SIS | 2022

51st Scientific Meeting of the Italian Statistical Society

Caserta, 22-24 June

Università degli Studi della Campania *Luigi Vanvitelli*

SIS Società Italiana di Statistica

www.unicampania.it

Book of the Short Papers

Editors: Antonio Balzanella, Matilde Bini, Carlo Cavicchia, Rosanna Verde

Città di Caserta

800 ANNI 1222·2022 Università degli Studi di Padova

Dipartimento di Scienze Statistiche

Stata

sas

Università degli Studi del Sannio Benevento

Pearson

PUBLISHED BY PEARSON

WWW.PEARSON.COM

ISBN 9788891932310

Wasserstein distance and applications to Bayesian nonparametrics

Distanza di Wasserstein e applicazioni in statistica bayesiana nonparametrica

Marta Catalano, Hugo Lavenant, Antonio Lijoi, Igor Prünster

Abstract Bayesian nonparametric models are able to learn complex distributional patterns in the data by leveraging on infinite-dimensional parameters, typically consisting in vectors of random measures. To perform a principled BNP model comparison one thus needs a measure of discrepancy between vectors of random measures. In recent works the authors have proposed two different metrics based on the Wasserstein distance. We here provide new perspectives to our findings, highlighting a universal relation between the two metrics.

Abstract *I modelli bayesiani nonparametrici (BNP) sono in grado in imparare assetti distribuzionali complessi nei dati facendo leva su parametri infinito-dimensionali, che tipicamente consistono in vettori di misure aleatorie. Per portare avanti un paragone tra modelli BNP basato su fondamenta solide, è dunque necessario definire una misura di discrepanza tra vettori di misure aleatorie. In alcuni lavori recenti gli autori hanno proposto due diverse metriche basate sulla distanza di Wasserstein. In questo contributo forniamo una nuove prospettiva che mette in evidenza una relazione universale tra le due metriche.*

Key words: Bayesian model comparison, Lévy measure, Wasserstein distance.

Marta Catalano
University of Warwick, UK, e-mail: marta.catalano@warwick.ac.uk

Hugo Lavenant
Bocconi University, Italy, e-mail: hugo.lavenant@bocconi.it

Antonio Lijoi
Bocconi University, Italy, e-mail: antonio.lijoi@bocconi.it

Igor Prünster
Bocconi University, Italy, e-mail: igor.pruenster@bocconi.it

1 Introduction

The Wasserstein distance provides a comprehensive way to quantify the comparison between two random objects. Its first definition traces back to [8], though it then appeared independently in many other scientific fields. As a result, one can find this *simple measure of discrepancy* (using Gini's words) under different names, such as Gini distance, coupling distance, Monge-Kantorovich distance, Earth Moving distance and Mallows distance; see [5, 12, 11] for reviews.
Random structures are the key to statistical modeling and inference, especially in a Bayesian framework where the parameter of the model is random as well. We argue that a principled way to perform Bayesian model comparison is then to measure the discrepancy between the random parameters of the models, by relying on the Wasserstein distance.
In this contribution we focus on Bayesian nonparametric (BNP) models, which are based on infinite-dimensional parameters typically consisting on vectors of random measures. For example, most BNP models for *partially exchangeable sequences* [4] are built as follows. First, one considers a vector of dependent random measures $\tilde{\mu} = (\tilde{\mu}_1, \ldots, \tilde{\mu}_d)$, such as the *completely random vectors* (CRVs) introduced in Section 3. Then, one models d groups of observations $\boldsymbol{X}^i = (X_1^i, \ldots, X_{n_i}^i)$ as conditionally independent given $\tilde{\mu}$ with

$$\boldsymbol{X}^i | \tilde{\mu} \overset{\text{ind}}{\sim} T(\tilde{\mu}_i)$$

for $i = 1, \ldots d$, where T is some transformation of the random measure.
To perform a BNP model comparison one thus needs the definition of a distance between CRVs, whose law is specified through a multivariate Lévy intensity. Following [3, 2, 1] we here describe two different ways of defining such a distance, one at the level of the random measures and one at the level of the underlying Lévy measures. Generalizing the findings in [1], we highlight that the two metrics are intimately connected and provide a universal relation between them that allows the practionner to exploit the best properties of each one.

2 Wasserstein distance

In its classical formulation, the Wasserstein distance provides the means for comparing two probability measures on a Polish space $\mathbb{X}$ with metric d. Recent developments allow to extend the comparison from probability measures (that is, measures with mass 1) to generic measures with possibly different masses. A particularly interesting scenario arises when the masses are infinite, as in the case of infinitely active Lévy measures. In this section we describe both the Wasserstein distance and its extended version.
We first recall the notion of coupling. Given two probability measures P^1, P^2 on a Polish space $(\mathbb{X}, d)$ we define a coupling to be the law of any random vector (X, Y)

on the product space $\mathbb{X} \times \mathbb{X}$ such that $X \sim P^1$ and $Y \sim P^2$. We denote by $\Gamma(P^1, P^2)$ the set of couplings between P^1 and P^2. A probability P^1 has finite p-th moment if for any $x \in \mathbb{X}$,

$$\int_{\mathbb{X}} d(s,x)^p \mathrm{d}P(s) < +\infty.$$

Definition 1. Let P^1 and P^2 be probability measures on $(\mathbb{X}, d)$ with finite p-th moment. The *Wasserstein distance* of order p between P^1 and P^2 is

$$\mathscr{W}_p(P^1, P^2)^p = \inf_{(X,Y) \in \Gamma(P^1,P^2)} \mathbb{E}(d(X,Y)^p).$$

We observe that the finiteness of the p-th moment guarantees the finiteness of the Wasserstein distance thanks to the triangular inequality of the metric d. A very common choice for $(\mathbb{X}, d)$ is $\mathbb{R}^d$ endowed with the Euclidean norm $\|\cdot\|$. In this case, P has finite p-th moment if and only if $\int \|x\|^p \mathrm{d}P(x)$ is finite.
Starting from this scenario we extend the definition to Lévy measures on $\Omega_d = [0, +\infty)^d \setminus \{\mathbf{0}\}$. This distance was studied in [7, 9, 1] for $p = 2$. To this end, we insist on an equivalent definition of coupling in terms of pushforward measures, which will be easier to generalize to Lévy measures. For a point $(\boldsymbol{s}, \boldsymbol{s}') \in \mathbb{X} \times \mathbb{X}$, we denote by $\pi_1(\boldsymbol{s}, \boldsymbol{s}') = \boldsymbol{s} \in \mathbb{X}$ and $\pi_2(\boldsymbol{s}, \boldsymbol{s}') = \boldsymbol{s}' \in \mathbb{X}$ its projections. For any measure spaces $\mathbb{X}^1$ and $\mathbb{X}^2$, if μ is a measure on $\mathbb{X}^1$ and $f : \mathbb{X}^1 \to \mathbb{X}^2$, the pushforward measure of μ by f is the measure on $\mathbb{X}^2$ defined by $(f_\# \mu)(A) = \mu(f^{-1}(A))$. Then an equivalent way of defining a coupling $\gamma \in C(P^1, P^2)$ is as a probability measure on $\mathbb{X} \times \mathbb{X}$ such that $\pi_{i\#}\gamma = P^i$ for $i = 1, 2$. Thus,

$$\mathscr{W}_p(P^1, P^2)^p = \inf_{\gamma \in \Gamma(P^1,P^2)} \int_{\mathbb{X} \times \mathbb{X}} d(x,y)^p \mathrm{d}\gamma(x,y).$$

We now extend the Wasserstein distance to Lévy measures on $\Omega_d = [0, +\infty)^d \setminus \{\mathbf{0}\}$ with p-th finite moment, that is, the set of positive Borel measures ν such that

$$\int_{\Omega_d} \|s\|^p \mathrm{d}\nu(s) < +\infty.$$

We observe that the finite moment condition does not prevent the mass from being infinite. Indeed, there can be an accumulation of infinite mass around the origin, as is the case with infinitely active Lévy measures.
Let ν^1, ν^2 two Lévy measures. An extended coupling γ between ν^1 and ν^2 is a Lévy measure on Ω_{2d} such that $\pi_{1\#}\gamma|_{\Omega_d} = \nu^1$ and $\pi_{2\#}\gamma|_{\Omega_d} = \nu^2$. We denote $\Gamma_*(\nu^1, \nu^2)$ the set of all extended couplings.

Definition 2. Let ν^1 and ν^2 be Lévy measures on $(\Omega_d, \|\cdot\|)$ with finite p-th moment. The *extended Wasserstein distance* of order p between ν^1 and ν^2 is

$$\mathscr{W}_{*,p}(\nu^1, \nu^2)^p = \inf_{\gamma \in \Gamma_*(\nu^1,\nu^2)} \iint_{\Omega_{2d}} \|\boldsymbol{s} - \boldsymbol{s}'\|^p \,\mathrm{d}\gamma(\boldsymbol{s}, \boldsymbol{s}'). \tag{1}$$

The main difference between a coupling and an extended coupling is that extended couplings are not defined on $\Omega_d \times \Omega_d$ but rather on Ω_{2d}, which is strictly larger since it also contains the axis $\{\mathbf{0}\} \times \Omega_d$ and $\Omega_d \times \{\mathbf{0}\}$. This is fundamental to ensure the compactness of extended couplings which, in turn, guarantees the existence of an *optimal extended coupling*, that is, a coupling that attains the minimum in the definition of extended Wasserstein distance.

3 Completely random vectors

The key component of Bayesian nonparametric models is the use of random measures to learn complex distributional patterns in the data. Among these, completely random measures stand out for combining inferential flexibility with analytical tractability. Quantifying the discrepancy between two completely random measures or their multivariate extension - completely random vectors - is key to establishing a principled approach to the comparisons between the induced Bayesian nonparametric models. This can be done through the Wasserstein distance at two different levels: either at the level of the random measures or at the level of their underlying Lévy intensity. We here describe these notion and extend a result by [1], which shows that these two levels are intimately connected.
Let $\mathscr{M}(\mathbb{X})$ denote the set of boundedly finite measures on $\mathbb{X}$ endowed with the weak$^\sharp$ topology [6]. We recall that a sequence of measures μ_n converges weakly$^\sharp$ to μ if and only if $\int f \mathrm{d}\mu_n \to \int f \mathrm{d}\mu$ for every bounded continuous f vanishing outside a bounded set. A random measure is a random element on $\mathscr{M}(\mathbb{X})$. Similarly, a d-dimensional random vector of measures $\tilde{\mu} = (\tilde{\mu}_1, \ldots, \tilde{\mu}_d)$ is a random element on the product space $\mathscr{M}(\mathbb{X})^d$.

Definition 3. A random vector of measures is a *completely random vector* (CRV) if for every finite collection of pairwise disjoint bounded sets $\{A_1, \cdots, A_n\}$ in $\mathscr{B}(\mathbb{X})$, $\{\tilde{\mu}(A_1), \cdots, \tilde{\mu}(A_n)\}$ are mutually independent.

The definition of completely random vector was first given by [2] and it extends the notion of completely random measure (CRM [10]) in a natural way. In particular, a CRM is a 1-dimensional CRV. Every CRV can be decomposed as the sum of three independent components, $\mu + \tilde{\mu}_f + \tilde{\mu}$ in distribution, where μ is a deterministic vector of measures, $\tilde{\mu}_f$ is a random vector of measures with fixed atoms and $\tilde{\mu}$ is a random vector of measures without fixed atoms. In the rest of the contribution we will focus on CRVs without fixed atoms. Remarkably, for every such CRV there exists a measure ρ on $\Omega_d \times \mathbb{X}$ such that

$$\mathrm{d}\tilde{\mu}(x) = \int_{\Omega_d} s \, \mathrm{d}\mathscr{N}(s,x)$$

in distribution, where $\mathscr{N}$ is a Poisson random measure with Lévy intensity ρ on $\Omega_d \times \mathbb{X}$. In particular, $\tilde{\mu}(A)$ is an infinite divisible distribution with Lévy measure

$\mathrm{d}\nu_A(s) = \int_A \mathrm{d}\rho(s,x)$ on Ω_d. Thus, to assign a probability distribution to a CRV it suffices to provide a set of Lévy measures $(\nu_A)_A$.
It is then clear that the comparison between CRVs can be based either at the level of the random measures or at the level of the Lévy measures. In the first case, given two CRVs $\tilde{\mu}^1$ and $\tilde{\mu}^2$ one defines

$$d_{\mathscr{W},p}(\tilde{\mu}^1, \tilde{\mu}^2) = \sup_A \mathscr{W}_p(\mathscr{L}(\tilde{\mu}^1(A)), \mathscr{L}(\tilde{\mu}^2(A))), \tag{2}$$

where $\mathscr{L}$ denotes the law of a random object. We observe that $\mathscr{L}(\tilde{\mu}^i(A))$ is a probability measure and thus $\mathscr{W}_p$ is the classical Wasserstein distance in Definition 1. In the second case one defines

$$d_{\mathscr{W}_*,p}(\tilde{\mu}^1, \tilde{\mu}^2) = \sup_A \mathscr{W}_{*,p}(\nu_A^1, \nu_A^2). \tag{3}$$

Here, ν_A^i is a Lévy measure on Ω_d and thus $\mathscr{W}_{*,p}$ is the extended Wasserstein distance in Definition 2.
The advantage for (2) is that, being defined directly on the the random object, it has a very interpretable definition. On the other hand the advantage of (3) is that, being defined directly on the Lévy measure, it is easier to compute. The next result generalizes the findings in [1] and can be proved following the same technique therein. We provide a universal relation between the two distances that allows to keep the best of the two worlds: we can now base the computations on (3) though maintaining the interpretability of (2).

Theorem 1. *Let $\tilde{\mu}^1, \tilde{\mu}^2$ be CRVs with Lévy measures $(\nu_A^1)_A$ and $(\nu_A^2)_A$, respectively, with finite p-th moment. Then for every Borel set A,*

$$\mathscr{W}_p(\mathscr{L}(\tilde{\mu}^1(A)), \mathscr{L}(\tilde{\mu}^2(A))) \leq \mathscr{W}_{*,p}(\nu_A^1, \nu_A^2).$$

References

1. Marta Catalano, Hugo Lavenant, Antonio Lijoi, and Igor Prünster. A Wasserstein index of dependence for random measures. *arXiv:2109.06646*, 2022+.
2. Marta Catalano, Antonio Lijoi, and Igor Prünster. Measuring dependence in the Wasserstein distance for Bayesian nonparametric models. *Ann. Statist.*, 49(5):2916–2947, 2021.
3. Marta Catalano, Antonio Lijoi, and Igor Prünster. Approximation of Bayesian models for time-to-event data. *Electron. J. Statist.*, 14(2):3366–3395, 2020.
4. Donato Michele Cifarelli and Eugenio Regazzini. Nonparametric statistical problems under partial exchangeability: The role of associative means. *Quaderni Istituto Matematica Finanziaria dell'Università di Torino Serie III*, 12:1–36, 1978.
5. Donato Michele Cifarelli and Eugenio Regazzini. On the centennial anniversary of Gini's theory of statistical relations. *Metron*, 75(2):227–242, August 2017.
6. Daryl J. Daley and David Vere-Jones. *An Introduction to the Theory of Point Processes: Volume I: Elementary Theory and Methods*. Probability and Its Applications. Springer, 2002.

7. Alessio Figalli and Nicola Gigli. A new transportation distance between non-negative measures, with applications to gradients flows with Dirichlet boundary conditions. *Journal de Mathèmatiques Pures et Appliquèes*, 94(2):107–130, 2010.
8. Corrado Gini. Di una misura delle relazioni tra le graduatorie di due caratteri. *Saggi monografici del Comune di Roma, Tip. Cecchini*, 1914.
9. Nestor Guillen, Chenchen Mou, and Andrzej Święch. Coupling Lévy measures and comparison principles for viscosity solutions. *Transactions of the American Mathematical Society*, 372(10):7327–70, 2019.
10. John F. C. Kingman. *Pacific J. Math.*, (21):59–78, 1967.
11. Svetlozar Rachev. The Monge-Kantorovich mass transference problem and its stochastic applications. *Theory of Probability & Its Applications*, 29(4):647–676, 1985.
12. Cedric Villani. *Optimal Transport: Old and New*. Grundlehren der mathematischen Wissenschaften. Springer Berlin Heidelberg, 2008.