Information and Inference: A Journal of the IMA (2019) Page 1 of 23
doi:10.1093/imaiai/drn000

On parameter estimation with the Wasserstein distance

ESPEN BERNTON*,
Department of Statistics, Harvard University, USA
*Corresponding author: ebernton@g.harvard.edu
PIERRE E. JACOB
Department of Statistics, Harvard University, USA
pjacob@g.harvard.edu
MATHIEU GERBER
School of Mathematics, University of Bristol, UK
mathieu.gerber@bristol.ac.uk
AND
CHRISTIAN P. ROBERT
CEREMADE, Université Paris-Dauphine, PSL Research University, France, and Department of Statistics, University of Warwick, UK
xian@ceremade.dauphine.fr

[Received on 18 February 2019]

Statistical inference can be performed by minimizing, over the parameter space, the Wasserstein distance between model distributions and the empirical distribution of the data. We study asymptotic properties of such minimum Wasserstein distance estimators, complementing results derived by Bassetti, Bodini and Regazzini in 2006. In particular, our results cover the misspecified setting, in which the data-generating process is not assumed to be part of the family of distributions described by the model. Our results are motivated by recent applications of minimum Wasserstein estimators to complex generative models. We discuss some difficulties arising in the numerical approximation of these estimators. Two of our numerical examples (g-and-k and sum of log-Normals) are taken from the literature on approximate Bayesian computation, and have likelihood functions that are not analytically tractable. Two other examples involve misspecified models.

Keywords:

Wasserstein distance, parameter inference, optimal transport, minimum distance estimation

1. Introduction

We consider a statistical estimation approach for parametric models that is based on minimizing the Wasserstein distance between the empirical distribution of the data and the model distributions (Belili et al., 1999; Bassetti et al., 2006). We study two different point estimators, where the first, called the minimum Wasserstein estimator (MWE), arises as the most important special case of the estimator introduced by Bassetti et al. (2006). The second, which we term the minimum expected Wasserstein estimator (MEWE), is better suited to numerical approximations.

We derive theoretical properties of the estimators, such as existence, measurability, and consistency,

in the misspecified setting. That is, we do not assume that the observations are generated from the working model. For one-dimensional data, we also study the convergence rate and asymptotic distribution of the minimum Wasserstein estimator of order 1, extending the work of Bassetti and Regazzini (2006) on location-scale models. Our proofs are based on epi-convergence (Rockafellar and Wets, 2009) and general results on minimum distance estimation (Pollard, 1980), and are as such different from those presented by Bassetti and coauthors.

There are two main motivations for developing these results. Firstly, recent advances in computational optimal transport have led to the application of minimum Wasserstein distance estimators in increasingly complicated settings, where the models are likely to be misspecified. For instance, Genevay et al. (2018) apply the MEWE in the tuning of image generation models, and Genevay et al. (2017) show that a version of the MEWE also appears in the popular Wasserstein GAN method (Arjovsky et al., 2017). This development has been driven by the advent of efficient numerical algorithms to approximate the Wasserstein distance (see e.g. Peyré and Cuturi, 2018; Cuturi, 2013; Benamou et al., 2015; Genevay et al., 2016; Ye et al., 2017; Li et al., 2018; Altschuler et al., 2018).

Secondly, minimum Wasserstein distance estimators, which are particular instances of minimum distance estimators (Basu et al., 2011), appear to be practical and robust alternatives to likelihood-based estimation in the setting of generative models. In these models, synthetic observations can be generated given a parameter, but the likelihood function and associated maximum likelihood estimators might be intractable (Gouriéroux et al., 1993; Marin et al., 2012; Bernton et al., 2019). Some comments on the comparison between the Wasserstein distance and other distances commonly used in minimum distance estimation are provided.

The rest of this paper is organized as follows: we review the definitions of minimum distance estimation, of the Wasserstein distance, and of the estimators of interest in the rest of this section. Theoretical results, whose proofs can be found in the supplementary materials, and some open questions are stated in Section 2. We briefly review computational strategies to compute the Wasserstein distance and the estimators in Section 3, before illustrating their behavior on various examples in Section 4. We conclude in Section 5. Code to reproduce the numerical results can be found at https://github.com/pierrejacob/winference.

1.1 *Notation*

Throughout this paper we consider a probability space $(\Omega, \mathscr{F}, \mathbb{P})$, with associated expectation operator $\mathbb{E}$, on which all the random variables are defined. The set of probability measures on a space $\mathscr{X}$ is denoted by $\mathscr{P}(\mathscr{X})$. The data take values in $\mathscr{Y}$, a subset of $\mathbb{R}^d$ for some $d \in \mathbb{N}$, and is endowed with the Borel σ-algebra. We observe $n \in \mathbb{N}$ data points, $y_{1:n} = y_1, \ldots, y_n$, that are distributed according to $\mu_\star^{(n)} \in \mathscr{P}(\mathscr{Y}^n)$. Let $\hat{\mu}_n = n^{-1} \sum_{i=1}^n \delta_{y_i}$, where δ_y is the Dirac distribution with mass on $y \in \mathscr{Y}$. We refer to $\hat{\mu}_n$ as the empirical distribution of $y_{1:n}$, even in settings where the observations are not i.i.d.

A model refers to a collection of distributions on $\mathscr{Y}^n$, denoted by $\mathscr{M}^{(n)} = \{\mu_\theta^{(n)} : \theta \in \mathscr{H}\} \subset \mathscr{P}(\mathscr{Y}^n)$, where $\mathscr{H} \subset \mathbb{R}^{d_\theta}$ is the parameter space, endowed with a distance $\rho_{\mathscr{H}}$ and of dimension $d_\theta \in \mathbb{N}$. However, we will often assume that the sequence of models $(\mathscr{M}^{(n)})_{n \geqslant 1}$ is such that, for every $\theta \in \mathscr{H}$, the sequence $(\hat{\mu}_{\theta,n})_{n \geqslant 1}$ of random probability measures on $\mathscr{Y}$ converges (in some sense) to a distribution $\mu_\theta \in \mathscr{P}(\mathscr{Y})$, where $\hat{\mu}_{\theta,n} = n^{-1} \sum_{i=1}^n \delta_{z_i}$ with $z_{1:n} \sim \mu_\theta^{(n)}$. Similarly, we will often assume that $\hat{\mu}_n$ converges to some distribution $\mu_\star \in \mathscr{P}(\mathscr{Y})$ as $n \to \infty$. Whenever the notation $\mu_\star$ and μ_θ is used, it is implicitly assumed that these objects exist. In such cases, we instead refer to $\mathscr{M} = \{\mu_\theta : \theta \in \mathscr{H}\} \subset \mathscr{P}(\mathscr{Y})$ as the model. We say that it is well-specified if there exists $\theta_\star \in \mathscr{H}$ such that $\mu_\star = \mu_{\theta_\star}$; otherwise it is

misspecified. Parameters are identifiable if $\theta = \theta'$ is implied by $\mu_\theta = \mu_{\theta'}$. The weak convergence of a sequence of measures μ_n to μ is denoted by $\mu_n \Rightarrow \mu$. The Kullback-Leibler (KL) divergence between μ and ν is defined as $\mathrm{KL}(\mu|\nu) = \int \log(d\mu/d\nu)d\mu$ if μ is absolutely continuous with respect to ν, and $+\infty$ otherwise.

1.2 *Minimum distance estimation*

Minimum distance estimation refers to the minimization, over the parameter $\theta \in \mathscr{H}$, of a distance between the empirical distribution $\hat{\mu}_n$ and the model distribution μ_θ (Wolfowitz, 1957; Basu et al., 2011). More formally, denoting by $\mathscr{D}$ a distance or divergence on $\mathscr{P}(\mathscr{Y})$, the associated minimum distance estimator (MDE) can be defined as

$$\hat{\theta}_n = \operatorname*{argmin}_{\theta \in \mathscr{H}} \mathscr{D}(\hat{\mu}_n, \mu_\theta). \tag{1.1}$$

In broad terms, the minimum distance estimation principle captures the idea of many statistical paradigms. For instance, the generalized method of moments (Hansen, 1982) consists in minimizing a discrepancy $\mathscr{D}$ defined as the weighted Euclidean distance between moments of $\hat{\mu}_n$ and μ_θ. In the empirical likelihood method (Owen, 2001), $\mathscr{D}$ is taken to be the KL divergence, and the model is supported strictly on the set of observed data and subject to moment conditions. The maximum likelihood estimator minimizes the KL divergence between $\mu_\star$ and μ_θ in the limit of the number of observations going to infinity.

However, it is worth noting that the definition in (1.1) precludes the naive application of some discrepancy measures. For instance, one could not directly choose $\mathscr{D}$ to be the KL divergence or the total variation distance, since for any model distribution μ_θ not supported solely on the observed data, they would evaluate to $+\infty$ and 1 respectively. To apply discrepancies of this kind, one would first need to build sample-based estimators of the underlying population quantity $\mathscr{D}(\mu_\star, \mu_\theta)$, assuming it is well-defined. Many such approaches have been studied in detail by Basu et al. (2011).

The computation of the minimum distance estimator might be intractable, especially in settings where it is assumed that one can simulate data from the model distribution but not evaluate its density. For such generative models, the following minimum expected distance estimator might be more computationally convenient:

$$\hat{\theta}_{n,m} = \operatorname*{argmin}_{\theta \in \mathscr{H}} \mathbb{E}_m \mathscr{D}(\hat{\mu}_n, \hat{\mu}_{\theta,m}), \tag{1.2}$$

where the expectation $\mathbb{E}_m$ is taken over the distribution of the sample $z_{1:m} \sim \mu_\theta^{(m)}$ giving rise to $\hat{\mu}_{\theta,m} = m^{-1}\sum_{i=1}^m \delta_{z_i}$. When n is fixed and m is large, or when $n = m$ and n is large, one might hope that the expectation is close to $\mathscr{D}(\hat{\mu}_n, \mu_\theta)$, and that the estimators $\hat{\theta}_n$ and $\hat{\theta}_{n,m}$ have similar properties. Inference techniques such as the method of simulated moments (McFadden, 1989) and indirect inference (Gouriéroux et al., 1993) often (implicitly) use estimators of this form, in which $\mathscr{D}$ defined as the weighted Euclidean distance between sample moments or summary statistics of $y_{1:n}$ and $z_{1:m}$, and the expectation in (1.2) is replaced with a Monte Carlo approximation.

1.3 *Minimum Wasserstein estimation*

In this paper, we focus on minimum distance estimation with the Wasserstein distance. Let ρ be a distance on the observation space $\mathscr{Y}$, and let $\mathscr{P}_p(\mathscr{Y})$ with $p \geqslant 1$ (e.g. $p = 1$ or 2) be the set of distributions

$\mu \in \mathscr{P}(\mathscr{Y})$ with finite p-th moment, i.e. there exists $y_0 \in \mathscr{Y}$ such that $\int_{\mathscr{Y}} \rho(y, y_0)^p d\mu(y) < \infty$. The p-Wasserstein distance, also called the Monge-Kantorovich, Mallows, or Gini distance, is a finite metric on $\mathscr{P}_p(\mathscr{Y})$, defined by the optimal transport problem

$$\mathscr{W}_p(\mu, \nu)^p = \inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathscr{Y}\times\mathscr{Y}} \rho(x,y)^p d\gamma(x,y), \tag{1.3}$$

where $\Gamma(\mu,\nu)$ is the set of probability measures on $\mathscr{Y} \times \mathscr{Y}$ with marginals μ and ν respectively; see Chapter 6 of Villani (2008) for a brief history of this distance and its central role in optimal transport.

A useful property of the Wasserstein distance is that it is well-defined for distributions with non-overlapping supports. This allows us to define the minimum Wasserstein estimator (MWE) of order p, denoted $\hat{\theta}_n$, by simply plugging $\mathscr{W}_p$ into (1.1) in place of $\mathscr{D}$. Some properties of the MWE have been studied in Bassetti et al. (2006), for well-specified models and i.i.d. data; we derive new results in Section 2.1 under weaker assumptions. We also propose the minimum expected Wasserstein estimator (MEWE), obtained by replacing $\mathscr{D}$ with $\mathscr{W}_p$ in (1.2) and denoted $\hat{\theta}_{n,m}$. We describe some of its theoretical properties in Section 2.2.

Variations of these estimators have recently been applied by for instance Arjovsky et al. (2017) and Genevay et al. (2018). In the settings they consider, the models are likely to be misspecified, and are supported on low-dimensional manifolds that might not overlap with the support of the data-generating mechanism. While the Wasserstein distance is well-defined in that case, the KL divergence or the total variation are not. This motivates the study of minimum Wasserstein estimators for these settings.

2. Theoretical results

We prove the existence, measurability, and consistency of the MWE and MEWE under weak assumptions, allowing the model to be misspecified and to produce data with certain types of dependencies. Under stronger assumptions, we study the rate of convergence and the asymptotic distribution of the MWE when $d = 1$ and $p = 1$. Throughout, we compare our results to those of Bassetti et al. (2006) and Bassetti and Regazzini (2006).

Informally, the consistency of the MWE and MEWE can be understood as follows. Under some conditions, we expect $\hat{\mu}_n$ to converge to $\mu_\star$, in the sense that $\mathscr{W}_p(\hat{\mu}_n, \mu_\star) \to 0$ as $n \to \infty$. Consequently, the minimum of $\theta \mapsto \mathscr{W}_p(\hat{\mu}_n, \mu_\theta)$ might converge to the minimum of $\theta \mapsto \mathscr{W}_p(\mu_\star, \mu_\theta)$, denoted by $\theta_\star$, assuming its existence and unicity. The same can be said for the minimum of $\theta \mapsto \mathbb{E}_m \mathscr{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m})$, provided $m \to \infty$ also. The parameter $\theta_\star$ is thus the limiting object of interest, also termed the estimand. Beyond its interpretation as the minimizer of $\theta \mapsto \mathscr{W}_p(\mu_\star, \mu_\theta)$, this parameter would coincide to the data-generating parameter if we assume that the data are generated from the model. In the misspecified case, note that $\theta_\star$ is not necessarily the parameter that minimizes $\mathrm{KL}(\mu_\star|\mu_\theta)$, which is the limit of the maximum likelihood estimator under standard regularity conditions.

2.1 *Minimum Wasserstein estimator*

2.1.1 *Existence, measurability, and consistency.* We first list assumptions on the data-generating process and on the model that are sufficient for the existence, measurability, and consistency for the MWE.

ASSUMPTION 2.1 The data-generating process is such that $\mathscr{W}_p(\hat{\mu}_n, \mu_\star) \to 0$, $\mathbb{P}$-almost surely as $n \to \infty$.

ASSUMPTION 2.2 The map $\theta \mapsto \mu_\theta$ is continuous in the sense that $\rho_{\mathscr{H}}(\theta_n, \theta) \to 0$ implies $\mu_{\theta_n} \Rightarrow \mu_\theta$ as $n \to \infty$.

ASSUMPTION 2.3 For some $\varepsilon > 0$, the set $B_\star(\varepsilon) = \{\theta \in \mathscr{H} : \mathscr{W}_p(\mu_\star, \mu_\theta) \leqslant \varepsilon_\star + \varepsilon\}$ is bounded, where $\varepsilon_\star = \inf_{\theta\in\mathscr{H}} \mathscr{W}_p(\mu_\star, \mu_\theta)$.

THEOREM 2.1 (Existence and consistency of the MWE) Under Assumptions 2.1-2.3, there exists a set $E \subset \Omega$ with $\mathbb{P}(E) = 1$ such that, for all $\omega \in E$, $\inf_{\theta\in\mathscr{H}} \mathscr{W}_p(\hat{\mu}_n(\omega), \mu_\theta) \to \inf_{\theta\in\mathscr{H}} \mathscr{W}_p(\mu_\star, \mu_\theta)$, and there exists $n(\omega)$ such that, for all $n \geqslant n(\omega)$, the sets $\operatorname{argmin}_{\theta\in\mathscr{H}} \mathscr{W}_p(\hat{\mu}_n(\omega), \mu_\theta)$ are non-empty and form a bounded sequence with

$$\limsup_{n\to\infty} \operatorname*{argmin}_{\theta\in\mathscr{H}} \mathscr{W}_p(\hat{\mu}_n(\omega), \mu_\theta) \subset \operatorname*{argmin}_{\theta\in\mathscr{H}} \mathscr{W}_p(\mu_\star, \mu_\theta).$$

For a generic function f, let $\varepsilon\text{-}\operatorname{argmin}_x f = \{x : f(x) \leqslant \varepsilon + \inf_x f\}$. Theorem 2.1 also holds if one replaces $\operatorname{argmin}_{\theta\in\mathscr{H}} \mathscr{W}_p(\hat{\mu}_n(\omega), \mu_\theta)$ with $\varepsilon_n\text{-}\operatorname{argmin}_{\theta\in\mathscr{H}} \mathscr{W}_p(\hat{\mu}_n(\omega), \mu_\theta)$, for any sequence ε_n converging to zero. If $\theta_\star = \operatorname{argmin}_{\theta\in\mathscr{H}} \mathscr{W}_p(\mu_\star, \mu_\theta)$ is unique, the result can be rephrased as $\hat{\theta}_n \to \theta_\star$ $\mathbb{P}$-almost surely.

The following theorem derives from a general result by Brown and Purves (1973) on the measurability of estimators defined as minimizers.

THEOREM 2.2 (Measurability of the MWE) Suppose that $\mathscr{H}$ is a σ-compact Borel measurable subset of $\mathbb{R}^{d_\theta}$. Under Assumption 2.2, for any $n \geqslant 1$ and $\varepsilon > 0$, there exists a Borel measurable function $\hat{\theta}_n : \Omega \to \mathscr{H}$ that satisfies

$$\hat{\theta}_n(\omega) \in \begin{cases} \operatorname{argmin}_{\theta\in\mathscr{H}} \mathscr{W}_p(\hat{\mu}_n(\omega), \mu_\theta) & \text{if this set is non-empty,} \\ \varepsilon\text{-}\operatorname{argmin}_{\theta\in\mathscr{H}} \mathscr{W}_p(\hat{\mu}_n(\omega), \mu_\theta) & \text{otherwise.} \end{cases}$$

Theorem 2.1 generalizes the results of Bassetti et al. (2006), where the model is assumed to be well-specified in the sense that $\mu_\star \in \mathscr{M}$. Moreover, Theorem 2.1 allows for data-generating processes which do not produce independent data points. For instance, if the data form a stationary and ergodic time series whose marginal distribution has finite p-th moments, then Assumption 2.1 still holds. These and other sufficient conditions for Assumption 2.1 to be satisfied are elaborated upon in the supplementary materials. Theorem 2.2 is only a minor generalization of the result in Bassetti et al. (2006), where it is assumed that for each $n \geqslant 1$, $\operatorname{argmin}_{\theta\in\mathscr{H}} \mathscr{W}_p(\hat{\mu}_n(\omega), \mu_\theta)$ is non-empty for almost every $\omega \in \Omega$. In the next section, this small modification also enables the direct application of results by Pollard (1980).

2.1.2 *Rate of convergence and asymptotic distribution* . Under conditions guaranteeing the consistency of the minimum Wasserstein estimator, we study its rate of convergence and asymptotic distribution in the case where $p = 1$, $\mathscr{Y} = \mathbb{R}$, $\rho(x,y) = |x-y|$. Under this setup, it can be shown that $\mathscr{W}_1(\mu,\nu) = \int_0^1 |F_\mu^{-1}(s) - F_\nu^{-1}(s)|ds = \int_{\mathbb{R}} |F_\mu(t) - F_\nu(t)|dt$, where F_μ and F_ν denote the cumulative distribution functions (CDFs) of μ and ν respectively (see e.g. Ambrosio et al., 2005, Theorem 6.0.2). Additionally, assume that $\mathscr{H}$ is endowed with a norm: $\rho_{\mathscr{H}}(\theta,\theta') = \|\theta - \theta'\|_{\mathscr{H}}$. We also require the assumption that $\theta_\star$ is "well-separated":

ASSUMPTION 2.4 For all $\varepsilon > 0$, there exists $\delta > 0$ such that

$$\inf_{\theta\in\mathscr{H}:\|\theta-\theta_\star\|_{\mathscr{H}}\geqslant\varepsilon} \mathscr{W}_1(\mu_{\theta_\star}, \mu_\theta) > \delta.$$

This assumption is commonly made in the asymptotic study of M-estimators (e.g. Chapter 5 of Van der Vaart, 2000); see also the supplementary materials. We focus on the setting in which the model is well-specified, but also discuss some extensions to the misspecified setting in Section 2.1.3.

Our approach to derive asymptotic distributions follows Pollard (1980). Let F_θ, $F_\star$ and F_n denote the CDFs of μ_θ, $\mu_\star$ and $\hat{\mu}_n$ respectively. Informally speaking, we show that $\sqrt{n}W_1(\hat{\mu}_n, \mu_\theta)$ can be approximated by $\int_{\mathbb{R}} |\sqrt{n}(F_n(t) - F_\star(t)) - \langle \sqrt{n}(\theta - \theta_\star), D_{\theta_\star}(t)\rangle| dt$ near $\theta_\star$, for some $D_{\theta_\star} \in (L_1(\mathbb{R}))^{d_\theta}$, with $\langle \theta, u \rangle = \sum_{i=1}^{d_\theta} \theta_i u_i$. Results by del Barrio et al. (1999) and Dede (2009) give conditions under which $\sqrt{n}(F_n - F_\star)$ converges to a zero mean Gaussian process $G_\star$ with known covariance structure, for both independent and certain classes of dependent data. Heuristically, the distribution of $\sqrt{n}(\hat{\theta}_n - \theta_\star)$ is then close to that of $\operatorname{argmin}_{u \in \mathscr{H}} \int_{\mathbb{R}} |G_\star(t) - \langle u, D_{\theta_\star}(t)\rangle| dt$. The required form of $D_{\theta_\star}$ is given in the following assumption:

ASSUMPTION 2.5 There exists a non-singular $D_{\theta_\star} \in (L_1(\mathbb{R}))^{d_\theta}$ such that

$$\int_{\mathbb{R}} |F_\theta(t) - F_{\theta_\star}(t) - \langle \theta - \theta_\star, D_{\theta_\star}(t)\rangle| dt = o(\|\theta - \theta_\star\|_{\mathscr{H}}),$$

as $\|\theta - \theta_\star\|_{\mathscr{H}} \to 0$.

To provide some intuition into the nature of the "derivative" $D_{\theta_\star}$, we consider the following simple example. Let $\mu_\theta = \mathscr{N}(\theta, 1)$ for $\theta \in \mathbb{R}$, and $\mu_\star = \mu_{\theta_\star}$ for some $\theta_\star$. By Taylor expanding $F_\theta(t) = \Phi(t - \theta)$ around $\theta_\star$ (for fixed t), Assumption 2.5 can be shown to hold with $D_{\theta_\star}(t) = -\varphi(t - \theta_\star)$, where Φ and φ denote the CDF and density of a standard Gaussian variable, respectively. Next, we state a result that holds for a well-specified model producing i.i.d. data, and analogous results for misspecified models and certain types of dependent processes can be found in the supplementary materials.

THEOREM 2.3 Suppose $Y_i \sim \mu_\star = \mu_{\theta_\star}$ i.i.d., with $\theta_\star$ in the interior of $\mathscr{H}$, and that $\int_0^\infty \sqrt{\mathbb{P}(|Y_0| > t)} dt < \infty$. Suppose that Assumptions 2.1-2.5 hold and that $\operatorname{argmin}_{u \in \mathscr{H}} \int_{\mathbb{R}} |G_\star(t) - \langle u, D_{\theta_\star}(t)\rangle| dt$ is almost surely unique. Then, the MWE with $p = 1$ satisfies

$$\sqrt{n}(\hat{\theta}_n - \theta_\star) \Rightarrow \operatorname*{argmin}_{u \in \mathscr{H}} \int_{\mathbb{R}} |G_\star(t) - \langle u, D_{\theta_\star}(t)\rangle| dt,$$

as $n \to \infty$, where $G_\star$ is a zero mean Gaussian process with $\mathbb{E} G_\star(s) G_\star(t) = \min\{F_\star(s), F_\star(t)\} - F_\star(s) F_\star(t)$.

A similar statement for $p = 2$ can potentially be derived by considering the results of del Barrio et al. (2005). The condition $\int_0^\infty \sqrt{\mathbb{P}(|Y_0| > t)} dt < \infty$ implies the existence of second moments, and is itself implied by the existence of moments of order $2 + \varepsilon$ for some $\varepsilon > 0$ (see e.g. Section 2.9 in Wellner and van der Vaart, 1996). The uniqueness assumption on the argmin in the limit can be relaxed by considering convergence to the entire set of minimizing values, as in Section 7 of Pollard (1980). Still, uniqueness can sometimes be established, using e.g. the results of Cheney and Wulbert (1969). This approach is taken by Bassetti and Regazzini (2006), who directly show that Theorem 2.3 holds when $\mathscr{M}$ is a location-scale family supported on a bounded open interval. The existence and form of $D_{\theta_\star}$ can in many cases be derived if the model is differentiable in quadratic mean (Le Cam, 1970), which is elaborated upon in the supplementary materials. There, one can also find results to verify Assumptions 2.1 and 2.4. It can in some cases potentially be easier to verify the assumptions for a reparameterization of θ, say $\varphi = r(\theta)$. Provided that the theorem holds for $\hat{\varphi}_n$ and that the inverse map r^{-1} is differentiable, the limiting distribution of $\hat{\theta}_n$ can be derived using a delta method argument.

Computing confidence intervals using the asymptotic distribution provided by Theorem 2.3 is hard, due in part to its dependence on unknown quantities. However, the existence of the limiting distribution is in itself sufficient to guarantee the asymptotic validity of appropriately constructed subsampling confidence intervals (Politis et al., 1999, Theorem 2.2.1). This also generalizes to settings with certain

kinds of dependent data. Under slightly stronger assumptions, the closely related m out of n bootstrap produces asymptotically valid confidence intervals as well (see Bickel and Sakov, 2008, and references therein). In the numerical experiments of Section 4, we find that the standard bootstrap (Efron and Tibshirani, 1994) works well in practice.

Theorem 2.3 also holds for approximations of the MWE, say $\tilde{\theta}_n$, provided that $\tilde{\theta}_n = \hat{\theta}_n + o_{\mathbb{P}}(1/\sqrt{n})$, as can be seen from its proof. In light of the convergence of the MEWE to the MWE as $m \to \infty$ established in Section 2.2, there exists a sequence $m(n)$ (depending on ω) such that the associated MEWE $\hat{\theta}_{n,m(n)}$ satisfies the conclusion of Theorem 2.3.

2.1.3 *Extensions.* Under slightly stronger assumptions, Theorem 2.3 can be extended to the misspecified setting. In particular, suppose that there exists a neighborhood N of $\theta_\star$ and a constant $c > 0$ such that for any $\theta \in N$, $\mathscr{W}_1(\mu_\theta, \mu_\star) \geqslant \mathscr{W}_1(\mu_{\theta_\star}, \mu_\star) + c\|\theta - \theta_\star\|_{\mathscr{H}}$. In the well-specified case, this property is implied by Assumption 2.5. Then, as elaborated upon in the supplementary materials, the minimum of $\theta \mapsto \mathscr{W}_1(\hat{\mu}_n, \mu_\theta)$ is attained on the set $\mathscr{S}_n = \{\theta : \|\theta - \theta_\star\|_{\mathscr{H}} \leqslant 4\mathscr{W}_1(\hat{\mu}_n, \mu_\star)/c\}$ with probability going to one. Since the conditions of Theorem 2.3 imply that $\mathscr{W}_1(\hat{\mu}_n, \mu_\star) = O_{\mathbb{P}}(1/\sqrt{n})$, this immediately implies that $\|\hat{\theta}_n - \theta_\star\|_{\mathscr{H}} = O_{\mathbb{P}}(1/\sqrt{n})$ also. In other words, the minimum Wasserstein estimator retains its rate of convergence in the misspecified case.

To find its asymptotic distribution, one can observe that with probability going to one, the map $\theta \mapsto \sqrt{n}\mathscr{W}_1(\hat{\mu}_n, \mu_\theta)$ can be approximated uniformly well over $\mathscr{S}_n$ by the map $\theta \mapsto \sqrt{n}\int_{\mathbb{R}} |F_n(t) - F_{\theta_\star}(t) - \langle \theta - \theta_\star, D_{\theta_\star}(t)\rangle| dt$, which similarly achieves its minimum on $\mathscr{S}_n$. Therefore, as n gets large, $\sqrt{n}(\hat{\theta}_n - \theta_\star)$ behaves like a minimum of $u \mapsto \int_{\mathbb{R}} |\sqrt{n}(F_n(t) - F_\star(t)) + \sqrt{n}(F_\star(t) - F_{\theta_\star}(t)) - \langle u, D_{\theta_\star}(t)\rangle| dt$. Under the conditions of Theorem 2.3, $\sqrt{n}(F_n - F_\star)$ converges to $G_\star$ in the sense of del Barrio et al. (1999). In turn, $\sqrt{n}(\hat{\theta}_n - \theta_\star)$ should be distributed as the minimizer(s) of $u \mapsto \int_{\mathbb{R}} |G_\star(t) + \sqrt{n}(F_\star(t) - F_{\theta_\star}(t)) - \langle u, D_{\theta_\star}(t)\rangle| dt$ as n grows. A technical complication arises since this function converges pointwise to infinity, and we therefore leave formal statements to the supplementary materials.

Extensions to cases with multivariate data are left for future research. It is unclear whether convergence to $\theta_\star$ will occur at the same $\sqrt{n}$ rate in higher dimensions. This is because $\mathbb{E}\mathscr{W}_p(\hat{\mu}_n, \mu_\star)$ is on the order of $n^{-1/d}$ whenever $\mu_\star$ is absolutely continuous with respect to the Lebesgue measure and $d > 2p$ (see e.g. Weed and Bach, 2019, and references therein). On the other hand, del Barrio and Loubes (2017) show, under some assumptions, that the 2-Wasserstein distance satisfies the following CLT:

$$\sqrt{n}\left(\mathscr{W}_2^2(\hat{\mu}_n, \mu_\theta) - \mathbb{E}\mathscr{W}_2^2(\hat{\mu}_n, \mu_\theta)\right) \Rightarrow \mathscr{N}\left(0, \sigma^2(\mu_\star, \mu_\theta)\right),$$

where $\sigma^2(\mu_\star, \mu_\theta)$ has a known form and the expectation is taken with respect to the observations $y_{1:n} \sim \mu_\star^{(n)}$. Similar results are expected to hold for other p also. It therefore seems likely that the distance(s) between the MWE and the minimizer(s) of $\theta \mapsto \mathbb{E}\mathscr{W}_2^2(\hat{\mu}_n, \mu_\theta)$ converges to zero at the standard $\sqrt{n}$ rate. If these speculations hold true, one could interpret them in terms of a bias-variance trade-off: the bias would appear to be on the order of $n^{-1/d}$, whereas the variance is on the order of $n^{-1/2}$. However, note that the function $\theta \mapsto \mathbb{E}\mathscr{W}_2^2(\hat{\mu}_n, \mu_\theta)$ depends only on population properties of $\mu_\star^{(n)}$. As such, it is a reasonable alternative to the objective function $\theta \mapsto \mathscr{W}_2^2(\mu_\star, \mu_\theta)$, and might still yield reasonable identification of the parameters. For instance, if the model is well-specified and Gaussian with θ being a location parameter, it seems likely that $\theta \mapsto \mathbb{E}\mathscr{W}_2^2(\hat{\mu}_n, \mu_\theta)$ is minimized at $\theta_\star$ for any n. It is therefore unclear whether the slow convergence rate of the bias would always be of practical concern.

2.2 *Minimum expected Wasserstein estimator*

2.2.1 *Existence, measurability, and consistency.* In order to show similar results for the MEWE as for the MWE, we introduce the following additional assumptions.

ASSUMPTION 2.6 For any $m \geqslant 1$, if $\rho_{\mathscr{H}}(\theta_n, \theta) \to 0$, then $\mu_{\theta_n}^{(m)} \Rightarrow \mu_{\theta}^{(m)}$ as $n \to \infty$.

ASSUMPTION 2.7 If $\rho_{\mathscr{H}}(\theta_n, \theta) \to 0$, then $\mathbb{E}_n \mathscr{W}_p(\mu_{\theta_n}, \hat{\mu}_{\theta_n,n}) \to 0$ as $n \to \infty$.

Assumption 2.6 is slightly stronger than Assumption 2.2, stating that we not only need weak convergence of the "model" distributions μ_θ, but also of the sample distributions $\mu_\theta^{(m)}$ for any $m \geqslant 1$. Assumption 2.7 is implied by $\sup_{\theta \in \mathscr{H}} \mathbb{E}_n \mathscr{W}_p(\mu_\theta, \hat{\mu}_{\theta,n}) \to 0$, which in turn might hold when $\mathscr{H}$ is compact and the inequalities in Fournier and Guillin (2015) hold.

In the next result, we prove an analogous version of Theorem 2.1 for the MEWE as $\min\{n, m\} \to \infty$. For simplicity, we write m as a function of n and require that $m(n) \to \infty$ as $n \to \infty$.

THEOREM 2.4 (Existence and consistency of the MEWE) Under Assumptions 2.1-2.3 and 2.6-2.7, there exists a set $E \subset \Omega$ with $\mathbb{P}(E) = 1$ such that, for all $\omega \in E$, $\inf_{\theta \in \mathscr{H}} \mathbb{E}_{m(n)} \mathscr{W}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta,m(n)}) \to \inf_{\theta \in \mathscr{H}} \mathscr{W}_p(\mu_\star, \mu_\theta)$, and there exists $n(\omega)$ such that, for all $n \geqslant n(\omega)$, the sets $\operatorname{argmin}_{\theta \in \mathscr{H}} \mathscr{W}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta,m(n)})$ are non-empty and form a bounded sequence with

$$\limsup_{n \to \infty} \operatorname*{argmin}_{\theta \in \mathscr{H}} \mathbb{E}_{m(n)} \mathscr{W}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta,m(n)}) \subset \operatorname*{argmin}_{\theta \in \mathscr{H}} \mathscr{W}_p(\mu_\star, \mu_\theta).$$

THEOREM 2.5 (Measurability of the MEWE) Suppose that $\mathscr{H}$ is a σ-compact Borel measurable subset of $\mathbb{R}^{d_\theta}$. Under Assumption 2.6, for any $n \geqslant 1$ and $m \geqslant 1$ and $\varepsilon > 0$, there exists a Borel measurable function $\hat{\theta}_{n,m} : \Omega \to \mathscr{H}$ that satisfies

$$\hat{\theta}_{n,m}(\omega) \in \begin{cases} \operatorname{argmin}_{\theta \in \mathscr{H}} \mathbb{E}_m \mathscr{W}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta,m}), & \text{if this set is non-empty,} \\ \varepsilon\text{-}\operatorname{argmin}_{\theta \in \mathscr{H}} \mathbb{E}_m \mathscr{W}_p(\hat{\mu}_n(\omega), \hat{\mu}_{\theta,m}), & \text{otherwise.} \end{cases}$$

The results above appear to be the first of their kind for the MEWE.

2.2.2 *Convergence to the MWE.* The next result considers the case where the data are fixed, while $m \to \infty$. It shows that the MEWE converges to the MWE, assuming the latter exists. Using the results of del Barrio and Loubes (2017) and references therein, one could potentially derive the rate of this convergence, which we leave for future work. We formulate the following additional assumption, in which the observed empirical distribution is kept fixed and $\varepsilon_n = \inf_{\theta \in \mathscr{H}} \mathscr{W}_p(\hat{\mu}_n, \mu_\theta)$.

ASSUMPTION 2.8 For some $\varepsilon > 0$, the set $B_n(\varepsilon) = \{\theta \in \mathscr{H} : \mathscr{W}_p(\hat{\mu}_n, \mu_\theta) \leqslant \varepsilon_n + \varepsilon\}$ is bounded.

THEOREM 2.6 (MEWE converges to MWE as $m \to \infty$) Under Assumptions 2.2 and 2.6-2.8, then $\inf_{\theta \in \mathscr{H}} \mathbb{E}_m \mathscr{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m}) \to \inf_{\theta \in \mathscr{H}} \mathscr{W}_p(\hat{\mu}_n, \mu_\theta)$, and there exists an $\hat{m}$ such that, for all $m \geqslant \hat{m}$, the sets $\operatorname{argmin}_{\theta \in \mathscr{H}} \mathbb{E}_m \mathscr{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m})$ are non-empty and form a bounded sequence with

$$\limsup_{m \to \infty} \operatorname*{argmin}_{\theta \in \mathscr{H}} \mathbb{E}_m \mathscr{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m}) \subset \operatorname*{argmin}_{\theta \in \mathscr{H}} \mathscr{W}_p(\hat{\mu}_n, \mu_\theta).$$

3. Computational aspects

3.1 *Computing the Wasserstein distance*

We recall some strategies to calculate or approximate the Wasserstein distance between empirical distributions. In the case where $\mathscr{Y} \subset \mathbb{R}$, the exact computation is cheap, as the main computational task reduces to sorting the samples. However, in dimensions $d > 1$, the cost is in general expensive, which has motivated a rich literature on fast approximations (Peyré and Cuturi, 2018). We will write $\mathscr{W}_p(y_{1:n}, z_{1:m})$ for $\mathscr{W}_p(\hat{\mu}_n, \hat{\nu}_m)$, where $\hat{\mu}_n$ and $\hat{\nu}_m$ stand for the empirical distributions $n^{-1}\sum_{i=1}^n \delta_{y_i}$ and $m^{-1}\sum_{i=1}^m \delta_{z_i}$. The Wasserstein distance then takes the form

$$\mathscr{W}_p(y_{1:n}, z_{1:m})^p = \inf_{\gamma \in \Gamma_{n,m}} \sum_{i=1}^n \sum_{j=1}^m \rho(y_i, z_j)^p \gamma_{ij} \tag{3.1}$$

where $\Gamma_{n,m}$ is the set of $n \times m$ matrices with non-negative entries, columns and rows resp. summing to m^{-1} and n^{-1}.

3.1.1 *Exact computation.* The formulation in (3.1) is a linear program, and can be solved with generic linear program solvers. However, specialized approaches can be more efficient. In the univariate case with $\rho(x,y) = |x-y|$, the optimal transport coupling can be found by sorting the vectors $y_{1:n}$ and $z_{1:m}$ to get the collections of order statistics $\{y_{(i)}\}_{i=1}^n$ and $\{z_{(j)}\}_{j=1}^m$. Suppose that $m = \ell n$ for some $\ell \geqslant 1$. Then, the p-Wasserstein distance in (3.1) can be expressed as

$$\mathscr{W}_p^p(y_{1:n}, z_{1:m}) = \frac{1}{m}\sum_{i=1}^n \sum_{j=1}^{\ell} |y_{(i)} - z_{(\ell(i-1)+j)}|^p, \tag{3.2}$$

which can be seen from the representation $\mathscr{W}_p^p(\hat{\mu}_n, \hat{\nu}_m) = \int_0^1 |F_{\mu,n}^{-1}(s) - F_{\nu,m}^{-1}(s)|^p ds$ (see e.g. Ambrosio et al., 2005, Theorem 6.0.2). The cost of the Wasserstein distance computation is thus of order $m \log m$ in the univariate setting. Note that, in some cases, the generation of m sorted observations can be done directly for a cost of order m, for instance by generating already-sorted uniforms and applying a quantile function (Devroye, 1985). It should also be noted that the expression $\mathscr{W}_p^p(\mu,\nu) = \int_0^1 |F_\mu^{-1}(s) - F_\nu^{-1}(s)|^p ds$, in combination with a numerical integrator, could be used whenever the quantile functions of μ and ν are known (as in the g-and-k example of Section 4.1). In that case one can directly target the MWE with a numerical optimizer, as an alternative to computing the MEWE. The same is true if the CDFs are available, using the expression $\mathscr{W}_1(\mu,\nu) = \int_{\mathbb{R}} |F_\mu(t) - F_\nu(t)| dt$ given in Section 2.1.2.

In multivariate settings, one can solve the problem in (3.1) using dual ascent methods (see e.g. Bertsimas and Tsitsiklis, 1997). This includes the Hungarian algorithm, applicable in the setting where $m = n$, at a cost of order n^3. Other algorithms have a cost of order $n^{2.5}\log(nC_n)$, with $C_n = \max_{1 \leqslant i,j \leqslant n} \rho(y_i, z_j)$, and can therefore be more efficient when C_n is small (Burkard et al., 2009, Section 4.1.3). A practical alternative is the short-list method, derived from the network simplex algorithm, presented by Gottschlich and Schuhmacher (2014) and implemented in the `transport` R package (Schuhmacher et al., 2017). In general, simplex algorithms come without guarantees of polynomial running times, but Gottschlich and Schuhmacher (2014) show empirically that their method tends to have sub-cubic cost. When the cost of computing the Wasserstein distance exactly gets prohibitively large, we can resort to various approximations.

3.1.2 *Approximations.* In parallel with its increasing popularity as an inferential tool in statistics and machine learning, there has been fast growth in the number of algorithms that approximate the

Wasserstein distance at reduced computational costs. The book of Peyré and Cuturi (2018) provides an overview of many such methods. In particular, they provide a thorough discussion of the method introduced by Cuturi (2013), which regularizes the optimization problem in (3.1) using an entropic constraint. Specifically, the regularized version of (3.1) reads: $\gamma^{\zeta} = \text{argmin}_{\gamma \in \Gamma_{n,m}} \sum_{i=1}^{n} \sum_{j=1}^{m} \rho(y_i, z_j)^p \gamma_{ij} + \zeta \sum_{i=1}^{n} \sum_{j=1}^{m} \gamma_{ij} \log \gamma_{ij}$, which includes a penalty on the entropy of γ. The regularized problem can be solved iteratively by Sinkhorn's algorithm (Cuturi, 2013) or iterative Bregman projections (Benamou et al., 2015) for a total cost of order nm. Define the dual-Sinkhorn divergence $S_p^{\zeta}(y_{1:n}, z_{1:m})^p = \sum_{i=1}^{n} \sum_{j=1}^{m} \rho(y_i, z_j)^p \gamma_{ij}^{\zeta}$. If ζ goes to zero, the dual-Sinkhorn divergence goes to the Wasserstein distance. If ζ goes to infinity, it converges to the energy distance (Ramdas et al., 2017). Other fast approximations of the Wasserstein distance include Altschuler et al. (2017, 2018); Ye et al. (2017); Li et al. (2018).

In the case where $n = m$, computing the Wasserstein distance can be viewed as an assignment problem, which leads to other specialized approaches. For instance, Puccetti (2017) proposes a greedy algorithm based on swaps in the assignment, for a cost of n^2 per iteration. When a cost of order nm or n^2 is too large, Bernton et al. (2019) propose a new distance generalizing the idea of sorting when $d > 1$. It consists in sorting samples according to their projection via the Hilbert space-filling curve and computing a distance analogous to the one in (3.2), for a computational cost of the order of $m \log m$. A similar idea underlies the sliced Wasserstein distance (Rabin et al., 2011; Bonneel et al., 2015), which can be estimated by projecting the data onto L random lines, and by averaging the Wasserstein distances computed in the associated one-dimensional spaces, for a total cost on the order of $Lm \log m$.

3.2 *Computing the estimators*

The exact computation of the MWE and MEWE is in general intractable. This is also true when $\mathscr{W}_p$ is substituted for any of its approximations mentioned above. However, we can envision various schemes to numerically approximate the estimators.

The calculation of the MEWE can be based on the Monte Carlo approximation of $\mathbb{E}_m \mathscr{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m})$ using synthetic samples generated given θ. Assume that a data set $z_{1:m}$ can be sampled from $\mu_\theta^{(m)}$ by setting $z_{1:m} = g_m(u, \theta)$, where g_m is a deterministic function of the parameter θ and u a random variable independent of θ. Then, the empirical mean $k^{-1} \sum_{i=1}^{k} \mathscr{W}_p(y_{1:n}, g_m(u^{(i)}, \theta))$, where the $u^{(i)}$ are i.i.d., is a natural estimate of $\mathbb{E}_m \mathscr{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m})$. In the limit $k \to \infty$, $k^{-1} \sum_{i=1}^{k} \mathscr{W}_p(y_{1:n}, g_m(u^{(i)}, \theta)) \to \mathbb{E}_m \mathscr{W}_p(\hat{\mu}_n, \hat{\mu}_{\theta,m})$ almost surely. Since this estimator is an average of i.i.d. random variables, the CLT indicates that the rate of convergence is $\sqrt{k}$. Moreover, this approximation is a deterministic function of θ, which can be optimized with standard methods. In turn, this optimization step can be placed within a Monte Carlo Expectation-Maximization (MCEM) algorithm (Wei and Tanner, 1990), which would alternate between optimization of θ and resampling of $u^{(i)}$. Convergence results for such algorithms, as both the number of iterations and the value of k go to infinity, are reviewed in Neath et al. (2013).

In practice, we are naturally constrained to finite values of m and k. The incremental cost of increasing k is typically lower than that of increasing m, due in part to the potential for parallelization when calculating the distances $\mathscr{W}_p(y_{1:n}, g_m(\theta, u^{(i)}))$ for a given θ, and in part to the algorithmic complexity in m, which is super-linear as described in the previous section. In the numerical experiments of Section 4, we found that $m = 10^4$ and $k = 20$ within a single iteration of MCEM yielded accurate estimators. That is, we draw $u^{(i)}$ for $i = 1, \ldots, k$ once and for all, and optimize over θ. We illustrate the effect of choosing different m and k in Section 4.3.

Several alternatives to the MCEM approach exist. An approach to computing the MEWE was proposed in Genevay et al. (2018) based on the Sinkhorn divergence approximation to the Wasserstein

distance. They derive gradients of $S_p^{\zeta}(y_{1:n}, g_m(u,\theta))$ with respect to θ while u is fixed, allowing for the application of stochastic gradient descent. In practice, the gradients can be computed with auto-differentiation. A method for computing the MWE was proposed by Chen and Li (2018), in which they pull back the 2-Wasserstein metric tensor in $\mathscr{P}_2(\mathscr{Y})$ to $\mathscr{H}$, under which $\mathscr{H}$ becomes a Riemannian manifold. In turn, this structure allows them to derive a novel gradient descent algorithm. Alternatively, in the spirit of Monte Carlo optimization, one can modify the sampling algorithms used for the approximate Bayesian computation (ABC) approach described by Bernton et al. (2019) to approximate the MEWE. This has the benefit of not requiring the synthetic data to be generated via a deterministic function g_m with fixed-dimensional arguments. Related discussions can be found in Wood (2010); Rubio et al. (2013).

4. Illustrations

In Sections 4.1 and 4.2, we compute the MEWE in two well-specified models with intractable likelihoods that produce i.i.d. data, taken from the ABC literature. We empirically estimate the coverage of bootstrap confidence intervals for the data-generating parameter. In Section 4.1, we also compute the MEWE in a setting where the data-generating process produces a time series. In Section 4.3, we compare the distribution of the MEWE with that of the maximum likelihood estimator (MLE) in a simple misspecified setting. We also investigate the effect of k and m on the distribution of the approximate MEWE. In Section 4.4, we highlight the robustness of this choice by considering a heavy-tailed data-generating process for which the MLE is not consistent. Throughout the numerical experiments, we have chosen $p = 1$, as this imposes minimal assumptions on the existence of moments of both the data-generating process and the model.

4.1 *Quantile "g-and-κ" distribution*

4.1.1 *Independent data* . The g-and-κ distribution (Tukey, 1977; Jorge and Boris, 1984) is defined in terms of its quantile function:

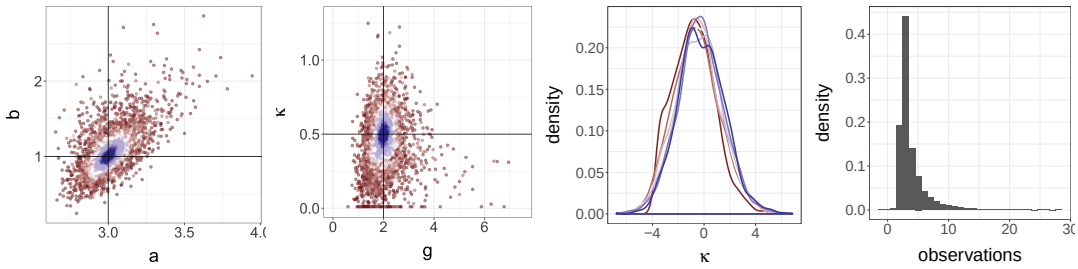
$$r \in (0,1) \mapsto a + b\left(1 + 0.8\frac{1-\exp(-gz(r))}{1+\exp(-gz(r))}\right)\left(1+z(r)^2\right)^{\kappa} z(r), \tag{4.1}$$

where $z(r)$ refers to the r-th quantile of the standard Normal distribution. The model is indexed by the parameter $\theta = (a,b,g,\kappa) \in [0,10]^4$, and we take $\mu_\star = \mu_{\theta_\star}$ with $\theta_\star = (3,1,2,0.5)$. The probability density function, and therefore the likelihood of the model, is analytically intractable; thus the model has become a standard benchmark for ABC methods (Sisson et al., 2018). Though, the likelihood can be estimated by numerically inverting and then differentiating the quantile function, as described in Rayner and MacGillivray (2002); Bernton et al. (2019).

Sampling i.i.d. variables from the g-and-κ distribution can be achieved straightforwardly by plugging independent standard Normals into (4.1) in place of $z(r)$. Therefore, the MEWE with large m can be computed to high precision. In Figure 1, we show the behavior of the MEWE with $p = 1$ and $m = 10^4$ for different numbers of observed data, and illustrate its concentration around the data-generating parameter $\theta_\star$. In computing the MEWE, we used $k = 20$ and only one iteration of MCEM. That is, we approximate the MEWE by sampling $k = 20$ independent $u^{(i)}$ random variables and minimize $\theta \mapsto k^{-1}\sum_{i=1}^{k} \mathscr{W}_p(y_{1:n}, g_m(u^{(i)},\theta))$ to form the estimator, using the `optim` function in `R` (R Core Team, 2015).

We check the coverage of bootstrap confidence intervals calculated for $\theta_\star = (3,1,2,0.5)$. We use the percentile bootstrap (Efron and Tibshirani, 1994) for data sets of size $n = 1,000$ and synthetic data sets of size $m = 10^4$, and calculate the MEWE with $k = 20$. We draw 400 data sets from the data-generating process, and 1,000 bootstrap data sets for each of these. The observed coverage rates of the resulting 0.95 confidence intervals were 0.928 for a, 0.945 for b, 0.960 for g, and 0.938 for κ. The coverage rates should approach 0.95 as $n \to \infty$, $m \to \infty$, and $k \to \infty$ within the MCEM algorithm. After a Bonferroni correction, the observed coverage of the confidence sets for $\theta_\star$ was 0.935.

As mentioned in Section 3.1.1, since the g-and-κ distribution has an explicit quantile function (insofar as the Normal quantile function can be considered explicit), one could instead directly estimate the Wasserstein distance between the g-and-k distribution and some empirical distribution using a representation of the distance in terms of an integral of the difference of quantile functions, combined with a numerical integrator.



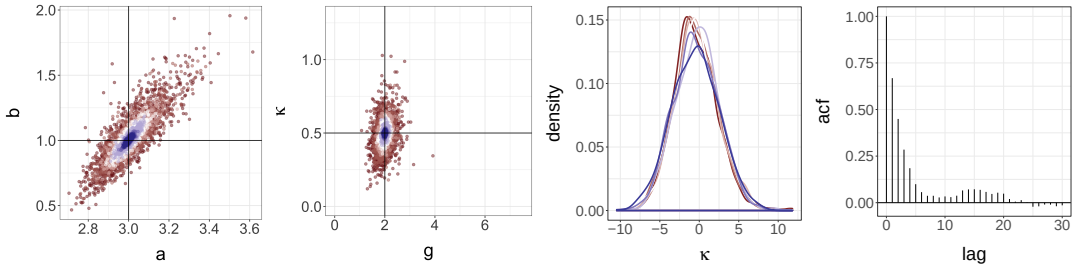
(a) MEWE: a vs b. (b) MEWE: g vs κ. (c) $\sqrt{n}$-scaled estim. of κ. (d) Histogram of data.

FIG. 1: Estimators in the well-specified g-and-κ model, as described in Section 4.1.1. Figures 1a and 1b show the MEWE's bivariate marginal sampling distributions for (a,b) and (g,κ) respectively, as n ranges from 50 to 10^4 (colors from red to white to blue as n increases). For each n, we plot $M = 1,000$ estimators based on independent data sets. Each estimator was computed with $p = 1$, $m = 10^4$, $k = 20$, and one iteration of MCEM. Note that for small data sizes ($n = 50$ and $n = 100$), the estimator occasionally appears to be on the boundary of the parameter space, which could mean that the optimization procedure failed to converge. The intersections of the black lines indicate data-generating parameters. Figure 1c shows the MEWE's marginal distribution for κ for the different levels of n, centered and rescaled by $\sqrt{n}$, illustrating the rate of convergence anticipated by Theorem 2.3. Figure 1d is a histogram of a data set generated with $\theta_\star = (3,1,2,0.5)$ and $n = 1,000$.

4.1.2 *Dependent data* . To illustrate the behavior of the estimator when the data-generating process produces dependent data, we also generated g-and-κ variables using Normals from an AR(1) process. Specifically, we let $x_0 \sim \mathcal{N}(0,1)$ and $x_t = \rho x_{t-1} + \eta_t$ for $t \geqslant 1$, where $\eta_t \sim \mathcal{N}(0, 1-\rho^2)$ independently, and $\rho = 0.75$. Hence, these variables are marginally distributed as $\mathcal{N}(0,1)$, but are positively correlated. To produce the observation y_t for each t, we plugged x_t into (4.1) in place of $z(r)$, using the same $\theta_\star$ as in the independent setting. The marginal distribution of the data are therefore the same as before, but the sequence of observations now forms a stationary and ergodic time series. This setting is covered by the theoretical results of Section 2; Assumption 2.1 holds with $\mu_\star = \mu_{\theta_\star}$. The model, as before, is taken to generate i.i.d. data.

To approximate the MEWE, we used the same computational approach as in the i.i.d. setting, with $p = 1$, $m = 10^4$, and $k = 20$. In Figure 2, we show that the MEWE appears to concentrate around $\theta_\star$ at

the same rate as in the i.i.d. setting, but that its asymptotic distribution has higher variance. Note that in Figure 2, the data sizes are 10 times larger than in the plots for the i.i.d. setting (Figure 1), as the correlation between the samples effectively reduces the sample size and makes the estimators poorly behaved when n is small.



(a) MEWE: a vs b. (b) MEWE: g vs κ. (c) $\sqrt{n}$-scaled estim. of κ. (d) ACF of data.

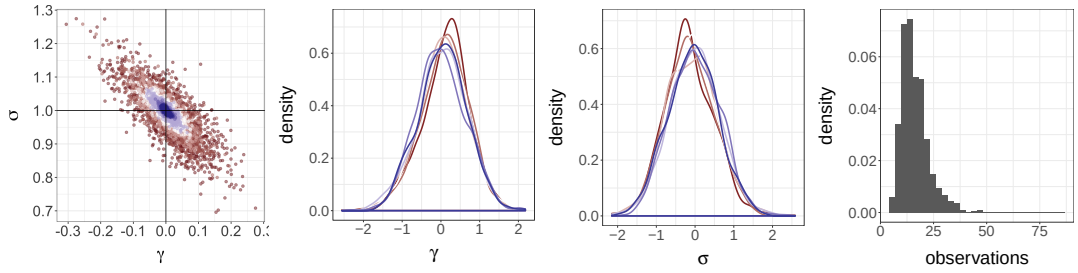
FIG. 2: Estimators in the g-and-κ model with dependent data, as described in Section 4.1.2. Figures 2a and 2b show the MEWE's bivariate marginal sampling distributions for (a,b) and (g,κ) respectively, as n ranges from 500 to 10^5 (colors from red to white to blue as n increases). Note that the sample sizes here are 10 times larger than in the plots for the i.i.d. setting. For each n, we plot $M = 1,000$ estimators based on independent data sets. Each estimator was computed with $p = 1$, $m = 10^4$, $k = 20$, and one iteration of MCEM. The intersections of the black lines indicate data-generating parameters. Figure 2c shows the MEWE's marginal distribution for κ for the different levels of n, centered and rescaled by $\sqrt{n}$, illustrating the rate of convergence anticipated by Theorem 2.3, but that the asymptotic variance is larger than in the i.i.d. case. Figure 2d shows the autocorrelation function of a data set generated with $\theta_\star = (3,1,2,0.5)$, $\rho = 0.75$, and $n = 1,000$.

4.2 *Sum of log-Normal random variables*

The distribution of the sum of log-Normal random variables appears in various settings (Fenton, 1960; Rodrigues et al., 2018), but no analytical formula is available for its probability density function, and thus the associated likelihood function is intractable. For a given positive integer L, $\gamma \in \mathbb{R}$ and $\sigma > 0$, the model generates an observation $y \in \mathbb{R}$ by sampling $x_1, \ldots, x_L \sim \mathcal{N}(\gamma, \sigma^2)$ independently, and defining $y = \sum_{\ell=1}^{L} \exp(x_\ell)$. Thus, sampling synthetic observations from the model is simple. We consider the task of estimating $\theta = (\gamma, \sigma)$ from data, fixing L to 10, and using the MEWE. We generate n observations independently using $\theta_\star = (0,1)$.

In Figure 3, we illustrate the behavior of the MEWE with $p = 1$ and $m = 10^4$ for different sizes of observed data n. The sampling distribution of the MEWE appears to concentrate around the data-generating parameter $\theta_\star$ at the $\sqrt{n}$ rate as n increases. In computing the MEWE, we used $k = 20$ and one iteration of MCEM as in the previous section.

We estimate the coverage of bootstrap confidence intervals calculated for $\theta_\star = (0,1)$. As before, we use the percentile bootstrap (Efron and Tibshirani, 1994) for data sets of size $n = 1,000$ and synthetic data sets of size $m = 10^4$, and calculate the MEWE with $k = 20$. We draw 400 data sets from the data-generating process, and 1,000 bootstrap data sets for each. The observed coverage rates were 0.945 and 0.940 for $\gamma_\star$ and $\sigma_\star$ respectively, which are close to the limiting 0.95 coverage rates. After a Bonferroni correction, the observed coverage of the confidence sets for $\theta_\star$ was 0.960.



(a) MEWE of (γ, σ). (b) $\sqrt{n}$-scaled estim. of γ. (c) $\sqrt{n}$-scaled estim. of σ. (d) Histogram of data.

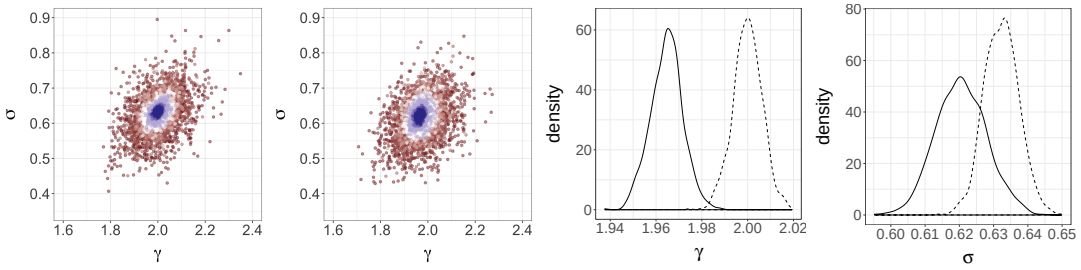
FIG. 3: Estimators in the well-specified sum of log-Normals model, as described in Section 4.2. Figure 3a shows the sampling distributions of the MEWE, as n ranges from 50 to 10^4 (colors from red to white to blue as n increases). For each n, we plot $M = 1,000$ estimators based on independent data sets. Each estimator was computed with $p = 1$, $m = 10^4$, $k = 20$, and one iteration of MCEM. The intersections of the black lines indicate data-generating parameters. Figures 3b and 3c show the MEWE's marginal distributions for the different levels of n, centered and rescaled by $\sqrt{n}$, illustrating the rate of convergence anticipated by Theorem 2.3. Figure 3d is a histogram of a data set generated with $\theta_\star = (0, 1)$ and $n = 1,000$.

4.3 *Gamma data fitted with a Normal model*

We now consider a misspecified setting. Let $\mu_\star$ be a Gamma$(10, 5)$ distribution (parametrized by shape and rate) and $\mathscr{M} = \{\mathcal{N}(\gamma, \sigma^2) : \gamma \in \mathbb{R}, \sigma > 0\}$. The Normal location-scale model is very simple, yet it is widely used in practice in the form of regression models. Figure 4 compares the sampling distributions of the maximum likelihood estimator and approximations of the MEWE of order 1, over $M = 1,000$ experiments, for different values of n. The MEWE converges at the same $\sqrt{n}$ rate as the MLE, albeit to a distribution that is centered at a different location. Therefore, despite both estimation techniques leading to similar values for γ and σ, the distributions of the estimators have very little overlap for large n, as observed in Figures 4c and 4d. For the MEWE, we have again used $m = 10^4$, $k = 20$, and one iteration of MCEM.

In Figure 5, we fix an observed data set of size $n = 100$, and compute $M = 500$ instances of the approximate MEWE for 8 different values of k and m, ranging from 1 to 1,000 and 10 to 10,000 respectively. In Figure 5a, we plot the estimators obtained for all the levels of k, given 4 different values of m. In Figure 5b, we plot the estimators obtained for all the levels of m, given 4 different values of k. The axis scales are different for each subplot. In both figures, black points correspond to the "true" MWE, calculated using a very large value of m ($m = 10^8$). For low values of m, the estimators might be significantly different from the MWE, as can be seen from the lower-right sub-plots of Figure 5b. When m increases, the estimators converge to the MWE. Increasing k reduces variation in the estimator. The changes in k and m had no significant impact on the number of evaluations of the objective required to locate the maximum using the `optim` function in `R` (R Core Team, 2015), which uses the Nelder–Mead simplex method (Nelder and Mead, 1965).

We check the coverage of bootstrap confidence intervals calculated for $\theta_\star$ (itself calculated using $n = m = 10^8$ and $k = 1$). As before, we use the percentile bootstrap (Efron and Tibshirani, 1994) for



(a) MLE of (γ, σ). (b) MEWE of (γ, σ). (c) Estimators of γ. (d) Estimators of σ.

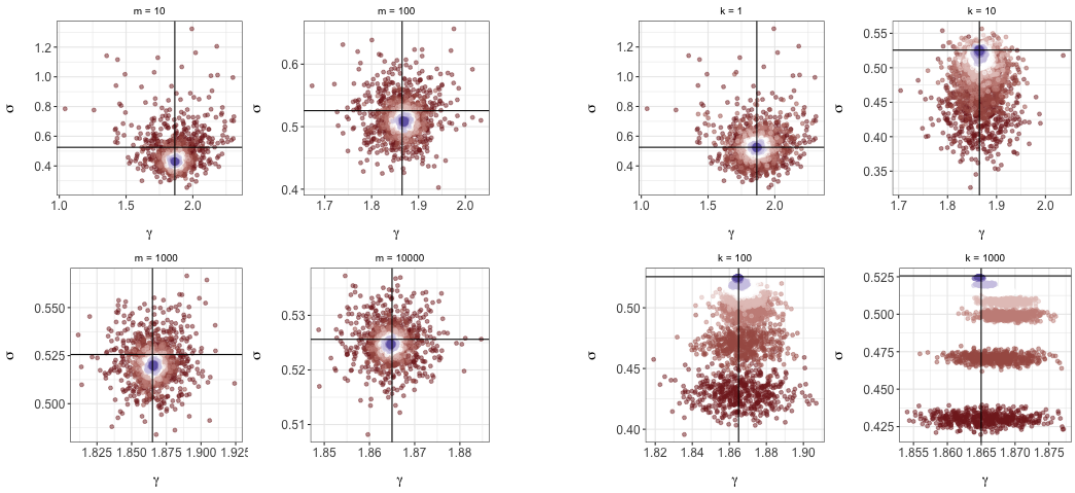
FIG. 4: Gamma data fitted with a Normal model, as described in Section 4.3. Figures 4a and 4b show the sampling distributions of the MLE and MEWE of order 1 respectively, as n ranges from 50 to 10^4 (colors from red to white to blue). Figures 4c and 4d show the marginal densities of the estimators of γ and σ respectively, for $n = 10^4$; the MLEs are shown in dashed lines and the MEWE in full lines. For the MEWE, we have used $m = 10^4$, $k = 20$ and one iteration of MCEM.

data sets of size $n = 1,000$ and synthetic data sets of size $m = 10^4$, and calculate the MEWE with $k = 20$. We draw 400 data sets from the data-generating process, and 1,000 bootstrap data sets for each of these. The observed coverage rates of the resulting 0.95 confidence intervals were 0.960 and 0.953 for $\gamma_\star$ and $\sigma_\star$ respectively. After a Bonferroni correction, the observed coverage rate of the confidence sets for $\theta_\star = (\gamma_\star, \sigma_\star)$ was 0.955.

4.4 *Cauchy data fitted with a Normal model*

Let $\mu_\star$ be Cauchy with median zero and scale one, and consider the model $\mathscr{M} = \{\mathcal{N}(\gamma, \sigma^2) : \gamma \in \mathbb{R}, \sigma > 0\}$. We explore the behavior of the MEWE of order 1, over $M = 1,000$ repeated experiments. Figure 6 shows its sampling distributions, for n ranging from 50 to 10^4. The marginal distribution of the estimator of γ concentrates around 0, the median of $\mu_\star$. The marginal distribution of the estimator of σ also concentrates to a value close to 2.2. The concentration appears to occur at rate $\sqrt{n}$, as shown by the marginal densities of the rescaled estimators of γ and σ in Figures 6a and 6b.

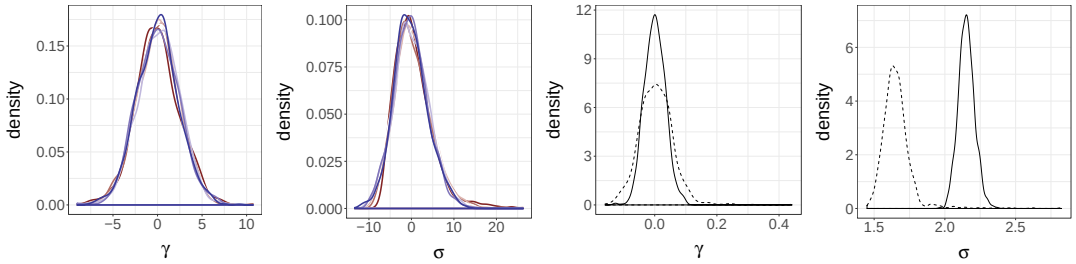
In this setting the maximum likelihood estimator would not converge as $n \to \infty$, as the maximum likelihood estimator for γ is the sample average, and the sample average of independent Cauchy variables is also Cauchy, with the same location and scale. As an alternative, we consider an estimator defined by minimizing a sample based estimator of the Kullback-Leibler divergence between μ_θ and $\mu_\star$. For the KL approximation we use the function `KL.divergence` in the `FNN` package (Beygelzimer et al., 2013), which approximates the KL divergence using ℓ-nearest neighbor estimates described in Boltz et al. (2009) (and using the default parameter $\ell = 5$). The resulting estimator is termed the minimum KL estimator (MKLE), and is a variation of the MDEs discussed by Basu et al. (2011). We compute it using the same approach as for the MEWE, using $k = 20$, $m = 10^4$, and one iteration of MCEM. For $n = 5,000$ the distributions of MEWEs and MKLEs are plotted in Figures 6c and 6d. Both estimators appear to be robust in the sense that they converge to well-defined limits, unlike the MLE approach. The estimators of γ are concentrated around 0, but the estimators of σ are concentrated around two different values: the MEWEs seem to concentrate around 2.15 and the MKLEs around 1.65. The marginal distributions of the MEWE appear to have slightly smaller variance than those of the MKLE.



(a) Approximate MEWE for increasing k (colors from red to white to blue as k increases), for different values of m.

(b) Approximate MEWE for increasing m (colors from red to white to blue as m increases), for different values of k.

FIG. 5: Gamma data with $n = 100$, fitted with a Normal model, as described in Section 4.3. MEWEs are obtained for different values of m (from 10 to $10,000$) and k (from 1 to $1,000$), using one iteration of MCEM, $M = 500$ times independently. The intersections of the black lines represent the location of the "exact" MWE computed with $n = m = 10^8$.



(a) $\sqrt{n}$-scaled estim. of γ. (b) $\sqrt{n}$-scaled estim. of σ. (c) Estimators of γ. (d) Estimators of σ.

FIG. 6: Cauchy data fitted with a Normal model, as described in Section 4.4. Marginals distributions of the MEWE of γ and σ, centered by $\theta^\star$ itself computed with $n = m = 10^8$, and rescaled by $\sqrt{n}$, are shown in Figures 6a and 6b. Figures 6c and 6d show the distributions of the MEWE for $n = 5{,}000$ (full lines), along with the distribution of an estimator obtained by minimizing an estimate of the Kullback–Leibler divergence (dashed lines).

Note that this example is not covered by the theoretical results of Section 2 since the Cauchy distribution does not have a finite first moment. Robustness properties of general minimum distance estimators are discussed in Parr and Schucany (1980), and of the MWE in location models in Bassetti and Regazzini (2006). In the location-scale model considered here, if the approximation of the MEWE is computed with $k = 1$ and $m = \ell n$ for some $\ell \geqslant 1$, it can be written

$$\operatorname*{argmin}_{\gamma,\sigma} \sum_{i=1}^{n} \sum_{j=1}^{\ell} |y_{(i)} - (\sigma x_{(\ell(i-1)+j)} + \gamma)|. \tag{4.2}$$

As such, the approximate MEWE can be seen as the coefficients in a median regression (Koenker and Hallock, 2001) of a vector $\tilde{Y}$ on a vector $\tilde{X}$, where $\tilde{Y}_{\ell(i-1)+1:\ell i} = y_{(i)}$ for each $i = 1, \ldots, n$, and $\tilde{X}$ contains the order statistics of an m-sample of $\mathcal{N}(0,1)$ random variables. Quantile regression is often presented as a robust alternative to linear regression in the presence of outliers, and further connections might explain the observed robustness of the MEWE with $p = 1$ in this example.

5. Discussion

The minimum Wasserstein (or Kantorovich) estimation approach (Bassetti et al., 2006) has received a renewed attention, due to recent advances in the field of computational optimal transport (Peyré and Cuturi, 2018), along with various applications in machine learning. In the broad context of generative models, these estimators present various appeals compared to maximum likelihood estimators. For instance, in Sections 4.1 and 4.2, we have observed the satisfactory behavior of minimum expected Wasserstein estimators in models where the likelihood function is not analytically available. In Sections 4.3 and 4.4 we have observed similarities and differences between MEWE and MLE in misspecified settings, illustrating some robustness properties of minimum Wasserstein estimation.

Minimum distance estimators were originally developed for obtaining almost surely convergent estimators (Wolfowitz, 1957), and we have showed that both the MWE and MEWE have this strong consistency property under mild conditions. We have also proved that the MWE converges to $\theta_\star$ at the optimal $\sqrt{n}$ convergence rate when the observations are univariate, and have derived its asymptotic distribution.

The generalization of this result to multivariate data is left for future research. Interestingly, given the known convergence properties of the Wasserstein distance, it seems reasonable to conjecture that the rate of the MWE depends (negatively) on the dimension of the observation space rather than that of the parameter space. Other topics for future research include a more general derivation of the limiting distributions of the estimators, whose existence is needed to justify the asymptotic coverage of subsampling confidence intervals, as well as the development of a better understanding of their robustness properties.

Acknowledgements

The bootstrap experiments were in part performed on the Odyssey cluster supported by the FAS Division of Science, Research Computing Group at Harvard University. Pierre E. Jacob acknowledges support from the National Science Foundation through grant DMS-1712872.

References

Altschuler, J., Bach, F., Rudi, A., and Weed, J. (2018). Massively scalable Sinkhorn distances via the Nyström method. *arXiv preprint arXiv:1812.05189*. 2, 10

Altschuler, J., Weed, J., and Rigollet, P. (2017). Near-linear time approximation algorithms for optimal transport via sinkhorn iteration. In *Advances in Neural Information Processing Systems*, pages 1964–1974. 10

Ambrosio, L., Gigli, N., and Savaré, G. (2005). *Gradient Flows in Metric Spaces and in the Space of Probability Measures*. Birkhäuser Verlag AG, Basel, second edition. 5, 9

Arjovsky, M., Chintala, S., and Bottou, L. (2017). Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 214–223. PMLR. 2, 4

Bassetti, F., Bodini, A., and Regazzini, E. (2006). On minimum Kantorovich distance estimators. *Statistics & probability letters*, 76(12):1298–1302. 1, 4, 5, 17

Bassetti, F. and Regazzini, E. (2006). Asymptotic properties and robustness of minimum dissimilarity estimators of location-scale parameters. *Theory of Probability and its Applications*, 50(2):171–186. 2, 4, 6, 17

Basu, A., Shioya, H., and Park, C. (2011). *Statistical Inference: the Minimum Distance Approach*. CRC Press. 2, 3, 15

Belili, N., Bensaï, A., and Heinich, H. (1999). Estimation based on the Kantorovich functional and the Lévy distance. *Comptes Rendus de l'Academie des Sciences Series I Mathematics*, 5(328):423–426. 1

Benamou, J.-D., Carlier, G., Cuturi, M., Nenna, L., and Peyré, G. (2015). Iterative Bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138. 2, 10

Bernton, E., Jacob, P. E., Gerber, M., and Robert, C. P. (2019). Approximate Bayesian computation with the Wasserstein distance. *To appear, Journal of the Royal Statistical Society: Series B*. 2, 10, 11

Bertsimas, D. and Tsitsiklis, J. N. (1997). *Introduction to Linear Optimization*. Athena Scientific. 9

Beygelzimer, A., Kakadet, S., Langford, J., Arya, S., Mount, D., and Li, S. (2013). FNN: fast nearest neighbor search algorithms and applications. *R package version*, 1. 15

Bickel, P. J. and Sakov, A. (2008). On the choice of m in the m out of n bootstrap and confidence bounds for extrema. *Statistica Sinica*, pages 967–985. 7

Boltz, S., Debreuve, E., and Barlaud, M. (2009). High-dimensional statistical measure for region-of-interest tracking. *IEEE Transactions on Image Processing*, 18(6):1266–1283. 15

Bonneel, N., Rabin, J., Peyré, G., and Pfister, H. (2015). Sliced and radon Wasserstein barycenters of measures. *Journal of Mathematical Imaging and Vision*, 51(1):22–45. 10

Brown, L. D. and Purves, R. (1973). Measurable selections of extrema. *Annals of Statistics*, 1(5):902–912. 5

Burkard, R., Dell'Amico, M., and Martello, S. (2009). *Assignment Problems*. Society for Industrial and Applied Mathematics (SIAM). 9

Chen, Y. and Li, W. (2018). Natural gradient in Wasserstein statistical manifold. *arXiv preprint arXiv:1805.08380*. 11

Cheney, E. W. and Wulbert, D. E. (1969). The existence and unicity of best approximations. *Mathematica Scandinavica*, 24:113–140. 6

Cuturi, M. (2013). Sinkhorn distances: lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2292–2300. 2, 10

Dede, S. (2009). An empirical central limit theorem in l^1 for stationary sequences. *Stochastic Processes and their Applications*, 119:3494 – 3515. 6

del Barrio, E., Giné, E., and Matrán, C. (1999). Central limit theorems for the Wasserstein distance between the empirical and the true distributions. *Annals of Probability*, pages 1009–1071. 6, 7

del Barrio, E., Giné, E., Utzet, F., et al. (2005). Asymptotics for l_2 functionals of the empirical quantile process, with applications to tests of fit based on weighted Wasserstein distances. *Bernoulli*, 11(1):131–189. 6

del Barrio, E. and Loubes, J.-M. (2017). Central limit theorems for empirical transportation cost in general dimension. *arXiv preprint arXiv:1705.01299*. 7, 8

Devroye, L. (1985). *Non-uniform random variate generation*. Springer-Verlag, New York. 9

Efron, B. and Tibshirani, R. J. (1994). *An Introduction to the Bootstrap*. CRC press. 7, 12, 13, 14

Fenton, L. (1960). The sum of log-Normal probability distributions in scatter transmission systems. *IRE Transactions on Communications Systems*, 8(1):57–67. 13

Fournier, N. and Guillin, A. (2015). On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162:707–738. 8

Genevay, A., Cuturi, M., Peyré, G., and Bach, F. (2016). Stochastic optimization for large-scale optimal transport. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3432–3440. 2

Genevay, A., Peyré, G., and Cuturi, M. (2017). GAN and VAE from an optimal transport point of view. *arXiv preprint arXiv:1706.01807*. 2

Genevay, A., Peyre, G., and Cuturi, M. (2018). Learning generative models with Sinkhorn divergences. In Storkey, A. and Perez-Cruz, F., editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1608–1617. 2, 4, 10

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(10):e110214. 9

Gouriéroux, C., Monfort, A., and Renault, E. (1993). Indirect inference. *Journal of Applied Econometrics*, 8:85–118. 2, 3

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica: Journal of the Econometric Society*, pages 1029–1054. 3

Jorge, M. and Boris, I. (1984). Some properties of the Tukey g and h family of distributions. *Communications in Statistics-Theory and Methods*, 13(3):353–369. 11

Koenker, R. and Hallock, K. F. (2001). Quantile regression. *Journal of Economic Perspectives*, 15(4):143–156. 17

Le Cam, L. (1970). On the assumptions used to prove asymptotic normality of maximum likelihood estimators. *Annals of Mathematical Statistics*, 41:802–828. 6

Li, W., Ryu, E. K., Osher, S., Yin, W., and Gangbo, W. (2018). A parallel method for Earth Movers distance. *Journal of Scientific Computing*, 75(1):182–197. 2, 10

Marin, J.-M., Pudlo, P., Robert, C. P., and Ryder, R. J. (2012). Approximate Bayesian computational methods. *Statistics and Computing*, 22(6):1167–1180. 2

McFadden, D. (1989). A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica*, 57(5):995–1026. 3

Neath, R. C. et al. (2013). On convergence properties of the Monte Carlo EM algorithm. In *Advances in Modern Statistical Theory and Applications: A Festschrift in Honor of Morris L. Eaton*, pages 43–62. Institute of Mathematical Statistics. 10

Nelder, J. A. and Mead, R. (1965). A simplex method for function minimization. *The Computer Journal*, 7(4):308–313. 14

Owen, A. B. (2001). *Empirical likelihood*. CRC press. 3

Parr, W. C. and Schucany, W. R. (1980). Minimum distance and robust estimation. *Journal of the American Statistical Association*, 75(371):616–624. 17

Peyré, G. and Cuturi, M. (2018). Computational Optimal Transport. *arXiv preprint arXiv:1803.00567*. 2, 9, 10, 17

Politis, D. N., Romano, J. P., and Wolf, M. (1999). *Subsampling*. Springer-Verlag New York. 6

Pollard, D. (1980). The minimum distance method of testing. *Metrika*, 27:43–70. 2, 5, 6

Puccetti, G. (2017). An algorithm to approximate the optimal expected inner product of two vectors with given marginals. *Journal of Mathematical Analysis and Applications*, 451(1):132–145. 10

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. 11, 14

Rabin, J., Peyré, G., Delon, J., and Bernot, M. (2011). Wasserstein barycenter and its application to texture mixing. In *International Conference on Scale Space and Variational Methods in Computer Vision*, pages 435–446. Springer. 10

Ramdas, A., Trillos, N. G., and Cuturi, M. (2017). On Wasserstein two-sample testing and related families of nonparametric tests. *Entropy*, 19(2):47. 10

Rayner, G. D. and MacGillivray, H. L. (2002). Numerical maximum likelihood estimation for the g-and-k and generalized g-and-h distributions. *Statistics and Computing*, 12(1):57–75. 11

Rockafellar, R. T. and Wets, R. J.-B. (2009). *Variational Analysis*, volume 317. Springer Science & Business Media. 2

Rodrigues, G., Prangle, D., and Sisson, S. (2018). Recalibration: A post-processing method for approximate Bayesian computation. *Computational Statistics & Data Analysis*, 126:53–66. 13

Rubio, F. J., Johansen, A. M., et al. (2013). A simple approach to maximum intractable likelihood estimation. *Electronic Journal of Statistics*, 7:1632–1654. 11

Schuhmacher, D., Bhre, B., Gottschlich, C., and Heinemann, F. (2017). `transport`: *Optimal Transport in Various Forms*. R package version 0.8-2. 9

Sisson, S. A., Fan, Y., and Beaumont, M. (2018). *Handbook of Approximate Bayesian Computation*. CRC Press. 11

Tukey, J. W. (1977). Modern techniques in data analysis. In *Proceedings of the NSF-Sponsored Regional Research Conference*, volume 7. Southern Massachusetts University. 11

Van der Vaart, A. W. (2000). *Asymptotic Statistics*. Cambridge university press. 5

Villani, C. (2008). *Optimal Transport, Old and New*. Springer-Verlag New York. 4

Weed, J. and Bach, F. (2019). Sharp asymptotic and finite-sample rates of convergence of empirical measures in Wasserstein distance. *To appear, Bernoulli*. 7

Wei, G. C. and Tanner, M. A. (1990). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, 85(411):699–704. 10

Wellner, J. A. and van der Vaart, A. W. (1996). *Weak Convergence and Empirical Processes*. Springer-Verlag New York. 6

Wolfowitz, J. (1957). The minimum distance method. *The Annals of Mathematical Statistics*, 28(1):75–88. 3, 17

Wood, S. N. (2010). Statistical inference for noisy nonlinear ecological dynamic systems. *Nature*, 466(7310):1102–1104. 11

Ye, J., Wang, J. Z., and Li, J. (2017). A simulated annealing based inexact oracle for Wasserstein loss minimization. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 3940–3948. PMLR. 2, 10

List of Figures