of convergence. Boissard and Gouic (2014) and Fournier and Guillin (2014) gave general deviation inequalities for the empirical Wasserstein distance on metric spaces. For a discussion in the light of our distributional limit results see Section 4.

Distributional limits give a natural perspective for practicable inference but despite considerable interest in the topic have remained elusive to a large extent. For measures on $\mathcal{X} = \mathbb{R}$ quite a complete theory is available (see Munk and Czado (1998), Freitag *et al.* (2007) and Freitag and Munk (2005) for $\mu_0 \neq \mu$ and for example Del Barrio *et al.* (1999), Samworth and Johnson (2005) and Del Barrio *et al.* (2005) for $\mu_0 = \mu$ as well as Mason (2016) and Bobkov and Ledoux (2014) for recent surveys). However, for $\mathcal{X} = \mathbb{R}^d$, $d \geqslant 2$, the only distributional result known to us is due to Rippl *et al.* (2015) for specific multivariate (elliptic) parametric classes of distributions, when the empirical measure is replaced by a parametric estimate. In the context of deformable models distributional results have been proved (Del Barrio *et al.*, 2015) for specific multi-dimensional parametric models which factor into one-dimensional parts.

The simple reason why the Wasserstein distance is so much easier to handle in the one-dimensional case is that in this case the optimal coupling attaining the infimum in expression (1) is known explicitly. In fact, the Wasserstein distance of order p between two measures on $\mathbb{R}$ then becomes the L^p-norm of the difference of their quantile functions (see Mallows (1972) for an early reference) and the analysis of empirical Wasserstein distances can be based on

quantile process theory. Beyond this case, explicit coupling results are known only for multivariate Gaussian and elliptic distributions (Gelbrich, 1990). A classical result of Ajtai *et al.* (1984) for the uniform distribution on $\mathcal{X} = [0, 1]^2$ suggests that, even in this simple case, distributional limits will have a complicated form if they exist at all. We shall elaborate on this thought in the discussion, in Section 4.

The Wasserstein distance on finite spaces has been considered recently by Gozlan *et al.* (2013) to derive entropy inequalities on graphs and by Erbar and Maas (2012) to define Ricci curvature for Markov chains on discrete spaces. To the best of our knowledge, empirical Wasserstein distances on finite spaces have been considered only by Samworth and Johnson (2004) in the special case of measures supported on $\mathbb{R}$. We shall show (Section 2.5) that our results extend theirs.

1.4.2. Directional Hadamard differentiability

We prove our distributional limit theorems by using the theory of parametric programming (Bonnans and Shapiro, 2013) which investigates how the optimal value and the optimal solutions of an optimization problem change when the objective function and the constraints are changed. Whereas differentiability properties of optimal values of linear programmes have been extremely well studied such results have, to the best of our knowledge, not yet been applied to the statistical analysis of Wasserstein distances.

It is well known that under certain conditions the optimal value of a mathematical programme is differentiable with respect to the constraints of the problem (Rockafellar, 1997; Gal *et al.*, 1997). However, the derivative will typically be non-linear. The appropriate concept for this is directional Hadamard differentiability (Shapiro, 1990). The derivative of the optimal value of a mathematical programme is typically again given as an extremal value.

Although the delta method for directional Hadamard derivatives has been known for a long time (Shapiro, 1991; Dümbgen, 1993), this notion scarcely appears in the statistical context (with some exceptions, such as Römisch (2004); see also Donoho and Liu (1988)). Recently, interest in the topic has evolved in econometrics (see Fang and Santos (2014) and references therein).

1.4.3. Organization of the paper

In Section 2 we give a comprehensive result on distributional limits for the Wasserstein distance for measures supported on finitely many points. In Section 3 we apply our methods to two data sets to highlight different aspects. In Section 4 we briefly address limitations and possible extensions of our work. In the on-line supplementary material we discuss the bootstrap for the Wasserstein distance and give some technical proofs.

2. Distributional limits

2.1. Main result

In this section we give a comprehensive result on distributional limits for the Wasserstein distance when the underlying population measures are supported on finitely many points $\mathcal{X} = \{x_1, \ldots, x_N\}$. We denote the inner product on the vector space $\mathbb{R}^{\mathcal{X}}$ by $\langle \mathbf{u}, \mathbf{u} \rangle' = \Sigma_{x \in \mathcal{X}} u_x u'_x$ for $\mathbf{u}, \mathbf{u}' \in \mathbb{R}^{\mathcal{X}}$.

Theorem 1. Let $p \geqslant 1$, $\mathbf{r}, \mathbf{s} \in \mathcal{P}_{\mathcal{X}}$ and $\hat{\mathbf{r}}_n$ and $\hat{\mathbf{s}}_m$ generated by IID samples $X_1, \ldots, X_n \sim \mathbf{r}$ and $Y_1, \ldots, Y_m \sim \mathbf{s}$ respectively. We define the convex sets

$$\Phi_p^* = \{\mathbf{u} \in \mathbb{R}^{\mathcal{X}} : u_x - u_{x'} \leqslant d^p(x, x'), \quad x, x' \in \mathcal{X}\},$$
$$\Phi_p^*(\mathbf{r}, \mathbf{s}) = \left\{ (\mathbf{u}, \mathbf{v}) \in \mathbb{R}^{\mathcal{X}} \times \mathbb{R}^{\mathcal{X}} : \begin{array}{l} \langle \mathbf{u}, \mathbf{r} \rangle + \langle \mathbf{v}, \mathbf{s} \rangle = W_p^p(\mathbf{r}, \mathbf{s}), \\ u_x + v_{x'} \leqslant d^p(x, x'),\ x, x' \in \mathcal{X} \end{array} \right\} \tag{3}$$

and the multinomial covariance matrix

$$\Sigma(\mathbf{r}) = \begin{pmatrix} r_{x_1}(1 - r_{x_1}) & -r_{x_1} r_{x_2} & \cdots & -r_{x_1} r_{x_N} \\ -r_{x_2} r_{x_1} & r_{x_2}(1 - r_{x_2}) & \cdots & -r_{x_2} r_{x_N} \\ \vdots & \vdots & \ddots & \vdots \\ -r_{x_N} r_{x_1} & -r_{x_N} r_{x_2} & \cdots & r_{x_N}(1 - r_{x_N}) \end{pmatrix} \tag{4}$$

such that with independent Gaussian random variables $\mathbf{G} \sim \mathcal{N}\{0, \Sigma(\mathbf{r})\}$ and $\mathbf{H} \sim \mathcal{N}\{0, \Sigma(\mathbf{s})\}$ we have the following results.

(a) (One sample—null hypothesis): with the sample size n approaching ∞, we have the weak convergence

$$n^{1/(2p)} W_p(\hat{r}_n, \mathbf{r}) \Rightarrow \Big(\max_{\mathbf{u} \in \Phi_p^*} \langle \mathbf{G}, \mathbf{u} \rangle \Big)^{1/p}. \tag{5}$$

(b) (One sample—alternative): with n approaching ∞, we have

$$n^{1/2} \{ W_p(\hat{\mathbf{r}}_n, \mathbf{s}) - W_p(\mathbf{r}, \mathbf{s}) \} \Rightarrow \frac{1}{p} W_p^{1-p}(\mathbf{r}, \mathbf{s}) \Big(\max_{(\mathbf{u}, \mathbf{v}) \in \Phi_p^*(\mathbf{r}, \mathbf{s})} \langle \mathbf{G}, \mathbf{u} \rangle \Big). \tag{6}$$

(c) (Two samples—null hypothesis): let $\rho_{n,m} = \{nm/(n+m)\}^{1/2}$. If $\mathbf{r} = \mathbf{s}$ and n and m are approaching ∞ such that $n \wedge m \to \infty$ and $m/(n+m) \to \lambda \in (0, 1)$ we have

$$\rho_{n,m}^{1/p} W_p(\hat{\mathbf{r}}_n, \hat{\mathbf{s}}_m) \Rightarrow \Big(\max_{\mathbf{u} \in \Phi_p^*} \langle \mathbf{G}\mathbf{u} \rangle \Big)^{1/p}. \tag{7}$$

(d) (Two samples—alternative): with n and m approaching ∞ such that $n \wedge m \to \infty$ and $m/(n+m) \to \lambda \in [0, 1]$

$$\rho_{n,m} \{ W_p(\hat{\mathbf{r}}_n, \hat{\mathbf{s}}_m) - W_p(\mathbf{r}, \mathbf{s}) \} \Rightarrow \frac{1}{p} W_p^{1-p}(\mathbf{r}, \mathbf{s}) \Big\{ \max_{(\mathbf{u}, \mathbf{v}) \in \Phi_p^*(\mathbf{r}, \mathbf{s})} \sqrt{\lambda} \langle \mathbf{G}, \mathbf{u} \rangle + \sqrt{(1 - \lambda)} \langle \mathbf{H}, \mathbf{v} \rangle \Big\}. \tag{8}$$

The sets Φ_p^* and $\Phi_p^*(\mathbf{r}, \mathbf{s})$ are (derived from) the dual solutions to the Wasserstein linear programme (see theorem 4 in Section 2.3). This result is valid for all probability measures with finite support, regardless of the (dimension of the) underlying space. In particular, it generalizes a result of Samworth and Johnson (2004), who considered a finite collection of point masses on the real line and $p = 2$. We shall reobtain their result as a special case in Section 2.5 when we give explicit expressions for the limit distribution when the metric d, which enters the limit law via the dual solutions Φ_p^* or $\Phi_p^*(\mathbf{r}, \mathbf{s})$, is given by a tree.

Remark 1. In our numerical experiments (see Section 3) we have found representation (8) to be numerically unstable when used to simulate from the limiting distribution under the alternative. We therefore give an alternative representation (1) in the on-line supplementary material as a one-dimensional optimization problem of a non-linear function (in contrast with a high dimensional linear programme shown here). Note that the limiting distribution under the null does not suffer from this problem and can be simulated from directly by using a linear programme solver.

The scaling rate in theorem 1 depends solely on p and is completely independent of the underlying space $\mathcal{X}$. This contrasts known bounds on the rate of convergence in the continuous case. We shall elaborate on the differences in the discussion. Typical choices are $p = 1, 2$. The faster scaling rate can be a reason to favour $p = 1$. In our numerical experiments, however, this advantage was frequently outweighed by larger quantiles of the limiting distribution.

Dümbgen (1993) showed that the naive n out of n bootstrap is inconsistent for functionals with a non-linear Hadamard derivative, but resampling fewer than n observations leads to a consistent bootstrap. Since we shall show in what follows that the Wasserstein distance belongs to this class of functionals, it is a direct consequence that the naive bootstrap fails for the Wasserstein distance (see section B in the on-line supplementary material for details) and that the following theorem holds.

Theorem 2. Let $\hat{\mathbf{r}}_n^*$ and $\hat{\mathbf{s}}_m^*$ be bootstrap versions of $\hat{r}_n$ and $\hat{s}_m$ that are obtained via resampling k observations with $k/n \to 0$ and $k/m \to 0$. Then, the plug-in bootstrap with $\hat{\mathbf{r}}_n^*$ and $\hat{\mathbf{s}}_m^*$ is consistent, i.e.

$$\sup_{f \in \mathrm{BL}_1(\mathbb{R})} E[f(\phi_p[\sqrt{k}\{(\hat{\mathbf{r}}_n^{**}, \hat{\mathbf{s}}_m^{**}) - (\hat{\mathbf{r}}_n, \hat{\mathbf{s}}_m)\}]) | X_1, \ldots, X_n, Y_1, \ldots, Y_m] - E[f[\rho_{n,m}\{W_p^p(\hat{r}_n, \hat{s}_m) - W_p^p(\mathbf{r}, \mathbf{s})\}]]$$

converges to 0 in probability.

In what follows we shall prove our main theorem 1 by

(a) introducing Hadamard directional differentiability, which does not require the derivative to be linear but still enables a delta method and
(b) showing that the map $(\mathbf{r}, \mathbf{s}) \mapsto W_p(\mathbf{r}, \mathbf{s})$ is differentiable in this sense.

2.2. Hadamard directional derivatives

In this section we follow Römisch (2004). A map f defined on a subset $D_f \subset \mathbb{R}^d$ with values in $\mathbb{R}$ is called *Hadamard directionally differentiable* at $\mathbf{u} \in \mathbb{R}^d$ if there is a map $f'_{\mathbf{u}} : \mathbb{R}^d \to \mathbb{R}$ such that

$$\lim_{n \to \infty} \frac{f(\mathbf{u} + t_n \mathbf{h}_n) - f(\mathbf{u})}{t_n} = f'_{\mathbf{u}}(\mathbf{h}) \tag{9}$$

for any $\mathbf{h} \in \mathbb{R}^d$ and for arbitrary sequences t_n converging to 0 from above and $\mathbf{h}_n$ converging to $\mathbf{h}$ such that $\mathbf{u} + t_n \mathbf{h}_n \in D_f$ for all $n \in \mathbb{N}$. In contrast with the usual notion of Hadamard differentiability (e.g. van der Vaart and Wellner (1996)) the derivative $\mathbf{h} \mapsto f'_{\mathbf{u}}(\mathbf{h})$ is *not* required to be linear. A prototypical example is the absolute value $f : \mathbb{R} \to \mathbb{R}$, $t \mapsto |t|$ which is not in the usual sense Hadamard differentiable at $t = 0$ but directionally differentiable with the non-linear derivative $t \mapsto |t|$.

Theorem 3 (Römisch (2004), theorem 1). Let f be a function that is defined on a subset F of $\mathbb{R}^d$ with values in $\mathbb{R}$, such that

(a) f is Hadamard directionally differentiable at $\mathbf{u} \in F$ with derivative $f'_{\mathbf{u}} : F \to \mathbb{R}$ and
(b) there is a sequence of $\mathbb{R}^d$-valued random variables X_n and a sequence of non-negative numbers $\rho_n \to \infty$ such that $\rho_n(X_n - \mathbf{u}) \Rightarrow X$ for some random variable X taking values in F.

Then, $\rho_n\{f(X_n) - f(\mathbf{u})\} \Rightarrow f'_{\mathbf{u}}(X)$.

2.3. Directional derivative of the Wasserstein distance

In this section we show that the functional $(\mathbf{r}, \mathbf{s}) \mapsto W_p^p(\mathbf{r}, \mathbf{s})$ is Hadamard directionally differentiable and give a formula for the derivative.

The *dual* programme (see Luenberger and Ye (2008), chapter 4), and also Kantorovich and Rubinstein (1958)) of the linear programme defining the Wasserstein distance (1) is given by

$$\max_{(\mathbf{u},\mathbf{v}) \in \mathbb{R}^{\mathcal{X}} \times \mathbb{R}^{\mathcal{X}}} \langle \mathbf{u}, \mathbf{r} \rangle + \langle \mathbf{s}, \mathbf{v} \rangle \qquad \text{subject to } u_x + v_{x'} \leqslant d^p(x, x') \quad \forall\, x, x' \in \mathcal{X}. \tag{10}$$

As noted above, the optimal value of the primal problem is $W_p^p(\mathbf{r}, \mathbf{s})$ and by standard duality theory of linear programmes (e.g. Luenberger and Ye (2008)) this is also the optimal value of the dual problem. Therefore, the set of optimal solutions to the dual problem is given by $\Phi_p^*(\mathbf{r}, \mathbf{s})$ as defined in expression (3).

Theorem 4. The functional $(\mathbf{r}, \mathbf{s}) \mapsto W_p^p(\mathbf{r}, \mathbf{s})$ is directionally Hadamard differentiable at all $(\mathbf{r}, \mathbf{s}) \in \mathcal{P}_{\mathcal{X}} \times \mathcal{P}_{\mathcal{X}}$ with derivative

$$(\mathbf{h}_1, \mathbf{h}_2) \mapsto \max_{(\mathbf{u},\mathbf{v}) \in \Phi_p^*(\mathbf{r},\mathbf{s})} -(\langle \mathbf{u}, \mathbf{h}_1 \rangle + \langle \mathbf{v}, \mathbf{h}_2 \rangle). \tag{11}$$

We can give a more explicit expression for the set $\Phi_p^*(\mathbf{r}, \mathbf{s})$ in the case $\mathbf{r} = \mathbf{s}$, when the optimal value of the primal and the dual problem is 0. Then, the condition $W_p^p(\mathbf{r}, \mathbf{s}) = \langle \mathbf{r}, \mathbf{u} \rangle + \langle \mathbf{s}, \mathbf{v} \rangle$ becomes $\langle \mathbf{r}, \mathbf{u} + \mathbf{v} \rangle = 0$. Since $u_x + v_{x'} \leqslant d^p(x, x')$ for all $x, x' \in \mathcal{X}$ implies that $\mathbf{u} + \mathbf{v} \leqslant 0$ this yields $\mathbf{u} = -\mathbf{v}$. This gives

$$\Phi_p^*(\mathbf{r}, \mathbf{r}) = \{(\mathbf{u}, -\mathbf{u}) \in \mathbb{R}^{\mathcal{X}} \times \mathbb{R}^{\mathcal{X}} : u_x - u_{x'} \leqslant d^p(x, x'), \quad x, x' \in \mathcal{X}\}$$

and the following more compact representation of the dual solutions in the case $\mathbf{r} = \mathbf{s}$, independent of $\mathbf{r}$:

$$\Phi_p^*(\mathbf{r}, \mathbf{r}) = \Phi_p^* \times (-\Phi_p^*). \tag{12}$$

2.4. Proof of theorem 1

(a) With the notation that was introduced in theorem 1, $n\hat{\mathbf{r}}_n$ is a sample of size n from a multinomial distribution with probabilities $\mathbf{r}$. Therefore, $(\hat{\mathbf{r}}_n - \mathbf{r})\sqrt{n} \Rightarrow \mathbf{G}$ as $n \to \infty$ (Wasserman (2011), theorem 14.6). The Hadamard derivative of the map $(\mathbf{r}, \mathbf{s}) \mapsto W_p^p(\mathbf{r}, \mathbf{s})$ as given in theorem 4 can now be used in the delta method from theorem 3. Together with the representation (12) of the set of dual solutions $\Phi_p^*(\mathbf{r}, \mathbf{s})$, this yields

$$W_p^p(\hat{\mathbf{r}}_n, \mathbf{r})\sqrt{n} \Rightarrow \max_{(\mathbf{u},\mathbf{v}) \in \Phi_p^*(\mathbf{r},\mathbf{r})} -\langle \mathbf{u}, \mathbf{G} \rangle \overset{\mathrm{D}}{\sim} \max_{\mathbf{u} \in \Phi_p^*} \langle \mathbf{u}, \mathbf{G} \rangle.$$

Here and in what follows $Z_1 \sim^{\mathrm{D}} Z_2$ means the distributional equality of the random variables Z_1 and Z_2. Applying to this the continuous mapping theorem with the map $t \mapsto t^{1/p}$ gives the assertion.

(b) Consider the map $(\mathbf{r}, \mathbf{s}) \mapsto W_p(\mathbf{r}, \mathbf{s}) = W_p^p(\mathbf{r}, \mathbf{s})^{1/p}$. By theorem 4 and the chain rule for Hadamard directional derivatives (Shapiro (1990), proposition 3.6), the Hadamard derivative of this map at $(\mathbf{r}, \mathbf{s})$ is given by

$$(\mathbf{h}_1, \mathbf{h}_2) \mapsto p^{-1} W_p^{1-p}(\mathbf{r}, \mathbf{s}) \Big\{ \max_{(\mathbf{u},\mathbf{v}) \in \Phi_p^*(\mathbf{r},\mathbf{s})} -(\langle \mathbf{u}, \mathbf{h}_1 \rangle + \langle \mathbf{v}, \mathbf{h}_2 \rangle) \Big\}. \tag{13}$$

An application of the delta method of theorem 3 concludes this part.

(c) Under the assumptions of theorem 1

$$\{(\hat{\mathbf{r}}_n, \hat{\mathbf{s}}_m) - (\mathbf{r}, \mathbf{s})\}\sqrt{\left(\frac{nm}{n+m}\right)} \Rightarrow (\mathbf{G}\sqrt{\lambda}, \mathbf{H}\sqrt{(1-\lambda)}). \tag{14}$$

(d) This part follows with the delta method from expressions (13) and (14).

For part (c) we use, as we did for part (a), the derivative given in theorem 4 and the continuous mapping theorem. The limit distribution is

$$\Big[\max_{(\mathbf{u},\mathbf{v})\in\Phi_p^*(\mathbf{r},\mathbf{s})}\{\langle\mathbf{G},\mathbf{u}\rangle\sqrt{\lambda} + \langle\mathbf{H},\mathbf{v}\rangle\sqrt{(1-\lambda)}\}\Big]^{1/p}.$$

If $\mathbf{r}=\mathbf{s}$ we have $(\mathbf{u},\mathbf{v})\in\Phi_p^*(\mathbf{r},\mathbf{s})$ if and only if $\mathbf{u}\in\Phi_p^*$ and $\mathbf{v}=-\mathbf{u}$, by expressions (12) and (3). Hence, with $\mathbf{G}\sim^{\mathrm{D}}\mathbf{H}$ we conclude that

$$\begin{aligned}\max_{(\mathbf{u},\mathbf{v})\in\Phi_p^*(\mathbf{r},\mathbf{s})}\{\langle\mathbf{G},\mathbf{u}\rangle\sqrt{\lambda} + \langle\mathbf{H},\mathbf{v}\rangle\sqrt{(1-\lambda)}\} &\overset{\mathrm{D}}{\sim} \max_{\mathbf{u}\in\Phi_p^*}\{\langle\mathbf{G},\mathbf{u}\rangle\sqrt{\lambda} - \langle\mathbf{H},\mathbf{u}\rangle\sqrt{(1-\lambda)}\}\\ &\overset{\mathrm{D}}{\sim} \max_{\mathbf{u}\in\Phi_p^*}\langle\mathbf{G},\mathbf{u}\rangle\sqrt{\{\lambda+(1-\lambda)\}}\\ &= \max_{\mathbf{u}\in\Phi_p^*}\langle\mathbf{G},\mathbf{u}\rangle.\end{aligned}$$

2.5. Explicit limiting distribution for tree metrics

Assume that the metric structure on $\mathcal{X}$ is given by a weighted tree, i.e. an undirected connected graph $\mathcal{T}=(\mathcal{X},E)$ with vertices $\mathcal{X}$ and edges $E\subset\mathcal{X}\times\mathcal{X}$ that contains no cycles. We assume that the edges are weighted by a function $w: E\to\mathbb{R}_{>0}$. For $x,x'\in\mathcal{X}$ let $e_1,\ldots,e_l\in E$ be the unique path in $\mathcal{T}$ joining x and x'; then the length of this path, $d_{\mathcal{T}}(x,x')=\Sigma_{j=1}^{l}w(e_j)$, defines a metric $d_{\mathcal{T}}$ on $\mathcal{X}$. Without imposing any further restriction on $\mathcal{T}$, we assume that it is rooted at $\mathrm{root}(\mathcal{T})\in\mathcal{X}$, say. Then, for $x\in\mathcal{X}$ and $x\neq\mathrm{root}(\mathcal{T})$, we may define $\mathrm{parent}(x)\in\mathcal{X}$ as the immediate neighbour of x in the unique path connecting x and $\mathrm{root}(\mathcal{T})$. We set $\mathrm{parent}\{\mathrm{root}(\mathcal{T})\}=\mathrm{root}(\mathcal{T})$. We also define $\mathrm{children}(x)$ as the set of vertices $x'\in\mathcal{X}$ such that there is a sequence $x'=x_1,\ldots,x_l=x\in\mathcal{X}$ with $\mathrm{parent}(x_j)=x_{j+1}$ for $j=1,\ldots,l-1$. With this definition $x\in\mathrm{children}(x)$. Additionally, define the linear operator $S_{\mathcal{T}}:\mathbb{R}^{\mathcal{X}}\to\mathbb{R}^{\mathcal{X}}$:

$$(S_{\mathcal{T}}\mathbf{u})_x = \sum_{x'\in\mathrm{children}(x)} u_{x'}.$$

Theorem 5. Let $p\geqslant 1$, $\mathbf{r}\in\mathcal{P}_{\mathcal{X}}$, defining a probability distribution on $\mathcal{X}$, and let the empirical measures $\hat{\mathbf{r}}_n$ and $\hat{\mathbf{s}}_m$ be generated by independent random variables $X_1,\ldots,X_n$ and $Y_1,\ldots,Y_m$ respectively, all drawn from $\mathbf{r}=\mathbf{s}$.

Then, with a Gaussian vector $\mathbf{G}\sim\mathcal{N}\{0,\Sigma(\mathbf{r})\}$ as defined in expression (4) we have the following results.

(a) (One sample): as $n\to\infty$,

$$n^{1/(2p)}W_p(\hat{\mathbf{r}}_n,\mathbf{r}) \Rightarrow \Big[\sum_{x\in\mathcal{X}}|(S_{\mathcal{T}}\mathbf{G})_x|d_{\mathcal{T}}\{x,\mathrm{parent}(x)\}^p\Big]^{1/p}. \tag{15}$$

(b) (Two samples): if $n\wedge m\to\infty$ and $n/(n+m)\to\lambda\in(0,1)$ we have

$$\left(\frac{nm}{n+m}\right)^{1/(2p)}W_p(\hat{\mathbf{r}}_n,\hat{\mathbf{s}}_m) \Rightarrow \Big[\sum_{x\in\mathcal{X}}|(S_{\mathcal{T}}\mathbf{G})_x|d_{\mathcal{T}}\{x,\mathrm{parent}(x)\}^p\Big]^{1/p}. \tag{16}$$

The proof of theorem 5 is given in the on-line supplementary material. Theorem 5 includes

the special case of a discrete measure on the real line, i.e. $\mathcal{X} \subset \mathbb{R}$ since, in this case, $\mathcal{X}$ can be regarded as a simple rooted tree consisting of only one branch.

Corollary 1 (Samworth and Johnson (2004), theorem 2.6). Let $\mathcal{X} = \{x_1 < \ldots < x_N\} \in \mathbb{R}$, $\mathbf{r} \in \mathcal{P}_{\mathcal{X}}$ and $\hat{\mathbf{r}}_n$ the empirical measure generated by IID random variables $X_1, \ldots, X_n \sim \mathbf{r}$. With $\bar{r}_j = \sum_{i=1}^{j} r_{x_i}$, for $j = 1, \ldots, N$ and B a standard Brownian bridge, we have, as $n \to \infty$,

$$n^{1/4} W_2(\hat{\mathbf{r}}_n, \mathbf{r}) \Rightarrow \left\{ \sum_{j=1}^{N-1} |B(\bar{r}_j)|(x_{j+1} - x_j)^2 \right\}^{1/2}. \tag{17}$$

3. Simulations and applications

The following numerical experiments were performed by using R (R Core Team, 2016). All computations of Wasserstein distances and optimal transport plans as well as their visualizations were performed with the R package `transport` (Schuhmacher *et al.*, 2014; Gottschlich and Schuhmacher, 2014). The code that was used for the computation of the limiting distributions is available as the R package `otinference` (Sommerfeld, 2017).

3.1. Speed of convergence

We investigate the speed of convergence to the limiting distribution in theorem 1 in the one-sample case under the null hypothesis. For this, we consider as ground space $\mathcal{X}$ a regular two-dimensional $L \times L$ grid with the Euclidean distance as the metric d and $L = 3, 5, 10$. We generate five random measures $\mathbf{r}$ on $\mathcal{X}$ as realizations of a Dirichlet random variable with concentration parameter $\boldsymbol{\alpha} = (\alpha, \ldots, \alpha) \in \mathbb{R}^{L \times L}$ for $\alpha = 1, 5, 10$. $\alpha = 1$ corresponds to a uniform distribution on the probability simplex. For each measure, we generate 20 000 realizations of $n^{1/(2p)} W_p(\hat{\mathbf{r}}_n, \mathbf{r})$ with $n\hat{\mathbf{r}}_n \sim \text{Multinom}(\mathbf{r})$ for $n = 10, 100, 1000$ and of the theoretical limiting distribution given in theorem 1. The Kolmogorov–Smirnov distance (i.e. the maximum absolute difference between their cumulative distribution functions) between these two samples (averaged over the five measures) is shown in Fig. 1. The experiment shows that the limiting distribution is a good approximation of the finite sample version even for small sample sizes. For the parameters considered the size of the ground space $N = L^2$ seems to slow the convergence only marginally. Similarly, the underlying measure seems to have no sizable effect on the speed of convergence as the dependence on the concentration parameter α demonstrates.

3.2. Testing the null: real and synthetic fingerprints

The generation and recognition of synthetic fingerprints is a topic of great interest in forensic science and current state of the art methods (Cappelli *et al.*, 2000) produce synthetic fingerprints that even human experts fail to recognize as such (Maltoni *et al.* (2009), page 292 and following feature). Recently, Gottschlich and Huckemann (2014) presented a method using the Wasserstein distance that can distinguish synthetic from real fingerprints with high accuracy. Their method is probabilistic in nature, since it is based on a hypothesized unknown distribution of certain features of the fingerprint. We use our distributional limits to assess the statistical significance of the differences.

3.2.1. Minutiae histograms

The basis for the comparison of fingerprints is so-called *minutiae*, which are key qualities in biometric identification based on fingerprints (Jain, 2007). They are certain characteristic features

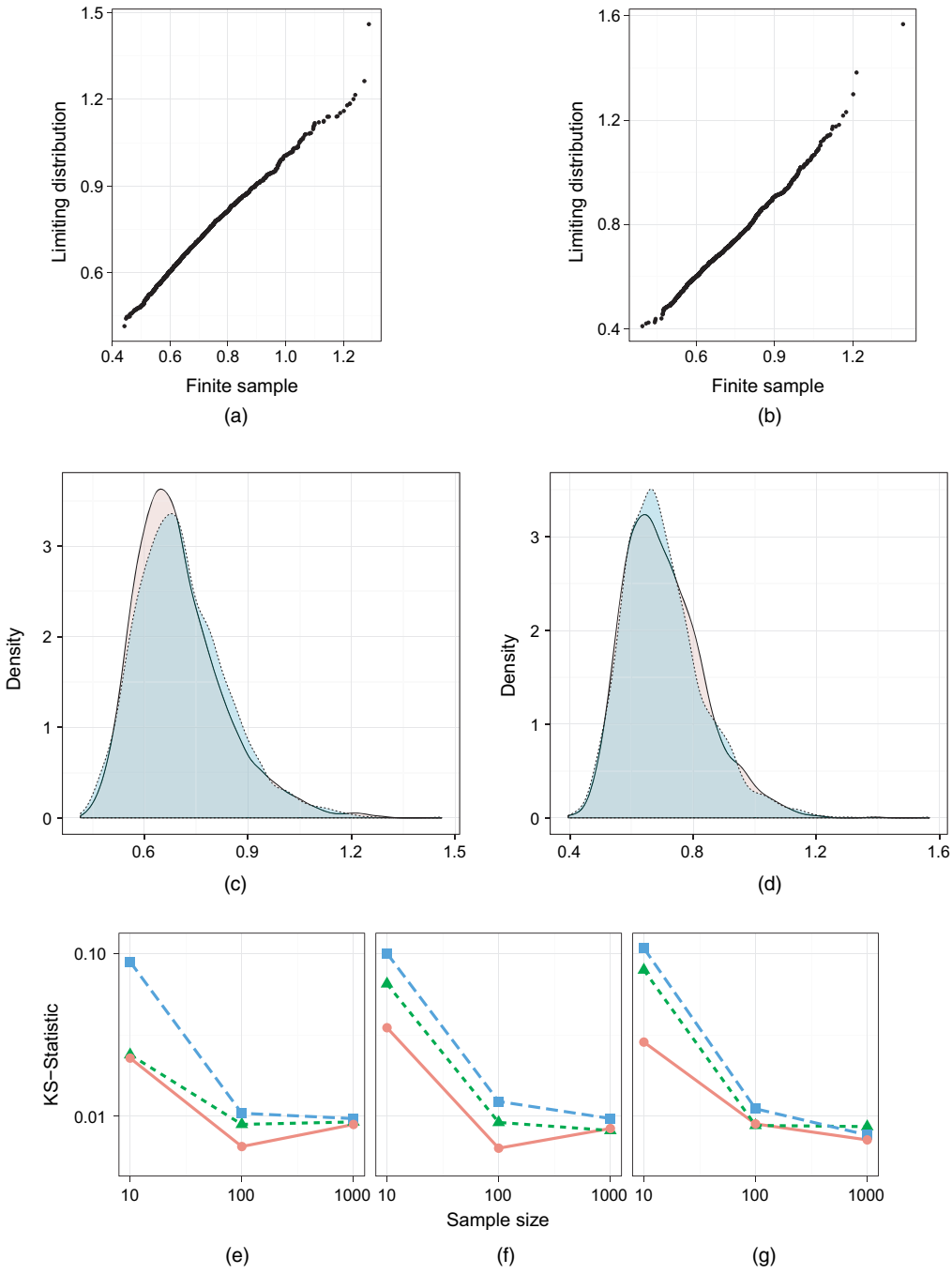

Fig. 1. Comparison of the finite sample distribution and the theoretical limiting distribution on a regular grid of length L for various sample sizes: (a), (b) *QQ*-plots, for sample sizes $n = 50$ and $n = 1000$ respectively, (c), (d) kernel density estimates (bandwidth, Silverman's rule of thumb (Silverman, 1986); ———, finite sample, · · · · · ·, limiting distribution) for $L = 10$ and sample sizes $n = 50$ and $n = 1000$ respectively, and (e), (f), (g) Kolmogorov–Smirnov statistic between the two distributions as a function of the sample size for $L = 3$ (●), 5 (▲), 10 (■) and for concentration parameters $\alpha = 1, 5, 10$ respectively

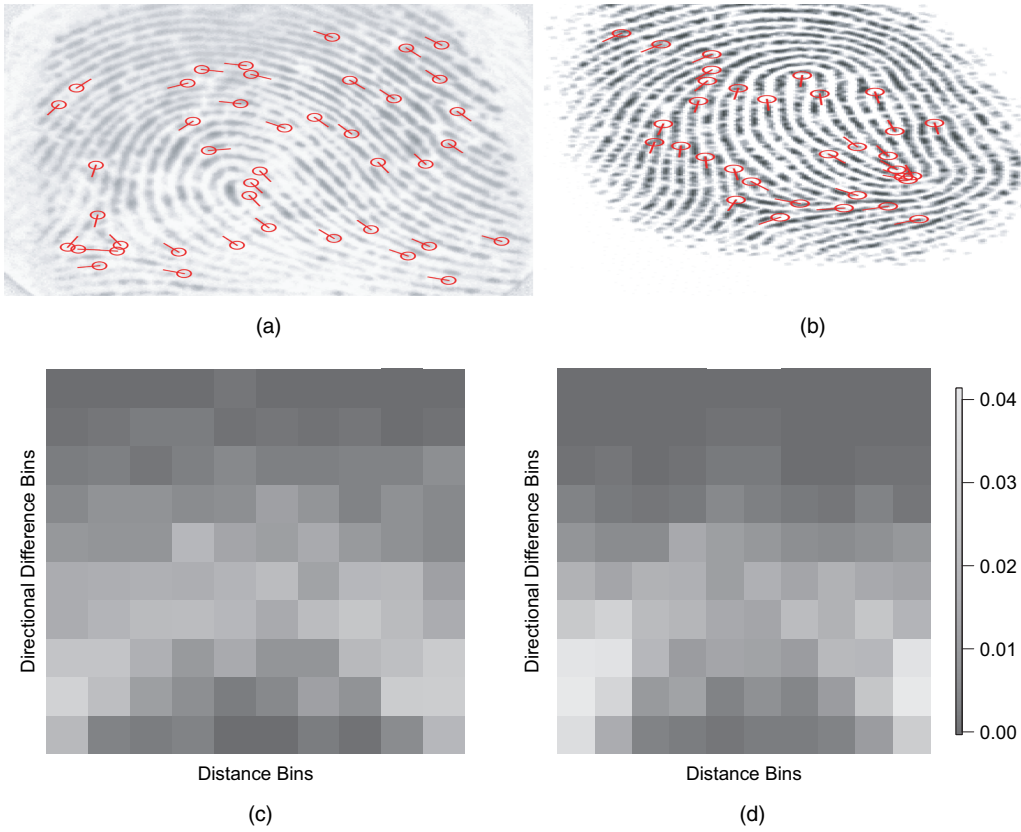

Fig. 2. Minutiae of (a) a real and (b) a synthetic fingerprint, and MHs of (c) real and (d) synthetic fingerprints

such as bifurcations of the line patterns of the fingerprint. Each of the *minutiae* have a location in the fingerprint and a direction such that it can be characterized by two real numbers and an angle. Fig. 2 shows a real and a synthetic fingerprint with their *minutiae*.

The recognition method of Gottschlich and Huckemann (2014) considers pairs of *minutiae* and records their distance and the difference between their angles. On the basis of these two values each *minutiae* pair is put in one of 100 bins arranged in a regular grid (10 directional by 10 distance bins) to obtain a so-called *minutiae* histogram (MH). On the basis of the binwise mean of MHs for several fingerprints to construct a typical MH, they found that the proximity in Wasserstein distance to these references is a good classifier for distinguishing real and synthetic fingerprints.

To assess the statistical significance of the difference in *minutiae* pair distributions, we consider fingerprints from the databases 1 and 4 of the fingerprint verification competition of 2002 (Maio *et al.*, 2002), containing 110 real and synthetic fingerprints respectively. From each database the *minutiae* were obtained by an automatic procedure using a commercial off-the-shelf programme. For each fingerprint we chose disjoint *minutiae* pairs at random to avoid the issue of pairs being dependent, yielding a total of 1917 and 1437 *minutiae* pairs from real and synthetic fingerprints respectively.

Whereas two-sample tests for univariate data are abundant and well studied there are no multivariate methods that could be considered standard in this setting. Therefore, we report

Table 1. Results of different two-sample tests for the difference in the distribution of MHs of real and fake fingerprints

	Results for the following tests:			
	Wasserstein	*Cross-match*	*Permutation*	*Kernel density estimate*
Raw	0.00	2.99×10^{-1}	1.00×10^{-3}	1.12×10^{-8}
Centred	4.00×10^{-4}	4.48×10^{-5}	1.00×10^{-3}	2.60×10^{-21}
Centred and scaled	2.54×10^{-2}	1.01×10^{-2}	1.71×10^{-1}	1.79×10^{-14}

on the findings of several tests from the literature for comparison with the Wasserstein-based method from expression (7). We tested the null hypothesis that the underlying distributions are equal for the uncentred, the centred and the centred and scaled (to variance 1) data to assess effects beyond first moments by using the following methods:

(a) comparing the empirical Wasserstein distance W_1 after binning on a regular 10×10 grid with the limiting distribution from theorem 1;
(b) a permutation test;
(c) the cross-match test proposed by Rosenbaum (2005);
(d) the kernel-based test (Anderson *et al.*, 1994) implemented in the R package `ks`.

Table 1 shows the resulting empirical distributions on a 10×10 grid and the p-values for the various tests. The differences are highly significant according to all tests, except the permutation test for the centred and scaled data. In this particular example at least, the Wasserstein-based test seems to be able to pick up differences in distributions (in the first moment and beyond) at least as well as current state of the art methods.

In addition to testing, the Wasserstein method provides us with an optimal transport plan, transforming one measure into the other. For the MHs under consideration this is illustrated in Fig. 3. This transport plan gives information beyond a simple test for equality as it highlights structural changes in the distribution. In this specific application it reveals how in the MH of synthetic fingerprints compared with the histogram of real fingerprints mass has been shifted from large and intermediate directional differences to smaller differences, in particular to small and large distances, and only to a lesser extent to intermediate distances. In conclusion we may say that synthetic fingerprints show smaller differences in the directions of *minutiae* and stronger clustering of *minutiae* distances around small and large values. Insight of this sort may lead to improved generation or detection of synthetic fingerprints.

3.3. Asymptotics under the alternative: metagenomics

Metagenomics studies microbial communities by analysing genetic material in an environmental sample such as a stool sample of a human. High throughput sequencing techniques no longer require cultivated cloned microbial cultures to perform sequencing. Instead, a sample with potentially many different species can be analysed directly and the abundance of each species in the sample can be recovered. The applications of this technique are countless and constantly growing. In particular, the composition of microbial communities in the human gut has been associated with obesity, inflammatory bowel disease and others (Turnbaugh *et al.*, 2007).

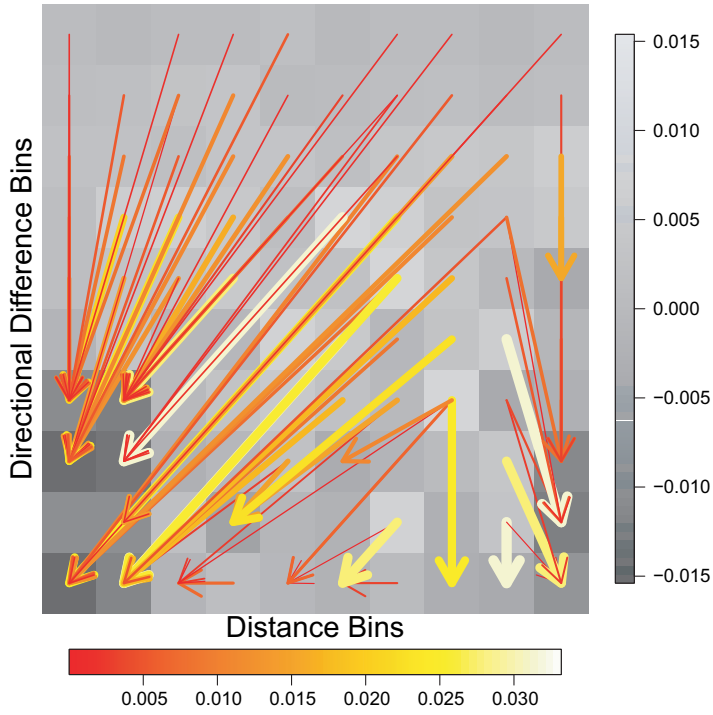

Fig. 3. Optimal transport plan between the MHs of real and fake fingerprints: the grey values indicate the magnitude of the difference of the two MHs and the arrows show the transport; the amount of mass transported is encoded in the colour and thickness of the arrows

The analysis of a sample with high throughput sequencing techniques yields several thousands to many hundreds of thousands of sequences. After elaborate preprocessing, these sequences are aligned to a reference database and clustered in *operational taxonomic units* (OTUs). These OTUs can be thought of (albeit omitting some biological detail) as the different species in the sample. For each OTU this analysis yields the number of sequences that are associated with it, i.e. how often this particular OTU was detected in the sample. Further, comparing the genetic sequences that are associated with an OTU yields a biologically meaningful measure of similarity between OTUs—and hence a distance. A metagenomic sample can therefore be regarded as a sample in a discrete metric space with OTUs being the points of the space. Comparing such samples representing microbial communities is of great interest (Kuczynski *et al.*, 2010). The Wasserstein distance has been recognized to provide valuable insight and to facilitate tests for equality of two communities (Evans and Matsen, 2012). This previous application, however, relies on a phylogenetic tree that is built on the OTUs and the distance is then measured in the tree. This additional preprocessing step involves many parameter choices and is unnecessary with our method.

A further drawback of the method of Evans and Matsen (2012) is that it allows only for testing the null hypothesis that two communities are equal. In practice, one frequently finds that natural variation is so high that even two samples from the same source taken at different times will be recognized as different. This raises the question whether variation within samples from the same source is smaller than the difference to samples of another source. Statistically speaking we are looking for confidence sets for differences which are assumed to be different from zero. This requires asymptotics under the alternative $\mathbf{r} \neq \mathbf{s}$, which is provided by theorem 1.

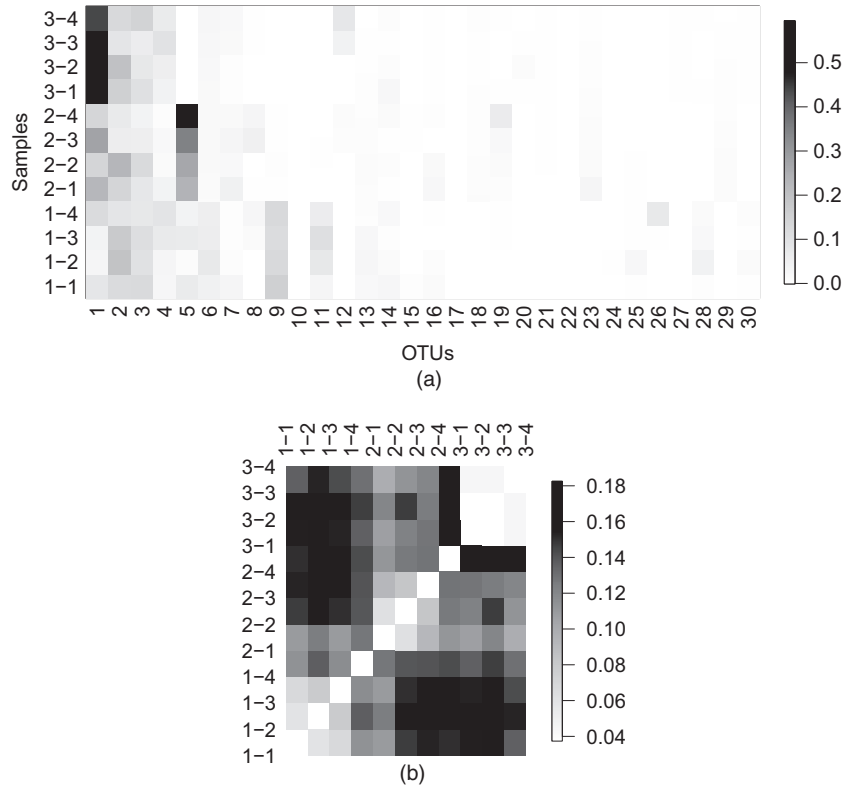

Fig. 4. (a) Relative abundances of the 30 first OTUs in the 12 samples and (b) Wasserstein distances of the microbial communities: here, *ij* is the *j*th sample of the *i*th person.

3.3.1. Data analysis

We consider part of the data of Costello *et al*. (2009). Four stool samples were taken from each of three people at different times. We used the preparation of these data by P. Schloss available from `https://www.mothur.org/w/images/d/d8/CostelloData.zip`. The reads were preprocessed with the program `mothur` (Schloss *et al*., 2009) using the procedure that was outlined in Schloss *et al*. (2011) and Schloss (2015). The relative abundances of the 30 most frequent OTUs and the Wasserstein 2 distances of the microbial communities are shown in Fig. 4. In this and all other figures we use $i - j$ to denote sample j of person i. Note that it is typical for these data that most of the mass is concentrated on a few OTUs.

The Wasserstein 2 distances for all 66 pairs and their 99% confidence intervals were computed by using the asymptotic distribution in theorem 1. The results are shown in Fig. 5. The entire analysis took less than 1 min on a standard laptop. The confidence intervals show that intrapersonal distances are in fact significantly smaller than interpersonal distances.

4. Discussion

We discuss limitations, possible extensions of the work presented and promising directions for future research.

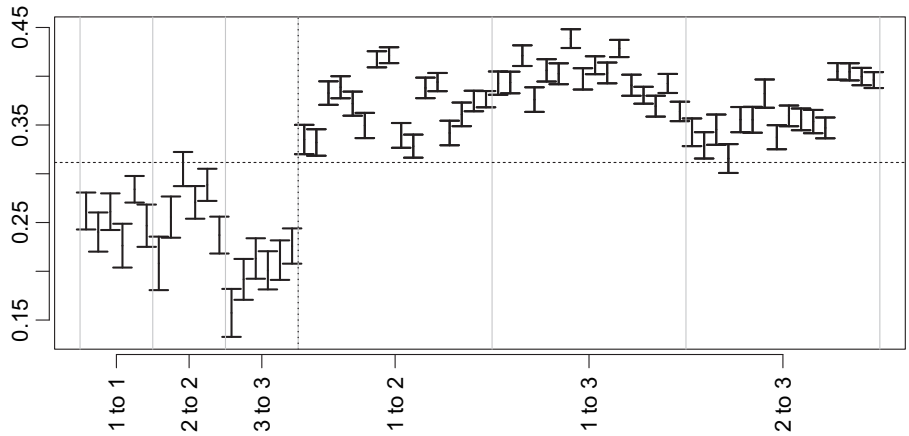

Fig. 5. Display of 95% confidence intervals of Wasserstein distances of microbial communities: the horizontal axis shows which person pair the distances belong to (separated by grey vertical lines) and the dotted vertical line separates intrapersonal (lower) from interpersonal (upper) distances

4.1. Beyond finite spaces I: rates in the finite and the continuous setting ($d = 1$)

The scaling rate in theorem 1 depends solely on p and is completely independent of the underlying space $\mathcal{X}$. This contrasts known bounds on the rate of convergence in the continuous case (see the references in Section 1), which exhibit a strong dependence on the dimension of the space and the moments of the distribution.

Under the null hypothesis (i.e. the two underlying population measures are equal) and when $\mathcal{X} = \mathbb{R}$ and $p = 2$, the scaling rate for a continuous distribution is known to be $n^{1/2}$, at least under additional tail conditions (see for example Del Barrio *et al.* (2005)). This means that in this case the scaling rate for a discrete distribution is slower (namely $n^{1/4}$). Under the alternative (different population measures) the scaling rate is $n^{1/2}$ and coincides in the discrete and the continuous case (see Munk and Czado (1998)).

4.2. Beyond finite spaces II: higher dimensions ($d \geqslant 2$)

For a continuous measure μ the Wasserstein distance is the solution of an infinite dimensional optimization problem. Although differentiability results also exist for such problems (e.g. Shapiro (1992)), there are strong indications that the argument that is presented here cannot carry over to the case for $d \geqslant 2$. This is most easily seen from the classical results of Ajtai *et al.* (1984). We consider the uniform distribution on the unit square. For two samples of size n independently drawn from this distribution, Ajtai *et al.* (1984) showed that there are constants C_1 and C_2 such that the Wasserstein 1 distance $\hat{W}_1^{(n)}$ between them satisfies

$$C_1 n^{-1/2} \log(n)^{1/2} \leqslant \hat{W}_1^{(n)} \leqslant C_2 n^{-1/2} \log(n)^{1/2}$$

with probability $1 - o(1)$. Hence, for $c_n \hat{W}_1^{(n)}$ to have a non-degenerate limit, we need $c_n = \sqrt{\{n/\log(n)\}}$. However, a common property of all delta methods is that they preserve the rate of convergence, which is not satisfied here.

4.3. Transport distances on trees

Complementing our theorem 5 a further result on transport distances on trees was proved by Evans and Matsen (2012) in the context of phylogenetic trees for the comparison of metagenomic samples (see also our application in Section 3). They pointed out that the Wasserstein 1 distance

on trees is equal to the so-called *weighted unifrac distance* which is very popular in genetics. Inspired by this distance they gave a formal generalization mimicking a cost exponent $p > 1$ and considered its asymptotic behaviour. However, as they remarked, these generalized expressions are no longer related (beyond a formal resemblance) to Wasserstein distances with cost exponent $p > 1$. Comparing the performance of their *ad hoc* metric and the true Wasserstein distance on trees that is under consideration here is an interesting topic for further research.

4.4. Bootstrap

We showed that, whereas the naive n out of n bootstrap fails for the Wasserstein distance (on-line supplementary material section B), the m out of n bootstrap is consistent. An interesting and challenging question is how m should be chosen.

4.5. Wasserstein barycentres

Barycentres in the Wasserstein space (Agueh and Carlier, 2011) have recently received much attention (Cuturi and Doucet, 2014; Del Barrio *et al*., 2015). We expect that the techniques that are developed here can be of use in providing a rigorous statistical theory (e.g. distributional limits). The same applies to geodesic principal component analysis in the Wasserstein space (Bigot *et al*., 2013; Seguy and Cuturi, 2015).

4.6. Alternative cost matrices and transport distances

Theorem 1 holds in very large generality for arbitrary cost matrices, including in particular the case of a cost matrix derived from a metric but using a cost exponent $p < 1$.

Beyond this obvious modification it seems worthwhile to extend the methodology of directional differentiability in conjunction with a delta method to other functionals related to optimal transport, e.g. entropically regularized (Cuturi, 2013) or sliced Wasserstein distances (Bonneel *et al*., 2015). This would require a careful investigation of the analytical properties of these quantities similarly to classical results for the Wasserstein distance.

Acknowledgements

The authors gratefully acknowledge support by Deutsche Forschungsgemeinschaft Research Training Group 2088, project A1. They thank L. Dümbgen, A. Hein, S. Huckemann, C. Gottschlich, D. Schuhmacher and R. Schultz for helpful discussions and C. Tameling for careful reading of the manuscript.

References

Agueh, M. and Carlier, G. (2011) Barycenters in the Wasserstein space. *SIAM J. Math. Anal.*, **43**, 904–924.

Agulló-Antolín, M., Cuesta-Albertos, J. A., Lescornel, H. and Loubes, J.-M. (2015) A parametric registration model for warped distributions with Wasserstein's distance. *J. Multiv. Anal.*, **135**, 117–130.

Ajtai, M., Komlós, J. and Tusnády, G. (1984) On optimal matchings. *Combinatorica*, **4**, 259–264.

Ambrosio, L. (2003) Lecture notes on optimal transport problems. In *Mathematical Aspects of Evolving Interfaces* (eds L. Ambrosio, K. Deckelnick, G. Dziuk, M. Mimura, V. A. Solonnikov and H. Mete Soner), pp. 1–52. New York: Springer.

Anderson, N. H., Hall, P. and Titterington, D. M. (1994) Two-sample test statistics for measuring discrepancies between two multivariate probability density functions using kernel-based density estimates. *J. Multiv. Anal.*, **50**, 41–54.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc.* B, **57**, 289–300.

Bickel, P. J. and Freedman, D. A. (1981) Some asymptotic theory for the bootstrap. *Ann. Statist.*, **9**, 1196–1217.
Bigot, J., Gouet, R., Klein, T. and López, A. (2013) Geodesic PCA in the Wasserstein space by convex PCA. *Ann. Inst. H. Poincaré Probab. Statist.*, **53**, 1–26.
Bobkov, S. and Ledoux, M. (2014) One-dimensional empirical measures, order statistics and Kantorovich transport distances. *Preprint*.
Boissard, E. and Gouic, T. L. (2014) On the mean speed of convergence of empirical and occupation measures in Wasserstein distance. *Ann. Inst. H. Poincaré Probab. Statist.*, **50**, 539–563.
Boissard, E., Gouic, T. L. and Loubes, J.-M. (2015) Distribution's template estimate with Wasserstein metrics. *Bernoulli*, **21**, 740–759.
Bonnans, J. F. and Shapiro, A. (2013) *Perturbation Analysis of Optimization Problems*. New York: Springer.
Bonneel, N., Rabin, J., Peyré, G. and Pfister, H. (2015) Sliced and Radon Wasserstein barycenters of measures. *J. Math. Imgng Visn*, **51**, 22–45.
Cappelli, R., Erol, A., Maio, D. and Maltoni, D. (2000) Synthetic fingerprint-image generation. In *Proc. 15th Int. Conf. Pattern Recognition*, vol. 3, pp. 471–474.
Costello, E. K., Lauber, C. L., Hamady, M., Fierer, N., Gordon, J. I. and Knight, R. (2009) Bacterial community variation in human body habitats across space and time. *Science*, **326**, 1694–1697.
Cuturi, M. (2013) Sinkhorn distances: lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems* (eds C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani and K. Q. Weinberger), pp. 2292–2300. Red Hook: Curran Associates.
Cuturi, M. and Doucet, A. (2014) Fast computation of Wasserstein barycenters. In *Proc. 31st Int. Conf. Machine Learning, Beijing*, pp. 685–693.
Del Barrio, E., Cuesta-Albertos, J. A., Matrán, C. and Rodríguez-Rodríguez, J. M. (1999) Tests of goodness of fit based on the L2-Wasserstein distance. *Ann. Statist.*, **27**, 1230–1239.
Del Barrio, E., Giné, E. and Utzet, F. (2005) Asymptotics for L2 functionals of the empirical quantile process, with applications to tests of fit based on weighted Wasserstein distances. *Bernoulli*, **11**, 131–189.
Del Barrio, E., Lescornel, H. and Loubes, J.-M. (2015) A statistical analysis of a deformation model with Wasserstein barycenters: estimation procedure and goodness of fit test. *Preprint arXiv:1508.06465*. Universidad de Valladolid, Valladolid.
Dobrushin, R. (1970) Prescribing a system of random variables by conditional distributions. *Theory Probab. Appl.*, **15**, 458–486.
Donoho, D. L. and Liu, R. C. (1988) Pathologies of some minimum distance estimators. *Ann. Statist.*, **16**, 587–608.
Dorea, C. C. Y. and Ferreira, D. B. (2012) Conditions for equivalence between Mallows distance and convergence to stable laws. *Acta Math. Hung.*, **134**, 1–11.
Dümbgen, L. (1993) On nondifferentiable functions and the bootstrap. *Probab. Theory Reltd Flds*, **95**, 125–140.
Dümbgen, L., Samworth, R. and Schuhmacher, D. (2011) Approximation by log-concave distributions, with applications to regression. *Ann. Statist.*, **39**, 702–730.
Erbar, M. and Maas, J. (2012) Ricci curvature of finite Markov chains via convexity of the entropy. *Arch. Ratnl Mech. Anal.*, **206**, 997–1038.
Evans, S. N. and Matsen, F. A. (2012) The phylogenetic Kantorovich–Rubinstein metric for environmental sequence samples. *J. R. Statist. Soc.* B, **74**, 569–592.
Fang, Z. and Santos, A. (2014) Inference on directionally differentiable functions. *Preprint arXiv:1404.3763*. Department of Economics, Kansas State University, Manhattan.
Fournier, N. and Guillin, A. (2014) On the rate of convergence in Wasserstein distance of the empirical measure. *Probab. Theory Reltd Flds*, **162**, 1–32.
Freitag, G., Czado, C. and Munk, A. (2007) A nonparametric test for similarity of marginals—with applications to the assessment of population bioequivalence. *J. Statist. Planng Inf.*, **137**, 697–711.
Freitag, G. and Munk, A. (2005) On Hadamard differentiability in k-sample semiparametric models—with applications to the assessment of structural relationships. *J. Multiv. Anal.*, **94**, 123–158.
Gal, T., Greenberg, H. J. and Hillier, F. S. (eds) (1997) *Advances in Sensitivity Analysis and Parametric Programming*, vol. 6. New York: Springer.
Gangbo, W. and McCann, R. J. (2000) Shape recognition via Wasserstein distance. *Q. Appl. Math.*, **58**, 705–737.
Gelbrich, M. (1990) On a formula for the L2 Wasserstein metric between measures on Euclidean and Hilbert spaces. *Math. Nachr.*, **147**, 185–203.
Gottschlich, C. and Huckemann, S. (2014) Separating the real from the synthetic: minutiae histograms as fingerprints of fingerprints. *Inst. Engng Technol. Biometr.*, **3**, 291–301.
Gottschlich, C. and Schuhmacher, D. (2014) The Shortlist method for fast computation of the earth mover's distance and finding optimal solutions to transportation problems. *PLOS ONE*, **9**, no. 10, article e110214.
Gozlan, N., Roberto, C., Samson, P.-M. and Tetali, P. (2013) Displacement convexity of entropy and related inequalities on graphs. *Probab. Theory Reltd Flds*, **160**, 47–94.
Gray, R. M. (1988) *Probability, Random Processes, and Ergodic Properties*. New York: Springer.
Halder, A. and Bhattacharya, R. (2011) Model validation: a probabilistic formulation. In *Proc. 50th Conf. Decision and Control and European Control Conf.*, pp. 1692–1697. New York: Institute of Electrical and Electronics Engineers.

Horowitz, J. and Karandikar, R. L. (1994) Mean rates of convergence of empirical measures in the Wasserstein metric. *J. Computnl Appl. Math.*, **55**, 261–273.
Jain, A. K. (2007) Technology: Biometric recognition. *Nature*, **449**, 38–40.
Johnson, O. and Samworth, R. (2005) Central limit theorem and convergence to stable laws in Mallows distance. *Bernoulli*, **11**, 829–845.
Jordan, R., Kinderlehrer, D. and Otto, F. (1998) The variational formulation of the Fokker–Planck Equation. *SIAM J. Math. Anal.*, **29**, 1–17.
Kantorovich, L. V. and Rubinstein, G. S. (1958) On a space of completely additive functions. *Vestn. Leningrad Univ.*, **13**, 52–59.
Kloeckner, B. R. (2013) A geometric study of Wasserstein spaces: ultrametrics. *Mathematika*, **61**, 1–17.
Kuczynski, J., Liu, Z., Lozupone, C., McDonald, D., Fierer, N. and Knight, R. (2010) Microbial community resemblance methods differ in their ability to detect biologically relevant patterns. *Nat. Meth.*, **7**, 813–819.
Luenberger, D. G. and Ye, Y. (2008) *Linear and Nonlinear Programming*. New York: Springer.
Maio, D., Maltoni, D., Cappelli, R., Wayman, J. L. and Jain, A. K. (2002) FVC2002: second fingerprint verification competition. In *Proc. 16th Int. Conf. Pattern Recognition*, vol. 3, pp. 811–814. New York: Institute of Electrical and Electronics Engineers.
Mallows, C. L. (1972) A note on asymptotic joint normality. *Ann. Math. Statist.*, **43**, 508–515.
Maltoni, D., Maio, D., Jain, A. K. and Prabhakar, S. (2009) *Handbook of Fingerprint Recognition*. New York: Springer.
Mason, D. M. (2016) A weighted approximation approach to the study of the empirical Wasserstein distance. In *High Dimensional Probability*, vol. VII, *The Cargèse Volume* (eds C. Houdré, D. M. Mason, P. Reynaud-Bouret and Jan Rosiński), pp. 137–154. Cham: Springer.
Munk, A. and Czado, C. (1998) Nonparametric validation of similar distributions and assessment of goodness of fit. *J. R. Statist. Soc.* B, **60**, 223–241.
Ni, K., Bresson, X., Chan, T. and Esedoglu, S. (2009) Local histogram based segmentation using the Wasserstein distance. *Int. J. Comput. Visn*, **84**, 97–111.
Orlova, D. Y., Zimmerman, N., Meehan, S., Meehan, C., Waters, J., Ghosn, E. E. B., Filatenkov, A., Kolyagin, G. A., Gernez, Y., Tsuda, S., Moore, W., Moss, R. B., Herzenberg, L. A. and Walther, G.. (2016) Earth mover's distance (EMD): a true metric for comparing biomarker expression levels in cell populations. *PLOS ONE*, **11**, no. 3, article e0151859.
Otto, F. (2001) The geometry of dissipative evolution equations: the porous medium equation. *Communs Partl Differentl Equns*, **26**, 101–174.
Oudre, L., Jakubowicz, J., Bianchi, P. and Simon, C. (2012) Classification of periodic activities using the Wasserstein distance. *IEEE Trans. Biomed. Engng*, **59**, 1610–1619.
Rachev, S. T. (1985) The Monge-Kantorovich mass transference problem and its stochastic applications. *Theory Probab. Appl.*, **29**, 647–676.
Rachev, S. T. and Rüschendorf, L. (1998) *Mass Transportation Problems*, vol. I, *Theory*. Berlin: Springer.
R Core Team (2016) *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.
Rippl, T., Munk, A. and Sturm, A. (2015) Limit laws of the empirical Wasserstein distance. *J. Multiv. Anal.*, **151**, 90–109.
Rockafellar, R. T. (1997) Directional differentiability of the optimal value function in a nonlinear programming problem. In *Advances in Sensitivity Analysis and Parametric Programming*, vol. 6 (eds T. Gal and H. J. Greenberg). New York: Springer.
Römisch, W. (2004) Delta method, infinite dimensional. In *Encyclopedia of Statistical Sciences*. New York: Wiley.
Rosenbaum, P. R. (2005) An exact distribution-free test comparing two multivariate distributions based on adjacency. *J. R. Statist. Soc.* B, **67**, 515–530.
Rubner, Y., Tomasi, C. and Guibas, L. J. (2000) The earth mover's distance as a metric for image retrieval. *Int. J. Comput. Visn*, **40**, 99–121.
Ruttenberg, B. E., Luna, G., Lewis, G. P., Fisher, S. K. and Singh, A. K. (2013) Quantifying spatial relationships from whole retinal images. *Bioinformatics*, **29**, 940–946.
Samworth, R. and Johnson, O. (2004) Convergence of the empirical process in Mallows distance, with an application to bootstrap performance. *Preprint arXiv:math/0406603*. Centre for Mathematical Sciences, Cambridge.
Samworth, R. and Johnson, O. (2005) The empirical process in Mallows distance, with application to goodness-of-fit tests. *Preprint arXiv:math/0504424*. Centre for Mathematical Sciences, Cambridge.
Schloss, P. D. (2015) Schloss lab 454 standard operating procedure. Department of Microbiology and Immunology, University of Michigan, Ann Arbor. (Available from `http://www.mothur.org/wiki/454_SOP`.)
Schloss, P. D., Gevers, D. and Westcott, S. L. (2011) Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PLOS ONE*, **6**, no. 12, article e27310.
Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., Oakley, B. B., Parks, D. H., Robinson, C. J., Sahl, J. W., Stres, B., Thallinger, G. G., Van Horn, D. J. and Weber, C. F. (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.*, **75**, 7537–7541.

Schuhmacher, D., Gottschlich, C. and Baehre, B. (2014) R-package transport: optimal transport in various forms. *R Package*. Institut für Mathematische Stochastik, Göttingen. (Available from `https://cran.r-project.org/package=transport`.)
Seguy, V. and Cuturi, M. (2015) Principal geodesic analysis for probability measures under the optimal transport metric. In *Advances in Neural Information Processing Systems*, pp. 3312–3320.
Shapiro, A. (1990) On concepts of directional differentiability. *J. Optimizn Theory Appl.*, **66**, 477–487.
Shapiro, A. (1991) Asymptotic analysis of stochastic programs. *Ann. Ops Res.*, **30**, 169–186.
Shapiro, A. (1992) Perturbation analysis of optimization problems in Banach spaces. *Numer. Functnl Anal. Optimizn*, **13**, 97–116.
Shorack, G. R. and Wellner, J. A. (1986) *Empirical Processes with Applications to Statistics*. New York: Wiley.
Silverman, B. W. (1986) *Density Estimation for Statistics and Data Analysis*. London: Chapman and Hall.
Sommerfeld, M. (2017) Otinference: inference for optimal transport. *R Package*. University of Göttingen, Göttingen. (Available from `https://cran.r-project.org/package=otinference`.)
Srivastava, S., Li, C. and Dunson, D. B. (2015) Scalable Bayes via barycenter in Wasserstein space. *Preprint arXiv:1508.05880*. Department of Statistics and Actuarial Science, University of Iowa, Iowa City.
Talagrand, M. (1992) Matching random samples in many dimensions. *Ann. Appl. Probab.*, **2**, 846–856.
Talagrand, M. (1994) The transportation cost from the uniform measure to the empirical measure in dimension ≥ 3. *Ann. Probab.*, **22**, 919–959.
Turnbaugh, P. J., Ley, R. E., Hamady, M., Fraser-Liggett, C. M., Knight, R. and Gordon, J. I. (2007) The human microbiome project. *Nature*, **449**, 804–810.
van der Vaart, A. W. and Wellner, J. A. (1996) *Weak Convergence*. New York: Springer.
Vasershtein, L. N. (1969) Markov processes over denumerable products of spaces describing large system of automata. *Probl. Pered. Inform.*, **5**, no. 3, 64–72.
Villani, C. (2003) *Topics in Optimal Transportation*. Providence: American Mathematical Society.
Villani, C. (2008) *Optimal Transport: Old and New*. New York: Springer.
Wasserman, L. (2011) *All of Statistics*. New York: Springer Science and Business Media.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article:

'Inference for empirical Wasserstein distances on finite spaces: supplementary material'.