Density Estimation of Multivariate Samples using Wasserstein Distance

ABSTRACT
Density estimation is a central topic in statistics and a fundamental task of machine learning. In this paper, we present an algorithm for approximating multivariate empirical densities with a piecewise constant distribution defined on a hyperrectangular-shaped partition of the domain. The piecewise constant distribution is constructed through a hierarchical bisection scheme, such that locally, the sample cannot be statistically distinguished from a uniform distribution. The Wasserstein distance has been used to measure the uniformity of the sample data points lying in each partition element. Since the resulting density estimator requires significantly less memory to be stored, it can be used in a situation where the information contained in a multivariate sample needs to be preserved, transferred or analysed.

KEYWORDS
Nonparametric density estimation; Wasserstein distance; piecewise constant distribution; multivariate histogram.

1. Introduction

Estimating the probability density function generating a data sample is a long-established concept and a fundamental topic in statistics, as it is a key issue in several problems of a broad range of disciplines such as physics, engineering, biology or economics. Nowadays, more than ever before, big data applications and simulation-driven approaches increasingly require scientists and researchers to analyse datasets with millions of entries. It is thus of primary importance to have procedures able to efficiently capture information contained in large samples. This paper introduces a nonparametric algorithm for estimating the empirical density with a piecewise constant distribution defined on a hyperrectangular-shaped partition of the domain. The algorithm starts with a trivial partition, a single box containing all the observations, and recursively grows it with bisections until the sample space is divided into regions where a stopping criterion is met. Our procedure design can be classified into adaptive partitioning density estimation methods [1] and framed within Density Estimation Trees, an approach formalized by [2] that is the analogue of Classification and Regression Trees [3] for density estimation.

Piecewise constant distributions are a flexible and concise class of distributions useful to construct summary structures for large data sets. They can approximate distributions with any shape, since the number of components scales with the complexity of the approximated distribution, and at the same time they can be represented or compressed very efficiently.

Our algorithm aims at generating a piecewise constant distribution such that data points inside hyperrectangles are sufficiently uniformly scattered and any further partitioning of the domain does not provide additional information about the underlying

density perspective. The uniformity is judged via a Wasserstein distance based hypothesis testing and a given hyperrectangle is not bisected when the hypothesis of uniformity is not rejected with a given significance level. The underlying reasoning of our procedure is comparable to [4] and [5].

Figure 1 outlines the process of the algorithm, which ceases to bisect the domain when the stopping rule is met in all hyperrectangles. To improve the algorithm implementation, the uniformity of the sample points within each hyperrectangle is initially tested only on marginals, and once this condition is satisfied, also at a joint level (see details below).

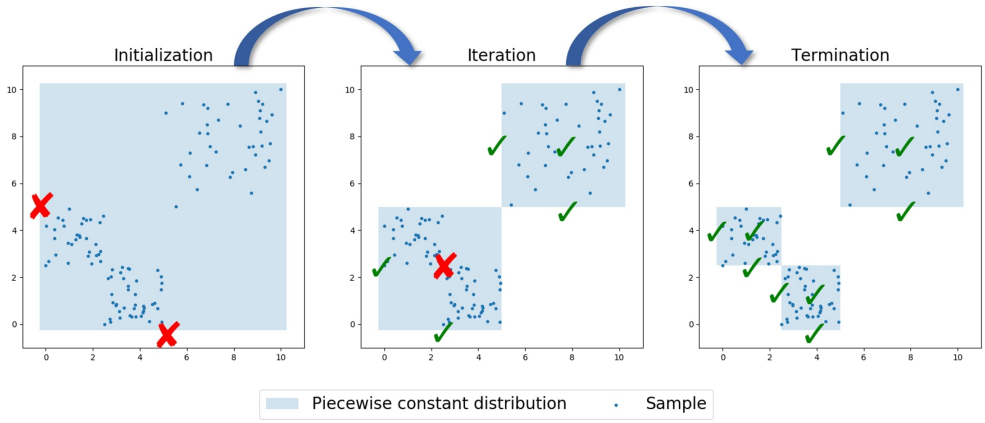

Figure 1. Sketch of algorithm phases: the domain is recursively partitioned until it is divided into regions where the stopping condition is met.

The Wasserstein distance, which arises from the idea of optimal transportation, has long been established as an important tool in probability theory and more recently has spread to both statistical theory and applications. Indeed, the Wasserstein distance is a powerful framework to compare two probability distributions and exhibits the distinctive ability to capture the geometry of the underlying space of the data, i.e. it incorporates a ground distance in comparison to other statistical distances, such as Kullback-Leibler, Hellinger, χ^2 or Total Variation, that, on the contrary, neglect how close two outcomes might be on the sample space. Moreover, the Wasserstein metric yields a map that specifies how to transform one probability distribution into the other. Last but not least, it can be applied to distributions with non-overlapping supports and compare two distributions even when one is discrete and the other is continuous.

However, the uptake of this probability distance as a statistical tool exhibits two major challenges. First, its distributional limits on spaces other than the real line are not fully known and fragmentary. Second, almost any application of the Wasserstein distance involves extensive computational effort. In this work we address both of these matters. Indeed, as mentioned above, the Wasserstein distance has been used as test statistic for verifying the uniformity hypothesis in a given hyperrectangle, which may contain a large multivariate sample. In the literature, the so-called L_2-Wassertein distance (the square root of the Wasserstein distance of order 2) has been adopted by [6] to introduce a goodness-of-fit hypothesis test between a fixed distribution and a location-scale family of probability distributions. To the knowledge of the authors other publications involving the topics of Wasserstein distance and hypothesis tests are [7] and [8]. The former introduced the Wasserstein distance in nonparametric two-

sample or homogeneity testing, the latter in uniformity and distributional property testing.

Piecewise constant distributions can be also seen as histograms with adaptive bandwidths for each dimension. Hence, our methodology can be considered as an estimation technique of multivariate histograms with data-dependent partitions [9]. Other publications involving histograms and Wasserstein distance are [10] and [11]. The former describes a strategy for constructing an optimal piecewise linear approximation of a univariate empirical distribution with a predetermined number of segments, using the Wasserstein distance of order 2 as goodness-of-fit measure. The latter presents a method to compute the Wasserstein distance of order 1 between a pair of two-dimensional histograms. Compared with these analyses, our algorithm can handle datasets with any dimension, the number of buckets is not fixed a priori, and the Wasserstein distance is a central element in the adaptive procedure for building the histogram.

This work, which has been inspired by [12] and represents an extension of their results to the multivariate setting, is organized as follows. In Section 2 we introduce piecewise constant distributions, the compatibility condition and other basic concepts which are essential to the work. Section 3, after defining the Wasserstein distance, presents the admissibility criteria, which determine the stopping rule of our partitioning algorithm, and details the characteristics of the Wasserstein distance based hypothesis test. In Section 4, the algorithm scheme for building a piecewise constant estimator is defined. Finally, in Section 5 illustrations and results of the algorithm run are presented; conclusion follows in Section 6.

2. Piecewise constant distributions

In this section, we define the class of random variables with a piecewise constant (PWC) distribution that is used as an approximation to the empirical distribution function of a given sample.

Definition 2.1. Consider a sample $(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)$ of real random vectors $\mathbf{X} : \Omega \to \mathbb{R}^d$, with common distribution function $F(t) = \mathbb{P}(\mathbf{X}_i \leq t)$, where $i = 1, ..., n$. The empirical cumulative distribution function $\hat{F} : \mathbb{R}^d \to [0, 1]$ is defined as:

$$\hat{F}(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\left\{\mathbf{X}_i \leq t\right\} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\left\{X_{i,1} \leq t_1, ..., X_{i,d} \leq t_d\right\}, \tag{1}$$

where $\mathbb{1}\{\ \}$ denotes the indicator function and $n \in \mathbb{N}$ is the sample size.

Definition 2.2. A hyperrectangle $Q_s \subset \mathbb{R}^d$ is the Cartesian product of d intervals:

$$Q_s = I_{s,1} \times I_{s,2}, ..., \times I_{s,d}, \tag{2}$$

where $I_{s,j} = (a_{s,j}, b_{s,j}]$ and $-\infty < a_{s,j} \leq b_{s,j} < +\infty$, for $j = 1, ..., d$, with the convention that $I_{s,j} = \{a_{s,j}\}$ if $a_{s,j} = b_{s,j}$.

Definition 2.3. Given a set of $S \in \mathbb{N}$ disjoint hyperrectangles $\mathcal{Q} = \{Q_s : s = 1, ..., S\}$ and probability weights $p = (p_s : s = 1, ..., S) \in \mathbb{R}^S_{\geq 0}$, such that $\sum_{s=1}^{S} p_s = 1$, a random vector $\mathbf{Y} : \Omega \to \mathbb{R}^d$ has a PWC distribution, $\mathbf{Y} \sim \mathrm{PWC}(p, \mathcal{Q})$, if its cumulative

distribution function $G : \mathbb{R}^d \to [0,1]$ can be written as

$$G(t) = \mathbb{P}(\mathbf{Y} \leq t) = \sum_{s=1}^{S} p_s \prod_{j=1}^{d} H_{I_{s,j}}(t_j), \tag{3}$$

where the function $H_I : \mathbb{R} \to [0,1]$ is defined as follows:

$$H_I(t) = \begin{cases} 0 & \text{if } t < a, \\ 1 & \text{if } b \leq t, \\ \dfrac{t-a}{b-a} & \text{else.} \end{cases} \tag{4}$$

Remark 2.4. Two special cases can be distinguished.

- Continuous case: In the case that $a_{s,j} \neq b_{s,j}$ for all j, in each s, $\mathbf{Y}$ is an absolutely continuous random vector and the probability density function $g : \mathbb{R}^d \to [0, \infty)$ can be expressed as

$$g(t) = \sum_{s=1}^{S} p_s \mathbb{1}\left\{t \in Q_s\right\} \frac{1}{\lambda(Q_s)}, \tag{5}$$

  where $\lambda(Q_s) = \prod_{j=1}^{d}(b_{s,j} - a_{s,j})$ denotes the d-volume of Q_s.
- Discrete case: If $a_{s,j} = b_{s,j}$ for all j, in each s, the hyperrectangles are points, and $\mathbf{Y}$ is a discrete random vector, whose probability mass function can be represented by

$$\mathbb{P}(\mathbf{Y} = t) = \begin{cases} p_s & \text{if } t = Q_s, \\ 0 & \text{otherwise.} \end{cases} \tag{6}$$

In the situation where only some, but not all hyperrectangles Q_s consist of points, $\mathbf{Y}$ is a mixed random vector. When $a_{s,j} = b_{s,j}$ for some s, the probability distribution does not have a density function.

Example 2.5. Figure 2 shows the cumulative distribution function G of a PWC distributed random vector for the continuous and the discrete cases.

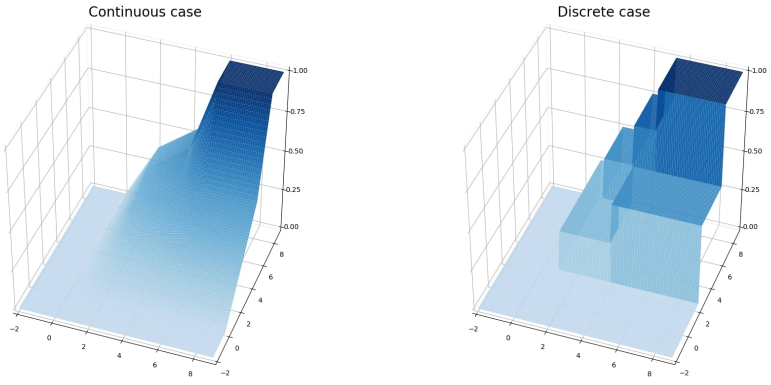

Figure 2. Cumulative distribution function of a two-dimensional PWC distributed random vector in the continuous case (plot on the left) and in the discrete case (plot on the right).

Definition 2.6. A PWC distribution is said to be *compatible* with the empirical distribution of $(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)$, if the following conditions hold:

a) Exist s such that $\mathbf{X}_i \in Q_s$, for all i,

b) $p_s = \dfrac{n_s}{n} = \dfrac{1}{n} \sum_{i=1}^{n} \mathbb{1}\left\{\mathbf{X}_i \in Q_s\right\}$, for all s.

The above requirements represent the *compatibility condition* between $\mathbf{Y} \sim \text{PWC}(p, \mathcal{Q})$ and $\mathbf{X} \sim \hat{F}$, i.e. for all s, $\mathbb{P}(\mathbf{X} \in Q_s) = \mathbb{P}(\mathbf{Y} \in Q_s)$, since p_s counts the elements of the sample that lie in Q_s.

Setting all p_s according to Definition 2.6 makes the PWC distribution a d-dimensional histogram [9].

Example 2.7. Figure 3 outlines the compatibility condition in $\mathbb{R}^2$ between $\mathbf{Y} \sim \text{PWC}(p, \mathcal{Q})$ and $\mathbf{X} \sim \hat{F}$, with $n = 12$. The former is a continuous random vector and bluish shaded areas indicate rectangles Q_s, $s = 1, ..., S$, where the density is positive; the latter has discrete support and its realizations are denoted by blue dots. Plot a): compatible. Plot b): not compatible, since there is i such that $\mathbf{x}_i \notin Q_s$ for all s, in fact two of the sample realizations lie within no hyperrectangle. Plot c): not compatible, since for example $p_4 \neq \dfrac{1}{n} \sum_{i=1}^{n} \mathbb{1}\left\{\mathbf{x}_i \in Q_4\right\} = 0$; according to the compatibility condition, given that none of the sample realizations is in rectangle Q_4, $\mathbf{Y}$ density should be equal to 0 in there.

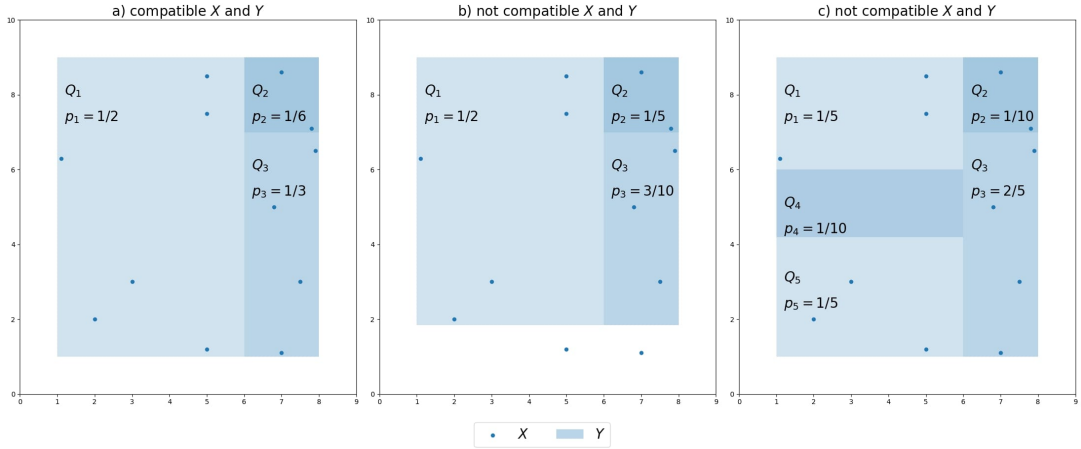

Figure 3. Illustration of the compatibility condition in $\mathbb{R}^2$ for a sample of $n = 12$ data points. In plot a) the PWC distribution is compatible with the observed sample. Conversely, in plot b) and plot c) the PWC distributions are not compatible.

Intuitively, a distribution is said to be PWC if it can be defined on a set of hyper-rectangles, on which the probability is constant.

Remark 2.8. It can be noted that if $T : \Omega \to \{1, 2, ..., S\}$ and $\mathbf{U}_s : \Omega \to Q_s$ are independent random variables, with $\mathbb{P}(T = s) = p_s$ and $\mathbf{U}_s \sim U_{Q_s}$, i.e. the (continuous) uniform random vector on Q_s, for $s = 1, 2, ..., S$, then $\mathbf{Y} \sim \text{PWC}(p, \mathcal{Q})$ can be considered as a mixture of uniform distributions:

$$\mathbf{Y} \overset{\text{d}}{=} \sum_{s=1}^{S} \mathbb{1}\{T = s\} \mathbf{U}_s, \tag{7}$$

where the mixing weights are $(p_s : s = 1, ..., S)$, above-mentioned. The representation of Formula (7) emphasizes that, conditioned on Q_s, $\mathbf{Y}$ is a uniform random vector on Q_s, i.e. $\mathbb{P}(\mathbf{Y} = y | \mathbf{Y} \in Q_s) = \mathbb{P}(\mathbf{U}_s = y)$.

This representation plays a major role in the design of our algorithm and points out a useful property of the class of PWC distributions: its moments can be computed analytically.

Lemma 2.9. *Let $\mathbf{Y} \sim PWC(p, \mathcal{Q})$ and $k > 0$, then the kth raw moment $\mathbb{E}\left[\mathbf{Y}^k\right]$ is expressed by:*

$$\begin{aligned} \mathbb{E}\left[\mathbf{Y}^k\right] &= \left(\mathbb{E}\left[Y_j^k\right] : j = 1, ..., d\right), \\ \mathbb{E}\left[Y_j^k\right] &= \sum_{s=1}^{S} \frac{p_s}{k+1} \frac{b_{s,j}^{k+1} - a_{s,j}^{k+1}}{b_{s,j} - a_{s,j}}, \end{aligned} \tag{8}$$

where $j = 1, ..., d$.

Proof. The result derives from the fact that Y_j can be considered as a mixture of

uniform distributions with mixing weights $(p_s : s = 1, ..., S)$:

$$\mathbb{E}\left[Y_j^k\right] = \mathbb{E}\left[\sum_{s=1}^{S} \mathbb{1}\left\{T = s\right\} U_{s,j}^k\right] = \mathbb{E}\left[\mathbb{E}\left[U_{s,j}|T\right]\right] = \sum_{s=1}^{S} \frac{p_s}{k+1} \frac{b_{s,j}^{k+1} - a_{s,j}^{k+1}}{b_{s,j} - a_{s,j}}, \quad (9)$$

where $U_{s,j} \sim U_{I_{s,j}}$. In fact, the kth raw moment for a convex combination of distributions is the convex combination of the kth raw moments, provided that they exist, of the component distributions. □

Lemma 2.10. *From Lemma 2.9 it can be noted that the expected value μ_j and the variance of Y_j boil down to:*

$$\mu_j = \sum_{s=1}^{S} p_s \frac{a_{s,j} + b_{s,j}}{2}, \quad (10)$$

and

$$Var(Y_j) = \mathbb{E}\left[Y_j^2\right] - \mu_j^2 = \sum_{s=1}^{S} p_s \frac{a_{s,j}^2 + b_{s,j}^2 + a_{s,j} b_{s,j}}{3} - \left(\sum_{s=1}^{S} p_s \frac{b_{s,j} + a_{s,j}}{2}\right)^2. \quad (11)$$

Lemma 2.11. *Furthermore, the covariance between the ith and jth marginals, is equal to:*

$$Cov(Y_i, Y_j) = \sum_{s=1}^{S} p_s \frac{(a_{s,i} + b_{s,i})(a_{s,j} + b_{s,j})}{4} - \sum_{s=1}^{S} p_s \frac{(a_{s,i} + b_{s,i})}{2} \sum_{s=1}^{S} p_s \frac{(a_{s,j} + b_{s,j})}{2}, \quad (12)$$

which shows that $\boldsymbol{Y}$ has a dependence structure, even though it is constructed from independent components.

Proof. The result derives from the so-called law of total covariance:

$$\begin{aligned} Cov(Y_i, Y_j) &= \mathbb{E}\left[Cov(Y_i, Y_j|T)\right] + Cov\left(\mathbb{E}\left[Y_i|T\right], \mathbb{E}\left[Y_j|T\right]\right) \\ &= \mathbb{E}\left[\mathbb{E}\left[Y_i|T\right]\mathbb{E}\left[Y_j|T\right]\right] - \mathbb{E}\left[\mathbb{E}\left[Y_i|T\right]\right]\mathbb{E}\left[\mathbb{E}\left[Y_j|T\right]\right] \\ &= \sum_{s=1}^{S} p_s \frac{(a_{s,i} + b_{s,i})(a_{s,j} + b_{s,j})}{4} - \sum_{s=1}^{S} p_s \frac{(a_{s,i} + b_{s,i})}{2} \sum_{s=1}^{S} p_s \frac{(a_{s,j} + b_{s,j})}{2}, \end{aligned} \quad (13)$$

where the term $\mathbb{E}\left[Cov(Y_i, Y_j|T)\right]$ is equal to zero because the component distributions have independent marginals. □

3. Wasserstein distance

We now introduce the Wasserstein distance, a function defined on a given metric space that allows us to quantify the proximity between two probability distributions. This distance function is the basis of the methodology that determines when a PWC distribution is an admissible approximation of an empirical distribution (see below).

Definition 3.1. Given a metric space $(\mathbb{R}^d, \mathrm{c})$, with metric c, for any two probability measures F, G on $\mathbb{R}^d$, the Wasserstein distance between F and G is defined by:

$$W(F,G) = \inf_{X\sim F,\, Y\sim G} \mathbb{E}\left[\mathrm{c}(X,Y)\right]. \tag{14}$$

The Wasserstein distance is thus the minimum expected distance among all pairs of random variables X and Y whose fixed marginal distributions are F and G respectively. Minimizers are called optimal transport plans or optimal couplings. For the sake of completeness we mention that Definition 3.1 holds true for Polish metric spaces and can be extended to the Wasserstein distance of order $q \in [1, \infty)$. Here we focus on the Wasserstein distance of order 1, given by Formula (14), and we restrict ourselves to the cases where c is a ℓ_p-norm, i.e. $\mathrm{c}(x,y) = \left(\sum_{j=1}^{d}(x_j - y_j)^p\right)^{1/p}$, due to its natural interpretation as the earth mover's distance [13], and implementation simplifications. An exhaustive dissertation on the topic, containing also some historical context, can be found in [14].

Definition 3.2. It can be shown that Formula (14) has the following dual expression:

$$W(F,G) = \sup_{\psi\in\Psi}\left\{\int \psi(x)dF(x) - \int \psi(x)dG(x)\right\}, \tag{15}$$

where Ψ denotes the set of all functions $\psi : \mathbb{R}^d \to \mathbb{R}$, such that $|\psi(y) - \psi(x)| \leq \mathrm{c}(x,y)$.

Equations (14) and (15) are equivalent (see Remark 6.5. in [14]).

3.1. Admissibility criteria

Admissibility criteria specify the conditions required for a PWC distribution to be an admissible approximation of $\hat{F}$. These rules are consistent with the essential assumption of piecewise constant distributions that their distribution conditioned on each hyperrectangle is a uniform, and arise from the fact that a uniform distribution on a hyperrectangle has two peculiar attributes: it has uniform marginals and these are mutually independent.

3.1.1. Marginal admissible approximation

Definition 3.3. Given a hyperrectangle Q_s, let $F_{s,j}$ and $\hat{F}_{s,j}$ denote, respectively, the cumulative distribution function of X_j in $I_{s,j}$, and the empirical cumulative distribution function of the sample projection on $I_{s,j}$. We define the null hypothesis

$$\mathbb{H}_0^{*\,s,j} : F_{s,j} = U_{I_{s,j}}, \tag{16}$$

and the test statistic

$$W(\hat{F}_{s,j}, U_{I_{s,j}}), \tag{17}$$

which is the Wasserstein distance in Q_s between the sample projection on the jth dimension, i.e the jth marginal, and the uniform density on $I_{s,j}$.

The above hypothesis test is aimed at verifying, using a Wasserstein distance based test statistic, that the jth margin of the sample contained in Q_s is uniformly spread over $I_{s,j}$.

Definition 3.4. We define $\mathbf{Y} \sim \text{PWC}(p, \mathcal{Q})$ to be a *marginal admissible approximation* of $\hat{F}$, if it is compatible and none of the null hypotheses $\left\{ {\mathbb{H}_0^*}^{s,j} : s = 1, ..., S, j = 1, ..., d \right\}$ is rejected. This means that with a significance level $\alpha \in [0, 1]$, for all s and j:

$$\mathbb{P}\left(W(\hat{F}_{s,j}, U_{I_{s,j}}) > w_{s,j} \mid {\mathbb{H}_0^*}^{s,j} \right) > \alpha \tag{18}$$

where $w_{s,j}$ denotes the observed value of the test statistic in $I_{s,j}$.

In other terms, Definition 3.4 states that $\mathbf{Y} \sim \text{PWC}(p, \mathcal{Q})$ is a marginal admissible approximation of $\hat{F}$ when $\hat{F}_{s,j}$ cannot be distinguished in a statistically significant manner from a uniform distribution on $I_{s,j}$, with a predefined significance level α, in each and every hyperrectangle Q_s and dimension j.

When $\mathbf{Y} \sim \text{PWC}(p, \mathcal{Q})$ is a marginal admissible approximation of $\hat{F}$, the first distinctive feature of uniform distributions is met: in each hyperrectangle, the margins of the sample contained in it do not significantly differ from the uniform distribution.

We introduce the following hypothesis test in addition to the one above.

3.1.2. Admissible approximation

Definition 3.5. Given a hyperrectangle Q_s, consider the random vector $\widetilde{\mathbf{U}}_s : Q_s \to C$, where $C = [0, 1]^d$, as the following transformation of $\mathbf{X}$ in Q_s:

$$\widetilde{\mathbf{U}}_s = \left(H_{I_{s,1}}(X_1), H_{I_{s,2}}(X_2), ..., H_{I_{s,d}}(X_d) \right). \tag{19}$$

Furthermore, let F_s and $\hat{F}_s$ denote, respectively, the cumulative distribution function of $\widetilde{\mathbf{U}}_s$ and the empirical cumulative distribution function of the transformed sample. We define the null hypothesis

$$\mathbb{H}_0^s : F_s = U_C, \tag{20}$$

and the test statistic

$$W(\hat{F}_s, U_C), \tag{21}$$

which is the Wasserstein distance between the transformed sample and the uniform density on set C.

The hypothesis test introduced by Definition 3.5 is aimed at verifying, using a Wasserstein distance based test statistic, that the transformed sample is uniformly spread over C.

Definition 3.6. We define $\mathbf{Y} \sim \text{PWC}(p, \mathcal{Q})$ to be an *admissible approximation* of $\hat{F}$, if it is a marginal admissible approximation, and none of the null hypotheses $\{\mathbb{H}_0^s : s = 1, ..., S\}$ is rejected. This means that with a significance level $\alpha \in [0, 1]$,

for all s:

$$\mathbb{P}\left(W(\hat{F}_s, U_C) > w_s \mid \mathbb{H}_0^s\right) > \alpha \tag{22}$$

where w_s denotes the observed value of the test statistic in Q_s.

In other terms, Definition 3.6 states that $\mathbf{Y} \sim \text{PWC}(p, \mathcal{Q})$ is an admissible approximation of $\hat{F}$ when the transformed sample cannot be distinguished in a statistically significant manner from a uniform distribution on C, with a predefined significance level α, in each and every hyperrectangle.

According to Definition 3.6, the marginal admissibility condition is required for a piecewise constant distribution to be also an admissible approximation of $\hat{F}$. Given this fact, in each Q_s, it stands to reason that sample margins are uniformly distributed on $I_{s,j}$, for all j, and consequently, by applying the transformation $H_{I_{s,j}}$, they are uniform on $[0, 1]$. Hence, the joint uniformity assumption can be tested against the transformed sample, which lies in C and retains the dependence structure of the original sample.

Intuitively, we want to detect distributions with uniform marginals at first, and secondarily exclude, between these, those that have not a joint uniform distribution. Evidently, not all distributions with uniform marginals have mutually independent marginals, and not all distributions with mutually independent marginals have uniform marginals.

Example 3.7. Figure 4 contrasts the marginal admissible and the admissible conditions in $\mathbb{R}^2$. $\mathbf{X}$ realizations are denoted by blue dots and the bluish shaded square indicates the support of $\mathbf{Y} \sim \text{PWC}(p, \mathcal{Q})$. Plot a) displays a sample uniformly spread out over the square. Conversely, plots b) and c) depict samples whose margins are close to uniforms, but joint distributions are not. In these cases, the PWC distribution is only a marginal admissible approximation of the empirical distribution.

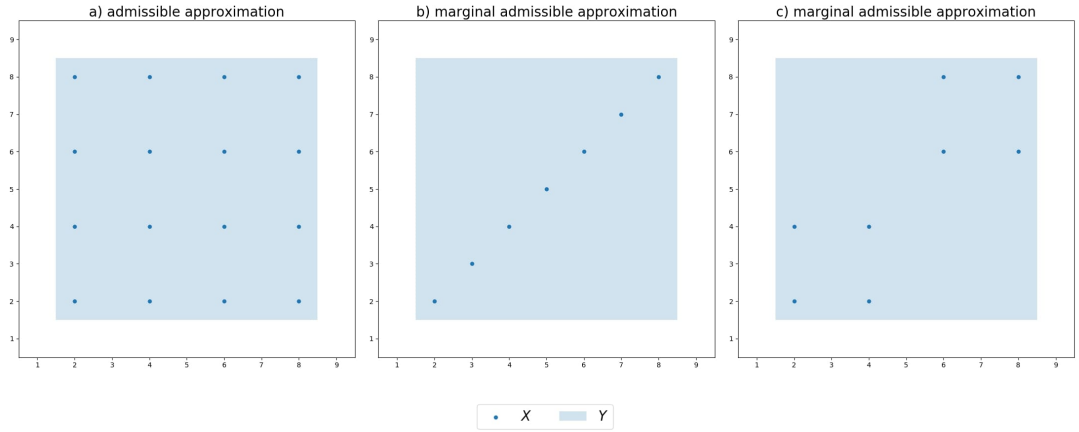

Figure 4. Illustration of the admissibility conditions in $\mathbb{R}^2$. Plot a) represents an admissible PWC distribution, whereas plots b) and c) show PWC distributions which are only marginal admissible.

Finally, it has to be noted that testing the marginal admissibility condition in a particular Q_s involves a set of statistical inferences simultaneously, and hence the multiple comparisons or multiple testing problem occurs, i.e. the more inferences are made, the more likely erroneous inferences are to occur. Different procedures are available in

statistics for adjusting p-values and controlling the so-called Family Wise Error Rate (FWER), namely the probability of at least one false positive (type I error). Whilst on one hand this approach seems appropriate, since a single significant p-value across $\mathbb{H}_0^{*\,s,j}$ establishes that Q_s is not marginal admissible, on the other hand it would increase the probability of false negative (type II error). In our context, we consider more important to ensure that non-uniformity is identified, rather than to limit uniformity to not being detected.

3.2. Wasserstein distance hypothesis testing

The Wasserstein distance has been receiving increasing attention from the research community and has found different utilization in statistics, including clustering [15] and PCA [16]. Nevertheless, practical applications remain tentative because its numerical calculation is very arduous, especially when $d > 1$: explicit coupling results are only known for multivariate Gaussian and elliptic distributions [17]. Recent publications have proposed a wide range of approaches to find efficient solvers that address the Wasserstein distance computation problem. When $d = 1$, by contrast, the Wasserstein distance has a closed-form expression and is easier to handle.

In the algorithm we propose, for practical purposes, we need an efficient and agile scheme: the overall computation can be extremely demanding, considering both the complexity of the Wasserstein distance calculation in itself, and the fact that it may need to be determined possibly numerous times. A decisive factor, in this respect, is that we verify the marginal admissibility condition at first, and the (joint) admissibility condition only after the former is already met. Hence, during the initial phase of the algorithm, we deal with Wassertein distance between one-dimensional distributions: this aspect considerably lightens the load of the calculation, especially when the number S of partitioning hyperrectangles is still limited and large samples could be situated within these. The algorithm successively moves to the hypothesis tests of Definition 3.5 regarding the joint uniformity when $\mathrm{PWC}(p, \mathcal{Q})$ is already a marginal admissible approximation and the initial sample should be sufficiently partitioned into smaller datasets within each bucket.

When testing $\mathbb{H}_0^{*\,s,j}$ and $\mathbb{H}_0^s$ hypotheses, the distribution of the test statistic under the null hypothesis entails the so-called empirical Wasserstein distance $W(\hat{F}, F)$, i.e. the distance between the empirical measure $\hat{F}$ of a sample drawn from F and F itself, and especially the case where F is a uniform density. It is a well-known consequence of the strong law of large numbers that if the first moment of F is finite, then $W(\hat{F}, F)$ converges to 0, almost surely, as the sample size approaches infinity (see [14], Cor. 6.11). However, getting hypothesis tests based on the (empirical) Wasserstein distance is severely hampered by a lack of inferential tools. Determining the exact rate of convergence and distributional limits, which give a genuine perspective for practicable inference, is the subject of a large body of literature, but despite the considerable interests in the topic, results have remained elusive and the problem of constructing confidence intervals for the Wasserstein distance is in general unsolved.

3.2.1. One-dimensional setting

In the one-dimensional setting, when c is any ℓ_p-norm, the optimal coupling attaining the minimum in (14) is known explicitly [18], and the Wasserstein distance can be

expressed in the following form:

$$W(F,G) = \int_0^1 |F^{-1}(t) - G^{-1}(t)|dt = \int_{-\infty}^{\infty} |F(x) - G(x)|dx, \tag{23}$$

where the last equality is obtained by the Fubini-Tonelli Theorem. The formula shows that the Wasserstein distance corresponds to the area between the two quantile functions or, equivalently, to the area between the two cumulative distribution functions of the two random variables.

In addition, for measures on $\mathbb{R}$, a rather complete theory regarding rates of convergence and distributional limit is available [19], and the following result applies for testing the marginal admissibility condition.

Theorem 3.8. *Consider a sequence of n independent random variables uniformly distributed on $I = (a, b]$, $a < b \in \mathbb{R}$, then as $n \to \infty$:*

$$\sqrt{n}\ \frac{W(\hat{F}, U_I)}{(b-a)} \xrightarrow{d} \int_0^1 |B(t)|dt \tag{24}$$

where $B(t), 0 \leq t \leq 1$ denotes a Brownian bridge process, that is, a centred Gaussian process with continuous sample paths and covariance $\mathbb{E}[B(s)B(t)] = \min\{s,t\} - st$.

Proof. The statement follows from Theorem 1.1 in [19] by substituting the quantile function associated to U_I and applying the scaling property of the Wasserstein distance [20], i.e. $W(\epsilon X, \epsilon Y) = |\epsilon| W(X,Y)$ for any scale $\epsilon \in \mathbb{R}$ and random variables X and Y. Here, for the sake of simplicity, we used random variables within the Wasserstein distance expression, in contrast with the notation of the paper. □

Theorem 3.8 provides us with a theoretical instrument for testing any $\mathbb{H}_0^{*\,s,j}$ hypothesis. The observed value of the test statistic, to compare with the critical one, is derived from the Wasserstein Distance between a uniform distribution on $I_{s,j}$ and the respective sample margin. In this circumstance, Formula (23) can be represented through a specific closed-form expression that can be derived from Theorem 3.3 in [12]. Given a sample $X \sim \hat{F}$ of size n, such that $a \leq X_{(1)} \leq ... \leq X_{(n)} \leq b$, and a uniform random variable on $I = (a, b]$, then $W(\hat{F}, U_I)$ is equal to:

$$W(\hat{F}, U_I) = \sum_{i=1}^{n+1} W_i^* = \sum_{i=1}^{n+1} \left[\overline{W}_i + \frac{1}{2} \left(\beta_i - \frac{\overline{W}_i}{\delta_i} \right) \cdot \max \left\{ \delta_i - \frac{\overline{W}_i}{\beta_i}, 0 \right\} \right], \tag{25}$$

where

$$\begin{aligned} \overline{W}_i &= \beta_i \left| \frac{i-1}{n} - \frac{X_{(i-1/2)} - a}{b-a} \right|, \\ \beta_i &= X_{(i)} - X_{(i-1)}, \\ \delta_i &= \frac{X_{(i-1/2)} - X_{(i-1)}}{b-a}, \end{aligned} \tag{26}$$

and $X_{(i)}, i = 1, ..., n$, are the order statistics of the sample. Note that we set $X_{(0)} = a$, $X_{(n+1)} = b$ and $X_{(i-1/2)} = (X_{(i)} + X_{(i-1)})/2$.

We have therefore the analytical tools required to check the marginal admissible condition.

Algorithm 3.9. Marginal admissible hypothesis testing.
For each hyperrectangle Q_s where the marginal admissibility condition is checked:

(1) For each dimension $j = 1, ..., d$, compute the observed test statistic $w_{s,j}$ and compare it with the respective α-level critical value w_α using Theorem 3.8 and Formula (25).
(2) If any $w_{s,j} \geq w_\alpha$, ${\mathbb{H}_0^*}^{s,j}$ is rejected, split Q_s and go back to (1), else the sample is considered to have uniform marginals.

3.2.2. Multidimensional setting

When testing $\mathbb{H}_0^s$ hypotheses we are in a multidimensional regime. However, as mentioned above, inferential tools for Wasserstein distances are elusive when $d \geq 2$, and hence complications arise for checking the joint admissibility condition. For measures on $\mathbb{R}^d$, there are, indeed, only few distributional results, none of which can be successfully employed in our framework (see e.g. [21,22]). With regard to the question of quantifying the rate of convergence of $W(\hat{F}, F)$, the major findings regarding our context are given by the works of [23], who considers the uniform distribution on the unit square, and [24,25], for the uniform distribution in higher dimensions on a d-dimensional unit cube.

The solution we propose to obtain critical values of the test statistic distribution under $\mathbb{H}_0^s$ is meant to guarantee a feasible implementation of the algorithm. It would be, in fact, possible to test the joint admissibility condition by obtaining, via simulation, an approximated distribution of the test statistic under the null hypothesis for each hyperrectangle Q_s. However, according to the authors, this *modus operandi* comes at too high a (computational) cost, which can undermine the whole scheme workability, especially when n_s is not a small value. The idea we propose combines two components: a reference simulated distribution of the test statistic under the null hypothesis and the results of rate of convergence for empirical Wasserstein distances concerning uniform densities. The procedure steps are detailed below.

Algorithm 3.10. Reference test statistic critical value. Before the algorithm is initiated, for $\alpha \in [0, 1]$ and $d \in \mathbb{Z}^+$:

(1) Draw N samples of size m from the uniform distribution on $C = [0, 1]^d$.
(2) Compute the Wasserstein distances $\mathcal{W} = (w_1, w_2, \ldots, w_N)$ between each sample and U_C.
(3) Determine the critical value w_α as the $(1 - \alpha)$-level empirical quantile of $\mathcal{W}$.

Algorithm 3.11. Admissible hypothesis testing. For each hyperrectangle Q_s where the admissibility condition is checked:

(1) Compute the observed test statistic w_s and scale the value with the appropriate order of convergence of the empirical Wasserstein distance (see below) for considering the actual number of data points n_s lying in Q_s.
(2) Compare the value w_s thus obtained with w_α; if $w_s \geq w_\alpha$, $\mathbb{H}_0^s$ is rejected and Q_s has to be split, else the sample is considered as uniform on Q_s.

With the above steps, an approximation of the α-level critical value of the test statistic under $\mathbb{H}_0^s$ is obtained, without simulating for each hyperrectangle sample size

n_s the test statistic distribution under the null hypothesis.

A heuristic rule for setting the sample size m of Algorithm 3.10 is $m = \lfloor \log n \rfloor$, where $\lfloor \cdot \rfloor$ is the floor function and log the natural logarithm. The reason for this is that m should be an arbitrary positive integer that allows for a relatively fast computation of the simulated distribution of the test statistic. The number of simulation N is determined in such a way that there is at least 90% confidence that the estimated quantile does not differ by more than 1% from the true value (see [26] Section 5.2).

The order of convergence adjustments, previously disclosed, are provided by [23] and [25]. In particular, for a uniform random vector defined on $C = [0, 1]^d$, where $d \geq 2$, the limiting behaviour of the empirical Wasserstein distance is given by

$$W(\hat{F}, U_C) = \begin{cases} \mathcal{O}(n^{-1/2} \log n^{1/2}), & \text{if } d = 2, \\ \mathcal{O}(n^{-1/d}), & \text{if } d > 2, \end{cases} \tag{27}$$

where $\mathcal{O}$ is the Big O(micron) notation and n is the sample size. The authors have confirmed through a simulation study the reliability of the above-mentioned approach.

It can be noted that, as the algorithm proceeds, the power of the hypothesis tests tends to diminish, i.e. the number of true positive correct inferences reduces. The reason for this is that the partitioning set $\mathcal{Q}$ grows in size and consequently the sample size n_s in its every element affected by the bisection technique decreases. This situation leads to a design less susceptible to overfitting, since makes it more likely for the algorithm to stop partitioning hyperrectangles.

With regard to the Wasserstein distances in Algorithm 3.10 and Algorithm 3.11, we make the computation a discrete problem: the uniform density is approximated with a quasi-random low-discrepancy sequence to evenly cover the d-dimensional hypercubic space. The number of elements in the sequence scales with d and is set equal to 10^d. When both measures involved are discretized (finite weighted sums of Dirac masses), the Wasserstein distance formulation fits into a discrete setting and its calculation becomes a Linear Assignment Problem [27]. The reason for modifying the original problem is the flexibility of discrete solvers: these consist of combinatorial optimization algorithms, based on the linear programming formulation, that work for any dimension d and for almost any ground metric c [28]. In order to accelerate this type of solvers, the entropic-regularized approach [29] adds an entropic penalization to the original optimal transport formulation and has been shown to be extremely efficient to approximate Wasserstein distances at a low computational cost. In this regard, one of the most-used method for solving the resulting regularized optimization problem is represented by the Sinkhorn algorithm and its recent refinements [30,31]. We adopted this approach, which has a complexity of $\mathcal{O}(n^2 \log n)$, i.e. in nearly linear time in the input size n^2 [30], for approximating the optimal transportation distance.

Example 3.12. Figure 5 illustrates approximation of the Wasserstein distance in $\mathbb{R}^2$, with ℓ_2-norm as ground metric, between a sample of size $n = 5$ and the uniform distribution on $[0, 1]^2$. The latter is represented with a quasi-random low-discrepancy sequence that covers the square area and is indicated with black dots. The approximated value of the Wasserstein distance is represented by the average length of the azure lines mapping the observations to the target points.

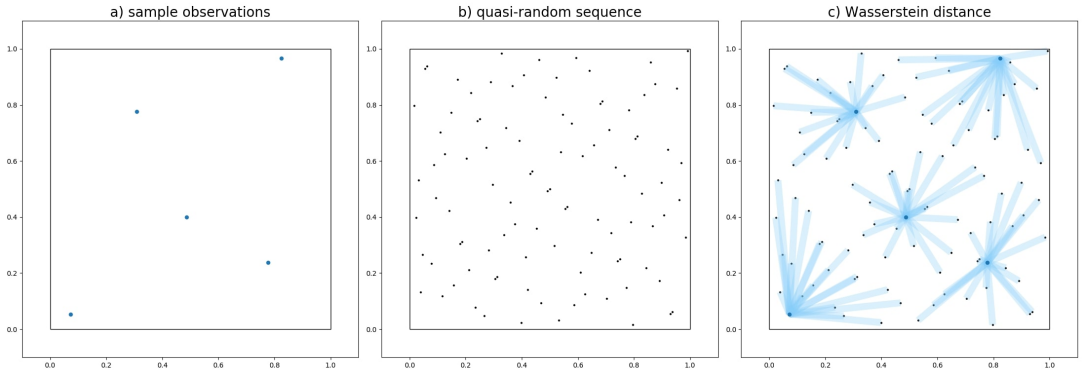

Figure 5. Illustration of the approximated Wasserstein distance between a sample of size $n = 5$ and the uniform distribution on the unit square, when the ground metric is the ℓ_2-norm. Plots a) and b) highlight the sample and the quasi-random low-discrepancy sequence serving as the uniform distribution, respectively. The optimal coupling associated to the (approximated) Wasserstein distance is displayed in plot c).

4. Algorithm

For obtaining an admissible PWC approximation of the sample we opted for a top-down algorithm with recursive layout that starts with a single axis-aligned hyperrectangle enclosing the entire observations, and builds a hierarchical partition by splitting an existing hyperrectangle into two non-overlapping ones. The recursive partitioning is repeated until the admissibility condition is met in each region of the partition.

The dimension along which the split is executed is chosen according to a specific bisection technique. As each bisection concerns only a single dimension, the regions in the resulting partition always have axis-parallel boundaries. In addition, since the bisection of each hyperrectangle is independent from the bisection of the other partition elements, the algorithm enables a high degree of parallelism.

Finally, it should be noted that the algorithm is compatible with the divide-and-conquer design paradigm: it works by repeatedly breaking down a problem into subproblems of the same type, until all of these come to a halt.

4.1. Initialization

The algorithm starts with a single box whose sides are parallel to the d coordinate axes and containing all the observations, and hence the piecewise constant distribution simply coincides with a uniform density.

Algorithm 4.1. Initialization procedure.

(1) Set $S = 1$, $\mathcal{Q} = \{Q_1\}, p = (1), \mathrm{PWC}(p, \mathcal{Q}) = U_{Q_1}$. Where

$$Q_1 = [\hat{\mathtt{a}}_1, \hat{\mathtt{b}}_1] \times [\hat{\mathtt{a}}_2, \hat{\mathtt{b}}_2] \times ... \times [\hat{\mathtt{a}}_d, \hat{\mathtt{b}}_d],$$

and $\hat{\mathtt{a}}_j$, $\hat{\mathtt{b}}_j$, for $j = 1, ..., d$ are given by

$$\begin{aligned} \hat{\mathtt{a}}_j &= \frac{nX_{(1)} - X_{(n)}}{n-1} = X_{(1)} - \frac{X_{(n)} - X_{(1)}}{n-1}, \\ \hat{\mathtt{b}}_j &= \frac{nX_{(n)} - X_{(1)}}{n-1} = X_{(n)} + \frac{X_{(n)} - X_{(1)}}{n-1}. \end{aligned} \tag{28}$$

$X_{(1)}$ and $X_{(n)}$ are, respectively, the first and last order statistics of the sample jth marginal.

Formula (28) represents the minimum-variance unbiased estimators for the two parameters $\mathtt{a}_j$ and $\mathtt{b}_j$ of a uniform on $[\mathtt{a}_j, \mathtt{b}_j]$ [32].

Thereafter, during algorithm iterations, the partition (and the PWC distribution consequently) is grown by splitting each partition member into two sub-hyperrectangles, until the stopping condition is met and PWC$(p, \mathcal{Q})$ is an admissible approximation of $\hat{F}$. The choice of using hyperrectangular shaped buckets is also driven by the data compression intent of the algorithm: this type of shape can be represented concisely, allowing a large number of buckets to be stored efficiently.

4.2. Bisection technique

The aim of the bisection technique is to build and shape the PWC distribution. The bisection scheme selects in each hyperrectangle Q_s, where either $\mathbb{H}_0^s$ null hypothesis or at least one of $\mathbb{H}_0^{*\,s,j}$ null hypotheses is rejected, the dimension to split and the relative split point. More specifically, it operates as follows.

Algorithm 4.2. Bisection technique.

(1) **Dimension selection**. In a given hyperrectangle Q_s to bisect, split the dimension in which the Wasserstein distance between $\hat{F}_{s,j}$ and $U_{I_{s,j}}$, i.e. the jth marginal distribution of the sample in Q_s and the uniform density on jth dimension range respectively, is associated to the smallest p-value. Namely:

$$j^* = \underset{j \in \{1,2,...,d\}}{\operatorname{argmax}} \sqrt{n_s}\ \frac{W(\hat{F}_{s,j}, U_{I_{s,j}})}{(b_{s,j} - a_{s,j})}. \tag{29}$$

(2) **Split point selection and bisection**. For the selected dimension j^*, let the index k denote:

$$k = \underset{i \in \{1,2,...,n+1\}}{\operatorname{argmax}} W_i^*, \tag{30}$$

where W_i^* is defined in Formula (25). Bisect at $(X_{(k)} + X_{(k-1)})/2$, where $X_{(k)}$ is the marginal k th-order statistic of the sample. By implication, the initial hyperrectangle Q_s is split in two non-overlapping hyperrectangles and S is augmented by one.

Both steps of the bisection technique relies on Formula (25). In particular, the split point selection phase uses the fact that above-mentioned formula is composed by $n + 1$ areas, each of which measures the vertical difference, occurring between two consecutive data points.

Example 4.3. Figure 6 exemplifies how the bisection procedure works. Each group of three plots depicts $\mathbf{X} \sim \hat{F}$ and $\mathbf{Y} \sim \text{PWC}(p, \mathcal{Q})$ random variables in a given Q_s (lower-left graph), and their marginal cumulative distribution functions (upper and lower-right graphs). In the right group, W_k^* and the selected split are highlighted.

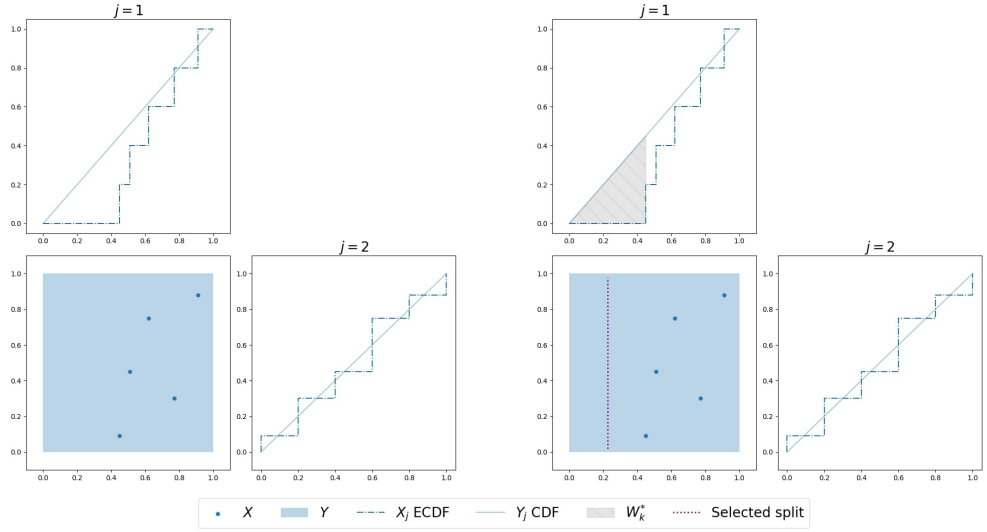

Figure 6. Illustration of the way in which bisection technique operates in $\mathbb{R}^2$. The algorithm is able to detect the dimension that is the most in need of partitioning ($j = 1$) and bisect the area where the largest vertical discrepancy from uniformity occurs.

4.3. Full Algorithm

In this section, after having presented the details of all scheme components, we resume the full algorithm.

Algorithm 4.4. Complete Algorithm.
Require: $\alpha \in [0, 1]$, ℓ_p-norm.
Input: an observed sample $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)$ of a d-dimensional random vector.
Output: an admissible PWC distribution.

(1) **Reference critical value:** Determine the test statistic α-level critical value using Algorithm 3.10.
(2) **Initialize:** Start the PWC distribution as stated by Algorithm 4.1.
(3) **First step:** Test the marginal admissible condition using Algorithm 3.9. Bisect $\mathcal{Q}$ elements that require it, according to Algorithm 4.2, until a marginally admissible PWC$(p, \mathcal{Q})$ is found.
(4) **Second step:** Test the admissible condition using Algorithm 3.11. Further bisect $\mathcal{Q}$ elements that require it, according to Algorithm 4.2, until an admissible PWC$(p, \mathcal{Q})$ is found.

Definition 4.5. An admissible PWC distribution resulting from Algorithm 4.4 is said to be a PWC estimator of the unknown probability distribution of the observed sample.

Lemma 4.6. *Given an empirical distribution $\hat{F}$ of a sample of size n, for a PWC*

estimator G, it holds that:

$$\lim_{\alpha \to 1} G(t) = \hat{F}(t). \tag{31}$$

Proof. As $\alpha \to 1$, by the definition of significance level, the probability of rejecting the null hypothesis approaches one. Therefore, from the algorithm construction, the partition grows until only a single point x_i is left in each hyperrectangle Q_i, for $i = 1, ..., n$. At this stage, in any Q_i, the null hypothesis is still rejected and the bisection technique continues to shrink all intervals $I_{i,j}$, $j = 1, ..., d$, by alternately splitting at either $(a_{i,j} + x_j)/2$ or $(x_j + b_{i,j})/2$. As a result, the non-empty hyperrectangles are reduced to the observations. □

5. Implementation and illustrations

The performance of our methodology has been investigated on datasets in two-, three- and nine-dimensional spaces. All experiments and analyses were run on a computer with an Intel® Core™ i7-6700HQ processor with 16GB RAM, running at 809.549 MHz, on Ubuntu Linux distribution version 18.04.2. An implementation of our algorithm in Python is available under the permissive free software MIT license. It can be obtained through the authors.

5.1. Two-dimensional space

In the first instance, we evaluate our methodology against a non-trivial two-dimensional distribution. In the literature, the same bivariate distribution has been originally adopted by [33] for the testing of their work. The reader should refer to the original paper for a more detailed explanation.

The distribution is defined as a random sample from any of 350 Gaussian distributions. The first 349 normal distributions are sampled with equal probability of 1/698 and have all variances of 0.3 and no covariance. The last distribution is also a normal distribution, but it is sampled with probability 349/698 and its dispersion matrix is defined by the covariance matrix of the means of the above-mentioned 349 distributions. The first 349 normal distributions are located in a manner that reproduces the (rotated) logo of the original paper lead author's home institution. The last component is centred on the origin, with a width and height that traverse the other component distributions and with principal axes parallel to the x-y axes. This produces a mixture distribution with a long-wave feature combined with a sophisticated structure of comparatively shortwave elements aligned with different axes.

Figure 7 outlines the application of our algorithm to approximate a sample of size 1 million drawn from the two-dimensional mixture model. Plot a) depicts the probability density function of the mixture distribution, plot b) illustrates the underlying sample and plot c) shows the resulting probability density function of the PWC estimator.

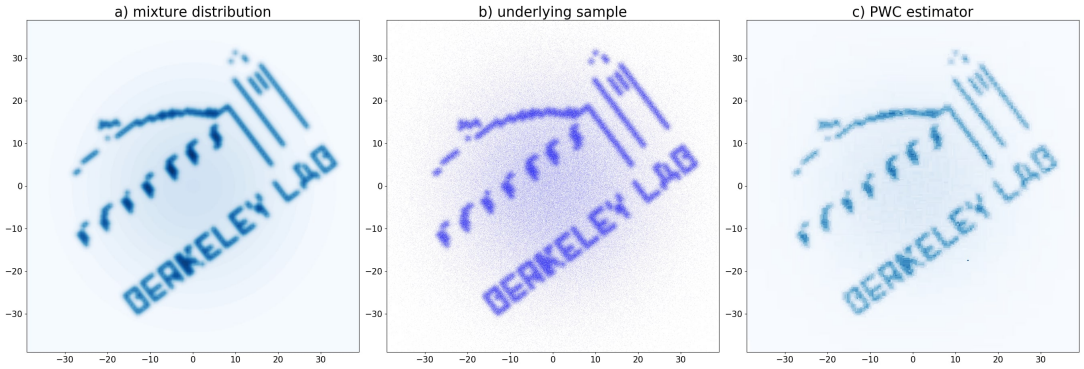

Figure 7. Illustration of the PWC estimator of a non-trivial mixture of Gaussian distributions in $\mathbb{R}^2$.

Figure 8 sets out in more detail part of the sample realizations (blue dots) and the partitioning of the domain forming the PWC estimator.

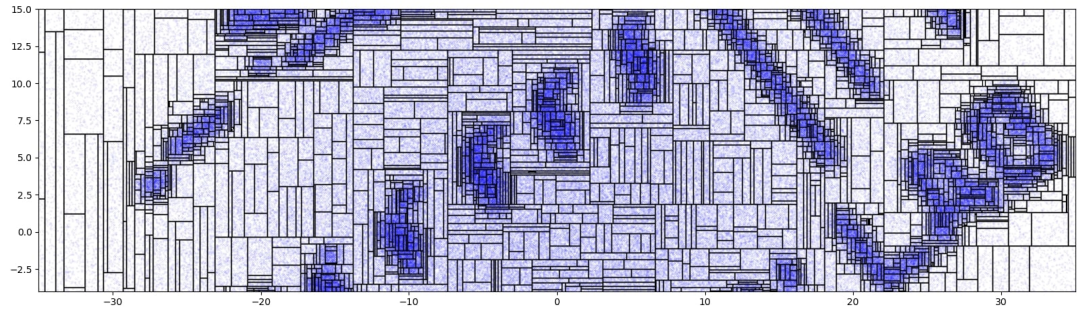

Figure 8. Detail of the PWC estimator partitioning rectangles dividing the domain and encapsulating the sample observations.

Table 1 summarizes some features regarding the PWC estimator such as the number of rectangles S partitioning the domain, the value of the parameter α and the selected ℓ_p-norm as ground distance.

Table 1. PWC estimator characteristics.

PWC estimator	
Number of rectangles	6880
Number of split along $j = 1$	3594
Number of split along $j = 2$	3509
Ground metric	ℓ_1-norm
α	0.05
Execution time (in sec.)	196.31
Sample size	1000000

A total of 7103 bisections have been performed by the algorithm on the starting trivial partition to obtain a PWC estimator that is equipped with 6880 hyperrectangles, containing on average approximately 145 data points (ranging from a minimum of a single data point to a maximum of 675 observations). This aspect exhibits one of the PWC estimator attributes: it is efficient in terms of information needed for storing purposes.

Table 2. Comparison of PWC estimator and sample statistics.

	Mean	Variance	Covariance	Skewness
Sample	(1.481, 0.403)	(238.119, 244.734)	44.856	(0.009, -0.102)
PWC estimator	(1.481, 0.402)	(237.487, 244.303)	45.404	(0.010, -0.102)

Table 2 matches PWC estimator moments against their empirical ones. It can be noted that the shape of the sample distribution is preserved and the PWC estimator, despite losing part of the information contained in the data, had characteristics very similar to the sample ones.

The impact on the PWC estimator of different values of α has been assessed on the same sample. When this parameter value increases, as stated by Lemma 4.6, the partition naturally becomes finer, and the final PWC estimator grows nearer to each single observation. Figure 9 shows how the PWC estimator changes by varying the significance level, in terms of execution time, number of rectangles, average number of points per rectangle, and total number of splits executed. With an increasing α, the total number of splits and the execution time escalate, the number of rectangles rises and tends to the number of observations, the average number of points decreases to the value 1.

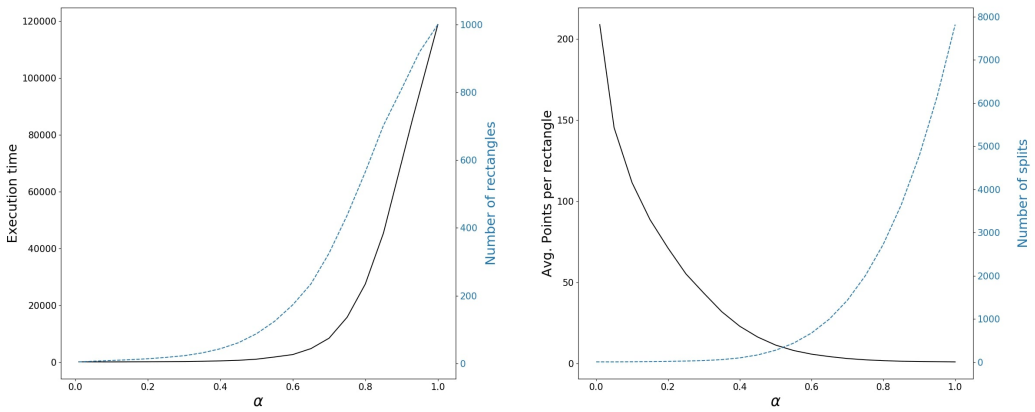

Figure 9. Algorithm sensitivity analysis of the variation of the significance level α. The left plot reports the execution time (in seconds) and the number of rectangles (in thousands) of the resulting PWC estimator. The right plot indicates the average numbers of data points per rectangle and the total number of splits executed (in thousands).

5.2. Three-dimensional spaces

Moreover, we considered a group of simulated datasets with known density to evaluate the ability of the PWC estimator to recover the underlying distributions in three dimensions. Similarly to [34], the following four samples, each of size 600 000, have been examined.

- Dataset 1 resembles a wall-like and a filament-like structure. The first and the second dimensions of the wall-like structure are both uniform on $[0, 100]$, the third dimension is drawn from a Gaussian distribution with mean 50 and variance 5. The filament-like structure, conversely, is created with a bivariate Gaussian distribution, with location (50, 50), variances 5 and 0 covariance, in the first and second coordinates, and a uniform distribution on $[0, 100]$ in the third dimension.

- Dataset 2 mirrors three wall-like structures. Each wall consists of uniform distributions on $[0, 100]^2$ and a Gaussian distribution. In one wall-like structure the Gaussian has mean 10 and variance 5, in the others it has mean 50 and variance 5.
- Dataset 3 is generated from a three-dimensional distribution with independent and identically distributed lognormal components, whose mean and variance are equal to 3 and 4, respectively.
- Dataset 4 contains points drawn from two trivariate Gaussian distributions. Each has independent and identically distributed marginals, one with locations 25 and variances 5, the other with location 65 and variance 20. To these is added a uniform noise on $[0, 100]^3$.

The three-dimensional scatter plots of the above-mentioned samples and the relative PWC estimators are displayed in Figure 10.

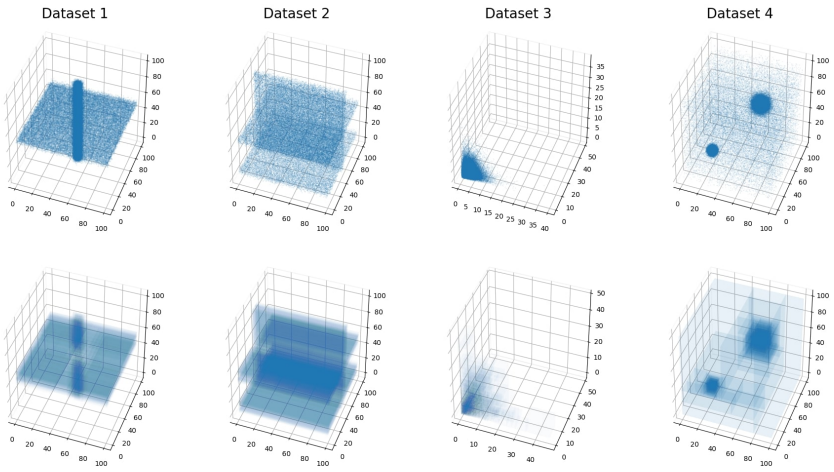

Figure 10. Scatter graphs of the simulated datasets 1-4 from left to right (top) and their corresponding PWC estimators (bottom).

Table 3 reports the characteristics of the PWC estimators for the four datasets. It can be noted that, as expected, in datasets 1 and 2 the dimensions with uniform components are affected by a lower number of bisections. In datasets 3 and 4 the number of splits are similar across the dimensions.

Table 3. PWC estimators characteristics for datasets 1-4.

	Dataset 1	Dataset 2	Dataset 3	Dataset 4
Number of rectangles	1793	1062	3968	2944
Number of split along $j = 1$	631	57	1373	984
Number of split along $j = 2$	583	653	1456	1040
Number of split along $j = 3$	683	379	1404	988
Ground metric	ℓ_1-norm	ℓ_1-norm	ℓ_1-norm	ℓ_1-norm
α	0.05	0.05	0.05	0.05
Execution time (in sec.)	783.41	341.36	283.14	1658.95
Sample size	600000	600000	600000	600000

5.3. Nine-dimensional spaces

Finally, our algorithm has been run on an observed galaxy sample drawn from the Sloan Digital Sky Survey (SDSS). The analysed SDSS dataset contains 108 070 observations from SDSS Skyserver DR12 database and consists of the following nine variables:

- *ra*: standard astronomical right ascension.
- *dec*: declination.
- *u*, *g*, *r*, *i*, *z*: camera filters from 1 to 5.
- *camcol*: the output of one camera column as part of the length of a strip observed in a single contiguous observing pass scan.
- *redshift*: the redshift phenomenon value.

The original data, together with a more detailed explanation of its features, is available at [35]. Table 4 summarizes some aspects of the resulting PWC estimator, and Table 5 highlights the number of splits occurred in each dimension. The dimension in which more splits have been executed by the algorithm is *camcol*, which is the only feature of the data with atoms.

Table 4. PWC estimator characteristics.

PWC estimator	
Number of rectangles	15295
Ground metric	ℓ_1-norm
α	0.05
Execution time (in sec.)	3751.76
Sample size	108070

Table 5. Number of splits per dimension.

Dimension	splits
ra	2104
dec	2104
u	4045
g	5703
r	6447
i	6396
z	5273
camcol	12869
redshift	3209

As shown in Table 6, which compares PWC estimator moments with their empirical counterparts, the PWC estimator is able to capture the information contained in the data also in the nine-dimensional case considered.

Table 6. Comparison of PWC estimator and sample statistics.

	ra	*dec*	*u*	*g*	*r*	*i*	*z*	*camcol*	*redshift*
					Mean				
Sample	228.175	55.589	21.727	20.137	19.148	18.528	18.333	3.561	0.505
PWC estimator	228.923	55.746	21.718	20.134	19.147	18.550	18.330	3.550	0.506
					Variance				
Sample	514.537	22.978	5.618	4.199	3.421	931.673	3.091	2.632	0.501
PWC estimator	523.600	23.317	5.907	4.303	3.516	545.172	3.208	2.349	0.508
					Skewness				
Sample	1.897	2.265	0.009	-0.410	-0.444	-327.105	-0.144	-0.034	2.745
PWC estimator	1.782	2.139	0.050	-0.401	-0.438	-342.802	-0.131	-0.022	2.766

Figure 11, at last, illustrates the dependence structure of the PWC estimator dimensions, as expressed by the pairwise correlations, and confronts it with the corresponding ones observed in the sample. Evident is that the PWC estimator, besides approximating the sample marginal behaviour, is also capable of recognizing the relationships intervening between the dimensions of the dataset.

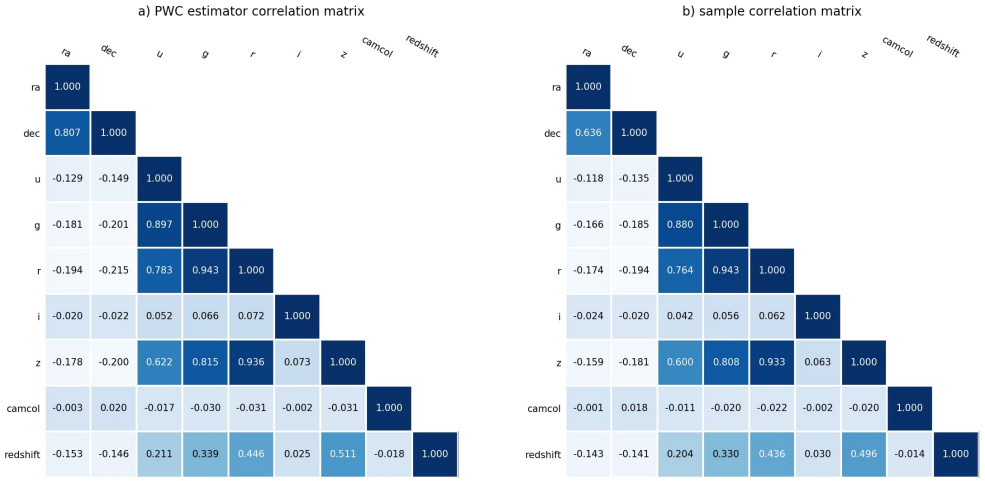

Figure 11. Comparison of the dependence structure: plot a) reports the entries of the PWC estimator correlation matrix, whereas plot b) indicates the same quantities calculated on the original data.

6. Conclusion

6.1. Discussion

As opposed to other piecewise constant density estimators with a tree structure ([2,4, 5]), our methodology centres on the idea of assessing uniformity within each partition element using a hypothesis test based on the Wasserstein distance. As a result, the learning of the tree is implicitly defined by the hypothesis test, and its significance level establishes the stopping rule controlling the partition growth process. On top of that, our algorithm is not prone to overfit the data, because, as the cardinality of the partition grows, the power of the hypothesis test in each new partition element, resulting from the bisection, decreases with lower sample size. Hence, in the calibration phase, there is no arbitrary threshold to select or error function to minimize, and it is not necessary to consider any techniques to penalize complexity and lessen the chance of overfitting, such as regularization, cross-validation, early stopping and pruning.

In addition, our Wasserstein distance based hypothesis test for assessing uniformity provides also a comprehensive method, in sense that it works for any dimension and for any type of distribution. Other candidates that would fit in a scheme similar to ours are the Kolmogorov-Smirnov (K-S) and Pearson's chi-squared (χ^2) goodness-of-fit tests. The former, however, poses non-trivial obstacles in more than one dimension, and there is currently no single approach which is universally applicable (see e.g. [36]). The χ^2 test, although it may theoretically be applied for testing any multivariate distribution, is sensitive to the set of non-overlapping bins chosen to reduce the observations to a set of counts. The greater the number of bins, the more accurately the local data behaviour will be quantified. Nevertheless, the test may give invalid results if not all expected frequencies are sufficiently large (greater than 5 is the usual rule of thumb) and, as a result, the test cannot be trusted in high dimensions [37]. On the other hand, when using a small number of bins, the test loses power. This, in our algorithm, would increase the chance of having hyperrectangles where data are not adequately uniform. The authors have compared the empirical rejection rates of the Wasserstein distance

based hypothesis test with K-S and χ^2 goodness-of-fit tests. Results are displayed in Appendix A and show the value of our approach.

Finally, referring to the comparison of density estimation methodologies carried out in [2] when introducing Density Estimation Trees (DET), we summarize some properties of our approach in Table 7.

Table 7. Qualitative characteristics of the PWC estimator.

Methodology	Accuracy	Interpretability			Adaptability		Speed	
		COD	VI	Rules	ABD	AWD	Calibration	Query
PWC estimator	medium	✓	✓	✓	✓	✓	slow $\mathcal{O}(n^2 \log(n))$	fast $\mathcal{O}(D)$

Despite the cost of having relatively less accuracy in prediction, our PWC estimator enjoys adaptability, interpretability and the efficient querying like other DET-based approaches [38]. Flexibility applies in terms of adaptability between dimensions (ABD) and within dimension (AWD). The former means that dimensions are treated differently according to their impact on the density; the latter implies that the estimator, in a given dimension, adjusts to the local behaviour of the observations. The estimator we present also benefits from interpretability since:

- It is able to detect clusters and outliers (COD).
- It provides variable importance (VI), i.e. identify dimensions that significantly affect the density.
- It produces rules for specifying subsets of the data which might represent a cluster or outliers.

Although the calibration phase is more onerous than other methodologies, this cost is compensated by the efficient queries. The computational cost for fitting the PWC estimator is dependent on the complexity of the algorithm adopted for computing the Wasserstein distance in the admissibility condition verification phase; in our case $\mathcal{O}(n^2 \log(n))$. The query time for the PWC estimator is $\mathcal{O}(D)$, the same of DET, where D is the depth of the diagram tree representation.

6.2. Summary

This paper introduces an algorithm that computes a piecewise constant estimator to approximate the underlying probability density of a multivariate sample, with possibly hundred of thousands or millions data points.

The PWC estimator is determined using a recursive procedure that generates a partition of the sample domain constituted by hyperrectangular regions where the sample is sufficiently uniformly distributed. Uniformity is assessed using a Wasserstein distance based hypothesis testing.

The algorithm is efficient since the resulting distribution can be concisely represented and requires significantly less memory than the original sample. Instead of storing all data, one can only know the estimate for each nonempty hyperrectangles, which are typically fewer in number than the sample size [39]. Moreover, because of the hierarchical and recursive bisection scheme, the PWC estimator can be conveniently represented through a tree diagram, in which the root is the starting bounding box, each node represents a bisection of the domain, and the leaves are the final partition elements associated to the resulting admissible PWC distribution.

Lastly, as highlighted in [12], using PWC distributions constitutes a favourable ap-

proach when information on empirical distributions should be preserved or transferred between systems, because of its memory and bandwidth efficiency, and because it does not distort shape or statistics of the sample. Therefore, our algorithm can be advantageous in applied environments where empirical distributions are repeatedly used and transferred among different users.

References

[1] Darbellay GA, Vajda I. Estimation of the information by an adaptive partitioning of the observation space. IEEE Transactions on Information Theory. 1999; 45(4): 1315–1321.

[2] Ram P, Gray AG. Density estimation trees. In Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD'11; 2011; 627–635, New York, NY, USA. ACM.

[3] Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and Regression Trees. Wadsworth and Brooks, Monterey, CA. 1984.

[4] Li D, Yang K, Wong WH. Density estimation via discrepancy based adaptive sequential partition. In Lee DD, Sugiyama M, Luxburg U V, Guyon I, Garnett R, editors, Advances in Neural InformationProcessing Systems 29. 2016; 1091–1099. Curran Associates, Inc.

[5] Meyer DW. Density estimation with distribution element trees. Statistics and Computing. 2018; 28(3):609–632.

[6] Del Barrio E, Giné E, Utzet F. Asymptotics for L_2 functionals of the empirical quantile process, with applications to tests of fit based on weighted Wasserstein distances. Bernoulli. 2005; 11(1):131–189.

[7] Ramdas A, Trillos NG, Cuturi M. On Wasserstein Two-Sample Testing and Related Families of Nonparametric Tests. Entropy. 2017; 19(2):47.

[8] Deng S, Li W, Wu X. Wasserstein identity testing. CoRR, abs/1710.10457. 2017.

[9] Klemelä J. Smoothing of Multivariate Data: Density Estimation and Visualization. Wiley. 2009.

[10] Irpino A, Romano E. Optimal histogram representation of large data sets: Fisher vs piecewiselinear approximation. Revue des Nouvelles Technologies de l'Information RNTI-E-9. 2007; 1:99–110.

[11] Bassetti F, Gualandi S, Veneroni M. On the Computation of Kantorovich-Wasserstein Distances between 2D-Histograms by Uncapacitated Minimum Cost Flows. ArXiv: 1804.00445. 2018.

[12] Arbenz P, Guevara-Alarcón W. Piecewise linear approximation of empirical distributions under a Wasserstein distance constraint. Journal of Statistical Computation and Simulation. 2018; 1–24.

[13] Rubner Y, Guibas L, Tomasi C. The earth movers distance, multi-dimensional scaling, and color-based image retrieval. Proceedings of the ARPA Image Understanding Workshop. 1997; 661–668.

[14] Villani C. Optimal Transport, Old and New. Springer-Verlag, Berlin. 2009.

[15] Staib M, Jegelka S. Wasserstein k-means++ for Cloud Regime Histogram Clustering. In Proceedings of the Seventh International Workshop on Climate Informatics: CI 2017. 2017.

[16] Bigot J, Gouet R, Klein T, López A. Geodesic PCA in the Wasserstein space by convex PCA. Annales de l'Institut Henri Poincaré, Probabilités et Statistiques. 2017; 53(1):1–26.

[17] Gelbrich M. On a Formula for the L_2 Wasserstein Metric between Measures on Euclidean and HilbertSpaces. Mathematische Nachrichten. 1990; 147(1):185–203.

[18] Mallows C. A note on asymptotic joint normality. Annals of Mathematical Statistics. 1972; 43:508–515.

[19] Del Barrio E, Giné E, Matrán C. Central limit theorems for the Wasserstein distance between the empirical and the true distributions. Annals of Probability. 1999; 27(2):1009–

1071.
[20] Bickel PJ, Freedman DA. Some asymptotic theory for the bootstrap. The Annals of Statistics. 1981; 1196–1217.
[21] Rippl T, Munk A, Sturm A. Limit laws of the empirical Wasserstein distance: Gaussian distributions. Journal of Multivariate Analysis. 2016; 151:90–109.
[22] Sommerfeld M, Munk A. Inference for empirical wasserstein distances on finite spaces. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2018; 80(1):219–238.
[23] Ajtai M, Komlós J, Tusnády G. On optimal matchings. Combinatorica. 1984; 4(4):259–264.
[24] Talagrand M. Matching random samples in many dimensions. Annals of Applied Probability. 1992; 2(4):846–856.
[25] Dobríc V, Yukich JE. Asymptotics for transportation cost in high dimensions. Journal of Theoretical Probability. 1995; 8(1):97–118.
[26] Meeker WQ, Hahn GJ, Escobar LA. Statistical Intervals: A Guide for Practitioners and Re-searchers. Wiley Series in Probability and Statistics. Wiley, 2nd edition. 2017.
[27] Burkard R, Dell'Amico M, Martello S. Assignment Problems. Society for Industrial and Applied Mathematics, Philadelphia. 2012.
[28] Schmitzer B. A Sparse multiscale algorithm for dense optimal transport. Journal of Mathematical Imaging and Vision. 2016; 56(2):238–259.
[29] Cuturi M. Sinkhorn distances: Lightspeed computation of optimal transport. Advances in Neural Information Processing Systems. 2013; 2292–2300.
[30] Altschuler J, Weed J, Rigollet P. Near-linear time approximation algorithms for optimal transportvia Sinkhorn iteration. Proceedings of NIPS. 2017.
[31] Abid BK, Gower R. Greedy stochastic algorithms for entropy-regularized optimal transport problems. 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, 1505–1512. PMLR. 2018.
[32] Gibbons JD, Litwin S. Simultaneous estimation of the unknown upper and lower limits in a two-parameter uniform distribution. Sankhyā: The Indian Journal of Statistics, Series B (1960-2002). 1974; 36(1):41–54.
[33] O'Brien TA, Kashinath K, Cavanaugh NR, Collins WD, O'Brien JP. A fast and objective multidimensional kernel density estimation method: fastkde. Computational Statistics & Data Analysis, 2016; 148–160.
[34] Ferdosi BJ, Buddelmeijer H, Trager SC, Wilkinson MHF Roerdink JBTM. Comparison of density estimation methods for astronomical datasets. Astronomy & Astrophysics, 2011; 531: A114.
[35] Saxena A, Sloan Digital Sky Survey DR12 Server Data. Retrieved July 2019 from https://www.kaggle.com/ashishsaxena2209/sloan-digital-sky-survey-dr12-server-data.
[36] Lopes HC, Reid R, Ivan R, Hobson P. The two-dimensional Kolmogorov-Smirnov test. XI International Workshop on Advanced Computing and Analysis Techniques in Physics Research, 2007.
[37] Maydeu-Olivares A, Garcia-Forero C. Goodness-of-fit testing. International Encyclopedia of Education. 2010;7:190–196.
[38] Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer-Verlag, New York. 2001.
[39] Györfi L, Kohler M, Krzyzak A, Walk H. A Distribution-Free Theory of Nonparametric Regression. Springer Series in Statistics. Springer-Verlag, New York. 2002.
[40] Liang JJ, Fang KT, Hickernell FJ, Li R. Testing multivariate uniformity and its applications. Mathematics of Computation, 2001; 70(223):337–355.
[41] Berrendero JR, Cuevas A, Vázquez-Grande F. Testing Multivariate Uniformity: The Distance-to-Boundary Method. The Canadian Journal of Statistics, 2006; 34(4):693–707.

Appendix A.

The performance of the Wasserstein distance based test, herein referred to as Wasserstein test, has been analysed through a simulation study, whose design is similar to those in [40,41]. We quantified the empirical type I error rate and the power of our test statistic, and compared with other procedures used to test the uniformity of random samples.

We evaluated rejection percentages along 5000 independent runs based on samples drawn both from the null and from the alternative hypothesis; this provides us with the empirical type I error rate (false positive) and the power (true positive) of the tests. The analysis considers increasing sample sizes of 25, 50, 100 and 500, and nominal significance levels of 0.01, 0.05 and 0.10. Simulation outputs are displayed in Tables A1, A2 and A3.

All the models have support on the hypercube $[0,1]^d$, where $d = 1, 2, 3$. The null hypotheses refer to samples drawn from the corresponding uniform measure. The alternative hypothesis models are devised for taking into account complementary aspects of non-uniformity: the one-dimensional case especially evaluates the ability to ascertain marginal deviations from uniformity; the multidimensional setting is useful to assess the capacity of detecting departures from independence, keeping the uniformity of marginals.

With regard to the one-dimensional setting, the Wasserstein test (WASS) has been compared with the K-S and the χ^2 goodness-of-fit tests, and the following alternative hypothesis models have been considered.

- The (univariate) uniform contamination models (UC): the observed sample is drawn from a mixture of type $(1-v)\,U_{[0,1]} + v\,U_{[1/2,1/2]}$. The parameter v has been set equal to 0.1 and 0.2 respectively.
- The arcsine distribution (AS).
- The beta models (B) with parameters (1.3, 1.3), (5, 2) and (0.8, 0.8) respectively.
- The truncated Standard Normal distribution (TZ).

In the multidimensional context, three statistics have been considered: the Wasserstein distances with ℓ_1-norm and ℓ_2-norm cost functions, and the χ^2 statistics. As for alternative distributions, we have checked the following distribution families:

- The (multivariate) uniform contamination models (UC): The observed sample is drawn from a mixture of type $(1-v)\,U_{[0,1]^d} + v\,U_{D_d}$, where D_d denotes a cube with the same centre as $[0,1]^d$ and measure $1/2$. The parameter v has been set equal to 0.1 and 0.2 respectively.
- Meta-type uniform distributions: These are the distributions obtained from transformed $\mathbb{R}^d$-supported elliptically distributed random variables. The transformation applied is the relevant probability integral transform, to guarantee that marginals are uniformly distributed. MUT is derived from a multivariate Student's random variable with 5 degrees of freedom, MUC is obtained from a Cauchy variable and finally MUN results from a multivariate Gaussian centred at the origin $N(0,\Sigma)$, where $\Sigma = (\sigma_{i,j})$, $\sigma_{i,i} = 1$, $\sigma_{i,j} = 0.5$ for $i \neq j$.
- The beta independent models (B): All the marginals are independent, identically distributed according to a beta models with parameters (1.3, 1.3) and (0.8, 0.8).

The uniform contamination and the beta models, with parameters (1.3, 1.3) and (0.8, 0.8), are considered in both experiments as they should represent circumstances of non-uniformity hard to detect. The former specifically represents a deviation from

the theoretical model associated with the presence of "inliers" [41]. Either one of them has independent marginals, as in the case of the null distribution.

The results of rejection rates simulation indicate that:

- The empirical type I errors of the Wasserstein test have converged to the significance levels.
- In both the one-dimensional and multidimensional setting, the Wasserstein test hardly detects the uniform contamination models.
- In the one-dimensional setting, the K-S test shows overall a better performance than the other two tests. Nevertheless, the Wasserstein test is the most powerful for the arcsine and the truncated normal distributions. In the multivariate case the Wasserstein test exhibited, in general, the highest power figures. The Wasserstein statistics with ℓ_1-norm and ℓ_2-norm cost functions exhibit a similar behaviour.
- In the multivariate setting the Wasserstein test is able to classify samples from multivariate distributions with uniform marginals. This, along with the favourable power revealed by Wasserstein test when $d = 1$, suggests that the process of testing marginal admissibility condition at first, and subsequently the joint admissibility condition, should succeed in picking out non-uniformly distributed sample.

Table A1. One-dimensional setting rejection rates; sample size n = 25, 50, 100, 500.

Test	α	U	$UC(0.1)$	$UC(0.2)$	AS	$B(1.3, 1.3)$	$B(5, 2)$	$B(0.8, 0.8)$	TZ
					$n = 25$				
WASS	0.01	0.0082	0.0066	0.0046	0.0600	0.0062	0.9928	0.0146	0.0244
WASS	0.05	0.0508	0.0428	0.0348	0.2356	0.0400	1.0000	0.0748	0.0950
WASS	0.10	0.0992	0.0872	0.0872	0.4024	0.0982	1.0000	0.1400	0.1610
K-S	0.01	0.0078	1.0000	1.0000	0.0694	1.0000	1.0000	1.0000	0.0168
K-S	0.05	0.0424	1.0000	1.0000	0.2044	1.0000	1.0000	1.0000	0.0722
K-S	0.10	0.0848	1.0000	1.0000	0.3360	1.0000	1.0000	1.0000	0.1324
χ^2	0.01	0.0078	0.0108	0.0154	0.2244	0.0182	0.9062	0.0164	0.0128
χ^2	0.05	0.0452	0.0554	0.0824	0.3688	0.0762	0.9910	0.0702	0.0640
χ^2	0.10	0.0930	0.1068	0.1434	0.4978	0.1402	0.9986	0.1274	0.1216
					n = 50				
WASS	0.01	0.0100	0.0070	0.0048	0.1426	0.0064	1.0000	0.0176	0.0442
WASS	0.05	0.0558	0.0466	0.0466	0.4748	0.0538	1.0000	0.0814	0.1632
WASS	0.10	0.1042	0.0922	0.1144	0.6754	0.1148	1.0000	0.1552	0.2508
K-S	0.01	0.0080	1.0000	1.0000	0.1476	1.0000	1.0000	1.0000	0.0362
K-S	0.05	0.0462	1.0000	1.0000	0.3952	1.0000	1.0000	1.0000	0.1270
K-S	0.10	0.0914	1.0000	1.0000	0.5574	1.0000	1.0000	1.0000	0.2088
χ^2	0.01	0.0118	0.0140	0.0256	0.5056	0.0242	0.9998	0.0262	0.0222
χ^2	0.05	0.0468	0.0620	0.0996	0.6572	0.0940	1.0000	0.0952	0.0856
χ^2	0.10	0.0998	0.1168	0.1888	0.7394	0.1810	1.0000	0.1758	0.1528
					n = 100				
WASS	0.01	0.0108	0.0070	0.0100	0.4904	0.0088	1.0000	0.0180	0.1118
WASS	0.05	0.0512	0.0472	0.0730	0.8618	0.0698	1.0000	0.1036	0.2802
WASS	0.10	0.0984	0.1054	0.1772	0.9484	0.1676	1.0000	0.1986	0.3914
K-S	0.01	0.0082	1.0000	1.0000	0.4014	1.0000	1.0000	1.0000	0.0852
K-S	0.05	0.0412	1.0000	1.0000	0.7522	1.0000	1.0000	1.0000	0.2300
K-S	0.10	0.0896	1.0000	1.0000	0.8872	1.0000	1.0000	1.0000	0.3404
χ^2	0.01	0.0094	0.0196	0.0526	0.8276	0.0406	1.0000	0.0468	0.0248
χ^2	0.05	0.0504	0.0738	0.1586	0.9192	0.1410	1.0000	0.1374	0.0918
χ^2	0.10	0.0982	0.1312	0.2506	0.9514	0.2284	1.0000	0.2278	0.1588
					n = 500				
WASS	0.01	0.0092	0.0118	0.1672	1.0000	0.1622	1.0000	0.1220	0.7192
WASS	0.05	0.0470	0.0930	0.6096	1.0000	0.6148	1.0000	0.4666	0.8838
WASS	0.10	0.0976	0.2172	0.8112	1.0000	0.8144	1.0000	0.6810	0.9334
K-S	0.01	0.0080	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.6220
K-S	0.05	0.0452	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.8324
K-S	0.10	0.0970	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	0.9012
χ^2	0.01	0.0100	0.0260	0.1806	1.000	0.1470	1.0000	0.2088	0.0536
χ^2	0.05	0.0502	0.0962	0.3878	1.000	0.3432	1.0000	0.4106	0.1720
χ^2	0.10	0.0932	0.1758	0.5206	1.000	0.4864	1.0000	0.5418	0.2824

Table A2. Two-dimensional setting rejection rates; sample size $n = 25$, 50, 100, 500.

Test	α	U	$UC(0.1)$	$UC(0.2)$	MUT	MUC	MUN	$B(1.3, 1.3)$	$B(0.8, 0.8)$
				$n = 25$					
WASS ℓ_1	0.01	0.0110	0.0084	0.0086	0.0400	0.0450	0.0418	0.0062	0.0244
WASS ℓ_1	0.05	0.0488	0.0402	0.0384	0.1738	0.2212	0.1790	0.0420	0.0986
WASS ℓ_1	0.10	0.0960	0.0838	0.0904	0.3132	0.3888	0.3224	0.0876	0.1674
WASS ℓ_2	0.01	0.0112	0.0090	0.0076	0.0674	0.0916	0.0666	0.0054	0.0252
WASS ℓ_2	0.05	0.0446	0.0430	0.0354	0.2346	0.2758	0.2316	0.0364	0.0944
WASS ℓ_2	0.10	0.0956	0.0920	0.0840	0.3684	0.4334	0.3744	0.0768	0.1644
χ^2	0.01	0.0084	0.0076	0.0080	0.0920	0.0798	0.0848	0.0088	0.0090
χ^2	0.05	0.0398	0.0386	0.0386	0.2262	0.2136	0.2288	0.0416	0.0432
χ^2	0.10	0.0864	0.0851	0.0846	0.3604	0.3406	0.3562	0.0934	0.0918
				n = 50					
WASS ℓ_1	0.01	0.0098	0.0068	0.0072	0.1396	0.2292	0.1452	0.0058	0.0272
WASS ℓ_1	0.05	0.0514	0.0412	0.0432	0.4765	0.6029	0.4658	0.0522	0.1112
WASS ℓ_1	0.10	0.1012	0.0916	0.0943	0.6632	0.7692	0.6502	0.1178	0.1981
WASS ℓ_2	0.01	0.0118	0.0066	0.0054	0.2402	0.3254	0.2384	0.0056	0.0262
WASS ℓ_2	0.05	0.0504	0.0402	0.0402	0.5466	0.6386	0.5408	0.0456	0.1036
WASS ℓ_2	0.10	0.1036	0.0902	0.0936	0.7094	0.7864	0.7064	0.1004	0.1936
χ^2	0.01	0.0114	0.0126	0.0134	0.1212	0.3762	0.0908	0.0202	0.0244
χ^2	0.05	0.0464	0.0485	0.0568	0.2612	0.5796	0.2218	0.0762	0.0716
χ^2	0.10	0.1162	0.1234	0.1346	0.4078	0.7273	0.3658	0.1664	0.1556
				n = 100					
WASS ℓ_1	0.01	0.0125	0.0114	0.0114	0.6475	0.8178	0.6384	0.0188	0.0436
WASS ℓ_1	0.05	0.0534	0.0482	0.0636	0.9163	0.9716	0.9120	0.1026	0.1581
WASS ℓ_1	0.10	0.1062	0.099	0.1256	0.9661	0.9913	0.9684	0.1978	0.2732
WASS ℓ_2	0.01	0.0122	0.0082	0.0104	0.7592	0.8572	0.7578	0.0158	0.0392
WASS ℓ_2	0.05	0.0576	0.0458	0.0542	0.9346	0.9744	0.9340	0.0862	0.1438
WASS ℓ_2	0.10	0.1106	0.096	0.1118	0.9713	0.9912	0.9766	0.1736	0.2576
χ^2	0.01	0.0182	0.0194	0.0310	0.2140	0.6534	0.1450	0.0381	0.0376
χ^2	0.05	0.0484	0.0492	0.0776	0.3410	0.7782	0.2608	0.0846	0.0898
χ^2	0.10	0.1256	0.1242	0.1722	0.4998	0.8748	0.4170	0.1774	0.1880
				n = 500					
WASS ℓ_1	0.01	0.0094	0.0118	0.0671	1.0000	1.0000	1.0000	0.4866	0.3618
WASS ℓ_1	0.05	0.0424	0.0656	0.2566	1.0000	1.0000	1.0000	0.8224	0.6874
WASS ℓ_1	0.10	0.0844	0.1302	0.4182	1.0000	1.0000	1.0000	0.9162	0.8268
WASS ℓ_2	0.01	0.0114	0.0138	0.0548	1.0000	1.0000	1.0000	0.4124	0.3151
WASS ℓ_2	0.05	0.0470	0.0648	0.2412	1.0000	1.0000	1.0000	0.7932	0.6684
WASS ℓ_2	0.10	0.0992	0.1386	0.4104	1.0000	1.0000	1.0000	0.9163	0.8226
χ^2	0.01	0.0176	0.0226	0.0280	0.3128	0.9218	0.1564	0.0344	0.0362
χ^2	0.05	0.0486	0.0554	0.0724	0.4736	0.9648	0.2928	0.0921	0.0982
χ^2	0.10	0.0798	0.0860	0.1084	0.5532	0.9774	0.3754	0.1286	0.1452

Table A3. Three-dimensional setting rejection rates; sample size n = 25, 50, 100, 500.

Test	α	U	$UC(0.1)$	$UC(0.2)$	MUT	MUC	MUN	$B(1.3, 1.3)$	$B(0.8, 0.8)$
					n = 25				
WASS ℓ_1	0.01	0.0108	0.0078	0.0065	0.2062	0.2916	0.2093	0.0034	0.0342
WASS ℓ_1	0.05	0.0488	0.0412	0.0372	0.4584	0.5732	0.4551	0.0243	0.1314
WASS ℓ_1	0.10	0.0948	0.0848	0.0766	0.6028	0.7158	0.5998	0.0584	0.2202
WASS ℓ_2	0.01	0.0108	0.0078	0.0066	0.2655	0.3522	0.2512	0.0028	0.0334
WASS ℓ_2	0.05	0.0492	0.0368	0.0322	0.5074	0.6128	0.5011	0.0214	0.1308
WASS ℓ_2	0.10	0.0986	0.0858	0.0722	0.6456	0.7414	0.6374	0.0492	0.2204
χ^2	0.01	0.0096	0.0092	0.0081	0.2748	0.2776	0.2728	0.0086	0.0068
χ^2	0.05	0.0410	0.0431	0.0458	0.4862	0.4902	0.4818	0.0486	0.0452
χ^2	0.10	0.0856	0.0871	0.0882	0.5942	0.5968	0.5828	0.0932	0.0872
					n = 50				
WASS ℓ_1	0.01	0.0098	0.0066	0.0054	0.6778	0.8174	0.6534	0.0058	0.0402
WASS ℓ_1	0.05	0.0472	0.0426	0.0346	0.8824	0.9572	0.8798	0.0355	0.1610
WASS ℓ_1	0.10	0.1008	0.0872	0.0761	0.9432	0.9812	0.9374	0.0782	0.2688
WASS ℓ_2	0.01	0.0102	0.0066	0.0078	0.7308	0.8358	0.7172	0.0036	0.0390
WASS ℓ_2	0.05	0.0496	0.0421	0.0346	0.9032	0.9614	0.9022	0.0288	0.1536
WASS ℓ_2	0.10	0.1016	0.0883	0.0771	0.9524	0.9814	0.9494	0.0648	0.2568
χ^2	0.01	0.0116	0.0112	0.0106	0.6532	0.7966	0.6252	0.0236	0.0202
χ^2	0.05	0.0480	0.0523	0.0506	0.8098	0.9074	0.7938	0.0826	0.0705
χ^2	0.10	0.0992	0.0962	0.0891	0.8766	0.9456	0.8666	0.1438	0.1222
					n = 100				
WASS ℓ_1	0.01	0.008	0.0094	0.0072	0.9908	0.9992	0.9894	0.0114	0.0624
WASS ℓ_1	0.05	0.0484	0.0444	0.0497	0.9992	1.0000	0.999	0.0776	0.2244
WASS ℓ_1	0.10	0.0934	0.0902	0.0864	1.0000	1.0000	1.0000	0.1606	0.3626
WASS ℓ_2	0.01	0.0088	0.0062	0.0084	0.9932	0.9992	0.9938	0.0086	0.0584
WASS ℓ_2	0.05	0.0498	0.0434	0.0394	0.9992	1.0000	0.9996	0.0578	0.2146
WASS ℓ_2	0.10	0.0954	0.086	0.0864	1.0000	1.0000	1.0000	0.1264	0.3454
χ^2	0.01	0.0124	0.0112	0.0116	0.9414	0.9975	0.9282	0.0432	0.0266
χ^2	0.05	0.0465	0.0524	0.0598	0.9842	0.9994	0.9796	0.1362	0.1046
χ^2	0.10	0.0976	0.1122	0.1192	0.9934	1.0000	0.9898	0.2242	0.1858
					n = 500				
WASS ℓ_1	0.01	0.0082	0.0108	0.0214	1.0000	1.0000	1.0000	0.5408	0.5759
WASS ℓ_1	0.05	0.0424	0.0538	0.1074	1.0000	1.0000	1.0000	0.8338	0.8436
WASS ℓ_1	0.10	0.0839	0.1028	0.1932	1.0000	1.0000	1.0000	0.9218	0.9274
WASS ℓ_2	0.01	0.0114	0.01205	0.0192	1.0000	1.0000	1.0000	0.4564	0.5343
WASS ℓ_2	0.05	0.0474	0.0574	0.0944	1.0000	1.0000	1.0000	0.8019	0.8322
WASS ℓ_2	0.10	0.0954	0.1124	0.1794	1.0000	1.0000	1.0000	0.9106	0.9184
χ^2	0.01	0.0136	0.0172	0.0330	1.000	1.000	1.000	0.1266	0.0876
χ^2	0.05	0.0571	0.0636	0.1122	1.000	1.000	1.000	0.2972	0.2378
χ^2	0.10	0.1006	0.1144	0.1818	1.000	1.000	1.000	0.4122	0.3483