Bernoulli **24**(4A), 2018, 3147–3179
https://doi.org/10.3150/17-BEJ957

Wide consensus aggregation in the Wasserstein space. Application to location-scatter families

PEDRO C. ÁLVAREZ-ESTEBAN[1,*], EUSTASIO DEL BARRIO[1,**], JUAN A. CUESTA-ALBERTOS[2,†] and CARLOS MATRÁN[1,‡]

[1]*Departamento de Estadística e Investigación Operativa. IMUVA, Universidad de Valladolid, Paseo de Belén, 7. 47011-VALLADOLID. Spain.*
E-mail: [*]*pedroc@eio.uva.es;* [**]*tasio@eio.uva.es;* [‡]*carlos.matran@uva.es*
[2]*Departamento de Matemáticas, Estadística y Computación, Universidad de Cantabria, Avda. de los Castros, 48. 39005-SANTANDER. Spain. E-mail:* [†]*cuestaj@unican.es*

We introduce a general theory for a consensus-based combination of estimations of probability measures. Potential applications include parallelized or distributed sampling schemes as well as variations on aggregation from resampling techniques like boosting or bagging. Taking into account the possibility of very discrepant estimations, instead of a full consensus we consider a "wide consensus" procedure. The approach is based on the consideration of trimmed barycenters in the Wasserstein space of probability measures. We provide general existence and consistency results as well as suitable properties of these robustified Fréchet means. In order to get quick applicability, we also include characterizations of barycenters of probabilities that belong to (non necessarily elliptical) location and scatter families. For these families, we provide an iterative algorithm for the effective computation of trimmed barycenters, based on a consistent algorithm for computing barycenters, guarantying applicability in a wide setting of statistical problems.

Keywords: impartial trimming; parallelized inference; robust aggregation; trimmed barycenter; trimmed distributions; Wasserstein distance; wide consensus

1. Introduction

Data that consists of samples composed of probability distributions are increasingly common. Examples include the distribution of a set of medical measurements in hospitals in a multicenter clinical trial or that of several economic magnitudes (income and age distribution, for instance) in different countries. Often these distributions are not directly observed, but some estimation is available. This paper introduces a new approach for the combination of several estimations of probabilities. Our goal is to provide a tool to combine available estimations to get a consensus-based global estimation. We recall that this goal has been largely pursued under different frameworks. Merging information, pooling estimation, aggregation estimation or meta-analysis, are expressions related with this common goal. The potential applications that we have in mind also include parallelized or distributed estimation schemes as well as those provided by resampling methods designed to improve unstable procedures or to provide approximate solutions through algorithms involving combinatorial complexity problems.

1350-7265

At present, statistical methodologies under a parallelized or distributed scheme are receiving growing interest. In fact, they constitute a basic statistical challenge in a world where we want to exploit massive data sets that could have been collected by different units or that exceed the size that would make their analysis on a single machine feasible. The need for aggregation methods becomes clear in the following two cases. One, when the different sets of data would be obtained, stored and even processed by the different units, perhaps using different experimental techniques. Another, associated to the "divide and conquer" principle, would include the combination of results obtained from the partition of the data set in smaller, tractable subsets. Note that the partition of the data, in this second category, is often performed based on computational convenience criterions, say by their storage location, or oldness in the data basis, hence essentially both categories share the same handicap: the hypothesis of homogeneity of the distributions corresponding to the different units seems to be excessively optimistic in practice.

Regarding the already mentioned resampling methods, since the introduction of bagging by Breiman [9], subagging, and other aggregating procedures have been introduced in the last years to improve the performance of estimators in different setups, including regression or classification (see, e.g., Bühlmann and Yu [13], Bühlmann [12] and Meinshausen and Bühlmann [32]). The aggregation is usually achieved just by averaging, but there are also other proposals like bragging (in [12]) or magging (in [32]), which aim at robust aggregation. In a different problem, the available algorithms for the obtention of some well known estimators (like the Minimum Volume Ellipsoid (MVE), Minimum Covariance Determinant (MCD) and several others) involve the use of a iterative procedure starting from many initial random choices, either for statistical or computational convenience, that result in a set of different estimations that must be combined to produce a better (or just computable) estimation (see, e.g., Woodruff and Rocke [43], Croux and Haesbroeck [16], or Rousseeuw and Driessen [38]). Depending on the intrinsic geometry of the estimated objects, the aggregation procedure may have to be based on some sort of non standard averaging technique.

Aggregation of a set of estimations of probabilities to provide a final estimation – the consensus – is analyzed in this paper under a novel point of view. We can get motivation for our goal from the following hypothetical situation. Consider a biomedical study to be carried out and processed by a network of hospitals. Each hospital will provide an estimation of the distribution of interest and the goal is to obtain a meta-estimation summarizing the estimations. This combination of information is sensitive to two different possible types of atypic or noisy data. First, the sample obtained in any hospital could have some contaminating data. Second, one or several hospitals could produce very atypical results when compared to the others simply because the patients in the influence zone of the hospital have very different (social, cultural, ethnic, nutritional) features. To handle this general setting we will assume that there exist k units, say $U_1, \ldots, U_k$, and that unit U_i will process a sample $x_1^i, \ldots, x_{n_i}^i$ of $\mathbb{R}^d$-valued data obtained from a distribution P_i. As the results of processing their associated samples, the units produce a new sample consisting in the estimations $\hat{P}_1, \ldots, \hat{P}_k$, perhaps given through the estimations of suitable parameters. Our goal will be to produce a consensus estimator from those obtained by the different units. However, since some units could process very contaminated batches, whose consideration would lead to large deviations from the mainstream model (if any), we will include the possibility of obtaining a wide consensus instead of a full consensus. In our scheme, the meaning of wide consensus must be understood as the possibility of avoiding the results of the most discrepant units, elaborating the consensus just from the remaining units.

We emphasize that our approach assigns a different status to the samples processed by the units (composed by points in $\mathbb{R}^d$) and to the meta-sample, of size k (composed by probability distributions on $\mathbb{R}^d$), provided by the units. The primitive samples have the usual meaning in Statistics and will be processed through more or less standard procedures, a task that will not be considered here, our object of interest being the sample of probability distributions. To work with this sample, we will make a careful use of the structure of these objects. To illustrate this point, let us consider a simple example involving estimation in a normal model. Our proposal aims at producing a normal distribution which is an optimal representation of k normal distributions, $\hat{P}_1, \ldots, \hat{P}_k$ in some sense. Note that a (weighted) average of probabilities is a probability, but the mixture of normal distributions that we could produce in this way is not normal and could be very far, in terms of shape, from the k normal distributions that we are trying to summarize, hence making a different aggregation procedure to be more convenient.

Our choice for the basic aggregation procedure is the Wasserstein barycenter, that we briefly describe next. We will work in the space $\mathcal{P}_2(\mathbb{R}^d)$ of probability measures on $\mathbb{R}^d$ with finite second order moment, endowed with the *L_2-Wasserstein distance*, $\mathcal{W}_2$, defined for $P, Q \in \mathcal{P}_2(\mathbb{R}^d)$ by

$$\mathcal{W}_2(P, Q) := \inf\big\{\big(\mathbf{E}\|U - V\|^2\big)^{1/2} : \mathcal{L}(U) = P, \mathcal{L}(V) = Q\big\}, \tag{1}$$

where we use $\mathcal{L}(X)$ to denote the distribution law of a r.v. X. Given a finite set of elements, $P_1, \ldots, P_k \in \mathcal{P}_2(\mathbb{R}^d)$ and positive weights $\lambda_1, \ldots, \lambda_k$, with $\sum_{i=1}^k \lambda_i = 1$, we would like to obtain a representative element for the whole set. Like the mean of a set of vectors, a barycenter or Fréchet mean in this space can be a good candidate and would be any probability, $\bar{P} \in \mathcal{P}_2(\mathbb{R}^d)$ satisfying

$$\sum_{i=1}^k \lambda_i \mathcal{W}_2^2(\bar{P}, P_i) = \inf\left\{\sum_{i=1}^k \lambda_i \mathcal{W}_2^2(P, P_i) : P \in \mathcal{P}_2\big(\mathbb{R}^d\big)\right\}. \tag{2}$$

Such a probability, when it exists, is called a ($\{\lambda_i\}_{i=1}^k$-*weighted*) *barycenter* of $\{P_i\}_{i=1}^k$. The consideration of barycenters in the Wasserstein setting has been initiated by Agueh and Carlier in [1], with several extensions in Boissard *et al.* [8], Pass [34], Bigot and Klein [7] and in Le Gouic and Loubes [31], where the concept has been extended to arbitrary (non-necessarily finite) families of probabilities (see Definition 2.2 below).

A *full consensus representation* of $P_1, \ldots, P_k$ would be the barycenter associated to equal weights $\lambda_i = 1/k, i = 1, \ldots, k$. Different weights would be more appropriate if, for example, some of the P_i's have been obtained from (or represents) a considerably larger population than some others. On the other hand, the possible existence in $P_1, \ldots, P_k$ of very discrepant representations (possibly due to highly contaminated batches as before), would justify a trimming or reweighting action. Rather than using the ($\{\lambda_i\}_{i=1}^k$-weighted) barycenter of $\{P_i\}_{i=1}^k$, with the original weights, the *wide consensus representation* of $P_1, \ldots, P_k$ with weights $\{\lambda_i\}_{i=1}^k$ or α-trimmed barycenter, $\bar{P}_\alpha$, is a solution, for suitable weights $(\bar{\lambda}_i^\alpha)_{i=1}^k$, of the following double minimization problem

$$\sum_{i=1}^k \bar{\lambda}_i^\alpha \mathcal{W}_2^2\big(\bar{P}^\alpha, P_i\big) = \inf\left\{\sum_{i=1}^k \lambda_i^* \mathcal{W}_2^2(P, P_i) : P \in \mathcal{P}_2\big(\mathbb{R}^d\big), \lambda_i^* \le \lambda_i, \sum_{i=1}^k \lambda_i^* = 1 - \alpha\right\}. \tag{3}$$

Trimming procedures are of frequent use in Robust Statistics to prevent the influence of atypical data in statistical analyses. In fact, a trimmed version of the Wasserstein distance for probabilities on the line was introduced, in the context of Goodness of Fit tests, by Munk and Czado [33] to avoid the effects of data in the tails. This approach was extended in some papers (see Álvarez-Esteban *et al.* [3] and references therein) to cover trimmings like that considered in (3) that are "impartial". This means that there are not a priori selected directions or zones for trimming, being the complete data set which will provide that information. Although often trimming is used with the meaning of deleting a part of the data, here we follow a more flexible approach as in Gordaliza [29], based on probability trimmings (see Definition 3.1 below) which allows to decrease the weight of some regions without completely removing them. We include in Section 5.2 a succinct account of basic results on probability trimmings and refer to Álvarez-Esteban *et al.* [2] for further details.

In this paper, we introduce the concept of trimmed barycenter for probabilities μ on the Wasserstein space of probabilities on $\mathbb{R}^d$ with finite second moment endowed with the metric $\mathcal{W}_2$, extending that of trimmed mean introduced in Rousseeuw [36] and Gordaliza [29]. Notice that no moment assumption is made on μ. Our setup covers the case of general Borel probability measures on $\mathcal{P}_2(\mathbb{R}^d)$ (of which (3) corresponds to the particular case of finitely supported measures, see Definitions 2.2 and 3.1 below). We provide existence and consistency results for trimmed barycenters. In particular, we prove a Strong Law of Large Numbers (Theorem 3.6) for trimmed barycenters in this space, to be denoted throughout by $W_2(\mathcal{P}_2(\mathbb{R}^d))$ (see (7) below).

As noted before, a desirable feature of any aggregation method is adaptation to the shape of the objects to be aggregated. Remarkably, this is the case for barycenters and trimmed barycenters in location and scatter families such as the Gaussian family. In particular, we show that the barycenter or trimmed barycenter of a probability on $\mathcal{P}_2(\mathbb{R}^d)$ supported in a (non-necessarily finite) set of probabilities belonging to a location scatter family also belongs to the family. We also provide a characterization of barycenters in location and scatter families in terms of a fixed point equation as well as some equivariance results for general barycenters and trimmed barycenters. Notice that suitability of the location and scatter families in the Wasserstein space has been considered by Chernozhukov *et al.* in [15] in relation with Monge-Kantorovich quantiles. Also, Rippl *et al.* [35] take advantage of the explicit expression of the Wasserstein distance between Gaussian distributions. In a similar spirit to that considered here, they substitute sampling distributions obtained from Gaussian distributions by Gaussian distributions with estimated parameters, and address the problem of the asymptotic behavior of the Wasserstein distance between empirical and theoretical distributions. Their analysis includes the two-sample setting, for independent samples, through the distance between normal distributions when the parameters are estimated from the respective samples.

Turning back to the statistical motivation of this work, we note that the applicability of barycenters or trimmed barycenters for data analysis will strongly depend on the availability of efficient algorithms for their computation. In this sense, we stress the fact that, in the multivariate setting, even the computation of the barycenter of a finite collection of normal distributions can be a hard task since no closed form expression for the barycenter is available. On the other hand, convexity of the map $\eta \mapsto \mathcal{W}_2^2(\eta, \nu)$ implies that Wasserstein barycenters are minimizers of a convex functional. This fact is at the basis of a fast algorithm just introduced in [4] for the approximate computation of barycenters, including the case of location and scatter families. Here we show how this can be used for the efficient computation of trimmed barycenters in these loca-

tion and scatter families. Also we must stress that our approach constitutes a technical keystone for the introduction of robust clustering in the Wasserstein space, opening new applications in that wide setting (see del Barrio *et al.* [24]).

The remaining sections of this paper are organized as follows. In Section 2, we give a quick account of notable results on Wasserstein distance and Wasserstein spaces including the main known results on barycenters. Section 3 introduces trimmed barycenters and provides the main results announced before. They include existence, consistency, equivariance, and characterizations in location-scatter families as well as relevant properties involving shapes and sizes. Section 4 discusses computational issues for barycenters and trimmed barycenters and presents an algorithm for the computation of trimmed barycenters in location and scatter families. It also includes some toy examples and an application to aggregation of MCD's solutions obtained by subsampling on a real data set. The analysis of this example includes hints on the possible selection of the trimming as well as information on the running times of the algorithms. Most of the technical details and proofs are deferred to Section 5.

We conclude this Introduction with some explanations on notation. Unless explicitly noted, probability measures are defined on the Borel σ-algebra of the (metric) space. The indicator function of a set, A, will be represented by I_A, while $\delta_{\{x\}}$ will denote Dirac's measure on x. We write ℓ_d for Lebesgue measure on $\mathbb{R}^d$ and $\mu \ll \nu$ to mean that μ is absolutely continuous with respect to ν. Weak convergence of probability measures will be denoted by $\to_w$. We assume that weights, $\lambda_1, \ldots, \lambda_k$, are positive numbers, $\lambda_i > 0, i = 1, \ldots, k$ such that $\sum_{i=1}^k \lambda_i = 1$. We will denote by $\mathcal{P}_{2,ac}(\mathbb{R}^d)$ the subset of absolutely continuous probabilities (with respect to ℓ_d) in $\mathcal{P}_2(\mathbb{R}^d)$. Given $P \in \mathcal{P}_2(\mathbb{R}^d)$ and $r > 0$, $B_{\mathcal{W}}(P, r)$ (resp. $\overline{B}_{\mathcal{W}}(P, r)$) will be the open (resp. closed) ball with center at P and radius r for the distance $\mathcal{W}_2$, while $B(x, r)$, where $x \in \mathbb{R}^d$, will refer to the open ball with center at x and radius r for the Euclidean distance on $\mathbb{R}^d$. Finally, we will say that the map T *transports* (pushes forward) the probability P to Q if Q is the image measure of P by T, namely, if $Q = P \circ T^{-1}$.

2. Barycenters in Wasserstein space

As noted in the Introduction, our proposal for wide consensus aggregation is based on Wasserstein metrics and barycenters in Wasserstein space. We refer to the books of Villani [41,42] for a complete and well documented view of the general theory on Wasserstein spaces and optimal transport and to the papers by Agueh and Carlier [1] and Le Gouic and Loubes [31] for barycenters. Here we include a brief introduction, continued in Section 5.1, with some relevant facts and necessary results for our presentation.

It is well known that the infimum in (1) is attained, that is, there exists a pair (X, Y), defined on some probability space, with $\mathcal{L}(X) = P$ and $\mathcal{L}(Y) = Q$ such that $\mathbf{E}\|X - Y\|^2 = \mathcal{W}_2^2(P, Q)$. Such a pair (X, Y) is called a $\mathcal{W}_2$-*optimal transportation plan* ($\mathcal{W}_2$-o.t.p.) for (P, Q), although the alternative terminology L_2-*optimal coupling* for (P, Q) is often used.

For probabilities on the real line, it is well known that the quantile functions associated to P and Q, denote them by F_P^{-1} and F_Q^{-1}, are a $\mathcal{W}_2$-o.t.p.,

$$\mathcal{W}_2(P, Q) = \left(\int_0^1 \left(F_P^{-1}(t) - F_Q^{-1}(t) \right)^2 dt \right)^{1/2}, \tag{4}$$

but for multivariate distributions there is no equivalent explicit expression to compute $\mathcal{W}_2(P, Q)$. A useful fact, that allows to focus on the case of centered probabilities is that if m_P, m_Q are the means of P and Q, and P^*, Q^* are the corresponding centered probabilities, then

$$\mathcal{W}_2^2(P, Q) = \|m_P - m_Q\|^2 + \mathcal{W}_2^2\big(P^*, Q^*\big).$$

In the late 1980s and early 1990s, a series of papers by Brenier [10,11], Cuesta-Albertos and Matrán [18] and Rüschendorf and Rachev [39] put the basis for the analysis of optimal transporting: under continuity assumptions on the probability P, the L_2-o.t.p. (X, Y) for (P, Q) can be represented as $(X, T(X))$ for some suitable map T. Moreover, this *optimal transport map* for (P, Q) coincides with the (essentially unique) cyclically monotone map transporting P to Q.

A very interesting consequence of the characterization of optimal transportation maps is that, independently of the initial distribution, some maps have the optimal transport property between any initial probability P and its transported probability. In particular, if A is a positive definite matrix (here and through the paper we assume that positive definiteness includes symmetry), then (X, AX) is a $\mathcal{W}_2$-o.t.p. independently of the law $\mathcal{L}(X)$. This fact allows to characterize the optimal transport maps between nonsingular normal distributions and yields some additional facts that we quote in the next result, a version of Theorem 2.1 in Cuesta-Albertos *et al.* [20], which, in turn, improves the original statement by Gelbrich [28].

Theorem 2.1. *Let P and Q be probabilities in $\mathcal{P}_2(\mathbb{R}^d)$ with means m_P, m_Q and covariance matrices Σ_P, Σ_Q. If Σ_P is assumed nonsingular, then*

$$\begin{aligned}\mathcal{W}_2^2(P, Q) &\geq \|m_P - m_Q\|^2 + \operatorname{trace}\big(\Sigma_P + \Sigma_Q - 2\big(\Sigma_P^{1/2}\Sigma_Q\Sigma_P^{1/2}\big)^{1/2}\big)\\ &= \mathcal{W}_2^2\big(N(m_P, \Sigma_P), N(m_Q, \Sigma_Q)\big).\end{aligned} \tag{5}$$

Moreover the equality holds if and only if the map $T(x) = (m_Q - m_P) + Ax$ transports P to Q (in particular if P and Q are Gaussian), where A, semidefinite positive, is defined by

$$A := \Sigma_P^{-1/2}\big(\Sigma_P^{1/2}\Sigma_Q\Sigma_P^{1/2}\big)^{1/2}\Sigma_P^{-1/2}, \tag{6}$$

The set $\mathcal{P}_2(\mathbb{R}^d)$ equipped with the $\mathcal{W}_2$-distance is a Polish space (separable and complete metric space) that is often called a *Wasserstein space* and denoted as $W_2(\mathbb{R}^d)$. We can also consider (through a definition of the distance similar to that in (1)) a Wasserstein-type space over other spaces, notably over $\mathcal{P}_2(\mathbb{R}^d)$ leading to $W_2(\mathcal{P}_2(\mathbb{R}^d))$. This space consists of the probability measures, μ, on $\mathcal{P}_2(\mathbb{R}^d)$ (equipped with the Borel σ-field associated to the distance $\mathcal{W}_2$) such that

$$\int_{\mathcal{P}_2(\mathbb{R}^d)} \mathcal{W}_2^2(P, Q)\mu(dP) < \infty, \qquad \text{for some (hence, for every) } Q \in \mathcal{P}_2\big(\mathbb{R}^d\big). \tag{7}$$

Wasserstein distance in this space will be denoted by $\mathcal{W}_{\mathcal{P}_2}$. It is worthwhile to stress that the Wasserstein metric on $W_2(\mathcal{P}_2(\mathbb{R}^d))$ inherits the good properties that it exhibits on $\mathcal{P}_2(\mathbb{R}^d)$ (see Section 5.1). The space $W_2(\mathcal{P}_2(\mathbb{R}^d))$ is in the basis of the (more abstract) framework considered in [31] to generalize (2) to this definition of barycenters.

Definition 2.2. If $\mu \in W_2(\mathcal{P}_2(\mathbb{R}^d))$, then a barycenter of μ is any probability $\bar{\mu} \in \mathcal{P}_2(\mathbb{R}^d)$ such that $\mathcal{W}^2_{\mathcal{P}_2}(\mu, \delta_{\{\bar{\mu}\}}) = \inf\{\mathcal{W}^2_{\mathcal{P}_2}(\mu, \delta_{\{Q\}}),\ Q \in \mathcal{P}_2(\mathbb{R}^d)\}$, that is:

$$\int_{\mathcal{P}_2(\mathbb{R}^d)} \mathcal{W}_2^2(P, \bar{\mu})\mu(dP) = \mathrm{Var}(\mu) := \inf\left\{\int_{\mathcal{P}_2(\mathbb{R}^d)} \mathcal{W}_2^2(P, Q)\mu(dP) : Q \in \mathcal{P}_2\big(\mathbb{R}^d\big)\right\}. \quad (8)$$

We use the notation $\mathrm{Var}(\mu)$ to stress the role of variance of μ played by this quantity. Note that (8) is the natural extension of the already considered barycenters of a finite set of probabilities $P_1, \ldots, P_k \in \mathcal{P}_2(\mathbb{R}^d)$ with weights $\lambda_1, \ldots, \lambda_k$.

It will be convenient to consider a generic probability space $(\Omega, \sigma, \mathbf{P})$ where a measurable random element with values in $\mathcal{P}_2(\mathbb{R}^d)$ (and distribution law μ) is defined. The image of a generic $\omega \in \Omega$ will be denoted as μ_ω. Then equation (8) becomes

$$\int_\Omega \mathcal{W}_2^2(\mu_\omega, \bar{\mu})\mathbf{P}(d\omega) = \inf\left\{\int_\Omega \mathcal{W}_2^2(\mu_\omega, Q)\mathbf{P}(d\omega) : Q \in \mathcal{P}_2\big(\mathbb{R}^d\big)\right\}. \quad (9)$$

Existence of barycenters in this setting has been proved in [31], as well as uniqueness under absolute continuity assumptions (in fact this follows easily from Theorem 2.9 in [2]). Barycenters in Wasserstein space enjoy some continuity properties. We refer to Proposition 5.3 and Theorems 5.4 and 5.5 (which are essentially contained in Theorems 2 and 3 in [31]).

We show next that barycenters in Wasserstein space satisfy an equivariance property with respect to similarity transformations, namely, linear transformations that preserve shape. We recall that these transformations include rotations, reflections, translations and scaling. A proof can be found in Section 5.3.

Proposition 2.3. *Let $\bar{\mu} \in \mathcal{P}_2(\mathbb{R}^d)$ be a barycenter of $\mu \in W_2(\mathcal{P}_2(\mathbb{R}^d))$, and let T be a similarity transformation on $\mathbb{R}^d$. If μ^* is defined as the probability in $W_2(\mathcal{P}_2(\mathbb{R}^d))$ given, through the notation above, by $\mu^*_\omega = \mu_\omega \circ T^{-1}$, then $\bar{\mu} \circ T^{-1}$ is a barycenter of μ^*.*

We close this section with some remarks on the computability of Wasserstein barycenters. In general, it shares the serious computational difficulties inherent to optimal transportation. Explicit expressions are available just for distributions on the real line, a fact that is quoted in the next result.

Proposition 2.4. *If $F_1^{-1}, \ldots, F_k^{-1}$ are the quantile functions associated to probabilities $P_1, \ldots, P_k$ on the real line, and $\lambda_1, \ldots, \lambda_k$ are positive weights with $\sum_{i=1}^k \lambda_i = 1$, then the barycenter of $\{P_i\}_{i=1}^k$ is the probability with quantile function $\sum_{i=1}^k \lambda_i F_i^{-1}$.*

From Proposition 2.4 we see that for k normal distributions, $N(m_i, \sigma_i^2), i = 1, \ldots, k$, on $\mathbb{R}$, the barycenter would be the normal law $N(\sum_{i=1}^k \lambda_i m_i, (\sum_{i=1}^k \lambda_i \sigma_i)^2)$. More generally, for multivariate normal distributions there is an interesting characterization for the barycenter that comes from Knott and Smith [30] (but see also Rüschendorf and Uckelmann [40] and [1]).

Theorem 2.5. *Let $P_i = N(m_i, \Sigma_i)$, $i = 1, \ldots, k$ be normal probabilities on $\mathbb{R}^d$ with positive definite covariances, and $\lambda_1, \ldots, \lambda_k$ positive weights with $\sum_{i=1}^k \lambda_i = 1$. Then the unique barycenter of $P_1, \ldots, P_k$ is the normal law $N(\bar{\mu}, \bar{\Sigma})$, where $\bar{m} = \sum_{i=1}^k \lambda_i m_i$ and $\bar{\Sigma}$ is the only positive definite root of the equation*

$$\sum_{i=1}^{k} \lambda_i \left(\Sigma^{1/2} \Sigma_i \Sigma^{1/2}\right)^{1/2} = \Sigma. \tag{10}$$

Later, in Theorem 3.10, we will generalize this result to probabilities in $W_2(\mathcal{P}_2(\mathbb{R}^d))$ supported in an arbitrary location-scatter family. We note that our proof is elementary and self-contained (in particular, it does not use Theorem 2.5 but only general principles of optimal transportation).

3. Trimmed barycenters

We introduce in this section our approach for a wide consensus representative of a sample $P_1, \ldots, P_k$ of probabilities, with given weights $\lambda_1, \ldots, \lambda_k$. It is based on considering a suitably trimmed subsample. The trimming procedure allows partial discarding of some probabilities, through a suitable reweigthing as in the following definition.

Definition 3.1. Given $0 \le \alpha < 1$ and P a probability on a measurable space (Ω, σ), we say that the probability P^*, also defined on σ, is an α-trimming of P if there exists a measurable function $\tau : \Omega \to \mathbb{R}$ such that $0 \le \tau(\omega) \le 1$ for every $\omega \in \Omega$ and $P^*(A) = \frac{1}{1-\alpha} \int_A \tau(\omega) P(d\omega)$ for every $A \in \sigma$. Such a function is often called an α-trimming function. The set of all α-trimmings of P will be denoted by $\mathcal{T}_\alpha(P)$.

Remark 3.2. A typical trimming function would be the indicator function of a set A with probability $P(A) = 1 - \alpha$. The trimmed probability being then the conditional probability given A. However, our definition even includes the consideration of P, itself, as a trimmed version of P, with associated trimming function $\tau = (1 - \alpha) I_\Omega$.

Since trimmed probabilities and trimming functions are associated in an essentially one to one way, the notation $\mathcal{T}_\alpha(P)$ will be indistinctly used for the set of all α-trimmings of P and for the set of the corresponding trimming functions.

Given $\alpha \in (0, 1)$, and a probability μ on $\mathcal{P}_2(\mathbb{R}^d)$, we look for a $\bar{\mu}^\alpha \in \mathcal{P}_2(\mathbb{R}^d)$ and a probability $\mu^\alpha \in \mathcal{T}_\alpha(\mu)$, with associated trimming function τ_μ^α, which satisfy

$$\int \mathcal{W}_2^2\left(P, \bar{\mu}^\alpha\right) \mu^\alpha(dP) = \mathrm{Var}_\alpha(\mu) := \inf_{\mu^* \in \mathcal{T}_\alpha(\mu), \nu \in \mathcal{P}_2(\mathbb{R}^d)} \int \mathcal{W}_2^2(P, \nu) \mu^*(dP) \tag{11}$$

or, equivalently, in terms of trimming functions,

$$\int \mathcal{W}_2^2\left(P, \bar{\mu}^\alpha\right) \tau_\mu^\alpha(P) \mu(dP) = (1 - \alpha)\, \mathrm{Var}_\alpha(\mu) = \inf_{\tau \in \mathcal{T}_\alpha(\mu), \nu \in \mathcal{P}_2(\mathbb{R}^d)} \int \mathcal{W}_2^2(P, \nu) \tau(P) \mu(dP).$$

Such a $\bar{\mu}^\alpha$ will be called *(α-)trimmed barycenter of* μ and τ_μ^α an *(α-)optimal trimming function*. Similarly to $\mathrm{Var}(\mu)$, the value $\mathrm{Var}_\alpha(\mu)$ will be called the *(α-)trimmed variance* of μ. As usually, the previous definitions apply to any $\mathcal{P}_2(\mathbb{R}^d)$-valued random variable, by identifying these concepts for a random variable with those of its probability distribution. The following theorem (proved in Section 5.4) guarantees the existence of trimmed barycenters.

Theorem 3.3. *Let $\alpha \in (0,1)$ and let μ be a probability defined on $\mathcal{P}_2(\mathbb{R}^d)$. Then, there exists an α-trimmed barycenter of μ, which we will denote as $\bar{\mu}^\alpha$.*

By considering as trimming function (with the corresponding normalizing factor), the indicator set of a large enough ball centered at $\delta_{\{0\}}$, it becomes obvious that the minimum value $\mathrm{Var}_\alpha(\mu)$ must be finite and (recall the definition of $W_2(\mathcal{P}_2(\mathbb{R}^d))$ in (7)) that the set $\mathcal{T}_\alpha(\mu)$ can be substituted by the subset $\mathcal{T}_\alpha(\mu) \cap W_2(\mathcal{P}_2(\mathbb{R}^d))$. Since every probability on $W_2(\mathcal{P}_2(\mathbb{R}^d))$ has a barycenter, obviously $\bar{\mu}^\alpha$ must be a barycenter of μ^α, which justifies the notation we are using. Furthermore, and similar to the impartially trimmed means, trimmed barycenters must simultaneously be the barycenter of the trimmed distribution and the center of its support. To formalize this fact, we define

$$r_\alpha(P) := \inf\big\{r > 0 : \mu\big[B_{\mathcal{W}}(P,r)\big] \geq 1-\alpha\big\}. \tag{12}$$

It trivially follows that if $r < r_\alpha(P)$, then $\mu[B_{\mathcal{W}}(P,r)] < 1-\alpha$ and

$$\mu\big[B_{\mathcal{W}}\big(P, r_\alpha(P)\big)\big] \leq 1-\alpha \leq \mu\big[\overline{B}_{\mathcal{W}}\big(P, r_\alpha(P)\big)\big].$$

This is the key to the following result.

Proposition 3.4. *Let $\alpha \in (0,1)$, $\nu \in \mathcal{P}_2(\mathbb{R}^d)$ and $\tau^* \in \mathcal{T}_\alpha(\mu)$ be such that*

$$I_{B_{\mathcal{W}}(\nu, r_\alpha(\nu))} \leq \tau^* \leq I_{\overline{B}_{\mathcal{W}}(\nu, r_\alpha(\nu))}, \tag{13}$$

then, for every $\tau \in \mathcal{T}_\alpha(\mu)$, we have

$$\int \mathcal{W}_2^2(P,\nu)\tau^*(P)\mu(dP) \leq \int \mathcal{W}_2^2(P,\nu)\tau(P)\mu(dP). \tag{14}$$

Proof. Let $\tau \in \mathcal{T}_\alpha(\mu)$ and consider the real r.v. $X(P) := \mathcal{W}_2^2(P,\nu)$. It is clear that the distribution of X, when we consider in $\mathcal{P}_2(\mathbb{R}^d)$ the probability μ trimmed through the trimming function τ^*, is stochastically smaller than that associated to any other τ. Therefore (14) holds. □

Note that equality in (14) is only possible if (13) happens for τ. Thus, the optimal trimming functions must satisfy (13) where ν must be a barycenter of the trimmed probability associated to τ^*. In other words, the optimal trimming functions are essentially defined by the indicator of a ball centered at a trimmed barycenter.

We turn now to consistency of trimmed barycenters. Theorem 3.5 (see Section 5.4 for a proof) guarantees it under weak consistency of the probability distributions. Note that, unlike in the case of (non trimmed) barycenters, it is not necessary that $\mathcal{W}_2(\mu_n, \mu) \to 0$, but it suffices to assume that $\mu_n \to_w \mu$.

Theorem 3.5. *Let $(\mu_n)_n, \mu$ be probabilities on $\mathcal{P}_2(\mathbb{R}^d)$ such that $\mu_n \to_w \mu$. For a fixed $\alpha \in (0,1)$, let $\bar{\mu}_n^\alpha$ be any trimmed barycenter of μ_n. Then the trimmed variances converge, namely, $\mathrm{Var}_\alpha(\mu_n) \to \mathrm{Var}_\alpha(\mu)$, the sequence $(\bar{\mu}_n^\alpha)_n$ is precompact for the $\mathcal{W}_2$ topology and any limit is a trimmed barycenter of μ. If μ has only one trimmed barycenter, $\bar{\mu}^\alpha$, then $\mathcal{W}_2(\bar{\mu}_n^\alpha, \bar{\mu}^\alpha) \to 0$.*

Repeating the argument that we use for law of large numbers for barycenters (Theorem 5.5), we obtain from Theorem 3.5 the corresponding one for trimmed barycenters. We state the result under the additional hypothesis of uniqueness of the trimmed barycenter of the probability law. This kind of assumption is quite common when showing consistency of centralization measures to avoid complicated or too simplistic statements with complicated proofs even on $\mathbb{R}^k$ (see, for instance, Cuesta-Albertos and Matrán [17]). If the μ-probability of the set of absolutely continuous probabilities in $\mathcal{P}_2(\mathbb{R}^d)$ is greater than $1-\alpha$, the support of every α-trimmed version of μ would contain absolutely continuous probabilities, thus it would have only one barycenter. Therefore, lack of uniqueness of the trimmed barycenter should be provoked by particular configurations of μ. For example, for the uniform distribution on $[0,1]$, every point in the set $[(1-\alpha)/2, (1+\alpha)/2]$ is an α-trimmed mean. Section 5 in García-Escudero *et al.* [27] treats this problem, although in practice it is quite rare to find distributions where uniqueness fails and, even then, the lack of uniqueness could be only due to an improper choice of α.

Theorem 3.6. *Assume that μ is a probability on the space $\mathcal{P}_2(\mathbb{R}^d)$ with a unique trimmed barycenter. If μ_n is the sample probability giving mass $1/n$ to the probabilities $P_1, \ldots, P_n$ obtained as independent realizations of μ, then the trimmed barycenters and variances are strongly consistent: $\bar{\mu}_n^\alpha \to_{a.s.} \bar{\mu}^\alpha$, and $\mathrm{Var}_\alpha(\mu_n) \to_{\text{a.s.}} \mathrm{Var}_\alpha(\mu)$.*

3.1. Location-scatter families

Computation of Wasserstein distances and of barycenters for probabilities on the real line can be done through the explicit characterizations given in (4) and Proposition 2.4. In the multivariate setting, Proposition 2.4 can be extended to probabilities that can be parameterized in terms of a location and a scatter matrix, generalizing the normal multivariate model.

Definition 3.7. Let $\mathcal{M}_{d\times d}^+$ be the set of $d \times d$ positive definite matrices and let $\mathbf{X}_0$ be a random vector with probability law $P_0 \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$. The set

$$\mathcal{F}(P_0) := \left\{ \mathcal{L}(A\mathbf{X}_0 + m) : A \in \mathcal{M}_{d\times d}^+, m \in \mathbb{R}^d \right\}$$

of probability laws induced by positive definite affine transformations from P_0 will be called a *location-scatter family*.

As an easy consequence of Theorem 2.1, any probability $P \in \mathcal{F}(P_0)$ can be optimally transported to any other $Q \in \mathcal{F}(P_0)$ through an affine transformation with positive definite matrix. Thus w.l.o.g. we can assume that the mean of P_0 is the vector $\bar{0}$ and its covariance matrix is I_d, the identity matrix. Also note that to make reference to a probability in $\mathcal{F}(P_0)$ we could use its

mean m, and the transformation A or alternatively $\Sigma = A^2$, the corresponding covariance matrix. We will use the second option to share the usual notation in the normal model. Therefore, $\mathbb{P}_{m,\Sigma}$ will denote the probability in $\mathcal{F}(P_0)$ with mean m and covariance matrix Σ.

In the statistical literature, a location and scatter family usually refers to an elliptical model. However, the families considered in this work under this denomination include the elliptical families, but also families induced by different shapes. For instance, if we take in $\mathbb{R}^2$ the probability P_0 whose marginals are independent standard normal and exponential, respectively, then the family $\mathcal{F}(P_0)$ is not elliptical. We also note that to address a confidence set problem, P_0 and the choice of any measurable set $\mathcal{M}_\gamma$ in $\mathbb{R}^d$, such that $P_0(\mathcal{M}_\gamma) = \gamma$, will play the role of shape of the reference set. A typical asymptotic pivotal function for a parameter $\theta \in \mathbb{R}^d$ has the structure $n^{1/2}\hat{V}_n^{-1/2}(\hat{\theta}_n - \theta)$, thus, if we approximately know its law, P_0, then the set $\{\hat{\theta}_n - n^{-1/2}\hat{V}_n^{1/2}x : x \in \mathcal{M}_\gamma\}$ would be an approximate confidence set of level γ. Therefore, the estimation of the location and scatter in the family $\mathcal{F}(P_0)$ produces a confidence set of the desired level, and a consensus based estimation would automatically produce a consensus confidence set for the parameter.

We show in Theorems 3.8 and 3.10 below that Wasserstein barycenters and trimmed barycenters of probabilities supported on a location-scatter family belong to the location-scatter family, or, in other words, that location-scatter families *are closed for barycenters*. Of course, the general equivariance result for similarity transformations (recall Proposition 2.3) remains true in the location-scatter setup. We also include a Gelbrich's type result showing that the dispersion in the $\mathcal{W}_2$-sense is minimized just when the probabilities belong to a common location-scatter family, in particular when all the probabilities are normal. The proof can be found in Section 5.5.

Theorem 3.8. *Let $\{P_i\}_{i=1}^k$ be probabilities in $\mathcal{P}_{2,ac}(\mathbb{R}^d)$ with means $m_i, i = 1, \ldots, k$, and non-singular covariance matrices $\Sigma_i, i = 1, \ldots, k$. Let $N_i = N(m_i, \Sigma_i), i = 1, \ldots, k$, be normal probability distributions on $\mathbb{R}^d$. Also let $P_0 \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$ and let us denote by $\mathbb{P}_{m,\Sigma}$ the probability in $\mathcal{F}(P_0)$ with mean m and covariance matrix Σ.*

Let us consider $\lambda_1, \ldots, \lambda_k$ positive weights with $\sum_{i=1}^k \lambda_i = 1$, and respectively denote by $\bar{P}$, $\bar{N}$ and $\bar{\mathbb{P}}$ the (unique) barycenters of $\{P_i\}_{i=1}^k$, $\{N_i\}_{i=1}^k$ and $\{\mathbb{P}_{m_i,\Sigma_i}\}_{i=1}^k$. Then we have:

$$\sum_{i=1}^k \lambda_i \mathcal{W}_2^2(P_i, \bar{P}) \geq \sum_{i=1}^k \lambda_i \mathcal{W}_2^2(\mathbb{P}_{m_i,\Sigma_i}, \bar{\mathbb{P}}) = \sum_{i=1}^k \lambda_i \mathcal{W}_2^2(N_i, \bar{N}). \tag{15}$$

Moreover, the inequality in (15) can be an equality only if the mean and covariance matrix of $\bar{P}$ coincide with those of $\bar{N}$ and the relation $\{P_i\}_{i=1}^k \subset \mathcal{F}(\bar{P})$ holds.

Remark 3.9. We stress the fact that Theorem 3.8 generalizes (with the same proof but adding some notational complexity) to any $\mu \in W_2(\mathcal{P}_2(\mathbb{R}^d))$ if, using the notation employed in (9), we assume that for every $\omega \in \Omega$, $\mu_\omega \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$ with mean m_ω and covariance matrix $\Sigma_\omega \in \mathcal{M}_{d\times d}^+$.

Theorem 3.10. *Let $P_0 \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$, and $\mu \in W_2(\mathcal{P}_2(\mathbb{R}^d))$. With the notation in Remark 3.9, assume that for every $\omega \in \Omega$, the probability $\mu_\omega \in \mathcal{F}(P_0)$. Then the unique barycenter, $\bar{\mu}$, of μ*

also belongs to $\mathcal{F}(P_0)$. *The mean of* $\bar{\mu}$ *is* $\bar{m} := \int m_\omega \mathbf{P}(d\omega)$, *and the covariance matrix,* $\bar{\Sigma}$, *is the only positive definite matrix satisfying*

$$\bar{\Sigma} = \int \left(\bar{\Sigma}^{1/2}\Sigma_\omega \bar{\Sigma}^{1/2}\right)^{1/2} \mathbf{P}(d\omega).$$

Once we know that a family is closed for barycenters, the property will be shared by the trimmed barycenters. This is motivated by the fact that trimmed versions of a probability μ have their supports contained in that of μ, and a trimmed barycenter is characterized as a barycenter of an optimal trimmed version of μ. Once a trimming function has been fixed, the uniqueness of the barycenter of absolutely continuous distributions, obtained in [14], leads also to the uniqueness of the trimmed barycenter associated to that trimmed version of μ. However, we cannot deduce uniqueness of the trimmed barycenters in an easy way. In fact, this is a hard problem even for trimmed means in euclidean spaces.

Corollary 3.11. *Assume that* $P_0 \in \mathcal{P}_{2,ac}$ *and* μ *a probability on* $\mathcal{P}_2(\mathbb{R}^d)$ *that is supported in* $\mathcal{F}(P_0)$. *Then, for every* $\alpha \in (0,1)$, *any trimmed barycenter of* μ *also belongs to* $\mathcal{F}(P_0)$. *Moreover, any optimal trimming function for* μ *uniquely determines a trimmed barycenter.*

Remark 3.12. We emphasize the importance of Corollary 3.11 that allows to search for trimmed barycenters of, say a random normal distribution, looking just to the means and covariance functions. Moreover, by Theorem 2.1, the distance between probabilities in $\mathcal{F}(P_0)$ is given by

$$\mathcal{W}_2^2(\mathbb{P}_{m_1,\Sigma_1}, \mathbb{P}_{m_2,\Sigma_2}) = \|m_1 - m_2\|^2 + \operatorname{trace}\left(\Sigma_1 + \Sigma_2 - 2\left(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2}\right)^{1/2}\right), \tag{16}$$

which allows computation of Wasserstein distances. With applications in view, these facts will be complemented with the proposal of a feasible algorithm for addressing the computation of the trimmed barycenter of a finite set of probabilities that belong to a location-scatter family and a given set of weights.

Once this theory has been developed it can be argued that (16) is just a combination of metrics: the Euclidean metric for the means plus another one between covariance matrices. Since the final product only involves distributions in $\mathcal{F}(P_0)$, which are fully determined by the location and the scatter parameters, the problem is parametric and some comparison with other specialized metrics to analyze these parameters could be in order. Focusing on the metric on the covariance matrices, Fréchet means related to several metrics on the space of symmetric positive definite matrices have been proposed in the literature. Among these metrics particular attention is deserved by the affine-invariant metrics and Log-Euclidean metrics, introduced by considerations that mainly arise from the image analysis framework (see Arsigny *et al.* [5]). In both cases, the associated Fréchet means can be considered as generalizations of the geometric mean, although the Log-Euclidean mean could be preferred by its easier computation. We should note that our choice of (16) is not guided by the search for a metric on this set of matrices, but it is rather the restriction of a metric on the set of all probabilities with finite second moment – a kind of L_2 space– with suitable properties already pointed out in the literature in different scenarios. We note also that the computation of Wasserstein barycenters can be efficiently done through the

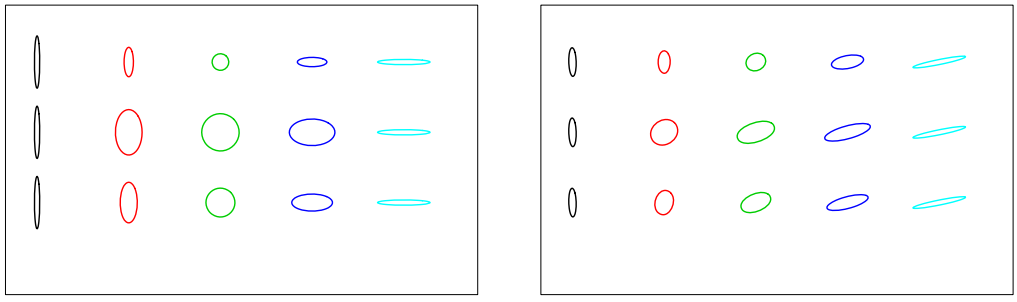

Figure 1. Each picture shows the effects of linear interpolation of two matrices corresponding to the weighted mean (middle row), barycenter (lower row) and Log-Euclidean mean (upper row), of the matrices represented by the black and cyan ellipses. From left to right we handle the weights 1, 0.75, 0.50, 0.25, 0 on the black one. Note the characteristic swelling effect associated to the weighted mean.

algorithm introduced in [4] and discussed in Section 4. Taking this into account, the comparison must rely on purely statistical arguments, like those involving the comparison between the mean and the geometric mean for real numbers. Any of them can be preferred for different tasks but, arguably, the usual mean is the preferred choice in most of the applications. To provide some illustrative idea of their relative behavior, in Figure 1 we resort to the comparison of the interpolation of two pairs of covariance matrices represented by the black and cyan ellipses in each picture. Notice that the average (the weighted mean of covariances) is included for reference. The upper, middle and lower rows are respectively associated to Log-Euclidean, average and barycenter approaches. The red, green and blue ellipses respectively represent the solutions associated to 0.75, 0.5 and 0.25 weights on the black covariance matrix (and 0.25, 0.5 and 0.75 on the cyan one). Additionally, we include in Figure 2 the density functions of three centered normal distributions accompanied by those associated to these approaches. For very similar standard

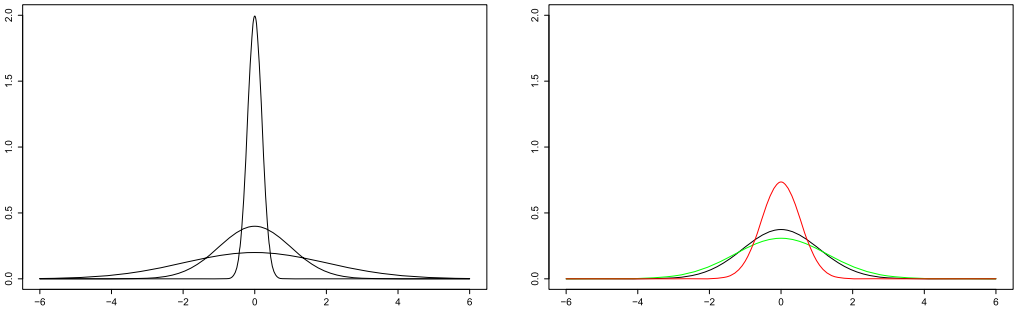

Figure 2. Left: Density functions corresponding to $N(0,\sigma_i)$ distributions for $\sigma_1 = 0.2, \sigma_2 = 1, \sigma_3 = 2$. Right: Normal density functions for $\sigma = (1/3\sum_{i=1}^3 \sigma_i^2)^{1/2} = 1.296$ (green), $\sigma = (\prod_{i=1}^3 \sigma_i)^{1/3} = 0.737$ (red) and $\sigma = 1/3\sum_{i=1}^3 \sigma_i = 1.067$ (black), respectively associated to the mean of variances and the geometric mean (or Log-Euclidean) of the variances, and the barycenter of the distributions.

deviations $\{\sigma_i\}_{i=1}^k$ and any associated weights $\{\lambda_j\}_{j=1}^k$, the three aggregation procedures would produce nearly the same result but, if this is not the case, the estimates can be very different.

An explanation for these different behaviors comes from Jensen's inequality. In the simplest one-dimensional case, these three averaging procedures result in standard deviations given by the left (Log-Euclidean), middle (Wasserstein barycenter) and right (weighted average of variances) terms in the following inequalities

$$\exp\left(\sum_{j=1}^k \lambda_j \log \sigma_j\right) \leq \sum_{j=1}^k \lambda_j \sigma_j \leq \sqrt{\sum_{j=1}^k \lambda_j \sigma_j^2}. \tag{17}$$

This shows that the standard deviation of the geometric mean is smaller than the average of the standard deviations which in turn is smaller than the standard deviation arising from the weighted mean of the variances. This gives some explanation to the swelling effect associated to the weighted mean. We also note that if we are willing to admit that the standard deviation $(\int |x|^2 P(dx))^{1/2}$ is a good measurement of the size of a centered distribution, P, then the Log-Euclidean mean results in summaries which are smaller than the average size of the objects to be summarized. In this sense, the Wasserstein barycenter provides the better choice between these alternatives.

For diagonal (in some basis) covariance matrices, this explains the intermediate size of the barycenter, avoiding the swelling effect of the mean of variances, but also the somewhat excessive decrease associated to the Log-Euclidean approach. In a location scatter model, for a finite collection $\{\mathbb{P}_{0,\Sigma_j}\}_{j=1}^k$ and weights $\{\lambda_j\}_{j=1}^k$, and if the principal directions of the Σ_j matrices are the same, then for some orthonormal matrix H, $\Sigma_j = H D_j H^t$, $j = 1, \ldots, k$ with $D_j = \mathrm{diag}(\sigma_{j1}^2, \ldots, \sigma_{jd}^2)$. If we denote by $\Sigma^*, \bar{\Sigma}, \widehat{\Sigma}$ the covariance matrices associated to the Log-Euclidean, Wasserstein barycenter and weighted average approaches, then also $\Sigma^* = H D^* H^t$, $\bar{\Sigma} = H \bar{D} H^t$, $\widehat{\Sigma} = H \widehat{D} H^t$, with $D^* = \mathrm{diag}(\sigma_1^{*2}, \ldots, \sigma_d^{*2})$, $\bar{D} = \mathrm{diag}(\bar{\sigma}_1^2, \ldots, \bar{\sigma}_d^2)$, $\widehat{D} = \mathrm{diag}(\widehat{\sigma}_1^2, \ldots, \widehat{\sigma}_d^2)$, which are related by $\sigma_j^* \leq \hat{\sigma}_j \leq \widehat{\sigma}_j$, $j = 1, \ldots, d$, from (17) because $\sigma_j^* = \exp(\sum_{i=1}^k \lambda_i \log \sigma_{ji})$, $\hat{\sigma}_j = \sum_{i=1}^k \lambda_i \sigma_{ji}$ and $\widehat{\sigma}_j{}^2 = \sum_{i=1}^k \lambda_i \sigma_{ji}{}^2$. Also note that in this case we obtain again that the "standard deviation" of the Barycenter is the weighted mean of the standard deviations.

Although the fact just noticed will not be true in full generality, we will show below that such weighted mean of standard deviations is an upper bound for the standard deviation of the barycenter. We would like to stress that this result will be proved for probabilities that do not necessarily belong to a location-scatter family. Even more, by Remark 3.4 in [4], the property is true even without the absolutely continuous assumption that we will impose here for a simpler argument.

Proposition 3.13. *Let $P_1, \ldots, P_k \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$ centered in mean, and $\lambda_1, \ldots, \lambda_k$ be positive weights adding one. If $\bar{P}$ is the associated barycenter, then*

$$\left(\int \|x\|^2 \bar{P}(dx)\right)^{1/2} \leq \sum_{j=1}^k \lambda_j \left(\int \|x\|^2 P_j(dx)\right)^{1/2}.$$

4. Computation of barycenters and trimmed barycenters

The characterization of trimmed barycenters given in Proposition 3.4 leads to consider the effective computation of barycenters as a first step in the obtention of trimmed barycenters. We recall for probabilities on $\mathbb{R}$ the characterization given in Proposition 2.4 in terms of quantiles. If μ is the probability on $\mathcal{P}_2(\mathbb{R})$ giving weights $\lambda_1, \ldots, \lambda_k$ to the probabilities $P_1, \ldots, P_k$, then the barycenter $\bar{\mu}$ is the distribution of the random variable $\sum_{j=1}^k \lambda_j F_j^{-1}$ (defined on the unit interval), thus denoting by m_j and σ_j^2 the mean and variance of P_j, and $\bar{m}$ and $\mathrm{Var}(\bar{\mu})$ those of $\bar{\mu}$:

$$\mathrm{Var}(\mu) = \sum_{j=1}^k \lambda_j (m_j - \bar{m})^2 + \sum_{j=1}^k \lambda_j \sigma_j^2 - \mathrm{Var}(\bar{\mu}). \tag{18}$$

When $P_1, \ldots, P_k$ belong to a common location-scale family, $\mathcal{F}(P_0)$, where P_0 has quantile function F_0^{-1} (with zero mean and variance 1), then $F_j^{-1} = m_j + \sigma_j F_0^{-1}$, $j = 1, \ldots, k$, and with $\bar{\sigma} := \sum_{j=1}^k \lambda_j \sigma_j$, (18) specializes to

$$\mathrm{Var}(\mu) = \sum_{j=1}^k \lambda_j (m_j - \bar{m})^2 + \sum_{j=1}^k \lambda_j (\sigma_j - \bar{\sigma})^2 = \sum_{j=1}^k \lambda_j \left(m_j^2 + \sigma_j^2\right) - \left(\bar{m}^2 + \bar{\sigma}^2\right).$$

In contrast, as previously noted, in the multivariate case closed expressions are only available just for situations essentially equivalent to several univariate cases. This is the case if, e.g. the probabilities share a common structure of dependence in some particular basis (see Section 2 in Cuesta-Albertos *et al.* [22] or Section 4 in [8]), or if they are radial transformations of a common probability law (see Section 3 in [22]). Turning to approximate computations, in recent times some papers addressed the goal of numerical computation of Wasserstein barycenters, see Cuturi and Doucet [23], Benamou *et al.* [6] or Carlier *et al.* [14]. In these cases, the approaches address the case of sample distributions or are based on the discretization of the problem through a fine grid and the use of suitable optimization procedures. Although their results allow to get good representations for the barycenter of distributions with very different shapes, the grid sizes for suitably approximating the distributions must be large and would strongly depend on the dimension making them highly time-consuming even in small dimensions and with a small number of distributions. Of course these procedures allow computation of barycenters, but regrettably, under trimming, the available methods to compute the trimmed barycenters (even for real random variables), like our Algorithm for the trimmed barycenter below, need several initializations and often require the iterative computation of several thousands of barycenters. This makes those algorithms based on discretizations to be, by now, inapplicable for our proposes. Fortunately, for one of the most important cases in multivariate statistics, namely the location-scatter families, a fast consistent procedure for approximating the numerical solution of equation (10) has recently been introduced in [4]. We give here a quick description of the procedure.

Assume that $P_1, \ldots, P_k \in \mathcal{P}_{2,ac}$ and the weights $\lambda_1, \ldots, \lambda_k$ are fixed. Given $\eta \in \mathcal{P}_2(\mathbb{R}^d)$, we consider the functional

$$V(\eta) := \sum_{i=1}^{k} \lambda_i \mathcal{W}_2^2(\eta, P_j),$$

looking for $\bar{P} \in \mathcal{P}_2(\mathbb{R}^d)$ such that

$$V(\bar{P}) = \min_{\eta \in \mathcal{P}_2(\mathbb{R}^d)} V(\eta).$$

If $\eta \in \mathcal{P}_{2,ac}$, we know that there exist optimal transport maps T_j from η to P_j. Assume that X is a random vector with law η, thus

$$\mathcal{L}\big(T_j(X)\big) = P_j, \quad \text{and} \quad \mathcal{W}_2^2(\eta, P_j) = \mathbf{E}\big\|X - T_j(X)\big\|^2, \qquad j = 1, \ldots, k.$$

With this notation we define

$$G(\eta) := \mathcal{L}\left(\sum_{j=1}^{k} \lambda_j T_j(X)\right),$$

to design a consistent, iterative procedure for the approximate computation of $\bar{P}$. Next, we collect some basic properties of G that show a link between the G transform and the barycenter problem.

Proposition 4.1. *If $\eta \in \mathcal{P}_{2,ac}$ then*

$$V(\eta) \geq V\big(G(\eta)\big) + \mathcal{W}_2^2\big(\eta, G(\eta)\big).$$

In particular, if the barycenter, $\bar{P}$, is absolutely continuous then $G(\bar{P}) = \bar{P}$.

We remark that the hypothesis of absolute continuity of $\bar{P}$ is required just to guarantee that $G(\bar{P})$ is defined. The theory developed for the location-scatter families, and particularly for normal distributions, allows to guarantee this in such cases. On the other hand, the conclusion of the proposition invites to consider an iterative process, starting from any $\eta_0 \in \mathcal{P}_{2,ac}$ and considering the sequence

$$\eta_{n+1} := G(\eta_n), \qquad n \geq 0. \tag{19}$$

We have proved the consistency of this iterative procedure in greater generality in [4], but for our present purposes it suffices that given in the following statement.

Theorem 4.2. *If $P_1, \ldots, P_k$ are nonsingular Gaussian distributions on $\mathbb{R}^d$ and the initial measure, η_0, is also a nonsingular Gaussian distribution, then the iteration defined by (19) is consistent, namely,*

$$\mathcal{W}_2(\eta_n, \bar{P}) \to 0,$$

as $n \to \infty$, where $\bar{P}$ is the (unique) barycenter of $P_1, \ldots, P_k$.

It is time to recall Theorem 3.10 on barycenters of location-scatter families. We know from it that, given positive definite matrices $\Sigma_1, \ldots, \Sigma_k$ there exists a unique positive definite matrix $\bar{\Sigma}$ solving (10).

Reading Theorem 4.2 just in terms of approximating the unique solution of (10), the conclusion becomes that if, starting from any positive definite matrix S_0, according to Theorem 2.1, we define

$$S_{n+1} = S_n^{-1/2}\left(\sum_{j=1}^{k} \lambda_j \left(S_n^{1/2}\Sigma_j S_n^{1/2}\right)^{1/2}\right)^2 S_n^{-1/2}, \tag{20}$$

then

$$\lim_{n\to\infty} S_n = \bar{\Sigma}.$$

Therefore the process leads to a consistent iterative method for approximating the solution of (10). The method is easily implemented and, in practice, shows a very good performance. We refer to [4] for further details.

The characterization of the distance between probabilities in the location-scatter family (16) leads to identical distances to those between normal laws with same location and covariance matrices. Therefore we can extend Theorems 2.5 and 4.2 in the following way.

Theorem 4.3. *If μ is the probability on $\mathcal{P}_2(\mathbb{R}^d)$ giving weights $\lambda_1, \ldots, \lambda_k$ to the probabilities $\mathbb{P}_{m_1,\Sigma_1}, \ldots, \mathbb{P}_{m_k,\Sigma_k} \in \mathcal{F}(P_0)$, a location-scatter family with $P_0 \in \mathcal{P}_{2,ac}(\mathbb{R}^d)$, then its barycenter is the probability $\mathbb{P}_{\bar{m},\bar{\Sigma}} \in \mathcal{F}(P_0)$, where $\bar{m} = \sum_{i=1}^k \lambda_i m_i$ and $\bar{\Sigma}$ is the only definite positive matrix satisfying equation (10). Moreover, $\bar{\Sigma}$ can be obtained as the limit of the sequence defined in (20). The variance of μ takes the value*

$$\begin{aligned}\mathrm{Var}(\mu) &= \sum_{j=1}^{k} \lambda_j \|m_j - \bar{m}\|^2 + \sum_{j=1}^{k} \lambda_j \,\mathrm{trace}(\Sigma_j - \bar{\Sigma}) \\ &= \sum_{j=1}^{k} \lambda_j \left(\|m_j\|^2 + \mathrm{trace}(\Sigma_j)\right) - \left(\|\bar{m}\|^2 + \mathrm{trace}(\bar{\Sigma})\right).\end{aligned}$$

Through Theorem 4.3 we can compute barycenters and variances for any finite set of probabilities and weights, once we know the corresponding locations $m_1, \ldots, m_k$ and covariance matrices $\Sigma_1, \ldots, \Sigma_k$. Moreover, the distances between probabilities are also easily computed through (16), which is valid for every location-scatter family. Therefore, Corollary 3.11 and the characterization of the best trimming functions given in Proposition 3.4 allow to search for a trimmed barycenter as the barycenter based on subsets of $P_1, \ldots, P_k$ with an accumulate weight of at least $1 - \alpha$ and minimum variance after normalizing the weights.

Next, we include an algorithm to obtain the trimmed barycenter of the probabilities $\mathbb{P}_{m_1,\Sigma_1}, \ldots, \mathbb{P}_{m_k,\Sigma_k} \in \mathcal{F}(P_0)$ with weights $\lambda_1, \ldots, \lambda_k$. It combines estimation and concentration steps, being an adaptation of usual algorithms for obtaining best (in some sense) trimmed regions, like the ones involved in the MCD or LTS robust estimators, with the necessary updates

of the distances and weights in each concentration step. Once an initial solution is provided, this kind of algorithm guarantees convergence through the estimation and concentration steps, but we must also consider the possibility of local optimizers, a fact that leads to consider random choices of initial candidates to be compared at the end. We simply emphasize the fact that this algorithm shares the good performance of the versions currently used in similar problems on estimation in the multivariate setting.

The algorithm.

0. Fix $n = 0$, and randomly choose initial candidates $\hat{m}_n$, $\hat{\Sigma}_n$ for the mean and the covariance matrix.
1. Compute the distances d_i^n between $\mathbb{P}_{\hat{m}_n,\hat{\Sigma}_n}$ and $\mathbb{P}_{m_i,\Sigma_i}$, $i = 1,\dots,k$, through (16).
2. Consider the permutation $((1),\dots,(k))$ such that $d^n_{(1)} \le \cdots \le d^n_{(k)}$.
3. Set $j_n = \inf\{j : \sum_{i\le j} \lambda_{(i)} \ge 1-\alpha\}$ and define the new weights:

$$\lambda^n_{(i)} = \begin{cases} \lambda_{(i)} & \text{if } i < j_n, \\ 1-\alpha-\displaystyle\sum_{i<j_n} \lambda^n_{(i)} & \text{if } i = j_n, \\ 0 & \text{if } i > j_n. \end{cases}$$

4. Since $\sum_{i=1}^k \lambda^n_{(i)} = 1-\alpha$, define $\lambda^n_{(i)} = (1-\alpha)^{-1}\lambda^n_{(i)}$, in order to have $\sum_{i=1}^k \lambda^n_{(i)} = 1$.
5. Using the updated weights, compute $\hat{m}_{n+1}$, the weighted mean of the means, and $\hat{\Sigma}_{n+1}$ through the recursive algorithm (20).
6. Iterate Steps 1 through 5 until convergence.
7. Compute the variance of the final trimmed sample of probabilities and weights.
8. Go to 0 and finalize after a moderate number of initial choices, reporting the barycenter producing the minimum variance.

As a toy illustration of the results of the computation of the barycenters (trimmed or not), we present now two examples, in which we handle 2-dimensional normal distributions, allowing a suitable visualization of the results. In these examples, we represent graphically a normal distribution with mean m and covariance matrix Σ by the set

$$\left\{x \in \mathbb{R}^2 : (x-m)^t \Sigma^{-1}(x-m) = 1\right\}.$$

Example 4.4. We have considered first the six normal distributions represented in the graph in the left hand side in Figure 3. We have computed the barycenter, and the $1/6$ and $2/6$ trimmed barycenters of these normal distributions. The results appear in the right hand side graphic. All three barycenters are normal distributions which are represented by the black, blue and red ellipses in the right hand side graphic in Figure 3.

The black ellipse is the non-trimmed barycenter. Trimming $\alpha = 1/6$ the barycenter is the blue ellipse, and the procedure trims the blue ellipse in the left graphic. The red ellipse shows the result of trimming $\alpha = 2/6$. In this case, the procedure trims the red and the blue ellipses in the left hand side graphic. Observe that the red ellipse lies in the middle of the four black ellipses in the left graphic showing a very similar shape.

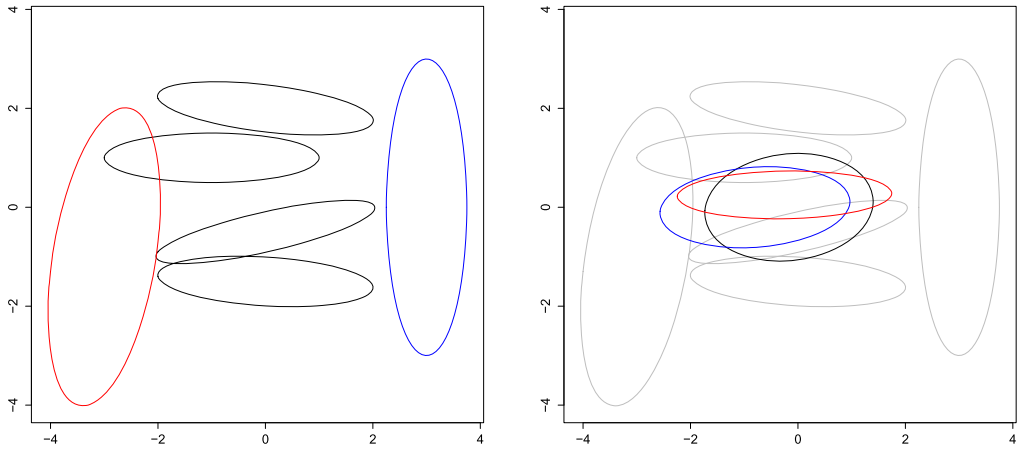

Figure 3. Computation of trimmed barycenters (right figure) of the ellipses shown in the figure in the left.

The previous result could have been anticipated because, according to (16), the decision of which distributions to trim depends on the shape and the location of the ellipses under consideration and, in this example, the colored ellipses have more different shapes and separated locations than the others. Because of this, we also show a not too big modification of this example which is shown in Figure 4. Here five ellipses coincide with the corresponding ones in Figure 3. However, the green ellipse in the left hand graph in this figure is one of the "horizontal" ellipses whose center has been moved two units along the ordinates axis. Now, it happens that the trimmed

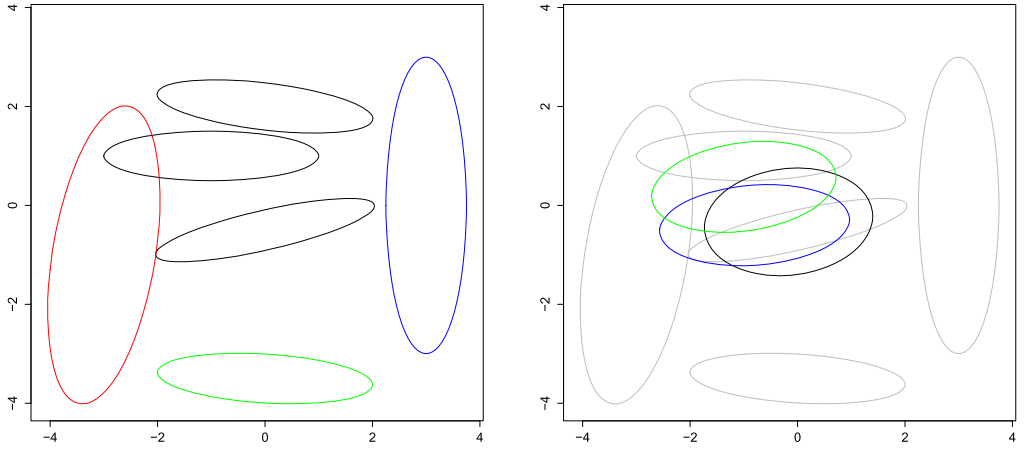

Figure 4. Computation of trimmed barycenters (right figure) of the ellipses shown in the figure in the left.

distribution when taking $\alpha = 1/6$ continues being the blue one, but when taking $\alpha = 2/6$ the procedure trims the blue and the green ellipses leaving the red one untrimmed.

Example 4.5. Let us assume that we are carrying out an experiment in $k = 100$ hospitals on a 2-dimensional r.v., that we are taking a sample with size $n = 100$ in each hospital, and that each hospital is sending only its own estimation of the mean and the covariance matrix based on the sample in its study.

Let us also assume that the population is divided in two subpopulations. The first subpopulation is composed by 90% of individuals and the distribution of the variable of interest in this subpopulation is standard normal, while the distribution in the second subpopulation is also normal, with the identity as covariance matrix and the mean at $(4, 4)$. The real goal of the study is the estimation of the parameters of the majority, the second subpopulation being considered as composed by outliers.

The statistician in charge of the experiment, being aware of these issues, decides that each hospital uses the Minimum Covariance Determinant method (MCD, proposed in Rousseeuw [37]), based on 80% of the points in its sample to estimate the mean and covariance matrix of the people in its area (similar results could be obtained through the procedure developed in Cuesta-Albertos *et al.* [19]), the reason to choose these estimators being that the probability of obtaining more than 20 outliers in a binomial sample with parameters $n = 100$ and $p = 0.1$ being 0.00081 and, as long as we obtain less than 20 outliers in a sample with size 100, the MCD method will give a fair estimation of the parameters in the main subpopulation.

However, it happens that, unknown to the statistician, the population is relatively heterogeneous, and that, in fact, the proportion of people in a given area belonging to the second subpopulation is chosen using a distribution Beta with parameters $(4, 36)$, which gives a global proportion of 0.1, but irregularly scattered.

We have made a simulation of this process resulting that 5 hospitals have got more than 20 outliers, leading to largely wrong estimations of the parameters. The results of this experiment appear in the left hand side graph in Figure 5. There, most estimations appear in grey, but a few of them have been drawn in black to give a general idea of the objects we have obtained in the first part of the process.

The right-hand side graph presents the area inside the square in the left hand side graph with some summarizing possibilities for the estimations shown in the left graph. Here the red ellipse represents the standard normal distribution (which can be considered as our target since this distribution produced most of the data in the analyzed samples). The green ellipse represents the normal distribution whose mean (resp. covariance matrix) is the sample mean of the estimated means (resp. covariance matrices). This estimator is not expected to be particularly good.

The magenta ellipse represents the (non-trimmed) barycenter. This estimation is affected by the anomalous estimations (but less that the previous one). The blue ellipse represents the 0.2-trimmed barycenter which, practically, matches the target.

Example 4.6. The Palomar Data is a data set considered in Rousseeuw and Driessen [38], consisting in astronomic measurements recorded at the California Institute of Technology within the Digitized Palomar Sky Survey. The set handled here, kindly shared by the authors, is the same analyzed in that paper, containing 132,402 observations in 6 variables. The analysis there

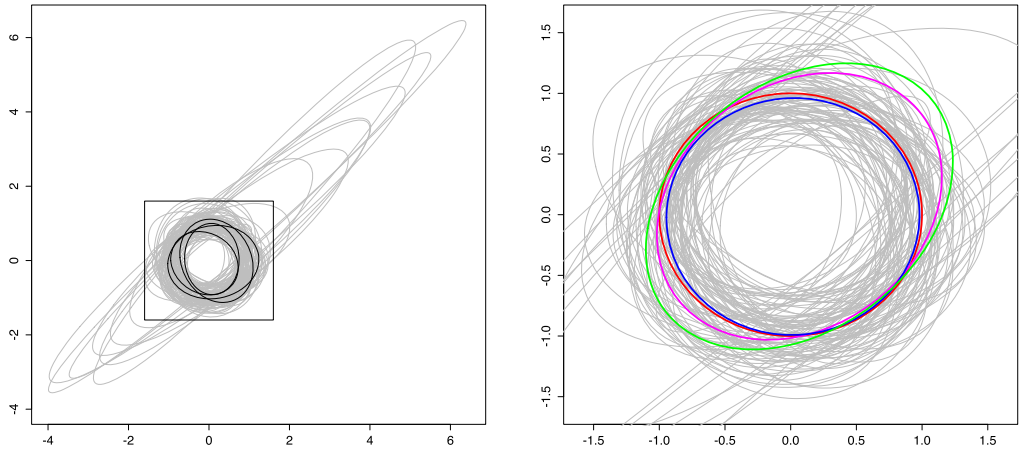

Figure 5. Computation of trimmed barycenters (right figure) of the ellipses shown in the left figure.

showed the interest of considering robust estimations of the covariance matrix and related metrics instead of the crude Mahalanobis distance, obtained through the sample covariance matrix. In fact, through a plot of MCD-based robust Mahalanobis distances, they found evidence on the existence of several groups in the data and, as a key part of the fast MCD algorithm for large data sets, introduced a pooling strategy on the initial subsets of the data leading to the better solutions. Our approach looks for the comparison between the MCD solution achieved for the whole data set and those provided from 100 randomly chosen subsamples of size 5000. Figure 6 is a plot on the two first variables (MAperF and csfF) of the data. It shows (gray) the 100 ellipses associated to the MCD's based on subsamples and that of the MCD based on the full sample (black dashed). It also includes the ellipses that result from several aggregations of the MCD's produced by the subsamples. The green one is just that associated to the mean of the 100 covariance matrices and centered in the mean of the 100 means estimations. In black, red, blue and magenta are represented the trimmed barycenters of the 100 MCD's respectively corresponding to the trimming levels $\alpha = 0.1, 0.2, 0.3, 0.4$. Figure 7 is the plot of trimmed variations vs. trimming levels associated to the 100 MCD's solutions.

Through these pictures we have a nice summary. From both figures, it becomes apparent that nearly 35% of the solutions correspond to ellipses centered around $(18500, 1000)$ with little variation within this group, while the remaining 65% are very similar to the MCD obtained with the complete sample. This implies that the right solutions should be selected when trimming, at least, that (35%) proportion. In agreement with the conclusions of the analysis carried in [38], such behavior would suggest the existence of at least two main bulks of data. Although most samples have a proportion of data coming from these bulks that justify the MCD based on the complete sample, small variations in these proportions would consistently produce a very different MCD. In this situation, aggregation methods based on simple average would typically produce bad solutions, while monitoring the trimmed barycenter solutions allows a well-founded, stable, "wide consensus" proposal.

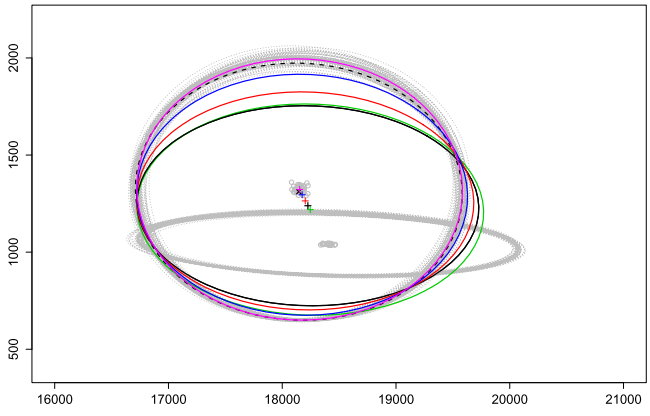

Figure 6. Graph summarizing the MCD's solutions obtained from 100 subsamples of size 5000 (gray) and those provided by several aggregations of the solutions and that given by the full Palomar Data set (black dashed). The green ellipse is associated to the mean of the solutions; those in black, red, blue and magenta correspond to trimmed barycenters with respective trimming sizes 0.1, 0.2, 0.3 and 0.4.

To give evidence of feasibility of the proposal, we give below some details on the execution times of the involved procedures. Computations have been carried on a MacBook Pro with a 4 Ghz processor Intel Core i7 and 16 Gb of RAM. The MCD's have been computed with TCLUST (available at the CRAN, see Fritz *et al.* [26]), an R application for model based robust clustering. The parameters for the solution based on the full sample were $k = 1$, alpha $= 0.5$, nstart $= 150$, restr.fact $=$ 1e10, iter.max $= 200$, equal.weights $= F$. The only change in these parameters for the subsamples was iter.max that was set to 100. The computations of trimmed barycenters have

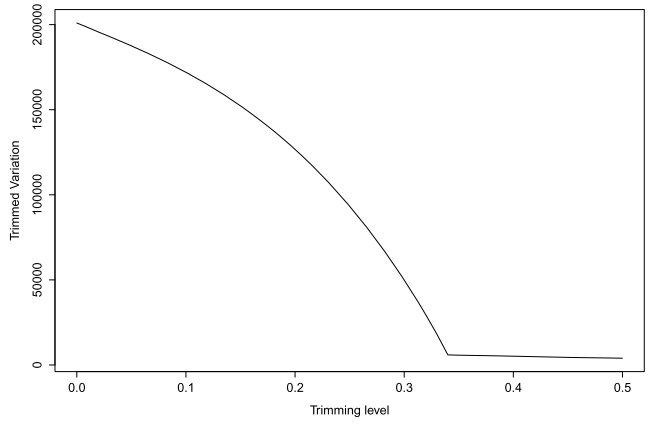

Figure 7. Plot showing the evolution of the trimmed variations vs trimming levels on the 100 MCD's solutions in the Palomar Data example.

been also carried into the R framework, with programs based on the algorithm presented in this section.

Runtimes in seconds: For the large MCD (sample of size 132,402) 120.497 sec; for 100 MCD's (on samples of size 5000) 45.125 sec; for the 0.3-trimmed barycenter of the 100 MCD's solutions 30.985 sec; for the (51) α-trimmed barycenters and trimmed variances (to produce the plot in Figure 7, $\alpha = k/100$ for $k = 0, \ldots, 50$) 1744.315 sec. Handling the MCD based on the complete sample as reference, the squares of the Wasserstein distances to the average solution and to those given by the trimmed barycenters for 0.1, 0.2, 0.3, 0.4 were respectively: 87,260.07, 71,459.66, 33,953.8, 6426.18, 357.25.

Repeating the whole process under the same conditions, but with subsample sizes of 10,000 instead of 5000, the only runtime that changed was the one corresponding to the 100 MCD's (on samples of size 10,000) 110.007 sec. The squares of the distances were now: 73,850.38, 54,857.57, 21,578.75, 1517.071, 175.90.

5. Technical details and proofs

5.1. Supplementary results on Wasserstein spaces

For ease of reference, we include in this section some relevant results on Wasserstein spaces for reference through the work. From a technical point of view a great deal of interest on the Wasserstein distance $\mathcal{W}_2$ comes from the fact that it metrizes the weak convergence of probabilities plus the convergence of their second order moments: Given $(P_n)_n \subset \mathcal{P}_2(\mathbb{R}^d)$ and $P \in \mathcal{P}_2(\mathbb{R}^d)$,

$$\begin{aligned} &\mathcal{W}_2(P_n, P) \to 0 \quad \text{if and only if} \\ &P_n \to_w P \quad \text{and} \quad \int_{\mathbb{R}^d} \|x\|^2 P_n(dx) \to \int_{\mathbb{R}^d} \|x\|^2 P(dx). \end{aligned} \tag{21}$$

More generally, the following theorem gives a very useful characterization (see, e.g., Theorem 7.12 in [42]) of the convergence in the space $W_2(\mathcal{P}_2(\mathbb{R}^d))$.

Theorem 5.1. *Let $(\mu_n)_n$ and μ be in $W_2(\mathcal{P}_2(\mathbb{R}^d))$, and consider the probability degenerated at zero, $\delta_{\{0\}}$ (that can be substituted by any other fixed probability in $\mathcal{P}_2(\mathbb{R}^d)$). Convergence $\mathcal{W}_{\mathcal{P}_2}(\mu_n, \mu) \to 0$ holds if and only if:*

$$\mu_n \to_w \mu \quad \text{and} \quad \lim_{R\to\infty} \limsup_{n\to\infty} \int_{\mathcal{W}_2(\delta_{\{0\}}, P)>R} \mathcal{W}_2^2(\delta_{\{0\}}, P)\mu_n(dP) = 0. \tag{22}$$

Proposition 5.2. *If the sequences $(\mu_n)_n, (\nu_n)_n$ in $W_2(\mathcal{P}_2(\mathbb{R}^d))$, verify $\mu_n \to_w \mu$ and $\nu_n \to_w \nu$, then $\mathcal{W}_{\mathcal{P}_2}(\mu, \nu) \le \liminf \mathcal{W}_{\mathcal{P}_2}(\mu_n, \nu_n)$. Moreover, if the convergences are in the sense showed in (22), then the convergence $\mathcal{W}_{\mathcal{P}_2}(\mu_n, \nu_n) \to \mathcal{W}_{\mathcal{P}_2}(\mu, \nu)$ holds.*

Note that the uniform integrability condition in (22) is similar to the uniform integrability condition of $\|x\|^2$ in (21).

Existence and continuity properties of barycenters in $W_2(\mathcal{P}_2(\mathbb{R}^d))$ are guaranteed by Proposition 5.3 and Theorem 5.4 to be stated next. They follow from the results in [31].

Proposition 5.3. *If $\mu \in W_2(\mathcal{P}_2(\mathbb{R}^d))$ and every P in the support of μ is absolutely continuous, then the barycenter of the random measure μ exists and it is unique.*

Theorem 5.4. *Let $(\mu_j)_{j=1}^{\infty} \subset W_2(\mathcal{P}_2(\mathbb{R}^d))$ and set $\bar{\mu}_j$ a barycenter of μ_j, for all $j = 1, \ldots$. Suppose that for some $\mu \in W_2(\mathcal{P}_2(\mathbb{R}^d))$ we have that $\mathcal{W}_{\mathcal{P}_2}(\mu, \mu_j) \to 0$. Then, the sequence $(\bar{\mu}_j)_{j=1}^{\infty}$ is precompact in $\mathcal{P}_2(\mathbb{R}^d)$ and any limit is a barycenter of μ.*

In particular, when the limit distribution μ has only one barycenter, this theorem ensures convergence in $\mathcal{P}_2(\mathbb{R}^d)$ of the barycenters to that of μ. In a sample setting, when the probability measures μ_n are the sample ones giving weight $1/n$ to the first n probabilities $P_1, \ldots, P_n$ obtained as realizations of the random probability measure $\mu \in W_2(\mathcal{P}_2(\mathbb{R}^d))$, by Varadarajan theorem, $\mu_n \to_w \mu$ almost surely. Now let us consider the probability degenerated at zero, $\delta_{\{0\}}$. Since the classical Strong Law of Large Numbers applied to the real i.i.d. random variables $\mathcal{W}_2^2(P_i, \delta_{\{0\}})$ gives

$$\int_{\mathcal{P}_2(\mathbb{R}^d)} \mathcal{W}_2^2(P, \delta_{\{0\}}) \mu_n(dP) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{W}_2^2(P_i, \delta_{\{0\}}) \to_{\text{a.s.}} \int_{\mathcal{P}_2(\mathbb{R}^d)} \mathcal{W}_2^2(P, \delta_{\{0\}}) \mu(dP),$$

the characterization in Theorem 5.1 of convergence in the $\mathcal{W}_{\mathcal{P}_2}$ sense, through Theorem 5.4, proves the following Strong Law of Large Numbers for barycenters.

Theorem 5.5. *Assume that $\mu \in W_2(\mathcal{P}_2(\mathbb{R}^d))$ and that the barycenter of μ is unique. If μ_n is the sample probability giving mass $1/n$ to the probabilities $P_1, \ldots, P_n$ obtained as independent realizations of μ, then the barycenters are consistent, i.e. $\mathcal{W}_2(\bar{\mu}_n, \bar{\mu}) \to_{\text{a.s.}} 0$.*

5.2. Overview on trimming

In this section, we recall some important properties of probability trimmings and obtain new results of interest in our current framework. In particular, we emphasize those connected with Wasserstein spaces and distances. We begin providing a list of statements arising from [2], that can be easily translated to the framework of Polish spaces (metrizable, separable and complete spaces).

Proposition 5.6. *Let P be a probability in any mesurable space (Ω, σ) and $\alpha \in [0, 1)$. The following statements are equivalent*:

(a) *The probability P^* is a trimmed version of P.*
(b) *P^* is absolutely continuous with respect to P, and $\frac{dP^*}{dP} \leq \frac{1}{1-\alpha}$.*
(c) *$(1-\alpha)P^*(A) \leq P(A)$ for every set $A \in \sigma$.*

Proposition 5.7. *Let P be a probability in any abstract space and T a measurable map taking values in a Polish space. If T transports P to Q, then for every α*

$$\mathcal{T}_\alpha(Q) = \big\{P^* \circ T^{-1} : P^* \in \mathcal{T}_\alpha(P)\big\}.$$

Proposition 5.8. *Let (E, d) be a Polish space and $\alpha \in (0, 1)$.*

(a) *If P is any probability measure on (E, d), then $\mathcal{T}_\alpha(P)$ is compact for the topology of weak convergence.*
(b) *If $\{P_n\}_n$ is a tight sequence of probabilities on (E, d) and $P_n^* \in \mathcal{T}_\alpha(P_n)$ for every n, then $\{P_n^*\}_n$ is tight. Moreover, if $P_n \to_w P$ and $P_n^* \to_w P^*$, then $P^* \in \mathcal{T}_\alpha(P)$.*

Proposition 5.9. *If $0 < \alpha < 1$ and $P \in W_2(\mathcal{P}_2(\mathbb{R}^d))$ or $P \in \mathcal{P}_2(\mathbb{R}^d)$, then $\mathcal{T}_\alpha(P)$ is compact in the $\mathcal{W}_2$ topology.*

Proof. The proof given in [2] for $P \in \mathcal{P}_2(\mathbb{R}^d)$ quickly extends to the case $P \in W_2(\mathcal{P}_2(\mathbb{R}^d))$ by handling the uniformly integrability condition in Theorem 5.1. □

Proposition 5.10. *Let $0 < \alpha < 1$, $\{P_n\}_n$ and P be probabilities on a Polish space (E, d), and assume that $P_n \to_w P$. Then, if $P^* \in \mathcal{T}_\alpha(P)$, there exists a sequence $\{P_n^*\}_n$ such that $P_n^* \in \mathcal{T}_\alpha(P_n)$, for all n, and $P_n^* \to_w P^*$.*

Proof. Use Skorohod's Representation Theorem (see, e.g., Theorem 11.7.2 in Dudley [25]), to obtain E-valued measurable maps $X, X_1, \ldots$ defined on a probability space $(\Omega, \sigma, \mathbf{P})$ such that $\mathcal{L}(X_n) = P_n$, $\mathcal{L}(X) = P$, and $X_n \to X$, $\mathbf{P}$-a.s.

By Proposition 5.7, $P^* \in \mathcal{T}_\alpha(P)$ can be represented as $P^* = \mathbf{P}^* \circ X^{-1}$ for some $\mathbf{P}^* \in \mathcal{T}_\alpha(\mathbf{P})$. By considering $P_n^* := \mathbf{P}^* \circ X_n^{-1}$, we obtain probabilities in $\mathcal{T}_\alpha(P_n)$, that obviously converge weakly to P^* because $X_n \to X$ also $\mathbf{P}^*$-a.s. □

Remark 5.11. Note that any kind of uniform integrability condition like the one in (22) verified for some sequence $\{P_n\}_n$ is automatically shared for any sequence $\{P_n^*\}_n$ such that $P_n^* \in \mathcal{T}_\alpha(P_n)$ for every n. Therefore Proposition 5.10 and (22) imply that if $P_n \to P$ in $\mathcal{W}_{\mathcal{P}_2}$, then the sequence $\{P_n^*\}_n$ is precompact in $\mathcal{W}_{\mathcal{P}_2}$ and any limit belongs to $\mathcal{T}_\alpha(P)$.

5.3. Proofs of Propositions 2.3 and 3.13

Proof of Proposition 2.3. A similarity transformation, T, can be expressed as a linear transformation $T = cA + b$, where c is a constant, A an orthogonal transformation and $b \in \mathbb{R}^d$. If (X, Y) is an $\mathcal{W}_2$-o.t.p. for the probabilities (P, Q), and (AX^*, AY^*) is an $\mathcal{W}_2$-o.t.p. for $(P \circ A^{-1}, Q \circ A^{-1})$ then we have

$$\begin{aligned}\mathcal{W}_2^2(P, Q) &= E\|X - Y\|^2 = E\|AX - AY\|^2 \geq \mathcal{W}_2^2\big(P \circ A^{-1}, Q \circ A^{-1}\big) \\ &= E\big\|AX^* - AY^*\big\|^2 = E\big\|X^* - Y^*\big\|^2 \geq \mathcal{W}_2^2(P, Q),\end{aligned}$$

hence $\mathcal{W}_2^2(P\circ A^{-1}, Q\circ A^{-1}) = \mathcal{W}_2^2(P,Q)$, and for T we easily obtain $\mathcal{W}_2^2(P\circ T^{-1}, Q\circ T^{-1}) = c^2\mathcal{W}_2^2(P,Q)$. Therefore, for every Q, we have

$$\begin{aligned}\int_\Omega \mathcal{W}_2^2\big(\mu_\omega\circ T^{-1}, \bar\mu\circ T^{-1}\big)\mathbf{P}(d\omega) &= c^2\int_\Omega \mathcal{W}_2^2(\mu_\omega,\bar\mu)\mathbf{P}(d\omega)\\ &\le c^2\int_\Omega \mathcal{W}_2^2(\mu_\omega, Q)\mathbf{P}(d\omega)\\ &= \int_\Omega \mathcal{W}_2^2\big(\mu_\omega\circ T^{-1}, Q\circ T^{-1}\big)\mathbf{P}(d\omega).\end{aligned}$$

Since T is invertible, denoting $S = T^{-1}$, every Q can be written as $Q = Q^*\circ S^{-1}$ for some Q^*, hence we deduce that

$$\begin{aligned}&\int_\Omega \mathcal{W}_2^2\big(\mu_\omega\circ T^{-1}, \bar\mu\circ T^{-1}\big)\mathbf{P}(d\omega)\\ &\quad\le \int_\Omega \mathcal{W}_2^2\big(\mu_\omega\circ T^{-1}, Q^*\big)\mathbf{P}(d\omega)\qquad \text{for every } Q^*\in\mathcal{P}_2\big(\mathbb{R}^d\big).\end{aligned}$$

□

Proof of Proposition 3.13. Let X be a random vector with $\mathcal{L}(X) = \bar P$, and T_j, $j = 1,\dots,k$ be optimal transport maps for $(\bar P, P_j)$. Denoting $X_j = T_j(X)$, we know that $\mathcal{L}(X_j) = P_j$ but also, by Proposition 4.1, $\bar P = \mathcal{L}(\sum_{j=1}^k \lambda_j X_j)$. Therefore, by Minkowski inequality, we have

$$\big(E\|X\|^2\big)^{1/2} = \left(E\left\|\sum_{j=1}^k \lambda_j X_j\right\|^2\right)^{1/2} \le \sum_{j=1}^k \big(E\|\lambda_j X_j\|^2\big)^{1/2} = \sum_{j=1}^k \lambda_j\big(E\|X_j\|^2\big)^{1/2}.$$

□

5.4. Existence and consistency of the trimmed barycenter

Let us begin noting that, under the additional assumption $\mu\in W_2(\mathcal{P}_2(\mathbb{R}^d))$, the results would easily follow from Theorem 5.4 and the compactness of the set $\mathcal{T}_\alpha(\mu)$ stated in Proposition 5.9. However, as stated in Theorem 3.3, that assumption is not needed at all.

Proof of Theorem 3.3. Recall definition (11) and assume that $\mu_n^*\in\mathcal{T}_\alpha(\mu)$ and $\nu_n\in\mathcal{P}_2(\mathbb{R}^d)$ verifying

$$\int \mathcal{W}_2^2(P,\nu_n)\mu_n^*(dP)\to \mathrm{Var}_\alpha(\mu). \tag{23}$$

We already know that $\mathrm{Var}_\alpha(\mu)$ is finite and that we can assume that every μ_n^* in the the minimizing sequence belongs to $W_2(\mathcal{P}_2(\mathbb{R}^d))$, hence the ν_n's can be chosen as their barycenters. Thus, we will take $\nu_n = \bar\mu_n^*$.

The next step is to show that the sequence $\{\int\|x\|^2\bar\mu_n^*(dx)\}_n$ as well as that of their associated radii $\{r_\alpha(\bar\mu_n^*)\}_n$ (defined in (12)) are bounded. For the sake of readability, we state this result as a lemma.

Lemma 5.12. *Let $\{Z_n\}_n$ be a sequence of r.v.'s defined on some probability space $(\Omega, \sigma, \mathbf{P})$ such that $\mathbf{P}_{Z_n} = \bar{\mu}_n^*$ for every $n \in \mathbb{N}$. Then, it happens that $M := \sup_n \mathbf{E}[\|Z_n\|^2] < \infty$. Moreover, the sequence $\{r_\alpha(\bar{\mu}_n^*)\}_n$ is bounded.*

Proof. Take $r_0 > 0$ such that $p := \mu[B_{\mathcal{W}}(\delta_{\{0\}}, r_0)] > \alpha$. Let us assume that there exists a subsequence such that $\lim_n \mathbf{E}(\|Z_{k_n}\|^2) = \infty$. For this subsequence, we have that if $P \in B_{\mathcal{W}}(\delta_{\{0\}}, r_0)$, then, since $\mathcal{W}_2$is a metric,

$$\mathcal{W}_2\big(P, \bar{\mu}_{k_n}^*\big) \geq \mathcal{W}_2\big(\delta_{\{0\}}, \bar{\mu}_{k_n}^*\big) - \mathcal{W}_2(P, \delta_{\{0\}}) \geq \big(\mathbf{E}\big(\|Z_{k_n}\|^2\big)\big)^{1/2} - r_0,$$

and, consequently, since $p > \alpha$,

$$\int \mathcal{W}_2^2\big(P, \bar{\mu}_{k_n}^*\big)\mu_{k_n}^*(dP) \geq \big(\big(\mathbf{E}\big(\|Z_{k_n}\|^2\big)\big)^{1/2} - r_0\big)^2(p - \alpha) \to \infty,$$

which contradicts the minimizing property (23) of the chosen sequence with $\mathrm{Var}_\alpha(\mu) < \infty$. Thus, the sequence $\{\mathbf{E}[\|Z_n\|^2]\}_n$ is bounded. Now, if $P \in B_{\mathcal{W}}(\delta_{\{0\}}, r_\alpha(\delta_{\{0\}}))$, then

$$\mathcal{W}_2\big(P, \bar{\mu}_n^*\big) \leq \mathcal{W}_2(P, \delta_{\{0\}}) + \mathcal{W}_2\big(\delta_{\{0\}}, \bar{\mu}_n^*\big) \leq r_\alpha(\delta_{\{0\}}) + \big(E\big(\|Z_n\|^2\big)\big)^{1/2},$$

which implies that the set $B_{\mathcal{W}}(\delta_{\{0\}}, r_\alpha(\delta_{\{0\}}))$ is a subset of the ball with center at $\bar{\mu}_n^*$ and radius $r_\alpha(\delta_{\{0\}}) + (E(\|Z_n\|^2))^{1/2}$, and therefore, $r_\alpha(\bar{\mu}_n^*) \leq r_\alpha(\delta_{\{0\}}) + \sup_m (E(\|Z_m\|^2))^{1/2}$ for every $n \in \mathbb{N}$. □

Returning to the proof of Theorem 3.3, note that, by the first result of the lemma, $\{\bar{\mu}_n^*\}_n$ is tight, so w.l.o.g. we can assume that it converges in distribution to some $\nu_0 \in \mathcal{P}_2(\mathbb{R}^d)$. Moreover, by the lemma, the supports of the associated trimmed probabilities μ_n^* are contained in a common ball $B_{\mathcal{W}}(\delta_{\{0\}}, M + \sup\{r_\alpha(\bar{\mu}_n^*), n \in \mathbb{N}\})$ in $\mathcal{P}_2(\mathbb{R}^d)$, thus we can also assume that it converges to some $\mu_0^* \in \mathcal{T}_\alpha(\mu)$ weakly and (by uniform integrability) in $\mathcal{W}_2$. This implies, by Theorem 5.4, that the limit ν_0 of the barycenters must be a barycenter of μ_0^* and that the convergence is also in $\mathcal{W}_2$.

By continuity of $\mathcal{W}_2$we have $\mathcal{W}_{\mathcal{P}_2}(\mu_n^*, \delta_{\{\bar{\mu}_n^*\}}) \to \mathcal{W}_{\mathcal{P}_2}(\mu_0^*, \delta_{\{\nu_0\}})$, leading also to

$$\lim_n \int \mathcal{W}_2^2\big(P, \bar{\mu}_n^*\big)\mu_n^*(dP) = \lim_n \mathcal{W}_2^2\big(\mu_n^*, \delta_{\{\bar{\mu}_n^*\}}\big) = \mathcal{W}_2^2\big(\mu_0^*, \delta_{\{\nu_0\}}\big) = \int \mathcal{W}_2^2(P, \nu_0)\mu_0^*(dP),$$

that shows that ν_0 is a trimmed barycenter of μ. □

An easy modification of this proof allows to guarantee a consistency result in the sense of Theorem 5.4 also without the integrability assumption.

Proof of Theorem 3.5. Since $\mu_n \to_w \mu$, we can choose a large enough $M > 0$ such that $\mu_n[B_{\mathcal{W}}(\delta_{\{0\}}, M)] > 1 - \alpha$ for every $n \in \mathbb{N}$. This implies that there exist trimmed versions $\mu_n^* \in \mathcal{T}_\alpha(\mu_n)$ with support contained in $B_{\mathcal{W}}(\delta_{\{0\}}, M)$. Therefore, we have that $\mathrm{Var}_\alpha(\mu_n) \leq \mathcal{W}_{\mathcal{P}_2}(\mu_n^*, \delta_{\{\delta_{\{0\}}\}}) \leq M$, and $\limsup_n \mathrm{Var}_\alpha(\mu_n) \leq M < \infty$.

From this point, we can repeat the proof of Lemma 5.12 to guarantee that the sequence of trimmed barycenters is contained in a large enough ball $B_{\mathcal{W}}(\delta_{\{0\}}, M)$ and that the sequence of associated radii $(r_\alpha(\bar{\mu}_n^*))_n$ is bounded. The argument at the end of the proof of Theorem 3.3 applies also here to prove that weakly convergent subsequences of trimmed barycenters must converge also in $\mathcal{W}_2$ and that the limits must be trimmed barycenters of the limit law μ. □

5.5. Proofs of Theorems 3.8 and 3.10

Recall that $\mathcal{F}(P_0) := \{\mathcal{L}(A\mathbf{X}_0 + m) : A \in \mathcal{M}_{d\times d}^+, m \in \mathbb{R}^d\}$ is the location-scatter family induced by positive definite affine transformations from the law $P_0 = \mathcal{L}(\mathbf{X}_0)$. We assume throughout that P_0 is absolutely continuous as an easy way to guarantee uniqueness of optimal transport maps and of barycenters, but much of the following analysis does not depend of this assumption. As we already noted, we can assume w.l.o.g. that P_0 has zero mean and covariance matrix I_d. The probabilities in $\mathcal{F}(P_0)$ are represented as $\mathbb{P}_{m,\Sigma}$, where m is the mean, and Σ the covariance matrix of the probability under consideration.

Relation (16) allows to extend Theorem 2.5 to any family $\mathcal{F}(P_0)$ in a simple way. However, we will give a direct proof. For this task, let us include the following proposition already obtained in Cuesta-Albertos *et al.* [21]. It will allow us to guarantee that barycenters of families of absolutely continuous probabilities in $\mathcal{P}_2(\mathbb{R}^d)$ cannot be degenerated on subspaces of dimension lower than d.

Proposition 5.13. *Let $P, Q \in \mathcal{P}_2(\mathbb{R}^d)$. Let us assume that $P \in \mathcal{P}_{2,ac}$ and that Q is supported on the subspace generated by the first q components of $\mathbb{R}^d$, with $q < d$. Denote by $T^{1,\dots,q}$ the $\mathcal{W}_2$ optimal map transporting the marginal probability, $P^{1,\dots,q}$, of P on that subspace to Q. Then the map $T(x_1, \dots, x_d) := T^{1,\dots,q}(x_1, \dots, x_q)$ is a $\mathcal{W}_2$ optimal map transporting P to Q.*

Proposition 5.14. *Let $\mu \in W_2(\mathcal{P}_2(\mathbb{R}^d))$ and, using the notation employed in (9), assume that for every $\omega \in \Omega$, the probability μ_ω is absolutely continuous. Then, the barycenter of μ cannot be supported on an affine subspace of dimension $q < d$.*

Proof. Let $\mu \in W_2(\mathcal{P}_2(\mathbb{R}^d))$, such that μ_ω is absolutely continuous for every $\omega \in \Omega$ and let m_ω be the mean of μ_ω. Under these conditions, existence and uniqueness of the barycenter are guaranteed by Proposition 5.3. Since it is trivial to show that the mean of the barycenter coincides with $\int_\Omega m_\omega \mathbf{P}(d\omega)$, we can simplify the problem by considering centered in mean distributions (that is, $m_\omega = 0$ for every ω) which remain absolutely continuous. Let $\bar{\mu}$ be the barycenter (with zero mean) of μ, so suppose that it is supported on a subspace (instead of a general affine subspace) of dimension $q < d$. We can assume, w.l.o.g., that $\bar{\mu}$ is supported on the subspace corresponding to the first q components. Let $\mu_\omega^{1,\dots,q}$ denote the marginal of μ_ω on this subspace. Since $\mu_\omega^{1,\dots,q} \ll \ell_q$, we know that there exists an optimal map $T_\omega^{1,\dots,q}$ transporting $\mu_\omega^{1,\dots,q}$ to $\bar{\mu}$. From the previous proposition, the map T_ω defined by $T_\omega(x_1, \dots, x_d) := T_\omega^{1,\dots,q}(x_1, \dots, x_q)$ is

an optimal map transporting μ_ω to $\bar{\mu}$. Therefore we have

$$\mathcal{W}_2^2(\mu_\omega, \bar{\mu}) = \mathcal{W}_2^2\big(\mu_\omega^{1,\dots,q}, \bar{\mu}\big) + \sum_{j=q+1}^{d} \mathcal{W}_2^2\big(\mu_\omega^j, \delta_{\{0\}}\big), \tag{24}$$

where μ_ω^j is the jth marginal of μ_ω.

Let us consider the probability $\mu^* := \bar{\mu} \times \bar{\mu}^{q+1} \times \cdots \times \bar{\mu}^d$ and denote by $\bar{\mu}^j$ the barycenter of the probability $\mu^j \in W_2(\mathcal{P}(\mathbb{R}))$, which is not degenerated because $\mu_\omega^j \ll \ell_1$ for every j (recall the comments preceding Theorem 2.5). Thus, from (24), we have

$$\begin{aligned}
\int_\Omega \mathcal{W}_2^2(\mu_\omega, \bar{\mu})\mathbf{P}(d\omega) &= \int_\Omega \mathcal{W}_2^2\big(\mu_\omega^{1,\dots,q}, \bar{\mu}\big)\mathbf{P}(d\omega) + \sum_{j=q+1}^{d} \int \mathcal{W}_2^2\big(\mu_\omega^j, \delta_{\{0\}}\big)\mathbf{P}(d\omega) \\
&> \int_\Omega \mathcal{W}_2^2\big(\mu_\omega^{1,\dots,q}, \bar{\mu}\big)\mathbf{P}(d\omega) + \sum_{j=q+1}^{d} \int_\Omega \mathcal{W}_2^2\big(\mu_\omega^j, \bar{\mu}^j\big)\mathbf{P}(d\omega) \\
&= \int_\Omega \mathcal{W}_2^2\big(\mu_\omega, \mu^*\big)\mathbf{P}(d\omega),
\end{aligned}$$

contradicting the character of barycenter of $\bar{\mu}$. □

Proof of Theorem 3.8. Let $P \in \mathcal{P}_2(\mathbb{R}^d)$ and let N be a normal law with the same mean and covariance matrix as P. From Gelbrich's bound (5), we have $\mathcal{W}_2^2(P_i, P) \geq \mathcal{W}_2^2(N_i, N)$ for $i = 1, \dots, k$, hence

$$\sum_{i=1}^{k} \lambda_i \mathcal{W}_2^2(P_i, P) \geq \sum_{i=1}^{k} \lambda_i \mathcal{W}_2^2(N_i, N) \geq \sum_{i=1}^{k} \lambda_i \mathcal{W}_2^2(N_i, \bar{N}). \tag{25}$$

Moreover, according to Theorem 2.1, equality in the first inequality is only possible if $P_i \in \mathcal{F}(P), i = 1, \dots, k$. On the other hand, let $\mathbb{P}^*$ be the probability law in $\mathcal{F}(P_0)$ with the same mean and covariance matrix as the barycenter $\bar{N}$ of $\{N_i\}_{i=1}^k$. Then we have

$$\sum_{i=1}^{k} \lambda_i \mathcal{W}_2^2(N_i, \bar{N}) = \sum_{i=1}^{k} \lambda_i \mathcal{W}_2^2\big(\mathbb{P}_{m_i, \Sigma_i}, \mathbb{P}^*\big) \geq \sum_{i=1}^{k} \lambda_i \mathcal{W}_2^2(\mathbb{P}_{m_i, \Sigma_i}, \bar{\mathbb{P}}).$$

Particularizing the first inequality in (25) for $P_i = \mathbb{P}_{m_i, \Sigma_i}, i = 1, \dots, k$ and $P = \bar{\mathbb{P}}$, the concatenation with the last chain of inequalities gives that a normal law with the same mean and covariance matrix as $\bar{\mathbb{P}}$ would be a barycenter for $\{N_i\}_{i=1}^k$. The uniqueness of this barycenter implies that $\bar{\mathbb{P}}$ and $\bar{N}$ must have the same mean and covariance matrix.

The proof ends by considering $P = \bar{P}$ in (25) because both equalities would imply that the mean and the covariance matrix of $\bar{P}$ must coincide with those of $\bar{N}$ and also that $\bar{P}$ can be obtained from every P_i through a positive definite transformation. By Proposition 5.14 these

covariance matrices must be nonsingular, thus the barycenters, in particular $\bar{P}$, must be also absolutely continuous and every P_i can be obtained from $\bar{P}$ through a positive definite affine transformation, thus $\{P_i\}_{i=1}^k \subset \mathcal{F}(\bar{P})$ holds. □

The following lemma can be proved through elementary arguments (see, e.g., equation (18) in [4]) and will be used in the proof of uniqueness involved in Theorem 3.10.

Lemma 5.15. *Let $\Sigma_i, i = 0, 1, 2$ be positive definite matrices and define*

$$\Sigma_{0,2} := \Sigma_0^{-1/2}\big(\Sigma_0^{1/2}\Sigma_2\Sigma_0^{1/2}\big)\Sigma_0^{-1/2}.$$

Let $X_i, i = 1, 2$ be random vectors on $\mathbb{R}^d$ with nonsingular respective laws $\mathbb{P}_{0,\Sigma_i} \in \mathcal{F}(P_0)$, $i = 1, 2$. Then the inequality

$$\mathcal{W}_2^2\big(N(0,\Sigma_1), N(0,\Sigma_2)\big) \geq \operatorname{trace}\big((I_d - \Sigma_{0,2})\Sigma_1\big) + \mathbf{E}\big(\|X_2\|^2 - X_2^t\Sigma_{0,2}^{-1}X_2\big)$$

holds. If $\Sigma_0 = \Sigma_1$ and $X_2 = \Sigma_{0,2}X_1$, then the inequality is an equality.

Proof of Theorem 3.10. The statement about the mean of $\bar{\mu}$ is already known, thus let us simplify the problem assuming that every μ_ω is centered in mean. By Proposition 5.14, $\bar{\mu}$ must be absolutely continuous, hence its covariance matrix $\bar{\Sigma}$ must be nonsingular. To simplify the notation, let us denote $\bar{P} = \mathbb{P}_{0,\bar{\Sigma}} \in \mathcal{F}(P_0)$. From Gelbrich's bound, we have

$$\int \mathcal{W}_2^2(\mu_\omega, \bar{\mu})\mathbf{P}(d\omega) \geq \int \mathcal{W}_2^2(\mu_\omega, \bar{P})\mathbf{P}(d\omega),$$

hence, by the uniqueness of the barycenter, $\bar{\mu} = \bar{P}$, and $\bar{\mu} \in \mathcal{F}(P_0)$. If we consider the optimal maps $\bar{T}_\omega$ transporting $\bar{\mu}$ to μ_ω, and define $\bar{T}(x) := \int \bar{T}_\omega(x)\mathbf{P}(d\omega)$, we have

$$\begin{aligned}
\int \mathcal{W}_2^2(\bar{\mu}, \mu_\omega)\mathbf{P}(d\omega) &= \int \left(\int \big\|x - \bar{T}_\omega(x)\big\|^2 \bar{\mu}(dx)\right)\mathbf{P}(d\omega) \\
&= \int \left(\int \big(\big\|x - \bar{T}(x)\big\|^2 + \big\|\bar{T}(x) - \bar{T}_\omega(x)\big\|^2\big)\mathbf{P}(d\omega)\right)\bar{\mu}(dx) \\
&\geq \int \left(\int \big\|\bar{T}(x) - \bar{T}_\omega(x)\big\|^2 \bar{\mu}(dx)\right)\mathbf{P}(d\omega) \\
&\geq \int \mathcal{W}_2^2\big(\bar{\mu}\circ\bar{T}^{-1}, \mu_\omega\big)\mathbf{P}(d\omega)
\end{aligned}$$

that (by the uniqueness) is possible only if $\bar{\mu}\circ\bar{T}^{-1} = \bar{\mu}$, that is, if $\bar{T}(x) = x$ $\bar{\mu}$-a.s.

To finalize, observe that the optimal transport maps $\bar{T}_\omega$ from $\bar{\mu}$ to μ_ω, being probabilities in $\mathcal{F}(P_0)$, take the form $\bar{\Sigma}^{-1/2}(\bar{\Sigma}^{1/2}\Sigma_\omega\bar{\Sigma}^{1/2})^{1/2}\bar{\Sigma}^{-1/2}$ (see (6)), therefore (since $\bar{\Sigma}$ is positive definite) the relation $\bar{T}(x) = x$ $\bar{\mu}$-a.s. is equivalent to

$$\bar{\Sigma} = \int \big(\bar{\Sigma}^{1/2}\Sigma_\omega\bar{\Sigma}^{1/2}\big)^{1/2}\mathbf{P}(d\omega).$$

This proves that $\bar{\Sigma}$ verifies the integral equation. To prove that the integral equation has only a positive definite solution, let $\hat{\Sigma}$ be any positive definite matrix and define

$$\Sigma_{0,\omega} := \hat{\Sigma}^{-1/2}\big(\hat{\Sigma}^{1/2}\Sigma_\omega\hat{\Sigma}^{1/2}\big)^{1/2}\hat{\Sigma}^{-1/2} \quad \text{and} \quad \hat{\Sigma}^* := \int \Sigma_{0,\omega}\mathbf{P}(d\omega).$$

If we apply Lemma 5.15 first to $\Sigma_0 = \hat{\Sigma}$, $\Sigma_1 = \Sigma$ and $\Sigma_2 = \Sigma_\omega$, later to $\Sigma_0 = \Sigma_1 = \hat{\Sigma}$ and $\Sigma_2 = \Sigma_\omega$, subtracting the results and integrating, we have that

$$\int \mathcal{W}_2^2(\mathbb{P}_{0,\Sigma}, \mathbb{P}_{0,\Sigma_\omega})\mathbf{P}(d\omega) - \int \mathcal{W}_2^2(\mathbb{P}_{0,\hat{\Sigma}}, \mathbb{P}_{0,\Sigma_\omega})\mathbf{P}(d\omega) \geq \operatorname{trace}\big(\big(I_d - \hat{\Sigma}^*\big)(\Sigma - \hat{\Sigma})\big).$$

Thus, if $\hat{\Sigma}$ is a solution of the integral equation, we would have that $\mathbb{P}_{0,\hat{\Sigma}}$ is the barycenter of μ, and the uniqueness of the barycenter gives that $\bar{\Sigma} = \hat{\Sigma}$. □

Acknowledgements

Research partially supported by the Spanish Ministerio de Economía y Competitividad y fondos FEDER, grants MTM2014-56235-C2-1-P, and MTM2014-56235-C2-2, and by Consejería de Educación de la Junta de Castilla y León, grant VA212U13

References

[1] Agueh, M. and Carlier, G. (2011). Barycenters in the Wasserstein space. *SIAM J. Math. Anal.* **43** 904–924. MR2801182

[2] Álvarez-Esteban, P.C., del Barrio, E., Cuesta-Albertos, J.A. and Matrán, C. (2011). Uniqueness and approximate computation of optimal incomplete transportation plans. *Ann. Inst. Henri Poincaré B, Probab. Stat.* **47** 358–375. MR2814414

[3] Álvarez-Esteban, P.C., del Barrio, E., Cuesta-Albertos, J.A. and Matrán, C. (2012). Similarity of samples and trimming. *Bernoulli* **18** 606–634. MR2922463

[4] Álvarez-Esteban, P.C., del Barrio, E., Cuesta-Albertos, J.A. and Matrán, C. (2016). A fixed-point approach to barycenters in Wasserstein space. *J. Math. Anal. Appl.* **441** 744–762. MR3491556

[5] Arsigny, V., Fillard, P., Pennec, X. and Ayache, N. (2006/2007). Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM J. Matrix Anal. Appl.* **29** 328–347. MR2288028

[6] Benamou, J.-D., Carlier, G., Cuturi, M., Nenna, L. and Peyré, G. (2015). Iterative Bregman projections for regularized transportation problems. *SIAM J. Sci. Comput.* **37** A1111–A1138. MR3340204

[7] Bigot, J. and Klein, T. (2015). Consistent estimation of a population barycenter in the Wasserstein space. Preprint. Available at arXiv:1212.2562v5.

[8] Boissard, E., Le Gouic, T. and Loubes, J.-M. (2015). Distribution's template estimate with Wasserstein metrics. *Bernoulli* **21** 740–759. MR3338645

[9] Breiman, L. (1996). Bagging predictors. *Mach. Learn.* **24** 123–140.

[10] Brenier, Y. (1987). Polar decomposition and increasing rearrangement of vector fields. *C. R. Acad. Sci. Paris Ser. I Math.* **305** 805–808.

[11] Brenier, Y. (1991). Polar factorization and monotone rearrangement of vector-valued functions. *Comm. Pure Appl. Math.* **44** 375–417. MR1100809

[12] Bühlmann, P. (2003). Bagging, subagging and bragging for improving some prediction algorithms. In *Recent Advances and Trends in Nonparametric Statistics* (M.G. Akritas and D.N. Politis, eds.) 19–34. Amsterdam: Elsevier. MR2498230

[13] Bühlmann, P. and Yu, B. (2002). Analyzing bagging. *Ann. Statist.* **30** 927–961. MR1926165

[14] Carlier, G., Oberman, A. and Oudet, E. (2015). Numerical methods for matching for teams and Wasserstein barycenters. *ESAIM Math. Model. Numer. Anal.* **49** 1621–1642. MR3423268

[15] Chernozhukov, V., Galichon, A., Hallin, M. and Henry, M. (2017). Monge-Kantorovich depth, quantiles, ranks and signs. *Ann. Statist.* **45** 223–256. MR3611491

[16] Croux, C. and Haesbroeck, G. (1997). An easy way to increase the finite-sample efficiency of the resampled minimum volume ellipsoid estimator. *Comput. Statist. Data Anal.* **25** 125–141. MR1468064

[17] Cuesta-Albertos, J.A. and Matrán, C. (1988). The strong law of large numbers for k-means and best possible nets of Banach valued random variables. *Probab. Theory Related Fields* **78** 523–534. MR0950345

[18] Cuesta, J.A. and Matrán, C. (1989). Notes on the Wasserstein metric in Hilbert spaces. *Ann. Probab.* **17** 1264–1276. MR1009457

[19] Cuesta-Albertos, J.A., Matrán, C. and Mayo-Íscar, A. (2008). Trimming and likelihood: Robust location and dispersion estimation in the elliptical model. *Ann. Statist.* **36** 2284–2318. MR2458188

[20] Cuesta-Albertos, J.A., Matrán-Bea, C. and Tuero-Díaz, A. (1996). On lower bounds for the L^2-Wasserstein metric in a Hilbert space. *J. Theoret. Probab.* **9** 263–283. MR1385397

[21] Cuesta-Albertos, J.A., Matrán Bea, C. and Rodríguez Rodríguez, J.M. (2002). Shape of a distribution through the L_2-Wasserstein distance. In *Distributions with Given Marginals and Statistical Modelling* (C.M. Cuadras, J. Fortiana and J.A. Rodríguez-Lallena, eds.) 51–61. Dordrecht: Kluwer Academic. MR2058979

[22] Cuesta-Albertos, J.A., Rüschendorf, L. and Tuero-Díaz, A. (1993). Optimal coupling of multivariate distributions and stochastic processes. *J. Multivariate Anal.* **46** 335–361. MR1240428

[23] Cuturi, M. and Doucet, A. (2014). Fast computation of Wasserstein barycenters. In *Proceedings of the* 31*st International Conference on Machine Learning*. JMLR: W&CP vol. 32.

[24] del Barrio, E., Cuesta-Albertos, J.A., Matrán, C. and Mayo-Íscar, A. (2016). Robust clustering tools based on optimal transportation. Preprint. Available at arXiv:1607.01179.

[25] Dudley, R.M. (1989). *Real Analysis and Probability*. Pacific Grove, CA: Wadsworth & Brooks. MR0982264

[26] Fritz, H., García-Escudero, L.A. and Mayo-Íscar, A. (2012). tclust: An R package for a trimming approach to cluster analysis. *J. Stat. Softw.* **47** 1–26.

[27] García-Escudero, L.A., Gordaliza, A. and Matrán, C. (1999). A central limit theorem for multivariate generalized trimmed k-means. *Ann. Statist.* **27** 1061–1079. MR1724041

[28] Gelbrich, M. (1990). On a formula for the L^2 Wasserstein metric between measures on Euclidean and Hilbert spaces. *Math. Nachr.* **147** 185–203. MR1127323

[29] Gordaliza, A. (1991). Best approximations to random variables based on trimming procedures. *J. Approx. Theory* **64** 162–180. MR1091467

[30] Knott, M. and Smith, C.S. (1994). On a generalization of cyclic monotonicity and distances among random vectors. *Linear Algebra Appl.* **199** 363–371. MR1274425

[31] Le Gouic, T. and Loubes, J.-M. (2015). Barycenter in Wasserstein spaces: Existence and consistency. *Probab. Theory Related Fields*. To appear. Available at hal-01163262v2.

[32] Meinshausen, N. and Bühlmann, P. (2014). Magging: maximin aggregation for inhomogeneous large-scale data. Available at arXiv:1409.2638v1.

[33] Munk, A. and Czado, C. (1998). Nonparametric validation of similar distributions and assessment of goodness of fit. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **60** 223–241. MR1625620
[34] Pass, B. (2013). Optimal transportation with infinitely many marginals. *J. Funct. Anal.* **264** 947–963. MR3004954
[35] Rippl, T., Munk, A. and Sturm, A. (2016). Limit laws of the empirical Wasserstein distance: Gaussian distributions. *J. Multivariate Anal.* **151** 90–109. MR3545279
[36] Rousseeuw, P. (1985). Multivariate estimation with high breakdown point. In *Mathematical Statistics and Applications*, *Vol. B* (*Bad Tatzmannsdorf*, 1983) (W. Grossman, G. Pflug, I. Vincze and W. Werttz, eds.) 283–297. Dordrecht: Reidel. MR0851060
[37] Rousseeuw, P.J. (1984). Least median of squares regression. *J. Amer. Statist. Assoc.* **79** 871–880. MR0770281
[38] Rousseeuw, P.J. and van Driessen, K. (1999). A fast algorithm for the minimum covariance determinant estimator. *Technometrics* **41** 212–223.
[39] Rüschendorf, L. and Rachev, S.T. (1990). A characterization of random variables with minimum L^2-distance. *J. Multivariate Anal.* **32** 48–54. MR1035606
[40] Rüschendorf, L. and Uckelmann, L. (2002). On the n-coupling problem. *J. Multivariate Anal.* **81** 242–258. MR1906379
[41] Villani, C. (2003). *Topics in Optimal Transportation*. *Graduate Studies in Mathematics* **58**. Providence, RI: Amer. Math. Soc. MR1964483
[42] Villani, C. (2009). *Optimal Transport*: *Old and New*. Berlin: Springer. MR2459454
[43] Woodruff, D.L. and Rocke, D.M. (1994). Computable robust estimation of multivariate location and shape in high dimension using compound estimators. *J. Amer. Statist. Assoc.* **89** 888–896. MR1294732

Received October 2016 and revised May 2017