

STATISTICAL INFERENCE FOR BURES–WASSERSTEIN BARYCENTERS

BY ALEXEY KROSHNIN¹, VLADIMIR SPOKOINY^{2,*} AND ALEXANDRA SUVORIKOVA^{2,†}

¹*Institute for Information Transmission Problems RAS, kroshnin@math.univ-lyon1.fr*

²*Weierstrass Institute for Applied Analysis and Stochastics, *spokoiny@wias-berlin.de; †suvorikova@wias-berlin.de*

In this work we introduce the concept of Bures–Wasserstein barycenter Q_* , that is essentially a Fréchet mean of some distribution \mathbb{P} supported on a subspace of positive semi-definite d -dimensional Hermitian operators $\mathbb{H}_+(d)$. We allow a barycenter to be constrained to some affine subspace of $\mathbb{H}_+(d)$, and we provide conditions ensuring its existence and uniqueness. We also investigate convergence and concentration properties of an empirical counterpart of Q_* in both Frobenius norm and Bures–Wasserstein distance, and explain, how the obtained results are connected to optimal transportation theory and can be applied to statistical inference in quantum mechanics.

1. Introduction. The space of finite-dimensional Hermitian operators is commonly applied for data representation. For instance, in quantum mechanics it is used for mathematical description of physical properties of a quantum system: the real-valued spectrum is associated to measurements observed in a physical experiment. Real-valued symmetric matrices are also widely used for description of systems in engineering applications, medical studies, neural sciences, evolutionary biology etcetera. Usually, one assumes a sample to be random; see, for example, [Álvarez-Esteban et al. \(2015\)](#), [Calsbeek and Goodnight \(2009\)](#), [del Barrio et al. \(2019b\)](#), [Gonzalez et al. \(2017\)](#), [Goodnight and Schwartz \(1997\)](#). Statistical characteristics of its distribution \mathbb{P} , such as mean and variance, are of interest for experimental design, and analysis of the results for further development of natural science models.

The current study focuses on the space of positive semi-definite Hermitian matrices $\mathbb{H}_+(d)$ and presents a theoretical analysis of aggregation methods of the relevant statistical information from data sets, for which the hypothesis of linearity might be violated. In this case, the widely-used Euclidean mean and variance are not sensitive enough to capture effects of interest. For instance, some data sets are described by probability measures belonging to some scale-location family, for example, [Álvarez-Esteban et al. \(2018\)](#), [Muzellec and Cuturi \(2018\)](#). The nonlinearity assumption requires an adaptation of the tools of classical statistical analysis. In order to capture nonlinear effects, we suggest to endow $\mathbb{H}_+(d)$ with the Bures–Wasserstein distance d_{BW} , originally introduced by [Bhatia, Jain and Lim \(2019\)](#). For a pair of positive semi-definite matrices $Q, S \in \mathbb{H}_+(d)$ the distance is defined as

$$(1.1) \quad d_{\text{BW}}^2(Q, S) = \text{tr } Q + \text{tr } S - 2 \text{tr}(Q^{1/2} S Q^{1/2})^{1/2}.$$

It is worth noting that being restricted to the sub-space of symmetric positive semidefinite matrices $\text{Sym}_+(d)$, d_{BW} turns into the 2-Wasserstein distance between two normal distributions. Let $\mathcal{N}(0, Q)$ and $\mathcal{N}(0, S)$ be two centred Gaussians. The 2-Wasserstein distance is

$$d_{\text{W}_2}^2(\mathcal{N}(0, Q), \mathcal{N}(0, S)) = \text{tr } Q + \text{tr } S - 2 \text{tr}(Q^{1/2} S Q^{1/2})^{1/2}.$$

Received February 2019; revised August 2020.

MSC2020 subject classifications. 60F05.

Key words and phrases. Bures–Wasserstein barycenter, Wasserstein barycenter, central limit theorem, concentration, Hermitian operators.

It is worth noting that a natural extension of the Gaussian case is the case of distributions coming from the same scale-location family (see, e.g., [Agueh and Carlier \(\(2011\), Section 6\)](#) or [Álvarez-Esteban et al. \(2018\)](#)).

In the last few years, the class of optimal transportation distances and in particular the 2-Wasserstein distance attract a lot of attention of the both mathematical and machine learning communities. The latter captures the geometrical similarities between objects coming from nonlinear spaces, see, for example, [Courty et al. \(2017\)](#), [Flamary et al. \(2018\)](#), [Montavon, Müller and Cuturi \(2016\)](#), while the recent advances in computations make the distance useful for the real-world problems ([Cuturi \(2013\)](#), [Gramfort, Peyré and Cuturi \(2015\)](#), [Uribe et al. \(2018\)](#)). For more information on the Wasserstein distance and optimal transportation theory in general, we recommend the excellent monograph by [Villani \(2009\)](#). The book by [Peyré, Cuturi et al. \(2019\)](#) provides a state-of-the-art survey of numerical methods and their applications in data sciences.

This study focuses on the following statistical setting. Let \mathbb{P} be a probability distribution supported on the set of nonnegatively definite Hermitian matrices $\mathbb{H}_+(d)$. Two important characteristics of \mathbb{P} are the Fréchet mean and variance. While the former is a “typical” representative of a data set in hand, the latter appears in the analysis of data variability; see, for example, [Del Barrio, Lescornel and Loubes \(2015\)](#). We briefly recall both concepts below. For an arbitrary point $Q \in \mathbb{H}_+(d)$, the Fréchet variance of \mathbb{P} is defined as

$$\mathcal{V}(Q) \stackrel{\text{def}}{=} \int_{\mathbb{H}_+(d)} d_{\text{BW}}^2(Q, S) d\mathbb{P}(S).$$

The Fréchet mean of \mathbb{P} is given by the set of global minimizers of the variance $\mathcal{V}(Q)$:

$$(1.2) \quad Q_* \in \underset{Q \in \mathbb{H}_+(d)}{\text{argmin}} \mathcal{V}(Q).$$

However, in some cases one might be interested in a minimizer belonging to an affine subspace of Hermitian operators $\mathbb{H}(d)$, $\mathbb{A} \subset \mathbb{H}(d)$:

$$(1.3) \quad Q_* \in \underset{Q \in \mathbb{H}_+(d) \cap \mathbb{A}}{\text{argmin}} \mathcal{V}(Q).$$

For instance, such a necessity arises when considering a random set of quantum density operators. Section 3.2 discusses this example in more detail. Note that the setting (1.3) covers the setting (1.2). So, without loss of generality, we further address only (1.3).

Obviously, the first crucial question concerns existence and uniqueness of Q_* . Theorem 2.1 presents the positive answers to both issues. This immediately allows us to define the global Fréchet variance of \mathbb{P} :

$$\mathcal{V}_* \stackrel{\text{def}}{=} \mathcal{V}(Q_*).$$

Given an i.i.d. sample $S_1, \dots, S_n, S_i \sim \mathbb{P}$, one constructs an empirical version of $\mathcal{V}(Q)$, for an arbitrary Q , as follows:

$$\mathcal{V}_n(Q) \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^n d_{\text{BW}}^2(Q, S_i).$$

The empirical Fréchet mean and the global empirical variance also exist and unique:

$$(1.4) \quad Q_n = \underset{Q \in \mathbb{H}_+(d) \cap \mathbb{A}}{\text{argmin}} \mathcal{V}_n(Q), \quad \mathcal{V}_n \stackrel{\text{def}}{=} \mathcal{V}_n(Q_n).$$

Both facts follow from Theorem 2.1.

This work studies the convergence of the estimators Q_n and \mathcal{V}_n and investigates the concentration properties of both objects. A discussion of the practical applicability of the obtained results is postponed to Section 3. There we explain the connection to optimal transportation theory and present a possible application to statistical analysis in quantum mechanics.

1.1. *Contribution of the present study.*

The central limit theorems for Q_n and \mathcal{V}_n . From now on we use bold symbols (e.g., \mathbf{A} , \mathbf{B}) to denote operators, whereas a classical font (e.g., A , B) stands for either matrices or vectors.

The first result of this study concerns the asymptotic normality of the approximation error of the population Fréchet mean by its empirical counterpart:

$$\sqrt{n}(Q_n - Q_*) \xrightarrow{w} \mathcal{N}(0, \mathbf{\Xi}),$$

where “ \xrightarrow{w} ” stands for the weak convergence, and $\mathbf{\Xi}$ is a covariance operator acting on a linear subspace $\mathbb{M} \subset \mathbb{H}(d)$ associated with the affine subspace \mathbb{A} . The result is derived under some suitable assumptions on the distribution \mathbb{P} introduced later, in Section 2.

At this point it is worth mentioning that the asymptotic normality of Q_n falls in the setting of the asymptotic normality of parametric M-estimations with a smooth and convex loss function $d_{\text{BW}}(\cdot, \cdot)$ defined over the convex set $\mathbb{A} \cap \mathbb{H}_+(d)$. The convexity and the smoothness of $d_{\text{BW}}(\cdot, \cdot)$ are validated by Lemma A.5 and Lemma A.6, respectively. However, the current study uses the proof techniques different from a verification of the standard assumptions on M-estimators. We discuss the issue in more detail in Section 2.4.

The above convergence result cannot be used directly for construction of asymptotic confidence sets, as it relies on the unknown covariance matrix $\mathbf{\Xi}$. However, Theorem 2.2 ensures that the covariance operator $\mathbf{\Xi}$ can be replaced by an empirical counterpart $\hat{\mathbf{\Xi}}_n$:

$$\sqrt{n} \hat{\mathbf{\Xi}}_n^{-1/2} (Q_n - Q_*) \xrightarrow{w} \mathcal{N}(0, \mathbf{I}),$$

where \mathbf{I} denotes the identity operator. Along with the asymptotic normality of $\sqrt{n}(Q_n - Q_*)$, we are interested in the limiting distribution of $\mathcal{L}(\sqrt{n}d_{\text{BW}}(Q_n, Q_*))$, where $\mathcal{L}(X)$ denotes the distribution of a random variable X . In what follows $\|A\|_F$ denotes the Frobenius norm of matrix A . Corollary 2.1 ensures

$$\mathcal{L}(\sqrt{n}d_{\text{BW}}(Q_n, Q_*)) \xrightarrow{w} \mathcal{L}(\|\xi\|_F),$$

where ξ is some normally distributed vector. The data-driven asymptotic confidence sets for $\sqrt{n}d_{\text{BW}}(Q_n, Q_*)$ are obtained by replacing ξ by its empirical counterpart ξ_n :

$$d_w(\mathcal{L}(\sqrt{n}d_{\text{BW}}(Q_n, Q_*)), \mathcal{L}(\|\xi_n\|_F)) \rightarrow 0,$$

where d_w is some metric inducing the weak convergence of measures. We also show the asymptotic normality of the approximation error of the variance \mathcal{V}_* by its empirical analogue \mathcal{V}_n (see Theorem 2.3):

$$\sqrt{n}(\mathcal{V}_n - \mathcal{V}_*) \xrightarrow{w} \mathcal{N}(0, \text{Var} d_{\text{BW}}^2(Q_*, S)).$$

All above-mentioned results are closely connected to the convergence of empirical 2-Wasserstein barycenters. For the sake of transparency we postpone a further discussion of this topic to Section 3.1.

The concentration of Q_n and \mathcal{V}_n . The technique of the proof of the central limit theorem, developed in the current study, appears to be suitable for an investigation of the concentration properties of Q_n and \mathcal{V}_n . To validate the concentration, we suppose the distribution \mathbb{P} to be sub-Gaussian; see Assumption 3. This assumption ensures the following bounds which hold with high probability:

$$\|Q_*^{-1/2} Q_n Q_*^{-1/2} - \mathbf{I}\|_F \leq \frac{C(\sqrt{m} + t)}{\sqrt{n}}, \quad d_{\text{BW}}(Q_n, Q_*) \leq \frac{C(\sqrt{m} + t)}{\sqrt{n}},$$

where m is the dimension of \mathbb{M} , $t \geq 0$, and C denotes a generic constant. For more details see Theorem 2.4 and Corollary 2.2, respectively. To the best of our knowledge, these results

appear to be novel. Along with concentration of the empirical barycenter, we investigate the concentration of the empirical variance \mathcal{V}_n , which holds with high probability:

$$|\mathcal{V}_n - \mathcal{V}_*| \leq \max\left(\frac{\mu t^2}{n}, \frac{\nu t}{\sqrt{n}}\right) + \frac{C(\sqrt{m} + t)^2}{n},$$

where μ and ν are some parameters depending on the distribution of $d_{\text{BW}}^2(Q_*, S)$, m is the dimension of \mathbb{M} , C stands for a generic constant, and the parameter is $t \geq 0$. The result is presented in Theorem 2.5. We discuss its relation to the existing results in Section 3.1.

The paper is organized as follows. Section 2 presents the obtained results in more detail. Section 3 illustrates the connection to other scientific problems. Finally, Section 4 suggests to use barycenters for replacement of the lost data. It contains a description of the idea, and experimental estimation of convergence rates for barycenters using both an artificial and a real data set. The latter one is related to the climate modelling.

2. Results. Following Bhatia, Jain and Lim (2019), we continue to investigate properties of $d_{\text{BW}}(Q, S)$. Further we present an alternative analytical expression for the distance. The result is well known for the case of real-valued symmetric matrices $Q, S \in \text{Sym}_+(d)$; see Olkin and Pukelsheim (1982). The proposition below extends it to the case of Hermitian matrices $\mathbb{H}_+(d)$.

PROPOSITION 2.1. *Let $Q, S \in \mathbb{H}_+(d)$ and $Q \succ 0$. Then (1.1) can be rewritten as*

$$d_{\text{BW}}^2(Q, S) = \|(T_Q^S - I)Q^{1/2}\|_F^2 = \text{tr}(T_Q^S - I)Q(T_Q^S - I),$$

where the optimal map from Q to S is

$$(2.1) \quad \begin{aligned} T_Q^S &\stackrel{\text{def}}{=} \underset{T: TQ^* = S}{\text{argmin}} \| (T - I)Q^{1/2} \|_F \\ &= S^{1/2}(S^{1/2}QS^{1/2})^{-1/2}S^{1/2} = Q^{-1/2}(Q^{1/2}SQ^{1/2})^{1/2}Q^{-1/2}. \end{aligned}$$

By $(S^{1/2}QS^{1/2})^{-1/2}$ we denote the pseudo-inverse matrix $((S^{1/2}QS^{1/2})^{1/2})^+$.

Note that being restricted to the sub-space $\text{Sym}_{++}(d)$, T_Q^S coincides with the optimal push-forward (also known as the optimal map) between two centred normal distributions $\mathcal{N}(0, Q)$ and $\mathcal{N}(0, S)$: $T_Q^S \# \mathcal{N}(0, Q) = \mathcal{N}(0, S)$. For more details on general optimal transportation maps see Brenier (1991), for a particular case of scale-location families one may refer to Álvarez-Esteban et al. (2018), Takatsu (2011). The differentiability of T_Q^S is one of the key ingredients in the proofs. It is validated in Lemma A.2. Note that for a particular choice of \mathbb{A} in (1.3), $\mathbb{A} = \text{Sym}_{++}(d)$, the differentiability of optimal transportation maps are proved in Rippl, Munk and Sturm (2016). Section A.2 is dedicated to the investigation of properties of T_Q^S and its differential dT_Q^S . Section A.3 investigates properties of d_{BW} . We highly recommend to at least look through these two sections for a better understanding of the tools used in the proofs of central limit theorems and concentrations.

2.1. *Existence and uniqueness of Q_* and Q_n .* Along with investigation of properties of the distance in hand, and before moving to more general questions, one should ask her- or himself, whether the Fréchet mean Q_* exists and, if so, is it unique or not. We further assume that \mathbb{A} has a nonempty intersection with the space of positive definite operators:

ASSUMPTION 1. Given the setting (1.3), we suppose an affine subspace $\mathbb{A} \subset \mathbb{H}(d)$ to be s.t. $\mathbb{H}_{++}(d) \cap \mathbb{A} \neq \emptyset$. By \mathbb{M} we denote the linear subspace of $\mathbb{H}(d)$ associated with \mathbb{A} , that is, the following representation holds: $\mathbb{A} = \{Q_0\} + \mathbb{M}$ for some $Q_0 \in \mathbb{H}(d)$.

Without loss of generality we assume that \mathbb{P} assigns positive probability to the space of positive definite Hermitian matrices $\mathbb{H}_{++}(d)$. We also suppose \mathbb{P} to be s.t. the spectrum of S is on average bounded away from infinity:

ASSUMPTION 2. Let data distribution \mathbb{P} , $S \sim \mathbb{P}$, be s.t.

$$\mathbb{P}(\mathbb{H}_{++}(d)) > 0, \quad \mathbb{E} \operatorname{tr} S < +\infty.$$

The next theorem ensures existence and uniqueness of the Fréchet mean introduced in (1.3).

THEOREM 2.1 (Existence and uniqueness of Fréchet mean Q_*). *Under Assumptions 1 and 2, there exists unique positive-definite barycenter Q_* of \mathbb{P} , $Q_* \succ 0$. Moreover, it is characterised as the unique solution of the equation*

$$(2.2) \quad \mathbf{\Pi}_{\mathbb{M}} \mathbb{E} T_Q^S = \mathbf{\Pi}_{\mathbb{M}} I, \quad Q \in \mathbb{H}_{++}(d),$$

where $\mathbf{\Pi}_{\mathbb{M}}$ is the orthogonal projector onto \mathbb{M} .

Note that for any fixed $Q \in \mathbb{H}_{++}(d)$, T_Q^S is a random variable because it is a continuous function of the random variable S . The equation (2.2) generalises the result for scale-location families in 2-Wasserstein space, presented in [Álvarez-Esteban et al. \(\(2015\), Theorem 3.10\)](#), and originally obtained for the Gaussian case in the seminal work [Agueh and Carlier \(\(2011\), Theorem 6.1\)](#). Namely, if $\mathbb{A} = \operatorname{Sym}_{++}(d)$, then Q_* exists and is the unique solution of a fixed-point equation:

$$Q = \mathbb{E}(Q^{1/2} S Q^{1/2})^{1/2}.$$

Note that it is similar to (2.2), as by multiplying the above equation from both sides by $Q^{-1/2}$ one obtains $\mathbb{E} T_Q^S = I$. Existence, uniqueness, and measurability of the estimator Q_n defined in (1.4) are a direct corollary of the above theorem. The proof of Theorem 2.1 is presented in Section A.4.

2.2. *Limiting distributions of $\sqrt{n}(Q_n - Q_*)$, $\sqrt{n}d_{\text{BW}}(Q_n, Q_*)$, and $\sqrt{n}(\mathcal{V}_n - \mathcal{V}_*)$.* Armed with the knowledge about properties of $d_{\text{BW}}(\cdot, \cdot)$, Q_* , and Q_n , we are now equipped enough to introduce the first main result of the current study. In what follows we denote the variance of optimal transportation map from the population barycenter Q_* to any $S \sim \mathbb{P}$ as

$$(2.3) \quad \operatorname{Var}(T_{Q_*}^S) = \mathbb{E}(T_{Q_*}^S - I) \otimes (T_{Q_*}^S - I) \quad \text{with } \mathbb{E} T_{Q_*}^S = I,$$

where \otimes stands for the tensor product. Theorem 2.2 presents the asymptotic convergence of Q_n to Q_* .

THEOREM 2.2 (Central limit theorem for the Fréchet mean). *Under Assumptions 1 and 2 the approximation error rate of the Fréchet mean Q_* by its empirical counterpart Q_n is*

$$(A) \quad \sqrt{n}(Q_n - Q_*) \xrightarrow{w} \mathcal{N}(0, \mathbf{\Xi}),$$

where $\mathbf{\Xi}$ is a self-adjoint linear operator acting from \mathbb{M} to \mathbb{M} defined in (A.7). Moreover, if $\operatorname{Var}(T_{Q_*}^S)$ is nondegenerated, then

$$(B) \quad \sqrt{n} \hat{\mathbf{\Xi}}_n^{-1/2} (Q_n - Q_*) \xrightarrow{w} \mathcal{N}(0, (\mathbf{I})_{\mathbb{M}}),$$

where $\hat{\mathbf{\Xi}}_n$ is a data-driven empirical counterpart of $\mathbf{\Xi}$ defined in (A.8).

REMARK 1. The notations $(\mathbf{A})_{\mathbb{M}}$ denote a linear operator associated to the restriction of a quadratic form \mathbf{A} to the subspace \mathbb{M} :

$$(\mathbf{A})_{\mathbb{M}}: \mathbb{M} \rightarrow \mathbb{M}, \quad X \mapsto \boldsymbol{\Pi}_{\mathbb{M}}\mathbf{A}(X).$$

We intentionally postpone the explicit definitions of $\boldsymbol{\Xi}$ and $\hat{\boldsymbol{\Xi}}_n$, because they require an introduction of many technical details. This would make the description of the main results less transparent. It is worth noting that the result (B) enables construction of data-driven asymptotic confidence sets. However, inversion of the empirical covariance might be a problem. For instance, numerical simulations show that $\hat{\boldsymbol{\Xi}}_n$ might be degenerated if \mathbb{P} is supported on a set of diagonal matrices. This immediately raises a question concerning introduction of a resampling approach which would make the computations tractable. We consider this as a subject for the further research.

The proof of the central limit theorem relies on the Fréchet differentiability of T_Q^S by the lower argument Q at the point Q_* :

$$T_{Q_n}^S = S^{1/2}(S^{1/2}Q_nS^{1/2})^{-1/2}S^{1/2}, \quad T_{Q_n}^S \approx T_{Q_*}^S + d\mathbf{T}_{Q_*}^S(Q_n - Q_*),$$

where $d\mathbf{T}_{Q_*}^S$ is a differential of T_Q^S at the point Q_* .

Since $\mathbb{H}_+(d)$ is endowed with the Bures–Wasserstein distance, the convergence properties of $d_{\text{BW}}(Q_n, Q_*)$ are also of great interest. The result is a corollary of the above theorem.

COROLLARY 2.1 (Asymptotic distribution of $d_{\text{BW}}(Q_n, Q_*)$). *Under conditions of Theorem 2.2 it holds*

$$\mathcal{L}(\sqrt{nd}_{\text{BW}}(Q_n, Q_*)) \xrightarrow{w} \mathcal{L}(\|Q_*^{1/2}d\mathbf{T}_{Q_*}^{Q_*}(Z)\|_F),$$

where $Z \in \mathbb{M} \subset \mathbb{H}(d)$ is a random matrix, $Z \sim \mathcal{N}(0, \boldsymbol{\Xi})$.

Moreover, replacing in the limiting distribution Q_* and Z by their empirical counterparts Q_n and $Z_n \sim \mathcal{N}(0, \hat{\boldsymbol{\Xi}}_n)$, $Z_n \in \mathbb{M}$, respectively, one obtains the following convergence

$$d_w(\mathcal{L}(\sqrt{nd}_{\text{BW}}(Q_n, Q_*)), \mathcal{L}(\|Q_n^{1/2}d\mathbf{T}_{Q_n}^{Q_n}(Z_n)\|_F)) \rightarrow 0,$$

where d_w is some metric inducing the weak convergence.

To illustrate the result, we consider the case of a diagonal $Q_* = \text{diag}(q_1, \dots, q_d)$. This setting admits the explicit form of the limiting distribution:

$$\mathcal{L}(\sqrt{nd}_{\text{BW}}(Q_n, Q_*)) \xrightarrow{w} \mathcal{L}\left(\sqrt{\sum_{i,j=1}^d \frac{Z_{ij}^2}{2(q_i + q_j)}}\right),$$

where $Z = (Z_{ij})_{i,j=1}^d$. This representation of the limiting distribution is derived in the proof of Corollary 2.1 which is based on the fact that

$$d_{\text{BW}}^2(Q_n, Q_*) = -\frac{1 + o_P(1)}{2} \langle d\mathbf{T}_{Q_*}^{Q_*}(Q_n - Q_*), Q_n - Q_* \rangle,$$

with $o_P(\cdot)$ being o -small in probability, and an explicit formula for $d\mathbf{T}_{Q_*}^S$ from Lemma A.2. We discuss the above approximation of $d_{\text{BW}}^2(Q_n, Q_*)$ in more detail later in Section 2.4.

The last result concerning convergence of empirical barycenter is the central limit theorem for the empirical variance \mathcal{V}_n .

THEOREM 2.3 (Central limit theorem for \mathcal{V}_n). *Let Assumptions 1 and 2 be fulfilled and $\mathbb{E}(\text{tr } S)^2 < \infty$. Then*

$$\sqrt{n}(\mathcal{V}_n - \mathcal{V}_*) \xrightarrow{w} \mathcal{N}(0, \text{Var } d_{\text{BW}}^2(Q_*, S)).$$

All proofs are collected in Section A.4. Section 4.1 illustrates the asymptotic behaviour of $\mathcal{L}(\sqrt{n}\|Q_n - Q_*\|_F)$, $\mathcal{L}(\sqrt{n}d_{\text{BW}}(Q_n, Q_*))$, and $\mathcal{L}(\sqrt{n}|\mathcal{V}_* - \mathcal{V}_n|)$.

2.3. Concentration of Q_n and \mathcal{V}_n . This section discusses the concentration properties of Q_n under the assumption of sub-Gaussianity of \mathbb{P} :

ASSUMPTION 3 (Sub-Gaussianity of $\sqrt{\text{tr } S}$). Let $\sqrt{\text{tr } S}$ be sub-Gaussian:

$$\mathbb{P}\{\sqrt{\text{tr } S} \geq t\} \leq B e^{-bt^2} \quad \text{for any } t \geq 0,$$

with some constants $B, b > 0$.

The first result concerns the concentration of $Q_*^{-1/2} Q_n Q_*^{-1/2}$ in Frobenius norm. This is a crucial step in the proof of concentration of $d_{\text{BW}}(Q_n, Q_*)$. From now on we denote the operator norm of a matrix A or an operator A as $\|A\|$, $\|A\|$, respectively. The notations $\lambda_{\min}(A)$, $\lambda_{\min}(A)$ denote their smallest eigenvalues.

THEOREM 2.4 (Concentration of $Q_*^{-1/2} Q_n Q_*^{-1/2}$ in F-norm). *Let Assumptions 1, 2, and 3 be fulfilled, then*

$$\mathbb{P}\left\{\|Q_*^{-1/2} Q_n Q_*^{-1/2} - I\|_F \geq \frac{c_Q}{\sqrt{n}}(\sqrt{m} + t)\right\} \leq 2me^{-nt_F} + e^{-t^2/2} + (1-p)^n$$

for any $t \geq 0$ and $n \geq c_Q^2(\sqrt{m} + t)^2$, where

$$m \stackrel{\text{def}}{=} \dim(\mathbb{M}), \quad p \stackrel{\text{def}}{=} \mathbb{P}(\mathbb{H}_{++}(d)),$$

$$c_Q \stackrel{\text{def}}{=} \frac{4\|Q_*\|\sigma_T}{\lambda_{\min}(F')}, \quad t_F \stackrel{\text{def}}{=} C \min\left(\frac{\lambda_{\min}(F')}{U \log^{1/2}(U/\sigma_F)}, \frac{\lambda_{\min}^2(F')}{\sigma_F^2}\right),$$

where the operator F' is defined in (B.3), constant σ_T comes from auxiliary Proposition B.2, constants σ_F and U are defined in auxiliary Proposition B.1, and C denotes a generic constant.

To make the result more transparent, we further discuss it in a less formal way. The proof is based on three steps, and each step yields a bounding term. The first step gives the term $2me^{-nt_F}$. It deals with the concentration of some auxiliary empirical operator F'_n defined in (B.2) in the vicinity of its population counterpart F' . These two operators are essentially a price to pay for moving from the space of optimal transportation maps T_Q^S to the space of barycenters. The concentration of F'_n is derived from a result by Koltchinskii (2011) which is presented in Proposition B.1. The constants σ_F and U appear due to this concentration. Some prior bounds on σ_F and U are obtained in Lemma B.3. The second step yields the term $e^{-t^2/2}$. It ensures the concentration of $\|\frac{1}{n} \sum_i T_{Q_*}^{S_i} - I\|_F$, and relies on the result by Hsu, Kakade and Zhang (2012). To make the text self-contained, we introduce it in Proposition B.2. The constant σ_T comes from a bound on $\|\frac{1}{n} \sum_i T_{Q_*}^{S_i} - I\|_F$. The last step yields the term $(1-p)^n$. It comes from the requirement on nondegeneracy of Q_n . In other words, a high degeneracy leads to a smaller p and, thus, to worse bounds.

The next result deals with the concentration of $d_{\text{BW}}(Q_n, Q_*)$. It is a corollary of the above theorem.

COROLLARY 2.2 (Concentration of Q_n in d_{BW} distance). *Under the conditions of Theorem 2.4 the following result holds:*

$$\mathbb{P}\left\{d_{\text{BW}}(Q_n, Q_*) \geq \frac{c_Q \|Q_*\|^{1/2}}{\sqrt{n}}(\sqrt{m} + t)\right\} \leq 2me^{-nt_F} + e^{-t^2/2} + (1 - p)^n.$$

The last important result of the current study describes the concentration properties of the empirical Fréchet variance \mathcal{V}_n .

THEOREM 2.5 (Concentration of \mathcal{V}_n). *Let Assumptions 1, 2, and 3 be fulfilled, then, in the notation of Theorem 2.4,*

$$\mathbb{P}\{|\mathcal{V}_n - \mathcal{V}_*| \geq z(\mu, \nu, d, n, t)\} \leq 2me^{-nt_F} + 3e^{-t^2/2} + (1 - p)^n$$

with

$$z(b, \nu, d, n, t) \stackrel{\text{def}}{=} \max\left(\frac{\mu t^2}{n}, \frac{\nu t}{\sqrt{n}}\right) + 3\frac{c_Q^2 \|F'\|}{n}(\sqrt{m} + t)^2.$$

A pair (ν, μ) is the parameters of a subexponential r.v. $d_{\text{BW}}^2(Q_*, S)$.

All the proofs are collected in Section B.

2.4. *Central limit theorem and asymptotic normality of M-estimators.* A possible approach to obtain the central limit theorem is to look at a more general result concerning the asymptotic normality of M-estimators. To make the text self-contained, we briefly recall the subject following Section 5.4 in the book by van de Geer (2006). Under the setting (1.4), $d_{\text{BW}}^2(Q, S)$ might be considered as a loss function parametrized by elements of the affine subspace, $Q \in \mathbb{A} \cap \mathbb{H}_+(d)$. Thus, the proof of the CLT for an empirical barycenter is equivalent to a validation of the following conditions.

(C1) There exists a function $\psi_Q: \mathbb{H}_+(d) \rightarrow \mathbb{H}(d)$ which is $L_2(\mathbb{P})$ -integrable, s.t.

$$\lim_{Q \rightarrow Q_*} \frac{|d_{\text{BW}}^2(Q, S) - d_{\text{BW}}^2(Q_*, S) - \langle \psi_{Q_*}(S), Q - Q_* \rangle|}{\|Q - Q_*\|} = 0.$$

(C2) As $Q \rightarrow Q_*$, it holds

$$\begin{aligned} & \int (d_{\text{BW}}^2(Q, S) - d_{\text{BW}}^2(Q_*, S)) d\mathbb{P}(S) \\ &= \frac{1}{2} \langle Q - Q_*, V(Q - Q_*) \rangle + o(\|Q - Q_*\|), \end{aligned}$$

where V is some positive definite operator.

(C3) Let $Q \neq Q_*$, and define $g_Q(S) \stackrel{\text{def}}{=} \frac{d_{\text{BW}}^2(Q, S) - d_{\text{BW}}^2(Q_*, S)}{\|Q - Q_*\|}$. Suppose that for some $\varepsilon > 0$, the class $\{g_Q(S), Q: \|Q - Q_*\| \leq \varepsilon\}$ has an envelope $G \in L_2(\mathbb{P})$ and that it is a Donsker class.

Lemma A.6 presents differentiability of the Bures–Wasserstein distance and ensures the following quadratic approximation. For any $Q \in \mathbb{H}_{++}(d)$ it holds

$$\begin{aligned} & - \frac{2}{(1 + \lambda_{\max}^{1/2}(Q_*^{-1/2} Q Q_*^{-1/2}))^2} \langle dT_{Q_*}^S(Q - Q_*), Q - Q_* \rangle \\ & \leq d_{\text{BW}}^2(Q, S) - d_{\text{BW}}^2(Q_*, S) + \langle T_{Q_*}^S - I, Q - Q_* \rangle \\ & \leq - \frac{2}{(1 + \lambda_{\min}^{1/2}(Q_*^{-1/2} Q Q_*^{-1/2}))^2} \langle dT_{Q_*}^S(Q - Q_*), Q - Q_* \rangle, \end{aligned}$$

where $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ stand for maximal and minimal eigenvalues of a matrix A , respectively. This approximation ensures Condition (C1) and (C2) to be fulfilled with $\psi_{Q_*}(S) = I - T_{Q_*}^S$, and $V = -\mathbb{E} dT_{Q_*}^S$, respectively. However, it is not clear how one should proceed with the validation of Condition (C3). On the other hand, the direct proof of CLT introduced in the present study is suitable for the proof of the concentration results.

3. Connection to other problems. In this section we explain a connection of obtained results to some other problems. Section 3.1 investigates the relation between the Bures–Wasserstein barycenter and the 2-Wasserstein barycenter of some scale-location family. Section 3.2 illustrates the idea of a barycenter restricted to an affine subspace $\mathbb{A} \subset \mathbb{H}(d)$.

3.1. *Connection to scale-location families of measures.* We first present the concept of a scale-location family of absolutely continuous measures supported on \mathbb{R}^d .

DEFINITION 3.1. Let $X \sim \mu$ be a random variable following a law $\mu \in \mathcal{P}_2^{\text{ac}}(\mathbb{R}^d)$, where $\mathcal{P}_2^{\text{ac}}(\mathbb{R}^d)$ is the set of absolutely continuous measures with a finite second moment. A set of all affine transformations of X is written as

$$\mathcal{SL}(\mu) \stackrel{\text{def}}{=} \{ \mathcal{L}(PX + p) : P \in \text{Sym}_+(d), p \in \mathbb{R}^d \}.$$

It is referred to as a scale-location family.

Scale-location families play an important role in modern data analysis and appear in many practical applications due to being user-friendly in terms of theoretical analysis and, at the same time, possessing high modelling power. For example, it is widely used in medical imaging Wassermann et al. (2010), modelling of molecular dynamic Gonzalez et al. (2017), clustering procedures del Barrio et al. (2019b), climate modelling Mallasto and Feragen (2017), embedding of complex objects in low-dimensional spaces Muzellec and Cuturi (2018), and so on.

A possible metric that takes into account nonlinearity of the underlying data-set is the 2-Wasserstein distance, d_{W_2} . Let μ_X, μ_Y be elements of $\mathcal{SL}(\mu)$, and let random variables sampled from μ_X and μ_Y be $X \sim \mu_X, Y \sim \mu_Y$, respectively. We denote their first and second moments as

$$(3.1) \quad \mathbb{E} X = m_X, \quad \mathbb{E} Y = m_Y, \quad \text{Var}(X) = S_X, \quad \text{Var}(Y) = S_Y.$$

It is a well-known fact that d_{W_2} between measures coming from the same scale-location family depends only on the first and second moments of the measures

$$d_{W_2}^2(\mu_X, \mu_Y) = \|m_X - m_Y\|^2 + d_{\text{BW}}^2(S_X, S_Y).$$

For more details on a general class of optimal transportation distances we recommend excellent books Ambrosio and Gigli (2013), Villani (2009).

Distribution over a scale-location family. In many cases we are interested in data sets coming from some scale-location family. Let \mathbb{P} be a probability measure supported on some $\mathcal{SL}(\mu)$. And let $(\Omega, \mathcal{F}, \mathbb{P})$ be a generic probability space, s.t. for any $\omega \in \Omega$ there exists an image $\mu_\omega \stackrel{\text{def}}{=} \mathcal{L}(P_\omega X + p_\omega)$, where $P_\omega \in \text{Sym}_+(d)$ is a scaling parameter and $p_\omega \in \mathbb{R}^d$ is a shift parameter. A randomly sampled measure μ_ω belongs to $\mathcal{SL}(\mu)$ by construction, and its first and second moments (m_ω, S_ω) are written as

$$m_\omega \stackrel{\text{def}}{=} P_\omega r + p_\omega, \quad S_\omega \stackrel{\text{def}}{=} P_\omega Q P_\omega^\top,$$

where the pair (r, Q) denote the first and the second moments of the template measure μ . The Fréchet variance of \mathbb{P} at any arbitrary point μ' is written as

$$\mathcal{V}(\mu') \stackrel{\text{def}}{=} \int_{\text{supp}(\mathbb{P})} d_{W_2}^2(\mu', \nu_\omega) \mathbb{P}(d\omega).$$

Given an i.i.d. sample ν_1, \dots, ν_n from \mathbb{P} , we define the empirical counterpart of $\mathcal{V}(\mu')$:

$$\mathcal{V}_n(\mu') \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^n d_{W_2}^2(\mu', \nu_i).$$

Then the population and the empirical barycenters μ_* and μ_n are

$$\mu_* = \underset{\mu \in \mathcal{P}_2(\mathbb{R}^d)}{\text{argmin}} \mathcal{V}(\mu), \quad \mu_n = \underset{\mu \in \mathcal{P}_2(\mathbb{R}^d)}{\text{argmin}} \mathcal{V}_n(\mu).$$

Note that μ_* and μ_n belong to $\mathcal{SL}(\mu)$ and are uniquely characterised by their first and second moments, (r_*, Q_*) and (r_n, Q_n) , respectively; see Theorem 3.10 in [Álvarez-Esteban et al. \(2015\)](#):

$$(3.2) \quad r_* = \int_{\text{supp}(\mathbb{P})} m_\omega \mathbb{P}(d\omega), \quad Q_* = \int_{\text{supp}(\mathbb{P})} (Q_*^{1/2} S_\omega Q_*^{1/2})^{1/2} \mathbb{P}(d\omega),$$

$$(3.3) \quad r_n = \frac{1}{n} \sum_{i=1}^n m_i, \quad Q_n = \frac{1}{n} \sum_{i=1}^n (Q_n^{1/2} S_i Q_n^{1/2})^{1/2}.$$

It is worth noting that the concept of Wasserstein barycenter presented originally in a seminal work by [Agueh and Carlier \(2011\)](#) becomes a topic of extensive scientific interest in the last few years. One of the main reasons is an introduction of computationally feasible procedures; see, for example, [Cuturi \(2013\)](#), [Dvurechensky et al. \(2018\)](#), [Kroshnin et al. \(2019\)](#), [Peyré, Cuturi et al. \(2019\)](#). There are also many works dedicated to investigation of theoretical properties of barycenters. A work of [Bigot and Klein \(2012\)](#) focuses on the convergence of a parametric class of barycenters, while [Bigot, Cazelles and Papadakis \(2016\)](#) investigate the asymptotic properties of the regularised barycenters. The paper [Le Gouic and Loubes \(2017\)](#) ensures the convergence of the Wasserstein barycenters. To the best of our knowledge, the most state-of-the-art result concerning the rates of convergence of an empirical 2-Wasserstein barycenter is obtained as a particular illustration of a more general result by [Le Gouic et al. \(2019\)](#). Namely, this work establishes fast rates of convergence for empirical barycenters over a large class of geodesic spaces with curvature bounds in the sense of Alexandrov. This work extends and completes the results by [Ahidar-Coutrix, Le Gouic and Paris \(2018\)](#). The latter paper provides the rates of convergence for empirical barycenters of the Borel probability measure on a metric space either under assumptions on weak curvature constraint of the underlying space or for a case of a nonnegatively curved space on which geodesics, emanating from a barycenter, can be extended. Corollary 2.1 extends the above results for $d_{W_2}(\mu_n, \mu_*)$ for the case of barycenters constrained to an affine subspace for measures coming from some scale-location family. The rate is of order $n^{-1/2}$.

The paper [Kroshnin \(2018\)](#) obtains an analogue of the law of large numbers for the case of an arbitrary cost function on some affine subspace \mathbb{A} .

The paper [Agueh and Carlier \(2017\)](#) introduces the CLT for an empirical barycenter for \mathbb{P} supported on a finite set of Gaussian measures. It is worth noting that the idea of the proof also relies on the differentiability of optimal transportation maps.

At this point, it is worth mentioning that there are some other works dealing with the central limit theorem for the Wasserstein distance, for example, [del Barrio and Loubes \(2019\)](#), [Rippl, Munk and Sturm \(2016\)](#). However, the setting in these works differs significantly from

what is done in the present study. The paper [Rippl, Munk and Sturm \(2016\)](#) derives the central limit theorems for the p -Wasserstein distance, $p \geq 1$, between empirical distributions sampled from Gaussians supported on \mathbb{R}^d . The work [del Barrio and Loubes \(2019\)](#) establishes the central limit theorem and the variance bounds for the 2-Wasserstein distance between an empirical measure and its true underlying counterpart on \mathbb{R}^d . A result, similar in spirit to [Theorem 2.5](#) is obtained in [del Barrio et al. \(2019a\)](#). However, the authors consider only the space of probability measures supported on the real line, $d = 1$, endowed with 2-Wasserstein distance. To the best of our knowledge, there are no results similar to the concentration [Theorem 2.4](#) and [Corollary 2.2](#) in the case of 2-Wasserstein distance.

3.2. Connection to quantum mechanics. The original Bures metric appears in quantum mechanics in relation to the fidelity measure between two quantum states and is used for the measurement of quantum entanglement [Dajka, Łuczka and Hänggi \(2011\)](#), [Marian and Marian \(2008\)](#). Let ρ and σ be two density operators. In essence a density matrix ρ is a Hermitian positive semi-definite operator with the unit trace, $\rho \in \mathbb{H}_+(d)$, $\text{tr } \rho = 1$. It is used as a possible way of description of statistical state of a quantum system. For an introduction to the density operators theory one may look [Fano \(1957\)](#). Let ρ and σ be two quantum states:

$$(3.4) \quad \rho, \sigma \in \mathbb{H}_+(d), \quad \text{tr } \rho = 1, \quad \text{tr } \sigma = 1.$$

Fidelity of these states is defined as $\mathcal{F}(\rho, \sigma) = (\text{tr } \sqrt{\rho^{1/2} \sigma \rho^{1/2}})^2$. It quantifies ‘‘closeness’’ of ρ and σ ; see [Jozsa \(1994\)](#). It is obvious, that in case of (3.4) the Bures–Wasserstein distance turns into Bures distance:

$$(3.5) \quad d_B^2(\rho, \sigma) = 2(1 - \mathcal{F}^{1/2}(\rho, \sigma)).$$

The rest of this section illustrates the idea of the barycenter restricted to some affine subspace \mathbb{A} . Given a random ensemble of density matrices, one is able to recovery its mean using averaging in the Euclidean sense. However, the Bures–Wasserstein barycenter suggests an alternative way to define the barycenter in terms of fidelity measure (3.5). We consider a following statistical setting. Let (Ω, F, \mathbb{P}) be some mechanism which generates quantum states ρ_ω . Given an i.i.d. sample ρ_1, \dots, ρ_n we write a population and an empirical variance of \mathbb{P} as

$$\mathcal{V}(\sigma) = \int_{\text{supp}(\mathbb{P})} d_{\text{BW}}^2(\sigma, \rho_\omega) \mathbb{P}(d\omega), \quad \mathcal{V}_n(\sigma) = \frac{1}{n} \sum_{i=1}^n d_{\text{BW}}^2(\sigma, \rho_i).$$

Then the population and the empirical barycenters belonging the class of all $d \times d$ -dimensional density operators are defined as

$$\rho_* = \underset{\sigma: \text{tr } \sigma = 1}{\text{argmin}} \mathcal{V}(\sigma), \quad \rho_n = \underset{\sigma: \text{tr } \sigma = 1}{\text{argmin}} \mathcal{V}_n(\sigma).$$

It can be easily shown, that by ‘‘taking the global Fréchet barycenter’’ or, in other words neglecting the condition $\text{tr } \sigma = 1$, we end up with the global barycenter, which is the solution of the fixed point equation which is already mentioned in [Section 2](#): $\rho = \int (\rho^{1/2} \rho_\omega \rho^{1/2})^{1/2} \mathbb{P}(d\omega)$. However, this is a contraction mapping. Thus $\text{tr } \rho_* < 1$, and ρ_* is not a density operator. In other words the condition $\text{tr } \sigma = 1$ ensures, that ρ_* and ρ_n also belong to the class of density operators. Taking into account the results obtained in [Section 2](#), ρ_n is in some sense a natural consistent estimator of ρ_* with the known rate of convergence and known deviation properties.

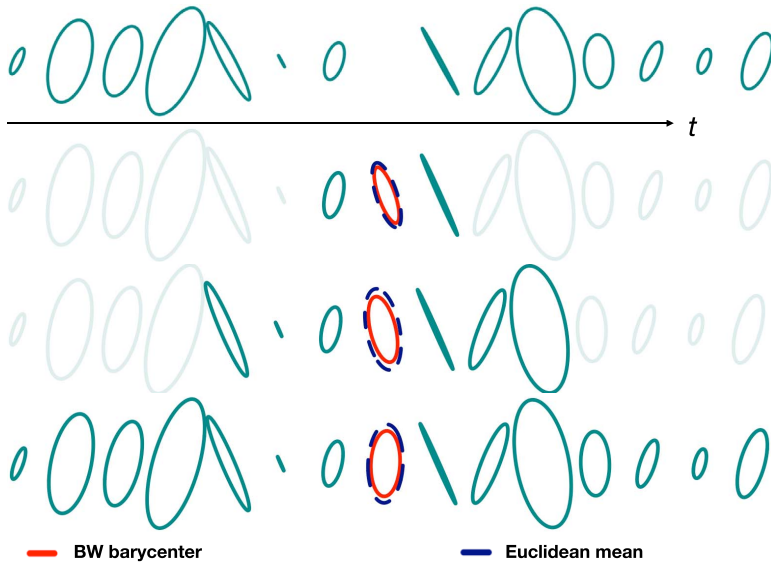


FIG. 1. *Interpolation of the lost data.*

4. Interpolation using empirical BW barycenters. We suggest to use the empirical Bures–Wasserstein barycenters for filling in gaps in data sets consisting of either measures coming from the same scale-location family, or from a family of Hermitian matrices. As a motivation we consider a data set related to the climate dynamics collected in Siberia (Russia) between 1930 and 2009, [Bulygina and Razuvaev \(2012\)](#), [Tatusko \(1974, 1990\)](#), where observations for the years 1934, 1938, 1942, 1948, and 1961 are lost. In this data set a behaviour of some quantities, such as a min/max daily temperatures during a year etcetera, is modelled using the Gaussian processes which parameters are estimated from the real measurements. We propose to replace the gaps in data with an an empirical barycenter constructed from available observations. To make the illustration more transparent, we consider a toy example for the case of two-dimensional covariance matrices. Each observed matrix is represented graphically by a two-dimensional ellipses. The upper panel at Figure 1 depicts a family of i.i.d. covariance matrices sampled consecutively over the discrete time t . The eighth observation is supposed to be missing. Three lower panels present a possible replacement constructed from two (2-d panel), six (3-d panel), and all available observations (4th panel). The observations used for data reconstruction are coloured in the dark green. The red ellipses correspond to the Bures–Wasserstein mean, while the blue ones depict the Euclidean mean. The difference in obtained results presented by three lower panels raises a question of a proper choice of number of observations used for missing data completion. Though being very interesting, this question is beyond the scope of the current study. Another question concerns the construction of nonasymptotic confidence sets for the estimators. For instance the work by [Ebert, Spokoiny and Suvorikova \(2017\)](#) suggests a suitable methodology based on multiplier bootstrap. However it considers only the case of commuting covariance matrices. Thus we consider this question as a matter for further research.

The next two sections provide some illustrations of the convergence rate of an empirical barycenter to the true one. To make the presentation complete, we also provide an illustration of the convergence of an empirical variance \mathcal{V}_n to \mathcal{V}_* .

4.1. *Simulated data.* In this section we consider a simulated data set. Covariance matrices are generated as follows. A matrix $\tilde{S}_k = A_k A_k^T$ is a d -dimensional matrix, where

Convergence of $\sqrt{n}\|Q_n - Q_*\|_F$

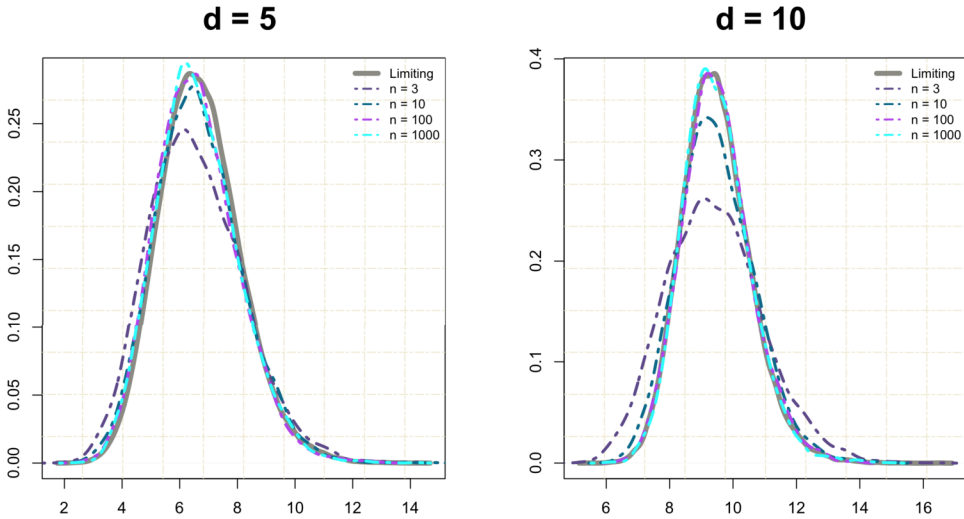


FIG. 2. Densities of $\mathcal{L}(\sqrt{n}\|Q_n - Q_*\|_F)$.

$A_k = (a_{ij}^k), a_{ij}^k \stackrel{iid}{\sim} \text{Unif}[0, 1] + 1$ for all i, j . To ensure that \tilde{S}_k is nondegenerated, we consider the orthogonal decomposition $\tilde{S}_k = \tilde{U}_k^* \tilde{\Lambda}_k \tilde{U}_k$, and replace $\tilde{\Lambda}_k$ by $\Lambda_k = \text{diag}(\lambda_1^k, \dots, \lambda_d^k)$ s.t. $\lambda_i^k \sim \text{Unif}[18, 22]$. Thus, an observed i.i.d. sample consists of matrices $S_k = \tilde{U}_k^* \Lambda_k \tilde{U}_k, k = 1, \dots, n$. In what follows, Q_n is a barycenter of the sample S_1, \dots, S_n .

Figure 2 illustrates the convergence of $\mathcal{L}(\sqrt{n}\|Q_n - Q_*\|_F)$ to $\mathcal{L}(\|Z\|_F)$ with $Z \sim \mathcal{N}(0, \Xi)$ presented in Theorem 2.2. Figure 3 depicts the convergence of the distribution $\mathcal{L}(\sqrt{nd}_{\text{BW}}(Q_n, Q_*))$ to $\mathcal{L}(\|Q_*^{1/2} dT_{Q_*}(Z)\|_F)$ obtained in Corollary 2.1. Finally, Figure 4 illustrates the convergence of density of $\mathcal{L}(\sqrt{n}(\mathcal{V}_* - \mathcal{V}_n))$ to the density of the Gaussian distribution $\mathcal{N}(0, \text{Var}(d_{\text{BW}}^2(Q_*, S)))$ validated by Theorem 2.3. The numerical experiments

Convergence of $\sqrt{nd}_{\text{BW}}(Q_n, Q_*)$

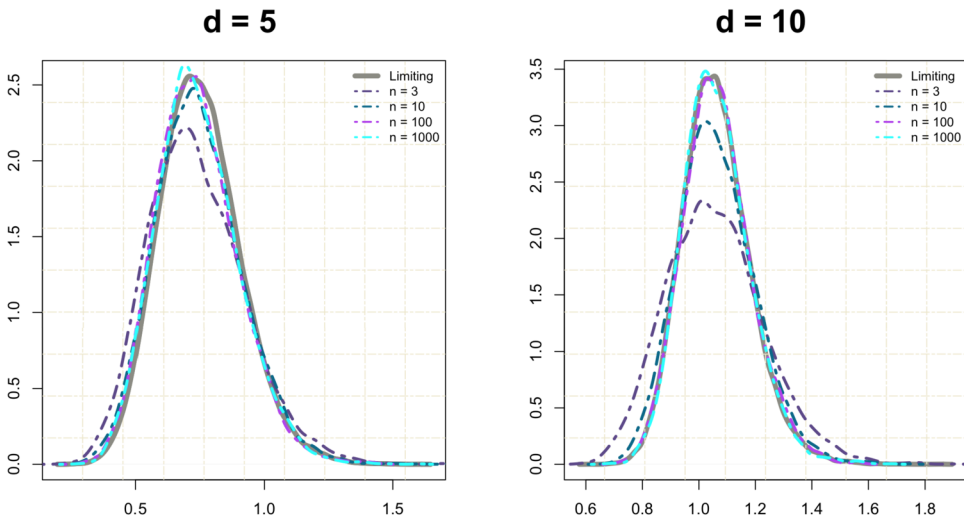


FIG. 3. Densities of $\mathcal{L}(\sqrt{nd}_{\text{BW}}(Q_n, Q_*))$.

Convergence of $\sqrt{n}(V_* - V_n)$

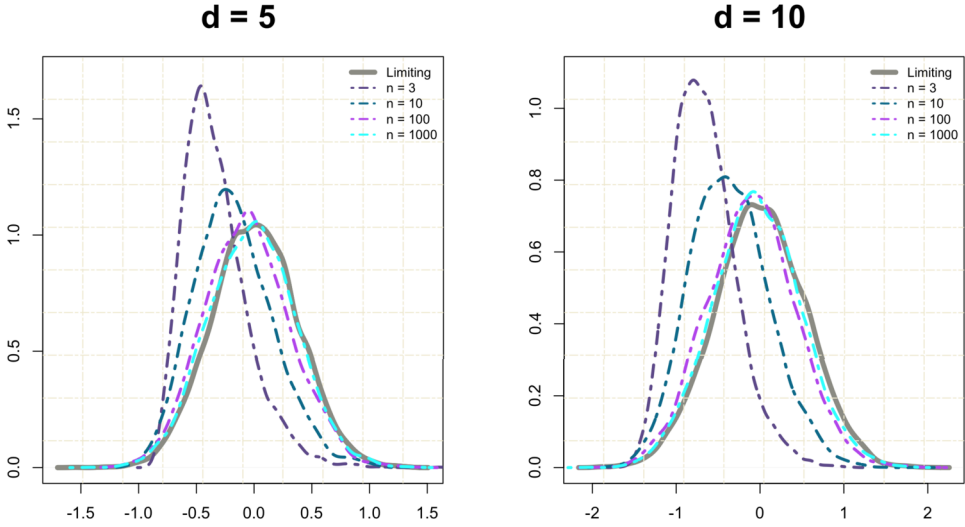


FIG. 4. Densities of $\mathcal{L}(\sqrt{n}(V_n - V_*))$.

are performed using R. The population barycenter Q_* was computed using a sample of 20,000 observed covariance matrices. A solid line depicts the density of a respective limiting distributions, while the dashed lines correspond to the densities of $\mathcal{L}(\sqrt{n}\|Q_n - Q_*\|_F)$, $\mathcal{L}(\sqrt{n}d_{\text{BW}}(Q_n, Q_*))$, or $\mathcal{L}(\sqrt{n}(V_* - V_n))$, respectively. We consider different sample sizes for calculation of an empirical Bures–Wasserstein barycenter Q_n with $n \in \{3, 10, 100, 1000\}$. Simulations were carried out for the dimensions $d = 5$ and $d = 10$.

4.2. *Data aggregation in climate modelling.* In this section we demonstrate the convergence rates for the climate-related data set. At first, we discuss the set in more detail. Following the original setting, we assume that the daily minimum temperatures within a year are described by a class of Gaussian processes. The temperature is measured at a set of 30 randomly sampled meteorological stations located in Siberia. Each Gaussian curve is obtained through the regression, and the maximum likelihood estimation, and is sampled in 50 points Mallasto and Feragen (2017). Thus, the observed data set \mathcal{D} consists of 71 Gaussian distributions:

$$\mathcal{D} = \{ \mathcal{N}(m_t, S_t), m_t \in \mathbb{R}^{50}, S_t \in \text{Sym}_{++}(50), t = 1, \dots, 71 \},$$

where $\mathcal{N}(m_t, S_t)$ is a Gaussian distribution related to a Gaussian process describing a t th year, $t = 1933, \dots, 2009$. The missing years in this data set are 1934, 1938, 1942, 1948 and 1961. This distribution is specified by a mean m_t and a covariance S_t . A Gaussian distribution $\mathcal{N}(r_*, Q_*)$ is the population Wasserstein barycenter of \mathcal{D} . It is characterised by the first and the second moments written as (r_*, Q_*)

$$r_* = \frac{1}{71} \sum_{t=1}^{71} m_t, \quad Q_* = \frac{1}{71} \sum_{t=1}^{71} (Q_*^{1/2} S_t Q_*^{1/2})^{1/2}.$$

A family of empirical barycenters $\mathcal{N}(r_n, Q_n)$ with parameters (r_n, Q_n) coming from (3.3) is constructed by means of re-sampling with replacement of the original data set.

Convergence of $\sqrt{n}\|Q_n - Q_*\|$

d = 50

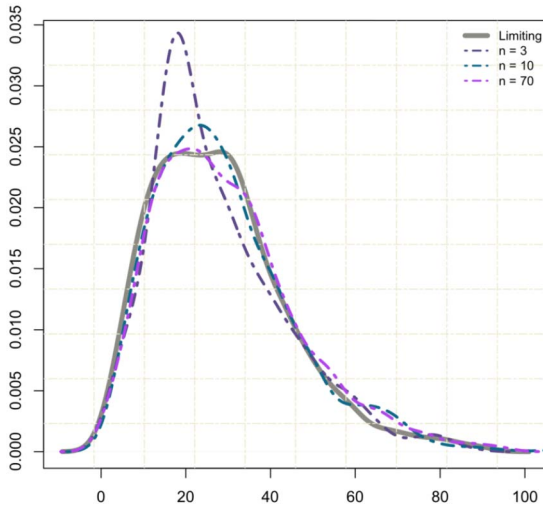


FIG. 5. $\mathcal{L}(\sqrt{n}\|Q_n - Q_*\|_F)$.

Figure 5 and Figure 6 present the convergence of densities of $\mathcal{L}(\sqrt{n}\|Q_n - Q_*\|_F)$ to $\mathcal{L}(\|Z\|_F)$ with $Z \sim \mathcal{N}(0, \Xi)$, and $\mathcal{L}(\sqrt{n}d_{\text{BW}}(Q_n, Q_*))$ to $\mathcal{L}(\|Q_*^{1/2}d\mathbf{T}_{Q_*}(Z)\|_F)$, respectively. The limiting distributions are depicted by solid lines, while the dashed ones stand for the densities computed for barycenters of n covariance matrices with $n \in \{3, 10, 70\}$.

APPENDIX A: PROOF OF CENTRAL LIMIT THEOREM

A.1. List of accepted notations. To make the presentation more transparent, we introduce a list of some used notations.

Convergence of $\sqrt{n}d_{\text{BW}}(Q_n, Q_*)$

d = 50

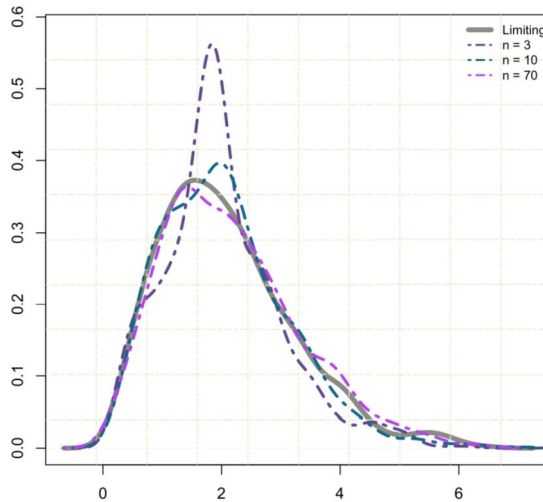


FIG. 6. $\mathcal{L}(\sqrt{n}d_{\text{BW}}(Q_n, Q_*))$.

A, B	Matrices or vectors
\mathbf{A}, \mathbf{B}	Operators
$\lambda_{\max}(\square), \lambda_{\min}(\square)$	Largest and smallest eigenvalue of an operator or a matrix
$(\square)_{\mathbb{M}}$	Restriction of a quadratic form to a subspace \mathbb{M}
$\ \square\ $	Operator norm
$\ \square\ _F$	Frobenius norm
$\ \square\ _1$	Schatten norm
$\ \square\ _{\psi_1}$	ψ_1 Orlicz norm
$\ \square\ _{\psi_2}$	ψ_2 Orlicz norm
$\kappa(\square) = \ \square\ \cdot \ \square^{-1}\ $	Condition number of an operator or a matrix
$\langle \square, \square \rangle$	Inner product associated to Frobenius norm
\otimes	Tensor product
$\mathcal{L}(X)$	Distribution of a r.v. X
\xrightarrow{w}	Weak convergence
$\xrightarrow{\text{a.s.}}$	A.s. convergence
$d_w(\square, \square)$	Metric inducing weak convergence
$o_P(\square)$	O-small in probability
$O_P(\square)$	O-big in probability

A.2. Properties of T_Q^S . PROOF OF PROPOSITION 2.1. First, we prove that optimal T is self-adjoint. Indeed, assume the opposite, then

$$Q^{1/2} T Q T^* Q^{1/2} = (Q^{1/2} T Q^{1/2})(Q^{1/2} T Q^{1/2})^* = Q^{1/2} S Q^{1/2}$$

and thus $\text{tr } Q^{1/2} T Q^{1/2} < \text{tr}(Q^{1/2} S Q^{1/2})^{1/2}$. Therefore

$$\begin{aligned} \text{tr}(T - I)Q(T^* - I) &= \text{tr } S + \text{tr } Q - 2 \text{tr } T Q = \text{tr } S + \text{tr } Q - 2 \text{tr } Q^{1/2} T Q^{1/2} \\ &> \text{tr } S + \text{tr } Q - 2 \text{tr}(Q^{1/2} S Q^{1/2})^{1/2} = d_{\text{BW}}^2(Q, S). \end{aligned}$$

If T is Hermitian but not positive semi-definite, then $Q^{1/2} T Q^{1/2} \preceq (Q^{1/2} S Q^{1/2})^{1/2}$, $Q^{1/2} T Q^{1/2} \neq (Q^{1/2} S Q^{1/2})^{1/2}$, hence again $\text{tr } Q^{1/2} T Q^{1/2} < \text{tr}(Q^{1/2} S Q^{1/2})^{1/2}$.

Finally, if $T \in \mathbb{H}_+(d)$, then it is straightforward to check that $T = T_Q^S$ given by (2.1) and

$$\text{tr}(T - I)Q(T^* - I) = \text{tr } S + \text{tr } Q - 2 \text{tr}(Q^{1/2} S Q^{1/2})^{1/2} = d_{\text{BW}}^2(Q, S). \quad \square$$

The proof of the central limit theorem mainly relies on the differentiability of the map (2.1). Lemma A.2 shows that T_Q^S can be linearised in the vicinity of Q :

$$T_{Q+X}^S = T_Q^S + dT_Q^S(X) + o(\|X\|),$$

where $dT_Q^S: \mathbb{H}(d) \rightarrow \mathbb{H}(d)$ is a self-adjoint negative-definite operator and $\|X\|$ stands for an operator norm of X . Properties of dT_Q^S are investigated in Lemma A.3. Let us introduce some notation: if $G(A)$ is a functional of a matrix A , then we denote its differential as $od_A G$.

LEMMA A.1. *Map $Q \mapsto g(Q) = Q^{1/2}$ is differentiable on $\mathbb{H}_{++}(d)$, and its differential is given by*

$$d_{\mathbf{Q}} g(X) = U^* \left(\frac{(U X U^*)_{ij}}{\sqrt{q_i} + \sqrt{q_j}} \right)_{i,j=1}^d U, \quad X \in \mathbb{H}(d),$$

where $Q = U^* \text{diag}(q)U$ is the eigenvalue decomposition.

PROOF. First, let us consider the map $P \mapsto f(P) = P^2$. It is smooth and its differential

$$d_P f(X) = PX + XP, \quad X \in \mathbb{H}(d)$$

is nondegenerated:

$$\langle d_P f(X), X \rangle = 2 \operatorname{tr} \overline{XPX} > 0, \quad X \neq 0,$$

whenever $P \in \mathbb{H}_{++}(d)$. From now on $\langle \cdot, \cdot \rangle$ denotes a scalar product associated to the Frobenius norm.

Now applying the inverse function theorem we obtain that the inverse map $g(\cdot)$ is also smooth and its differential enjoys the following equation

$$X = (d_P f|_{P=Q^{1/2}})(d_Q g(X)) = Q^{1/2} d_Q g(X) + d_Q g(X) Q^{1/2},$$

thus

$$\begin{aligned} UXU^* &= (\operatorname{diag}(q))^{1/2} U d_Q g(X) U^* + U d_Q g(X) U^* (\operatorname{diag}(q))^{1/2}, \\ (UXU^*)_{ij} &= (\sqrt{q_i} + \sqrt{q_j})(U d_Q g(X) U^*)_{ij}, \quad 1 \leq i, j \leq d, \end{aligned}$$

and

$$d_Q g(X) = U^* \left(\frac{(UXU^*)_{ij}}{\sqrt{q_i} + \sqrt{q_j}} \right)_{i,j=1}^d U. \quad \square$$

LEMMA A.2 (Fréchet-differentiability of the map T_Q^S). *For any $S \in \mathbb{H}_+(d)$ the map T_Q^S can be linearised in the vicinity of $Q \in \mathbb{H}_{++}(d)$ as*

$$T_{\tilde{Q}}^S = T_Q^S + dT_Q^S(\tilde{Q} - Q) + o(\|\tilde{Q} - Q\|) \quad \text{as } \tilde{Q} \rightarrow Q,$$

where

$$(A.1) \quad dT_Q^S(X) \stackrel{\text{def}}{=} -S^{1/2} U^* \Lambda^{-1/2} \delta \Lambda^{-1/2} U S^{1/2}, \quad X \in \mathbb{H}(d),$$

$U^* \Lambda U$ is an eigenvalue decomposition of $S^{1/2} Q S^{1/2}$

$$U^* \Lambda U = S^{1/2} Q S^{1/2}, \quad U^* U = U U^* = I, \quad \Lambda = \operatorname{diag}(\lambda_1, \dots, \lambda_{\operatorname{rank}(S)}, 0, \dots, 0),$$

$$\Lambda^{-1/2} = (\Lambda^{1/2})^+ = \operatorname{diag}(\lambda_1^{-1/2}, \dots, \lambda_{\operatorname{rank}(S)}^{-1/2}, 0, \dots, 0),$$

$$\delta = (\delta_{ij})_{i,j=1}^d, \quad \delta_{ij} = \begin{cases} \frac{\Delta_{ij}}{\sqrt{\lambda_i} + \sqrt{\lambda_j}} & i, j \leq \operatorname{rank}(S), \\ 0 & \text{otherwise,} \end{cases} \quad \Delta = U S^{1/2} X S^{1/2} U^*.$$

PROOF. The proof mainly relies on the differentiation of the pseudo-inverse term $(S^{1/2} Q S^{1/2})^{-1/2}$, as long as

$$dT_Q^S(X) = S^{1/2} d_Q (S^{1/2} Q S^{1/2})^{-1/2}(X) S^{1/2}.$$

Obviously we can consider only restriction to $\operatorname{range}(S)$ and therefore assume w.l.o.g. $S \succ 0$. As $(S^{1/2}(Q + X)S^{1/2})^{-1/2} = U^*(\Lambda + \Delta)^{-1/2}U$, by Lemma A.1 and von Neumann series expansion we obtain for infinitesimal $X \in \mathbb{H}(d)$ and corresponding Δ that

$$\begin{aligned} (\Lambda + \Delta)^{-1/2} &= (\Lambda^{1/2} + \delta + o(\|\Delta\|))^{-1} \\ &= (\Lambda^{1/4}(I + \Lambda^{-1/4}\delta\Lambda^{-1/4} + o(\|\Delta\|))\Lambda^{1/4})^{-1} \\ &= \Lambda^{-1/4}(I - \Lambda^{-1/4}\delta\Lambda^{-1/4} + o(\|\Delta\|))\Lambda^{-1/4} \\ &= \Lambda^{-1/2} - \Lambda^{-1/2}\delta\Lambda^{-1/2} + o(\|\Delta\|). \end{aligned}$$

Then the differential $d_{\mathbf{Q}}(S^{1/2}QS^{1/2})^{-1/2}(X)$ is written as

$$d_{\mathbf{Q}}(S^{1/2}QS^{1/2})^{-1/2}(X) = -U^* \Lambda^{-1/2} \delta \Lambda^{-1/2} U.$$

Therefore,

$$T_{\mathbf{Q}+X}^S = T_{\mathbf{Q}}^S + d\mathbf{T}_{\mathbf{Q}}^S(X) + o(\|X\|),$$

where $d\mathbf{T}_{\mathbf{Q}}^S(X)$ is defined by (A.1). \square

Lemmas A.3 and A.4 are technical and explore properties of $d\mathbf{T}_{\mathbf{Q}}^S$.

LEMMA A.3. *For any $S \in \mathbb{H}_+(d)$, $\mathbf{Q} \in \mathbb{H}_{++}(d)$, the properties of operator $d\mathbf{T}_{\mathbf{Q}}^S$ defined in (A.1) are following:*

- (I) *it is self-adjoint;*
- (II) *it is negative semi-definite;*
- (III) *it enjoys the following bounds:*

$$\begin{aligned}
 -\langle d\mathbf{T}_{\mathbf{Q}}^S(X), X \rangle &\leq \frac{\lambda_{\max}^{1/2}(S^{1/2}QS^{1/2})}{2} \|Q^{-1/2}XQ^{-1/2}\|_F^2, \\
 -\langle d\mathbf{T}_{\mathbf{Q}}^S(X), X \rangle &\geq \frac{\lambda_{\min}^{1/2}(S^{1/2}QS^{1/2})}{2} \|Q^{-1/2}XQ^{-1/2}\|_F^2.
 \end{aligned}$$

(IV) *it is homogeneous w.r.t. \mathbf{Q} with degree $-\frac{3}{2}$ and w.r.t. S with degree $\frac{1}{2}$, that is, $d\mathbf{T}_{a\mathbf{Q}}^S = a^{-3/2}d\mathbf{T}_{\mathbf{Q}}^S$ and $d\mathbf{T}_{\mathbf{Q}}^{aS} = a^{1/2}d\mathbf{T}_{\mathbf{Q}}^S$ for any $a > 0$;*

(V) *it is monotone w.r.t. $S^{1/2}QS^{1/2}$ (once range S is fixed): $d\mathbf{T}_{\mathbf{Q}_0}^{S_0} \preceq d\mathbf{T}_{\mathbf{Q}_1}^{S_1}$ in the sense of self-adjoint operators on $\mathbb{H}(d)$ whenever $S_0^{1/2}\mathbf{Q}_0S_0^{1/2} \preceq S_1^{1/2}\mathbf{Q}_1S_1^{1/2}$ and $\text{range}(S_0) = \text{range}(S_1)$; in particular, $d\mathbf{T}_{\mathbf{Q}}^S$ is monotone w.r.t. $\mathbf{Q} \in \mathbb{H}_{++}(d)$ for fixed S .*

PROOF. Slightly changing notation, we rewrite (A.1) as

$$d\mathbf{T}_{\mathbf{Q}}^S(X) = -S^{1/2}U^* \Lambda^{-1/2} \delta^X \Lambda^{-1/2} US^{1/2},$$

where matrices U and Λ come from Lemma A.2 and

$$\delta^X = (\delta_{ij}^X)_{i,j=1}^d, \quad \delta_{ij}^X = \frac{\Delta_{ij}^X}{\sqrt{\lambda_i} + \sqrt{\lambda_j}}, \quad \Delta^X = US^{1/2}XS^{1/2}U^*.$$

(I) *Self-adjointness.* Consider a scalar product

$$\begin{aligned}
 \langle d\mathbf{T}_{\mathbf{Q}}^S(X), Y \rangle &= \text{tr}(d\mathbf{T}_{\mathbf{Q}}^S(X)Y) = -\text{tr}(S^{1/2}U^* \Lambda^{-1/2} \delta^X \Lambda^{-1/2} US^{1/2}Y) \\
 &= -\text{tr}(\Lambda^{-1/2} \delta^X \Lambda^{-1/2} US^{1/2}YS^{1/2}U^*).
 \end{aligned}$$

We now introduce the following notation:

$$\Delta^Y \stackrel{\text{def}}{=} US^{1/2}YS^{1/2}U^*.$$

Then the above equality can be continued as follows:

$$\begin{aligned}
 &-\text{tr}(\Lambda^{-1/2} \delta^X \Lambda^{-1/2} US^{1/2}YS^{1/2}U^*) \\
 &= -\text{tr}(\Lambda^{-1/2} \delta^X \Lambda^{-1/2} \Delta^Y) \\
 &= -\sum_{i,j=1}^r \frac{\delta_{ij}^X}{\sqrt{\lambda_i \lambda_j}} \Delta_{ij}^Y = -\sum_{i,j=1}^r \frac{\Delta_{ij}^X \Delta_{ij}^Y}{\sqrt{\lambda_i \lambda_j} (\sqrt{\lambda_i} + \sqrt{\lambda_j})} \\
 &= \text{tr}(d\mathbf{T}_{\mathbf{Q}}^S(Y)X) = \text{tr}(Xd\mathbf{T}_{\mathbf{Q}}^S(Y)) = \langle X, d\mathbf{T}_{\mathbf{Q}}^S(Y) \rangle,
 \end{aligned}$$

where $r = \text{rank}(S)$. Thus the operator is self-adjoint.

(II) *Boundedness and (III) eigenvalues.* Denoting Δ^X by Δ (i.e., now $\Delta = US^{1/2}X \times S^{1/2}U^*$) and taking into account the above expansion of an inner product, one obtains

$$(A.2) \quad -\langle d\mathbf{T}_Q^S(X), X \rangle = \sum_{i,j=1}^r \frac{\Delta_{ij}^2}{\sqrt{\lambda_i \lambda_j}(\sqrt{\lambda_i} + \sqrt{\lambda_j})} = \sum_{i,j=1}^r \left(\frac{\Delta_{ij}}{\sqrt{\lambda_i \lambda_j}} \right)^2 \frac{\sqrt{\lambda_i \lambda_j}}{\sqrt{\lambda_i} + \sqrt{\lambda_j}}.$$

Note that the function $f(\lambda_i, \lambda_j) \stackrel{\text{def}}{=} \frac{\sqrt{\lambda_i \lambda_j}}{\sqrt{\lambda_i} + \sqrt{\lambda_j}}$ is monotonously increasing in both arguments λ_i and λ_j , thus

$$(A.3) \quad \max_{i,j} f(\lambda_i, \lambda_j) = \frac{\lambda_{\max}^{1/2}(\Lambda)}{2}, \quad \min_{i,j} f(\lambda_i, \lambda_j) = \frac{\lambda_{\min}^{1/2}(\Lambda)}{2}.$$

For the sake of simplicity we introduce a new variable

$$\zeta \stackrel{\text{def}}{=} Q^{-1/2} X Q^{-1/2},$$

its Frobenius norm is written as

$$\|\zeta\|_F^2 = \text{tr}(X Q^{-1} X Q^{-1}).$$

Moreover, the following inequality for trace holds:

$$\begin{aligned} \text{tr}(X Q^{-1} X Q^{-1}) &\geq \text{tr}(\Pi_S X \Pi_S Q^{-1} \Pi_S X \Pi_S Q^{-1} \Pi_S) \\ &= \text{tr}(\Delta \Lambda^+ \Delta \Lambda^+) = \|\Lambda^{-1/2} \Delta \Lambda^{-1/2}\|_F^2 = \sum_{i,j=1}^r \frac{\Delta_{ij}^2}{\lambda_i \lambda_j}. \end{aligned}$$

Here Π_S is the orthogonal projector onto the range of S .

Then combining (A.2) with (A.3), the upper and lower bounds can be obtained as follows:

$$\begin{aligned} -\langle d\mathbf{T}_Q^S(X), X \rangle &\leq \max_{i,j} f(\lambda_i, \lambda_j) \sum_{i,j=1}^r \left(\frac{\Delta_{ij}}{\sqrt{\lambda_i \lambda_j}} \right)^2 \leq \frac{\lambda_{\max}^{1/2}(\Lambda)}{2} \|\zeta\|_F^2, \\ -\langle d\mathbf{T}_Q^S(X), X \rangle &\geq \min_{i,j} f(\lambda_i, \lambda_j) \sum_{i,j=1}^r \left(\frac{\Delta_{ij}}{\sqrt{\lambda_i \lambda_j}} \right)^2 = \frac{\lambda_{\min}^{1/2}(\Lambda)}{2} \|\zeta\|_F^2. \end{aligned}$$

Note that if S is degenerated, the lower bound becomes trivial.

(IV) *Homogeneity and (V) monotonicity.* Homogeneity follows directly from definition (A.1). Now we prove monotonicity. As the range of $S^{1/2} Q S^{1/2}$ is fixed, we may assume $S > 0$. Consider

$$\begin{aligned} \langle d\mathbf{T}_Q^S(X), X \rangle &= \text{tr}(S^{1/2} U^* \Lambda^{-1/2} \delta \Lambda^{-1/2} U S^{1/2}, X) \\ &= \langle U^* \Lambda^{-1/2} \delta \Lambda^{-1/2} U, S^{1/2} X S^{1/2} \rangle \\ &= \langle d_Q(S^{1/2} Q S^{1/2})^{-1/2}(X), S^{1/2} X S^{1/2} \rangle \\ &= \langle d_M M^{-1/2}(S^{1/2} X S^{1/2}), S^{1/2} X S^{1/2} \rangle, \end{aligned}$$

with replacement $M = S^{1/2} Q S^{1/2}$ to be change of variables. As long as X is supposed to be fixed, it is enough to show that the differential $d_M M^{-1/2}$ is monotone in M . Notice that the

operator $(d_M M^{-1/2})^{-1}$ at point M is equal to the differential of the inverse map $P \mapsto P^{-2}$ at point $P = M^{-1/2}$:

$$d_M M^{-1/2} = (d_P P^{-2}|_{P=M^{-1/2}})^{-1}.$$

In turn, $d_P P^{-2}$ can be expressed as

$$d_P P^{-2}(X) = -P^{-1}(P^{-1}X + XP^{-1})P^{-1},$$

the right part of the above equation is self-adjoint, negative-definite and

$$\langle -P^{-1}(P^{-1}X + XP^{-1})P^{-1}, X \rangle = -2 \operatorname{tr} P^{-2} X P^{-1} X.$$

Choose $M_1 \succcurlyeq M_0 \succ 0$ (thus $M_1^{1/2} \succcurlyeq M_0^{1/2}$) and let $P_i = M_i^{-1/2}$ for $i = 0, 1$. Then for any fixed $X \in \mathbb{H}(d)$

$$-\operatorname{tr} P_1^{-2} X P_1^{-1} X = -\operatorname{tr} M_1 X M_1^{1/2} X \leq -\operatorname{tr} M_0 X M_0^{1/2} X = -\operatorname{tr} P_0^{-2} X P_0^{-1} X,$$

that is, $d_P P^{-2}|_{P_1} \preccurlyeq d_P P^{-2}|_{P_0}$ and hence for the differential of $M \mapsto M^{-1/2}$ the inverse inequality holds: $d_M M^{-1/2}|_{M_0} \preccurlyeq d_M M^{-1/2}|_{M_1}$. This entails monotonicity of $d\mathbf{T}_Q^S$. \square

COROLLARY A.1. *Under conditions of Lemma A.3, it holds*

$$\lambda_{\max}(-d\mathbf{T}_Q^S) \leq \frac{\lambda_{\max}^{1/2}(S^{1/2}QS^{1/2})}{2\lambda_{\min}^2(Q)}, \quad \lambda_{\min}(-d\mathbf{T}_Q^S) \geq \frac{\lambda_{\min}^{1/2}(S^{1/2}QS^{1/2})}{2\lambda_{\max}^2(Q)}.$$

PROOF. Item (III) from the above lemma ensures that

$$-\langle d\mathbf{T}_Q^S(X), X \rangle \leq \frac{\lambda_{\max}^{1/2}(S^{1/2}QS^{1/2})}{2} \|Q^{-1/2}XQ^{-1/2}\|_F^2 \leq \frac{\lambda_{\max}^{1/2}(S^{1/2}QS^{1/2})}{2\lambda_{\min}^2(Q)} \|X\|_F^2.$$

The second bound is proved in a similar way. \square

COROLLARY A.2. *We define a following rescaled operator*

$$(A.4) \quad d\mathbf{t}_Q^S(\zeta) \stackrel{\text{def}}{=} Q^{1/2}d\mathbf{T}_Q^S(Q^{1/2}\zeta Q^{1/2})Q^{1/2}, \quad \zeta \in \mathbb{H}(d).$$

Then, a following bound on its eigenvalues holds:

$$\begin{aligned} \lambda_{\min}(-d\mathbf{t}_Q^S) &= \frac{1}{2}\lambda_{\min}^{1/2}(S^{1/2}QS^{1/2}), \\ \lambda_{\max}(-d\mathbf{t}_Q^S) &= \frac{1}{2}\lambda_{\max}^{1/2}(S^{1/2}QS^{1/2}). \end{aligned}$$

PROOF. Notice that inequalities

$$\begin{aligned} \lambda_{\min}(-d\mathbf{t}_Q^S) &\geq \frac{1}{2}\lambda_{\min}^{1/2}(S^{1/2}QS^{1/2}), \\ \lambda_{\max}(-d\mathbf{t}_Q^S) &\leq \frac{1}{2}\lambda_{\max}^{1/2}(S^{1/2}QS^{1/2}), \end{aligned}$$

are a trivial consequence of Lemma A.3(III). Now defining for any $1 \leq k \leq \operatorname{rank}(S)$

$$\Delta_{ij}^k = \begin{cases} 1 & i = j = k, \\ 0 & \text{otherwise,} \end{cases} \quad X^k = S^{-1/2}U\Delta^kU^*S^{-1/2}, \quad \zeta^k = Q^{-1/2}X^kQ^{-1/2}$$

we obtain from (A.2) that

$$-\langle d\mathbf{t}_Q^S(\zeta^k), \zeta^k \rangle = -\langle d\mathbf{T}_Q^S(X^k), X^k \rangle = \frac{\lambda_k^{1/2}}{2} \|\zeta^k\|_F^2.$$

Therefore, the above inequalities are sharp. \square

LEMMA A.4. For any $Q_0, Q_1 \in \mathbb{H}_{++}(d)$, $S \in \mathbb{H}_+(d)$ consider

$$(A.5) \quad Q_t \stackrel{\text{def}}{=} (1-t)Q_0 + tQ_1, \quad Q' \stackrel{\text{def}}{=} Q_0^{-1/2}Q_1Q_0^{-1/2}.$$

Then

$$(I) \quad \begin{aligned} \frac{2}{\lambda_{\min}(Q') + \lambda_{\min}^{1/2}(Q')} dT_{Q_0}^S &\preceq \int_0^1 dT_{Q_t}^S dt \\ &\preceq \frac{2}{\lambda_{\max}(Q') + \lambda_{\max}^{1/2}(Q')} dT_{Q_0}^S \\ &\preceq \frac{1}{1 + 3\|Q' - I\|/4} dT_{Q_0}^S. \end{aligned}$$

Moreover, if $\|Q' - I\| < 1$, then

$$(II) \quad \int_0^1 dT_{Q_t}^S dt \succeq \frac{1}{1 - \|Q' - I\|} dT_{Q_0}^S.$$

REMARK 2. The above inequality might seem confusing due to the fact that $\lambda_{\min}(\cdot) \leq \lambda_{\max}(\cdot)$, however this is explained by the fact that dT_Q^S is *negative* definite.

PROOF. Notice that

$$((1-t) + t\lambda_{\min}(Q'))Q_0 \preceq Q_t = Q_0^{1/2}((1-t)I + tQ')Q_0^{1/2} \preceq ((1-t) + t\lambda_{\max}(Q'))Q_0.$$

Monotonicity and homogeneity with degree $-\frac{3}{2}$ of dT_Q^S (see Lemma A.3) yield

$$\begin{aligned} dT_{Q_t}^S &\preceq dT_{((1-t)+t\lambda_{\max}(Q'))Q_0}^S \\ &= ((1-t) + t\lambda_{\max}(Q'))^{-3/2} dT_{Q_0}^S \end{aligned}$$

and

$$\begin{aligned} dT_{Q_t}^S &\succeq dT_{((1-t)+t\lambda_{\min}(Q'))Q_0}^S \\ &= ((1-t) + t\lambda_{\min}(Q'))^{-3/2} dT_{Q_0}^S. \end{aligned}$$

Therefore,

$$\begin{aligned} \int_0^1 dT_{Q_t}^S dt &\preceq dT_{Q_0}^S \int_0^1 ((1-t) + t\lambda_{\max}(Q'))^{-3/2} dt \\ &= \frac{2}{\lambda_{\max}(Q') + \lambda_{\max}^{1/2}(Q')} dT_{Q_0}^S \end{aligned}$$

and respectively,

$$\int_0^1 dT_{Q_t}^S dt \succeq \frac{2}{\lambda_{\min}(Q') + \lambda_{\min}^{1/2}(Q')} dT_{Q_0}^S.$$

The inequality (II) follows from the fact that

$$\lambda_{\min}(Q') \geq 1 - \|Q' - I\|, \quad \lambda_{\max}(Q') \leq 1 + \|Q' - I\|,$$

and inequalities

$$\begin{aligned} \sqrt{1+x} &\leq 1 + \frac{x}{2} \quad \text{for } x \geq 0, \\ \sqrt{1-x} &\geq 1 - x \quad \text{for } 0 \leq x \leq 1. \end{aligned}$$

□

A.3. Properties of $d_{\text{BW}}(Q, S)$. The next lemma ensures strict convexity of $d_{\text{BW}}(Q, S)$. In essence, the proof mainly relies on Theorem 7 in [Bhatia, Jain and Lim \(2019\)](#).

LEMMA A.5. *For any $S \in \mathbb{H}_+(d)$ a function $Q \mapsto d_{\text{BW}}^2(Q, S)$ is convex on $\mathbb{H}_+(d)$. Moreover, if $S \succ 0$, then it is strictly convex.*

PROOF. According to [Bhatia, Jain and Lim \(\(2019\), Theorem 7\)](#) a function $h(X) = \text{tr } X^{1/2}$ is strictly concave on $\mathbb{H}_+(d)$, hence the function

$$Q \mapsto d_{\text{BW}}^2(Q, S) = \text{tr } S + \text{tr } Q - 2 \text{tr}(S^{1/2} Q S^{1/2})^{1/2}$$

is convex on $\mathbb{H}_+(d)$ for any positive semi-definite S . Moreover, if $S \succ 0$, then $Q \mapsto S^{1/2} Q S^{1/2}$ is an injective linear map, and therefore $d_{\text{BW}}^2(Q, S)$ is strictly convex. \square

Further we introduce differentiability of $d_{\text{BW}}^2(Q, S)$ and provides its quadratic approximation.

LEMMA A.6. *For any $Q \in \mathbb{H}_{++}(d)$, $S \in \mathbb{H}_+(d)$ the function $d_{\text{BW}}^2(Q, S)$ is twice differentiable in Q with*

$$\begin{aligned} d_Q d_{\text{BW}}^2(Q, S)(X) &= \langle I - T_Q^S, X \rangle, \quad X \in \mathbb{H}(d), \\ d_Q^2 d_{\text{BW}}^2(Q, S)(X, Y) &= -\langle X, dT_Q^S(Y) \rangle, \quad X, Y \in \mathbb{H}(d). \end{aligned}$$

Moreover, the following quadratic approximation holds: for any $Q_0, Q_1 \in \mathbb{H}_{++}(d)$,

$$\begin{aligned} & - \frac{2}{(1 + \lambda_{\max}^{1/2}(Q'))^2} \langle dT_{Q_0}^S(Q_1 - Q_0), Q_1 - Q_0 \rangle \\ & \leq d_{\text{BW}}^2(Q_1, S) - d_{\text{BW}}^2(Q_0, S) + \langle T_{Q_0}^S - I, Q_1 - Q_0 \rangle \\ & \leq - \frac{2}{(1 + \lambda_{\min}^{1/2}(Q'))^2} \langle dT_{Q_0}^S(Q_1 - Q_0), Q_1 - Q_0 \rangle \end{aligned}$$

with Q' defined in (A.5).

PROOF. Note that

$$d_Q (S^{1/2} Q S^{1/2})^{1/2}(X) = U^* \delta U,$$

where δ comes from Lemma A.2. Furthermore, Lemma A.1 implies that

$$\begin{aligned} d_Q \text{tr}(S^{1/2} Q S^{1/2})^{1/2}(X) &= \text{tr } d_Q (S^{1/2} Q S^{1/2})^{1/2}(X) = \text{tr } \delta \\ &= \sum_{i=1}^{\text{rank}(S)} \frac{\Delta_{ii}}{2\sqrt{\lambda_i}} = \frac{1}{2} \text{tr } \Delta \Lambda^{-1/2} \\ &= \frac{1}{2} \text{tr } S^{1/2} X S^{1/2} (S^{1/2} Q S^{1/2})^{-1/2} = \frac{1}{2} \langle T_Q^S, X \rangle. \end{aligned}$$

Consequently, $d_{\text{BW}}^2(Q, S)$ is differentiable, and

$$d_Q d_{\text{BW}}^2(Q, S)(X) = \text{tr } X - 2 d_Q \text{tr}(S^{1/2} Q S^{1/2})^{1/2}(X) = \langle I - T_Q^S, X \rangle.$$

Applying Lemma A.2 one obtains

$$d_Q^2 d_{\text{BW}}^2(Q, S)(X, Y) = d_Q \langle I - T_Q^S, X \rangle = -\langle dT_Q^S(Y), X \rangle(Y).$$

Quadratic approximation. Let $Q_0, Q_1 \in \mathbb{H}_{++}(d)$, $Q_t \stackrel{\text{def}}{=} (1-t)Q_0 + tQ_1$, $t \in [0, 1]$. The Taylor expansion in the integral form applied to $d_{\text{BW}}^2(Q_t, S)$ implies

$$\begin{aligned} d_{\text{BW}}^2(Q_1, S) &= d_{\text{BW}}^2(Q_0, S) + \langle I - T_{Q_0}^S, Q_1 - Q_0 \rangle \\ &\quad + \int_0^1 (1-t) \langle -dT_{Q_t}^S(Q_1 - Q_0), Q_1 - Q_0 \rangle dt \\ &= d_{\text{BW}}^2(Q_0, S) - \langle T_{Q_0}^S - I, Q_1 - Q_0 \rangle \\ &\quad - \left\langle \left[\int_0^1 (1-t) dT_{Q_t}^S dt \right] (Q_1 - Q_0), Q_1 - Q_0 \right\rangle. \end{aligned}$$

Following the same ideas as in the proof of Lemma A.4 one obtains that

$$\begin{aligned} \int_0^1 (1-t) dT_{Q_t}^S dt &\preceq \int_0^1 (1-t) ((1-t) + t\lambda_{\max}(Q'))^{-3/2} dT_{Q_0}^S dt \\ &= \frac{2}{(1 + \lambda_{\max}^{1/2}(Q'))^2} dT_{Q_0}^S \end{aligned}$$

and

$$\int_0^1 (1-t) dT_{Q_t}^S dt \succeq \frac{2}{(1 + \lambda_{\min}^{1/2}(Q'))^2} dT_{Q_0}^S.$$

Thus

$$\begin{aligned} & - \frac{2}{(1 + \lambda_{\max}^{1/2}(Q'))^2} \langle dT_{Q_0}^S(Q_1 - Q_0), Q_1 - Q_0 \rangle \\ & \leq d_{\text{BW}}^2(Q_1, S) - d_{\text{BW}}^2(Q_0, S) + \langle T_{Q_0}^S - I, Q_1 - Q_0 \rangle \\ & \leq - \frac{2}{(1 + \lambda_{\min}^{1/2}(Q'))^2} \langle dT_{Q_0}^S(Q_1 - Q_0), Q_1 - Q_0 \rangle. \quad \square \end{aligned}$$

A.4. Central limit theorem for Q_n and \mathcal{V}_n . First let us prove uniqueness and positive-definiteness of Bures–Wasserstein barycenter.

PROOF OF THEOREM 2.1. By Assumption 2, $\mathcal{V}(0)$ is bounded:

$$\mathcal{V}(0) = \mathbb{E} d_{\text{BW}}^2(0, S) = \mathbb{E} \text{tr} S < \infty.$$

Moreover, $d_{\text{BW}}(Q, S) \rightarrow \infty$ as $\|Q\| \rightarrow \infty$. This implies $\mathcal{V}(Q) \rightarrow \infty$ as $\|Q\| \rightarrow \infty$. Thus, any minimizing sequence for $\mathcal{V}(\cdot)$ is bounded. This observation allows us to use the compactness argument. As $\mathcal{V}(\cdot)$ is continuous, this implies existence of a barycenter Q_* by the compactness argument.

In the case of $\mathbb{P}(\mathbb{H}_{++}(d)) > 0$ applying Lemma A.5 we obtain strict convexity of the integral

$$Q \mapsto \mathbb{E} d_{\text{BW}}^2(Q, S) = \mathcal{V}(Q), \quad Q \in \mathbb{H}_+(d),$$

and therefore, uniqueness of the minimizer Q_* .

To prove that $Q_* \succ 0$ consider arbitrary degenerated $Q_0 \in \mathbb{H}_+(d) \cap \mathbb{A}$, $Q_1 \in \mathbb{H}_{++}(d) \cap \mathbb{A}$ (which exists by Assumption 1) and $S \in \mathbb{H}_{++}(d)$. Let us define $Q_t = (1-t)Q_0 + tQ_1 \in \mathbb{A}$. We are going to show, that

$$\frac{d}{dt} d_{\text{BW}}^2(Q_t, S) = \langle I - T_{Q_t}^S, Q_1 - Q_0 \rangle \rightarrow -\infty \quad \text{as } t \rightarrow 0.$$

To prove this convergence, we consider the following eigen-decomposition $S^{1/2}Q_0S^{1/2} = U^* \Lambda U$, $\Lambda = \text{diag}(\lambda_1, \dots, \lambda_r, 0, \dots, 0)$, where $r = \text{rank}(Q_0)$. We denote as $C = U S^{1/2} Q_1 \times S^{1/2} U^*$, and write it in a block form:

$$C = \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix}, \quad C_{11} \in \mathbb{H}_{++}(r), C_{12} = C_{21}^* \in \mathbb{C}^{r \times (d-r)}, C_{22} \in \mathbb{H}_{++}(d-r).$$

Thus, for all Q_t the following representation holds (see Section A.5.5, paragraph *Inverse of block matrix* in [Boyd and Vandenberghe \(2004\)](#)):

$$\begin{aligned} U(S^{1/2}Q_tS^{1/2})^{-1}U^* &= ((1-t)\Lambda + tC)^{-1} \\ &= \begin{pmatrix} E_t^{-1} + t^2E_t^{-1}C_{12}S_t^{-1}C_{21}E_t^{-1} & -tE_t^{-1}C_{12}S_t^{-1} \\ -tS_t^{-1}C_{21}E_t^{-1} & S_t^{-1} \end{pmatrix}, \end{aligned}$$

where $E_t = (1-t)\Lambda_{11} + tC_{11}$, $S_t = tC_{22} - t^2C_{21}E_t^{-1}C_{12}$, with $\Lambda_{11} = \text{diag}(\lambda_1, \dots, \lambda_r)$. When $t \rightarrow 0$, $E_t \rightarrow \Lambda_{11} > 0$, $\frac{S_t}{t} \rightarrow C_{22} > 0$. This yields

$$tU(S^{1/2}Q_tS^{1/2})^{-1}U^* \rightarrow \begin{pmatrix} 0 & 0 \\ 0 & C_{22}^{-1} \end{pmatrix},$$

and

$$\sqrt{t}U(S^{1/2}Q_tS^{1/2})^{-1/2}U^* \rightarrow \begin{pmatrix} 0 & 0 \\ 0 & C_{22}^{-1/2} \end{pmatrix}.$$

Therefore,

$$\begin{aligned} \sqrt{t}\langle T_{Q_t}^S, Q_0 \rangle &= \sqrt{t}\langle (S^{1/2}Q_tS^{1/2})^{-1/2}, S^{1/2}Q_0S^{1/2} \rangle \\ &= \langle \sqrt{t}U(S^{1/2}Q_tS^{1/2})^{-1/2}U^*, US^{1/2}Q_0S^{1/2}U^* \rangle \\ &\rightarrow \left\langle \begin{pmatrix} 0 & 0 \\ 0 & C_{22}^{-1/2} \end{pmatrix}, \begin{pmatrix} \Lambda_{11} & 0 \\ 0 & 0 \end{pmatrix} \right\rangle = 0. \end{aligned}$$

In the same way one can obtain

$$\sqrt{t}\langle T_{Q_t}^S, Q_1 \rangle \rightarrow \left\langle \begin{pmatrix} 0 & 0 \\ 0 & C_{22}^{-1/2} \end{pmatrix}, \begin{pmatrix} C_{11} & C_{12} \\ C_{21} & C_{22} \end{pmatrix} \right\rangle = \text{tr} C_{22}^{1/2} > 0 \quad \text{as } t \rightarrow 0.$$

Consequently,

$$\frac{d}{dt}d_{\text{BW}}^2(Q_t, S) = \langle I - T_{Q_t}^S, Q_1 - Q_0 \rangle = \text{tr} Q_1 - \text{tr} Q_0 - \frac{\text{tr} C_{22}^{1/2} + o(1)}{\sqrt{t}} \rightarrow -\infty.$$

By Assumption 1 it holds $\mathbb{P}(\mathbb{H}_{++}(d)) > 0$. Further, since $d_{\text{BW}}^2(Q, S)$ is convex, its directional derivatives are bounded by difference quotients, thus one can apply the Leibniz integral rule for a Lebesgue-integrable function. This yields the following equality:

$$\frac{d}{dt}\mathcal{V}(Q_t) = \mathbb{E} \frac{d}{dt}d_{\text{BW}}^2(Q_t, S) \rightarrow -\infty \quad \text{as } t \rightarrow 0,$$

thus Q_0 cannot be a barycenter of \mathbb{P} . This yields $Q_* > 0$.

Since $\mathcal{V}(\cdot)$ is convex and barycenter of \mathbb{P} is positive-definite and unique, it is characterized as a stationary point of Fréchet variation on subspace \mathbb{A} , that is, as a solution to equation

$$\mathbf{\Pi}_{\mathbb{M}} \nabla \mathcal{V}(Q) = \mathbf{\Pi}_{\mathbb{M}}(I - \mathbb{E} T_Q^S) = 0, \quad Q \in \mathbb{A} \cap \mathbb{H}_{++}(d),$$

as required. The first equality follows from Lemma A.6. \square

The proof of CLT relies on covariance operators on the space of optimal transportation maps and on the space of covariance matrices.

Covariance operator on the space of optimal maps. Consider $T_i \stackrel{\text{def}}{=} T_{Q_*}^{S_i}$ with $\mathbb{E} T_i = I$, and $T_i^n \stackrel{\text{def}}{=} T_{Q_n}^{S_i}$. We define a covariance Σ of T_i , its empirical counterpart Σ_n , and its data-driven estimator $\hat{\Sigma}_n$ as follows:

$$(A.6) \quad \begin{aligned} \Sigma &\stackrel{\text{def}}{=} \mathbb{E}(T_i - I) \otimes (T_i - I), & \Sigma_n &\stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^n (T_i - I) \otimes (T_i - I), \\ \hat{\Sigma}_n &\stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^n (T_i^n - I) \otimes (T_i^n - I). \end{aligned}$$

Covariance operators on the space of covariance matrices. Let Q_n be an empirical barycenter. The covariance of Q_n and its empirical counterpart are defined as

$$(A.7) \quad \Xi \stackrel{\text{def}}{=} F^{-1}(\Sigma)_{\mathbb{M}} F^{-1}, \quad \Xi : \mathbb{M} \rightarrow \mathbb{M},$$

$$(A.8) \quad \hat{\Xi}_n \stackrel{\text{def}}{=} \hat{F}_n^{-1}(\hat{\Sigma}_n)_{\mathbb{M}} \hat{F}_n^{-1}, \quad \hat{\Xi}_n : \mathbb{M} \rightarrow \mathbb{M},$$

where

$$(A.9) \quad F \stackrel{\text{def}}{=} -\mathbb{E}(dT_{Q_*}^S)_{\mathbb{M}}, \quad F_n \stackrel{\text{def}}{=} -\frac{1}{n} \sum_{i=1}^n (dT_{Q_*}^{S_i})_{\mathbb{M}},$$

$$(A.10) \quad \hat{F}_n \stackrel{\text{def}}{=} -\frac{1}{n} \sum_{i=1}^n (dT_{Q_n}^{S_i})_{\mathbb{M}}.$$

Now we are almost ready to prove the central limit theorem for the empirical barycenter Q_n (Theorem 2.2). Another key object which appears in the proofs quite often is a rescaled empirical barycenter:

$$(A.11) \quad Q'_n \stackrel{\text{def}}{=} Q_*^{-1/2} Q_n Q_*^{-1/2}.$$

For the sake of transparency we provide below a complete statement.

THEOREM (Central limit theorem for the covariance of empirical barycenter). *The approximation error rate of the Fréchet mean Q_* by its empirical counterpart Q_n is*

$$(A) \quad \sqrt{n}(Q_n - Q_*) \xrightarrow{w} \mathcal{N}(0, \Xi).$$

Moreover, if $(\Sigma)_{\mathbb{M}}$ is nondegenerated, then

$$(B) \quad \sqrt{n} \hat{\Xi}_n^{-1/2} (Q_n - Q_*) \xrightarrow{w} \mathcal{N}(0, (I)_{\mathbb{M}}).$$

PROOF OF THEOREM 2.2. The proof consists of two parts: proofs of (A) and (B).

Proof of (A). As $\mathcal{V}_n(\cdot)$ are convex functions, they a.s. uniformly converge to a strictly convex function $\mathcal{V}(\cdot)$ on any compact set by the uniform law of large numbers. Therefore, their minimizers also converge $Q_n \xrightarrow{\text{a.s.}} Q_*$; see, for example, van de Geer ((2006), Lemma 5.2.2). In particular, $\mathbb{P}(Q_n > 0) \rightarrow 1$, with $n \rightarrow \infty$. The expansion from Lemma A.2 at Q_* implies

$$(A.12) \quad T_i^n = T_i + \int_0^1 dT_{Q_t}^{S_i} (Q_n - Q_*) dt,$$

where $Q_t = (1 - t)Q_* + tQ_n$. Note that the condition for Q_n being a barycenter is $\Pi_{\mathbb{M}}(\frac{1}{n} \sum_i T_i^n - I) = 0$. This fact together with averaging of (A.12) over i give

$$(A.13) \quad \Pi_{\mathbb{M}} I = \Pi_{\mathbb{M}} \bar{T}_n - \alpha_n(Q_n - Q_*),$$

where

$$(A.14) \quad \bar{T}_n \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^n T_i, \quad \alpha_n \stackrel{\text{def}}{=} -\frac{1}{n} \sum_i \int_0^1 (dT_{Q'_t}^{S_i})_{\mathbb{M}} dt.$$

According to Lemma A.4

$$\frac{2}{\lambda_{\max}(Q'_n) + \lambda_{\max}^{1/2}(Q'_n)} \mathbf{F}_n \preceq \alpha_n \preceq \frac{2}{\lambda_{\min}(Q'_n) + \lambda_{\min}^{1/2}(Q'_n)} \mathbf{F}_n,$$

where \mathbf{F}_n is defined in (A.9), and Q'_n comes from (A.11). Recall that \mathbf{F} introduced in (A.9) is a population counterpart of \mathbf{F}_n . This operator is correctly defined since by Lemma A.3 one can show that it is self-adjoint, positive definite and bounded:

$$\|\mathbf{F}\| \leq \mathbb{E} \|dT_{Q_*}^S\| \leq \mathbb{E} \frac{\|S^{1/2} Q_* S^{1/2}\|}{2\lambda_{\min}^2(Q_*)} < \infty.$$

This bound follows directly from Corollary A.1.

Since by the law of large numbers $\mathbf{F}_n \xrightarrow{\text{a.s.}} \mathbf{F}$ and $Q'_n \xrightarrow{\text{a.s.}} I$, it holds that $\lambda_{\min}(Q'_n) \xrightarrow{\text{a.s.}} 1$ and $\lambda_{\max}(Q'_n) \xrightarrow{\text{a.s.}} 1$, thus $\alpha_n \xrightarrow{\text{a.s.}} \mathbf{F}$. Therefore we obtain from (A.13)

$$(A.15) \quad Q_n = Q_* + \alpha_n^{-1} \Pi_{\mathbb{M}}(\bar{T}_n - I)$$

$$(A.16) \quad = Q_* + \mathbf{F}^{-1} \Pi_{\mathbb{M}}(\bar{T}_n - I) + o_P(\|\Pi_{\mathbb{M}}(\bar{T}_n - I)\|),$$

where \mathbf{F}^{-1} is a bounded linear operator, because $dT_{Q_*}^S$ is negative definite for any $S > 0$ by Lemma A.3. The result (A) follows immediately from the CLT for $\Pi_{\mathbb{M}}\bar{T}_n$.

Proof of (B). Note that result (A) is equivalent to the fact, that

$$\sqrt{n} \Xi^{-1/2} (Q_n - Q_*) \xrightarrow{w} \mathcal{N}(0, (I)_{\mathbb{M}}).$$

To ensure convergence of $\hat{\Xi}_n \xrightarrow{\text{a.s.}} \Xi$ we need to show that:

- (a) $\hat{\Sigma}_n \xrightarrow{\text{a.s.}} \Sigma$ (follows from Lemma B.2, a.s. consistency of Q'_n , and the LLN);
- (b) $\hat{F}_n \xrightarrow{\text{a.s.}} \mathbf{F}$.

Consider

$$dT_{Q_n}^S \preceq dT_{\lambda_{\max}(Q'_n)Q_*}^S = (\lambda_{\max}(Q'_n))^{-3/2} dT_{Q_*}^S,$$

$$dT_{Q_n}^S \succeq dT_{\lambda_{\min}(Q'_n)Q_*}^S = (\lambda_{\min}(Q'_n))^{-3/2} dT_{Q_*}^S,$$

where the inequalities come from monotonicity of dT_Q^S (see (V) in Lemma A.3), and the fact that and bounds $\lambda_{\min}(Q'_n)Q_* \preceq Q_n \preceq \lambda_{\max}(Q'_n)Q_*$. The equalities hold due to homogeneity of dT_Q^S with degree $-\frac{3}{2}$ (see (IV) in Lemma A.3). This naturally leads to the following bounds:

$$\frac{1}{\lambda_{\max}^{3/2}(Q'_n)} \mathbf{F}_n \preceq \hat{F}_n \preceq \frac{1}{\lambda_{\min}^{3/2}(Q'_n)} \mathbf{F}_n.$$

Since $Q'_n \xrightarrow{\text{a.s.}} I$ and $\mathbf{F}_n \xrightarrow{\text{a.s.}} \mathbf{F}$, this implies $\hat{F}_n \xrightarrow{\text{a.s.}} \mathbf{F}$ due to the following continuity property: $\lambda_{\max}(Q'_n) \leq 1 + \|Q'_n - I\|$ and $\lambda_{\min}(Q'_n) \geq 1 - \|Q'_n - I\|$.

The above results ensure the validity of substitution Ξ by $\hat{\Xi}_n$. This yields (B). \square

The asymptotic convergence results for $d_{\text{BW}}(Q_n, Q_*)$ is a straightforward corollary of the above theorem. Here is the proof.

PROOF OF COROLLARY 2.1. Since $Q_n \xrightarrow{\text{a.s.}} Q_*$, Lemma A.6 implies

$$d_{\text{BW}}^2(Q_n, Q_*) = -\frac{1 + o_P(1)}{2} \langle d\mathbf{T}_{Q_*}^{Q_*}(Q_n - Q_*), Q_n - Q_* \rangle.$$

Without loss of generality we can consider case $Q_* = \text{diag}(q_1, \dots, q_d)$, thus Lemma A.3 implies (notice that $\Lambda = Q_*^2$ and $\Delta = Q_*^{1/2} X Q_*^{1/2}$)

$$\begin{aligned} -\langle d\mathbf{T}_{Q_*}^{Q_*}(X), X \rangle &= \sum_{i,j=1}^d \frac{X_{ij}}{q_i + q_j} X_{ij} = \sum_{i,j=1}^d (q_i + q_j) \left(\frac{X_{ij}}{q_i + q_j} \right)^2 \\ &= 2 \sum_{i,j=1}^d \left(\sqrt{q_i} \frac{X_{ij}}{q_i + q_j} \right)^2 = 2 \| Q_*^{1/2} d\mathbf{T}_{Q_*}^{Q_*}(X) \|_F^2. \end{aligned}$$

By Theorem 2.2 $\sqrt{n}(Q_n - Q_*)$ is asymptotically normal and centred, therefore

$$\mathcal{L}(\sqrt{n}d_{\text{BW}}(Q_n, Q_*)) \xrightarrow{w} \mathcal{L}(\| Q_*^{1/2} d\mathbf{T}_{Q_*}^{Q_*}(Z) \|_F),$$

where $Z \in \mathbb{M} \subset \mathbb{H}(d)$ and $Z \sim \mathcal{N}(0, \Xi)$.

Note, that $Q_n \xrightarrow{\text{a.s.}} Q_*$, $\hat{\Xi}_n \xrightarrow{\text{a.s.}} \Xi$, and $d\mathbf{T}_{Q_n}^{Q_n} \xrightarrow{\text{a.s.}} d\mathbf{T}_{Q_*}^{Q_*}$. The last result follows from Lemma A.3(IV, V), and can be validated using the same framework as in the proof of (B) in Theorem 2.2. Note, that $\lambda_{\min}(Q'_n)Q_* \preceq Q_n \preceq \lambda_{\max}(Q'_n)Q_*$, with Q'_n coming from (A.11). Then

$$\begin{aligned} d\mathbf{T}_{Q_n}^{Q_n} &\preceq d\mathbf{T}_{\lambda_{\max}(Q'_n)Q_*}^{\lambda_{\max}(Q'_n)Q_*} = \frac{1}{\lambda_{\max}(Q'_n)} d\mathbf{T}_{Q_*}^{Q_*} \rightarrow d\mathbf{T}_{Q_*}^{Q_*}, \\ d\mathbf{T}_{Q_n}^{Q_n} &\succeq d\mathbf{T}_{\lambda_{\min}(Q'_n)Q_*}^{\lambda_{\min}(Q'_n)Q_*} = \frac{1}{\lambda_{\min}(Q'_n)} d\mathbf{T}_{Q_*}^{Q_*} \rightarrow d\mathbf{T}_{Q_*}^{Q_*}, \end{aligned}$$

where the inequalities comes from monotonicity (see (V) in Lemma A.3). The equalities hold due to homogeneity (see (IV) in Lemma A.3). Furthermore, $\lambda_{\max}(Q'_n) \leq 1 + \|Q'_n - I\|$ and $\lambda_{\min}(Q'_n) \geq 1 - \|Q'_n - I\|$. This yields

$$\mathcal{L}(\| Q_n^{1/2} d\mathbf{T}_{Q_n}^{Q_n}(Z_n) \|_F) \xrightarrow{w} \mathcal{L}(\| Q_*^{1/2} d\mathbf{T}_{Q_*}^{Q_*}(Z) \|_F),$$

where $Z_n \sim \mathcal{N}(0, \hat{\Xi}_n)$. This, in turn, entails

$$d_w(\mathcal{L}(\sqrt{n}d_{\text{BW}}(Q_n, Q_*)), \mathcal{L}(\| Q_n^{1/2} d\mathbf{T}_{Q_n}^{Q_n}(Z_n) \|_F)) \rightarrow 0,$$

where d_w is some metric inducing the weak convergence of the measures. \square

Finally, we are ready to prove Theorem 2.3.

PROOF OF THEOREM 2.3. By definition the empirical Fréchet variance is

$$\mathcal{V}_n(Q) = \frac{1}{n} \sum_{i=1}^n d_{\text{BW}}^2(Q, S_i).$$

Lemma A.6 ensures the following bound on $\mathcal{V}_n(Q_*) - \mathcal{V}_n(Q_n)$:

$$0 \leq \mathcal{V}_n(Q_*) - \mathcal{V}_n(Q_n) \leq \frac{2}{(1 + \lambda_{\min}^{1/2}(Q'_n))^2} \langle F_n(Q_n - Q_*), Q_n - Q_* \rangle$$

with $Q'_n \stackrel{\text{def}}{=} Q_*^{-1/2} Q_n Q_*^{-1/2}$. The above quadratic bound together with $Q_n \rightarrow Q_*$, $F_n \rightarrow F$ and $\sqrt{n}(Q_n - Q_*) \xrightarrow{w} \mathcal{N}(0, \Xi)$ yield

$$\mathcal{V}_n(Q_n) - \mathcal{V}(Q_*) = \mathcal{V}_n(Q_*) - \mathcal{V}(Q_*) + O_P\left(\frac{1}{n}\right).$$

On the other hand, by the classical central limit theorem we obtain

$$\begin{aligned} \sqrt{n}(\mathcal{V}_n(Q_*) - \mathcal{V}(Q_*)) &= \sqrt{n}\left(\frac{1}{n} \sum_i d_{\text{BW}}^2(Q_*, S_i) - \mathbb{E} d_{\text{BW}}^2(Q_*, S)\right) \\ &\xrightarrow{w} \mathcal{N}(0, \text{Var} d_{\text{BW}}^2(Q_*, S)). \end{aligned} \quad \square$$

APPENDIX B: CONCENTRATIONS OF Q_n AND \mathcal{V}_n

B.1. Concentration of Q_n . The next lemma is a key ingredient in the proof of the concentration result for Q_n .

LEMMA B.1. Consider

$$(B.1) \quad \eta_n \stackrel{\text{def}}{=} \frac{1}{\lambda_{\min}(F'_n)} \|Q_*^{1/2} \Pi_{\mathbb{M}}(\bar{T}_n - I) Q_*^{1/2}\|_F,$$

where

$$(B.2) \quad F'_n(X) \stackrel{\text{def}}{=} Q_*^{1/2} F_n(Q_*^{1/2} X Q_*^{1/2}) Q_*^{1/2}, \quad X \in \{Q_*^{-1/2} Y Q_*^{-1/2} \mid Y \in \mathbb{M}\}.$$

Then

$$\|Q_*^{-1/2} Q_n Q_*^{-1/2} - I\|_F \leq \frac{\eta_n}{1 - \frac{3}{4}\eta_n}$$

whenever $\eta_n < \frac{4}{3}$ and $Q_n > 0$.

PROOF. Let us define $Q_t \stackrel{\text{def}}{=} t Q_n + (1 - t) Q_*$ for $t \in [0, 1]$, and Q'_n defined in (A.11). Due to Lemmas A.3 and A.4 we have for any $S \in \mathbb{H}_+(d)$

$$\begin{aligned} \langle \Pi_{\mathbb{M}}(T_{Q_*}^S - T_{Q_n}^S), Q_n - Q_* \rangle &= \langle T_{Q_*}^S - T_{Q_n}^S, Q_n - Q_* \rangle \\ &= \int_0^1 \langle -dT_{Q_t}^S(Q_n - Q_*), Q_n - Q_* \rangle dt \\ &\geq \frac{1}{1 + \frac{3}{4}\|Q'_n - I\|} \langle -dT_{Q_*}^S(Q_n - Q_*), Q_n - Q_* \rangle. \end{aligned}$$

Therefore,

$$\begin{aligned} \langle \Pi_{\mathbb{M}}(\bar{T}_n - I), Q_n - Q_* \rangle &\geq \frac{1}{1 + \frac{3}{4}\|Q'_n - I\|} \langle F_n(Q_n - Q_*), Q_n - Q_* \rangle \\ &= \frac{1}{1 + \frac{3}{4}\|Q'_n - I\|} \langle F'_n(Q'_n - I), Q'_n - I \rangle \\ &\geq \frac{\lambda_{\min}(F'_n)}{1 + \frac{3}{4}\|Q'_n - I\|} \|Q'_n - I\|_F^2. \end{aligned}$$

At the same time,

$$\begin{aligned} \langle \mathbf{\Pi}_{\mathbb{M}}(\bar{T}_n - I), Q_n - Q_* \rangle &= \langle Q_*^{1/2} \mathbf{\Pi}_{\mathbb{M}}(\bar{T}_n - I) Q_*^{1/2}, Q'_n - I \rangle \\ &\leq \|Q_*^{1/2} \mathbf{\Pi}_{\mathbb{M}}(\bar{T}_n - I) Q_*^{1/2}\|_F \|Q'_n - I\|_F. \end{aligned}$$

Hence

$$\|Q'_n - I\|_F \leq \frac{1 + \frac{3}{4} \|Q'_n - I\|}{\lambda_{\min}(\mathbf{F}'_n)} \|Q_*^{1/2} \mathbf{\Pi}_{\mathbb{M}}(\bar{T}_n - I) Q_*^{1/2}\|_F = \left(1 + \frac{3}{4} \|Q'_n - I\|\right) \eta_n.$$

Rewriting the inequality above we obtain

$$\|Q'_n - I\|_F \leq \frac{\eta_n}{1 - \frac{3}{4} \eta_n}$$

provided that $\eta_n < \frac{4}{3}$. \square

Before proving concentration results, we define operator $\mathbf{F}'(X)$ as follows:

$$(B.3) \quad \mathbf{F}'(X) \stackrel{\text{def}}{=} Q_*^{1/2} \mathbf{F}(Q_*^{1/2} X Q_*^{1/2}) Q_*^{1/2} \quad \text{for } X \in \{Q_*^{-1/2} Y Q_*^{-1/2} \mid Y \in \mathbb{M}\}.$$

PROOF OF THEOREM 2.4. Let t_n be s.t. the following upper bound on $\gamma_n(t_n)$ from Proposition B.1 holds:

$$(B.4) \quad \gamma_n(t_n) \leq \frac{1}{2} \lambda_{\min}(\mathbf{F}').$$

It is easy to see that this condition is fulfilled for $t_n = nt_F - \log(m)$ under a proper choice of generic constant in definition of t_F . Then with probability at least $1 - 2me^{-nt_F}$ the following bound holds:

$$\lambda_{\min}(\mathbf{F}'_n) \geq \lambda_{\min}(\mathbf{F}') - \|\mathbf{F}'_n - \mathbf{F}'\| \geq \frac{1}{2} \lambda_{\min}(\mathbf{F}'),$$

with \mathbf{F}'_n to be defined in (B.2). The above facts together with definition of η_n (B.1) yield

$$\begin{aligned} \eta_n &\stackrel{\text{def}}{=} \frac{\|Q_*^{1/2} \mathbf{\Pi}_{\mathbb{M}}(\bar{T}_n - I) Q_*^{1/2}\|_F}{\lambda_{\min}(\mathbf{F}'_n)} \\ &\leq \frac{2\|Q_*\|}{\lambda_{\min}(\mathbf{F}')} \|\mathbf{\Pi}_{\mathbb{M}}(\bar{T}_n - I)\|_F = \frac{c_Q}{2\sigma_T} \|\mathbf{\Pi}_{\mathbb{M}}(\bar{T}_n - I)\|_F. \end{aligned}$$

Combining the above bounds with Proposition B.2, we obtain

$$\mathbb{P}\left\{\eta_n \geq \frac{c_Q}{2\sqrt{n}}(\sqrt{m} + t)\right\} \leq 2me^{-nt_F} + e^{-t^2/2}.$$

Now it follows from Lemma B.1 that

$$\begin{aligned} \mathbb{P}\left\{\|Q'_n - I\|_F \geq \frac{c_Q}{\sqrt{n}}(\sqrt{m} + t)\right\} &\leq \mathbb{P}\left\{2\eta_n \geq \frac{c_Q}{\sqrt{n}}(\sqrt{m} + t)\right\} + \mathbb{P}\{Q_n \not\asymp 0\} \\ &\leq 2me^{-nt_F} + e^{-t^2/2} + (1 - p)^n, \end{aligned}$$

whenever $\frac{c_Q}{2\sqrt{n}}(\sqrt{m} + t) \leq \frac{2}{3}$. Here we used that $Q_n \succ 0$ if at least one of matrices S_1, \dots, S_n is nondegenerated. Here $Q \not\asymp 0$ means that a matrix Q is not positive definite. \square

PROOF OF COROLLARY 2.2. To prove this result we use Lemma A.6 and choose $Q_0 = S = Q_*$, $Q_1 = Q_n$. Thus we obtain

$$\begin{aligned} d_{\text{BW}}^2(Q_n, Q_*) &\leq -\frac{2}{(1 + \lambda_{\min}^{1/2}(Q'_n))^2} \langle d\mathbf{T}_{Q_*}^{Q_*}(Q_n - Q_*), Q_n - Q_* \rangle \\ &\stackrel{\text{Def. (A.4)}}{=} \frac{2}{(1 + \lambda_{\min}^{1/2}(Q'_n))^2} \langle -d\mathbf{t}_{Q_*}^{Q_*}(Q'_n - I), Q'_n - I \rangle \\ &\leq 2\lambda_{\max}(-d\mathbf{t}_{Q_*}^{Q_*}) \|Q'_n - I\|_F^2 \stackrel{\text{C. A.2}}{=} \lambda_{\max}(Q_*) \|Q'_n - I\|_F^2, \end{aligned}$$

with Q'_n coming from (A.11). Hence by Theorem 2.4

$$d_{\text{BW}}(Q_n, Q_*) \leq \|Q_*\|^{1/2} \frac{c_Q}{\sqrt{n}} (\sqrt{m} + t)$$

with probability at least $1 - 2me^{-ntF} - e^{-t^2/2} - (1 - p)^n$. \square

B.2. Concentration of \mathcal{V}_n .

PROOF OF THEOREM 2.5. Following the proof of Theorem 2.3 we consider $\mathcal{V}_n(Q_*) - \mathcal{V}_n(Q_n)$:

$$\begin{aligned} (B.5) \quad 0 \leq \mathcal{V}_n(Q_*) - \mathcal{V}_n(Q_n) &\leq \frac{2}{(1 + \lambda_{\min}^{1/2}(Q'_n))^2} \langle \mathbf{F}_n(Q_n - Q_*), Q_n - Q_* \rangle \\ &= \frac{2}{(1 + \lambda_{\min}^{1/2}(Q'_n))^2} \langle \mathbf{F}'_n(Q'_n - I), Q'_n - I \rangle \\ &\leq 2\|\mathbf{F}'_n\| \cdot \|Q'_n - I\|_F^2, \end{aligned}$$

with \mathbf{F}'_n to be defined in (B.2), and Q'_n in (A.11). Following the proof of Theorem 2.4, we obtain that with $\mathbb{P} \geq 1 - 2me^{-tFn} - e^{-t^2/2} - (1 - p)^n$ the following upper bounds hold:

$$\|Q'_n - I\|_F \leq \frac{c_Q}{\sqrt{n}} (\sqrt{m} + t), \quad \|\mathbf{F}'_n - \mathbf{F}'\| \leq \frac{1}{2} \lambda_{\min}(\mathbf{F}'),$$

with \mathbf{F}' coming from (B.3). Thus

$$\|\mathbf{F}'_n\| \leq \|\mathbf{F}'\| + \|\mathbf{F}'_n - \mathbf{F}'\| \leq \frac{3}{2} \|\mathbf{F}'\|$$

and consequently

$$0 \leq \mathcal{V}_n(Q_*) - \mathcal{V}_n(Q_n) \leq 3\|\mathbf{F}'\| \frac{c_Q^2}{n} (\sqrt{m} + t)^2.$$

Now we consider a difference $\mathcal{V}_n(Q_*) - \mathcal{V}(Q_*)$. According to Assumption 3 S, and therefore $d_{\text{BW}}^2(Q_*, S)$, are subexponential r.v. with some parameters (ν, μ) . Then Lemma B.4 ensures

$$|\mathcal{V}_n(Q_*) - \mathcal{V}(Q_*)| \leq \max\left(\frac{2\mu t'}{n}, \nu\left(\frac{2t'}{n}\right)^{1/2}\right)$$

with probability $1 - 2e^{-t'}$. Combining two above bounds we obtain

$$|\mathcal{V}_n(Q_n) - \mathcal{V}(Q_*)| \leq \max\left(\frac{2\mu t'}{n}, \nu\sqrt{\frac{2t'}{n}}\right) + 3\|\mathbf{F}'\| \frac{c_Q^2}{n} (\sqrt{m} + t)^2$$

with probability

$$\mathbb{P} \geq 1 - 2e^{-t'} - 2me^{-nt_F} - e^{-t^2/2} - (1 - p)^n.$$

Choosing $t' = t^2/2$, we get

$$\begin{aligned} \mathbb{P} \left\{ |\mathcal{V}_n(Q_n) - \mathcal{V}(Q_*)| \geq \max\left(\frac{\mu t^2}{n}, \frac{vt}{\sqrt{n}}\right) + 3\|\mathbf{F}'\| \frac{c_Q^2}{n} (\sqrt{m} + t)^2 \right\} \\ \leq 2me^{-nt_F} + 3e^{-t^2/2} + (1 - p)^n. \end{aligned} \quad \square$$

B.3. Auxiliary results.

LEMMA B.2. *Let $\|Q'_n - I\| \leq \frac{1}{2}$, with Q'_n coming from (A.11); then*

$$\|\hat{\Sigma}_n - \Sigma_n\|_1 \leq \beta_n \left[2 \left(\frac{1}{n} \sum_i \|T_i - I\|_F^2 \right)^{1/2} + \beta_n \right],$$

where

$$\beta_n \stackrel{\text{def}}{=} \kappa(Q_*) \left(\frac{\frac{1}{n} \sum_i \|S_i\|}{\|Q_*\|} \right)^{1/2} \|Q'_n - I\|_F,$$

where $\kappa(Q_*) = \|Q_*\| \cdot \|Q_*^{-1}\|$ is the condition number of matrix Q_* and $\|A\|_1$ is 1-Schatten (nuclear) norm of an operator A .

PROOF. Note, that for any $(T_i^n - I) \otimes (T_i^n - I)$ the following decomposition holds

$$\begin{aligned} (T_i^n - I) \otimes (T_i^n - I) \\ = (T_i - I) \otimes (T_i - I) + (T_i^n - T_i) \otimes (T_i - I) \\ + (T_i - I) \otimes (T_i^n - T_i) + (T_i^n - T_i) \otimes (T_i^n - T_i). \end{aligned}$$

Summing over i yields

$$\begin{aligned} \hat{\Sigma}_n - \Sigma_n &= \frac{1}{n} \sum_i (T_i^n - T_i) \otimes (T_i - I) \\ \text{(B.6)} \quad &+ \frac{1}{n} \sum_i (T_i - I) \otimes (T_i^n - T_i) + \frac{1}{n} \sum_i (T_i^n - T_i) \otimes (T_i^n - T_i). \end{aligned}$$

Note, that each

$$\|(T_i^n - T_i) \otimes (T_i - I)\|_1 \leq \|T_i^n - T_i\|_F \|T_i - I\|_F.$$

Lemmas A.3(III) and A.4 yield

$$\begin{aligned} \|T_i^n - T_i\|_F &\leq \frac{1}{1 - \|Q'_n - I\|} \|\mathbf{d}T_{Q_*}^{S_i}(Q_n - Q_*)\|_F \\ &\leq 2\|Q_*^{-1/2} \mathbf{d}T_{Q_*}^{S_i}(Q'_n - I) Q_*^{-1/2}\|_F \leq 2 \frac{\lambda_{\max}(\mathbf{d}T_{Q_*}^{S_i})}{\lambda_{\min}(Q_*)} \|Q'_n - I\|_F \\ &\leq \frac{\lambda_{\max}^{1/2}(S_i^{1/2} Q_* S_i^{1/2})}{\lambda_{\min}(Q_*)} \|Q'_n - I\|_F \leq \kappa(Q_*) \left(\frac{\|S_i\|}{\|Q_*\|} \right)^{1/2} \|Q'_n - I\|_F, \end{aligned}$$

where \mathbf{dt}_Q^S is defined in (A.4). Hence $\frac{1}{n} \sum_i \|T_i^n - T_i\|_F^2 \leq \beta_n^2$. The above expression together with (B.6) and the Cauchy–Schwarz inequality lead to the upper bound on $\|\hat{\Sigma}_n - \Sigma_n\|_1$:

$$\begin{aligned} \|\hat{\Sigma}_n - \Sigma_n\|_1 &\leq \frac{2}{n} \sum_i \|T_i - I\|_F \|T_i^n - T_i\|_F + \frac{1}{n} \sum_i \|T_i^n - T_i\|_F^2 \\ &\leq 2\beta_n \left(\frac{1}{n} \sum_i \|T_i - I\|_F^2 \right)^{1/2} + \beta_n^2. \end{aligned}$$

□

Further we present concentration of F_n around F . Denote as $\|X\|_{\psi_2}$ an Orlicz norm with Young function $\psi_2(x) = e^{x^2} - 1$, that is,

$$\|X\|_{\psi_2} \stackrel{\text{def}}{=} \inf\{c > 0 : \mathbb{E} \psi_2(|X|/c) \leq 1\}.$$

Then sub-Gaussianity of a r.v. X is equivalent to $\|X\|_{\psi_2} < \infty$ and it ensures

$$\text{Var}(X) \leq \sqrt{2} \|X\|_{\psi_2}.$$

PROPOSITION B.1 (Concentration of F'_n , Proposition 2 in Koltchinskii (2011)). *Let F'_n , F' , and \mathbf{dt}_Q^S be defined as (B.2), (B.3) and (A.4), respectively. There exists a constant $C > 0$, s.t. for all $t > 0$ it holds with probability at least $1 - e^{-t}$*

$$\|F'_n - F'\| \leq \gamma_n(t), \quad \gamma_n(t) \stackrel{\text{def}}{=} C \max\left(\sigma_F \sqrt{\frac{t + \log(2m)}{n}}, U \sqrt{\log\left(\frac{U}{\sigma_F}\right) \frac{t + \log(2m)}{n}} \right),$$

where $\sigma_F^2 \stackrel{\text{def}}{=} \|\mathbb{E}(\mathbf{dt}_{Q_*}^S - F')^2\|$, $U \stackrel{\text{def}}{=} \|\mathbf{dt}_{Q_*}^S - F'\|_{\psi_2}$.

LEMMA B.3. *The size of the above constants can be estimated as follows:*

$$\sigma_F \leq \frac{\|Q_*\|^{1/2}}{2} (\mathbb{E}\|S\|)^{1/2}, \quad U \leq \frac{3}{2} \|Q_*\|^{1/2} \|\|S\|\|_{\psi_1}^{1/2},$$

where $\psi_1(x) = e^x - 1$ is a Young function.

PROOF. By Corollary A.2 we obtain

$$\sigma_F^2 \stackrel{\text{def}}{=} \|\mathbb{E}(\mathbf{dt}_{Q_*}^S - F')^2\| \leq \mathbb{E}\|\mathbf{dt}_{Q_*}^S\|^2 \leq \frac{\|Q_*\|}{4} \mathbb{E}\|S\|$$

and (due to properties of Orlicz norm)

$$\begin{aligned} U \stackrel{\text{def}}{=} \|\mathbf{dt}_{Q_*}^S - F'\|_{\psi_2} &\leq \frac{\|F'\|}{\sqrt{\ln 2}} + \|\mathbf{dt}_{Q_*}^S\|_{\psi_2} \\ &\leq \frac{\|Q_*\|^{1/2}}{2} [2\mathbb{E}\|S\|^{1/2} + \|\|S\|\|_{\psi_2}^{1/2}] \\ &\leq \frac{\|Q_*\|^{1/2}}{2} [2(\mathbb{E}\|S\|)^{1/2} + \|\|S\|\|_{\psi_1}^{1/2}] \\ &\leq \frac{3}{2} \|Q_*\|^{1/2} \|\|S\|\|_{\psi_1}^{1/2}. \end{aligned}$$

□

The next proposition ensures the concentration of \bar{T}_n .

PROPOSITION B.2 (Concentration of \bar{T}_n ; Hsu, Kakade and Zhang (2012), Theorem 1). Under Assumption 3 it holds

$$\mathbb{P}\left\{\|\Pi_{\mathbb{M}}(\bar{T}_n - I)\|_F \geq \frac{\sigma_T}{\sqrt{n}}(\sqrt{m} + t)\right\} \leq e^{-t^2/2} \quad \text{for any } t \geq 0.$$

LEMMA B.4 (Subexponential tail bounds). Suppose that X is subexponential with parameters v, b . Then

$$\mathbb{P}\{X \geq \mathbb{E}X + t\} \leq \begin{cases} \exp\left(-\frac{t^2}{2v^2}\right) & \text{if } 0 \leq t \leq \frac{v^2}{b}, \\ \exp\left(-\frac{t}{2b}\right) & \text{if } t \geq \frac{v^2}{b}. \end{cases}$$

Acknowledgments. Alexey Kroshnin is also affiliated with National Research University Higher School of Economics, Russia, and Université Claude Bernard Lyon 1, France.

Vladimir Spokoiny is also affiliated with Institute for Information Transmission Problems RAS, Russia, and National Research University Higher School of Economics, Russia.

Alexandra Suvorikova is also affiliated with Institute for Information Transmission Problems RAS, Russia.

First of all, we would like to thank the anonymous referee who encouraged us consider in more detail a connection of Bures–Wasserstein barycenters to M-estimators. We are grateful to all referees who made a lot of useful comments and remarks which significantly improved the quality of the presentations.

The article was prepared within the framework of the HSE University Basic Research Program.

Funding. The work of Alexandra Suvorikova and Alexey Kroshnin in Sections 2 and 3 is funded by Russian Science Foundation grant No. 18-71-10108.

The work of Vladimir Spokoiny in Section 4 is supported by the Russian Science Foundation grant No. 19-71-30020. Vladimir Spokoiny thanks the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy—The Berlin Mathematics Research Center MATH+ (EXC-2046/1, project ID: 390685689).

REFERENCES

- AGUEH, M. and CARLIER, G. (2011). Barycenters in the Wasserstein space. *SIAM J. Math. Anal.* **43** 904–924. MR2801182 <https://doi.org/10.1137/100805741>
- AGUEH, M. and CARLIER, G. (2017). Vers un théorème de la limite centrale dans l'espace de Wasserstein? *C. R. Math. Acad. Sci. Paris* **355** 812–818. MR3673057 <https://doi.org/10.1016/j.crma.2017.05.010>
- AHIDAR-COUTRIX, A., LE GOUIC, T. and PARIS, Q. (2020). On the rate of convergence of empirical barycentres in metric spaces: Curvature, convexity and extendible geodesics. Preprint. Available at [arXiv:1806.02740](https://arxiv.org/abs/1806.02740).
- ÁLVAREZ-ESTEBAN, P. C., DEL BARRIO, E., CUESTA-ALBERTOS, J. A. and MATRÁN, C. (2015). Wide consensus for parallelized inference. ArXiv e-prints.
- ÁLVAREZ-ESTEBAN, P. C., DEL BARRIO, E., CUESTA-ALBERTOS, J. A. and MATRÁN, C. (2018). Wide consensus aggregation in the Wasserstein space. Application to location-scatter families. *Bernoulli* **24** 3147–3179. MR3779713 <https://doi.org/10.3150/17-BEJ957>
- AMBROSIO, L. and GIGLI, N. (2013). A user's guide to optimal transport. In *Modelling and Optimisation of Flows on Networks. Lecture Notes in Math.* **2062** 1–155. Springer, Heidelberg. MR3050280 https://doi.org/10.1007/978-3-642-32160-3_1
- BHATIA, R., JAIN, T. and LIM, Y. (2019). On the Bures–Wasserstein distance between positive definite matrices. *Expo. Math.* **37** 165–191. MR3992484 <https://doi.org/10.1016/j.exmath.2018.01.002>
- BIGOT, J., CAZELLES, E. and PAPADAKIS, N. (2016). Penalized barycenters in the Wasserstein space. Available at [arXiv:1606.01025](https://arxiv.org/abs/1606.01025).

- BIGOT, J. and KLEIN, T. (2012). Consistent estimation of a population barycenter in the Wasserstein space. ArXiv e-prints.
- BOYD, S. and VANDENBERGHE, L. (2004). *Convex Optimization*. Cambridge Univ. Press, Cambridge. MR2061575 <https://doi.org/10.1017/CBO9780511804441>
- BRENIER, Y. (1991). Polar factorization and monotone rearrangement of vector-valued functions. *Comm. Pure Appl. Math.* **44** 375–417. MR1100809 <https://doi.org/10.1002/cpa.3160440402>
- BULYGINA, O. N. and RAZUVAEV, V. N. (2012). Daily temperature and precipitation data for 518 Russian meteorological stations. Technical report, ESS-DIVE (Environmental System Science Data Infrastructure for a Virtual Ecosystem), Oak Ridge National Lab. (ORNL), Oak Ridge, TN (United States).
- CALSBECK, B. and GOODNIGHT, C. J. (2009). Empirical comparison of G matrix test statistics: Finding biologically relevant change. *Evolution* **63** 2627–2635.
- COURTY, N., FLAMARY, R., TUIA, D. and RAKOTOMAMONJY, A. (2017). Optimal transport for domain adaptation. *IEEE Trans. Pattern Anal. Mach. Intell.* **39** 1853–1865. <https://doi.org/10.1109/TPAMI.2016.2615921>
- CUTURI, M. (2013). Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems* 2292–2300.
- DAJKA, J., ŁUCZKA, J. and HÄNGGI, P. (2011). Distance between quantum states in the presence of initial qubit-environment correlations: A comparative study. *Phys. Rev. A* **84** 032120.
- DEL BARRIO, E. and LOUBES, J.-M. (2019). Central limit theorems for empirical transportation cost in general dimension. *Ann. Probab.* **47** 926–951. MR3916938 <https://doi.org/10.1214/18-AOP1275>
- DEL BARRIO, E., GORDALIZA, P., LESCORNEL, H. and LOUBES, J.-M. (2019a). Central limit theorem and bootstrap procedure for Wasserstein’s variations with an application to structural relationships between distributions. *J. Multivariate Anal.* **169** 341–362. MR3875604 <https://doi.org/10.1016/j.jmva.2018.09.014>
- DEL BARRIO, E., CUESTA-ALBERTOS, J. A., MATRÁN, C. and MAYO-ÍSCAR, A. (2019b). Robust clustering tools based on optimal transportation. *Stat. Comput.* **29** 139–160. MR3905545 <https://doi.org/10.1007/s11222-018-9800-z>
- DEL BARRIO, E., LESCORNEL, H. and LOUBES, J.-M. (2015). A statistical analysis of a deformation model with Wasserstein barycenters: Estimation procedure and goodness of fit test. Preprint. Available at arXiv:1508.06465.
- DVURECHENSKY, P., DVINSKIKH, D., GASNIKOV, A., URIBE, C. A. and NEDIĆ, A. (2018). Decentralize and randomize: Faster algorithm for Wasserstein barycenters. In *Advances in Neural Information Processing Systems, Vol. 31, NeurIPS 2018* (S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi and R. Garnett, eds.) 10783–10793. Curran Associates, Red Hook. Available at arXiv:1806.03915.
- EBERT, J., SPOKOINY, V. and SUVORIKOVA, A. (2017). Construction of non-asymptotic confidence sets in 2-Wasserstein space. Preprint. Available at arXiv:1703.03658.
- FANO, U. (1957). Description of states in quantum mechanics by density matrix and operator techniques. *Rev. Modern Phys.* **29** 74–93. MR0085885 <https://doi.org/10.1103/revmodphys.29.74>
- FLAMARY, R., CUTURI, M., COURTY, N. and RAKOTOMAMONJY, A. (2018). Wasserstein discriminant analysis. *Mach. Learn.* **107** 1923–1945. MR3854402 <https://doi.org/10.1007/s10994-018-5717-1>
- GONZALEZ, O., PASI, M., PETKEVIČIŪTĖ, D., GLOWACKI, J. and MADDOCKS, J. H. (2017). Absolute versus relative entropy parameter estimation in a coarse-grain model of DNA. *Multiscale Model. Simul.* **15** 1073–1107. MR3672368 <https://doi.org/10.1137/16M1086091>
- GOODNIGHT, C. J. and SCHWARTZ, J. M. (1997). A bootstrap comparison of genetic covariance matrices. *Biometrics* 1026–1039.
- GRAMFORT, A., PEYRÉ, G. and CUTURI, M. (2015). Fast optimal transport averaging of neuroimaging data. In *International Conference on Information Processing in Medical Imaging* 261–272. Springer, Berlin.
- HSU, D., KAKADE, S. M. and ZHANG, T. (2012). A tail inequality for quadratic forms of subgaussian random vectors. *Electron. Commun. Probab.* **17** no. 52, 6. MR2994877 <https://doi.org/10.1214/ECP.v17-2079>
- JOZSA, R. (1994). Fidelity for mixed quantum states. *J. Modern Opt.* **41** 2315–2323. MR1312945 <https://doi.org/10.1080/09500349414552171>
- KOLTCHINSKII, V. (2011). Von Neumann entropy penalization and low-rank matrix estimation. *Ann. Statist.* **39** 2936–2973. MR3012397 <https://doi.org/10.1214/11-AOS926>
- KROSHNIN, A. (2018). Fréchet barycenters in the Monge–Kantorovich spaces. *J. Convex Anal.* **25** 1371–1395. MR3818572
- KROSHNIN, A., TUPITSA, N., DVINSKIKH, D., DVURECHENSKY, P., GASNIKOV, A. and URIBE, C. (2019). On the complexity of approximating Wasserstein barycenters. In *Proceedings of the 36th International Conference on Machine Learning* (K. Chaudhuri and R. Salakhutdinov, eds.). *Proceedings of Machine Learning Research* **97** 3530–3540. Available at arXiv:1901.08686.
- LE GOUIC, T. and LOUBES, J.-M. (2017). Existence and consistency of Wasserstein barycenters. *Probab. Theory Related Fields* **168** 901–917. MR3663634 <https://doi.org/10.1007/s00440-016-0727-z>

- LE GOUIC, T., PARIS, Q., RIGOLLET, P. and STROMME, A. J. (2019). Fast convergence of empirical barycenters in Alexandrov spaces and the Wasserstein space. Preprint. Available at [arXiv:1908.00828](https://arxiv.org/abs/1908.00828).
- MALLASTO, A. and FERAGEN, A. (2017). Learning from uncertain curves: The 2-Wasserstein metric for Gaussian processes. In *Advances in Neural Information Processing Systems, Vol. 30* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan and R. Garnett, eds.) 5660–5670. Curran Associates, Red Hook.
- MARIAN, P. and MARIAN, T. A. (2008). Bures distance as a measure of entanglement for symmetric two-mode Gaussian states. *Phys. Rev. A* **77** 062319.
- MONTAVON, G., MÜLLER, K.-R. and CUTURI, M. (2016). Wasserstein training of restricted Boltzmann machines. In *Advances in Neural Information Processing Systems* 3718–3726.
- MUZELLEC, B. and CUTURI, M. (2018). Generalizing point embeddings using the Wasserstein space of elliptical distributions. Preprint. Available at [arXiv:1805.07594](https://arxiv.org/abs/1805.07594).
- OLKIN, I. and PUKELSHEIM, F. (1982). The distance between two random vectors with given dispersion matrices. *Linear Algebra Appl.* **48** 257–263. [MR0683223 https://doi.org/10.1016/0024-3795\(82\)90112-4](https://doi.org/10.1016/0024-3795(82)90112-4)
- PEYRÉ, G., CUTURI, M. et al. (2019). Computational optimal transport. *Found. Trends Mach. Learn.* **11** 355–607.
- RIPPL, T., MUNK, A. and STURM, A. (2016). Limit laws of the empirical Wasserstein distance: Gaussian distributions. *J. Multivariate Anal.* **151** 90–109. [MR3545279 https://doi.org/10.1016/j.jmva.2016.06.005](https://doi.org/10.1016/j.jmva.2016.06.005)
- TAKATSU, A. (2011). Wasserstein geometry of Gaussian measures. *Osaka J. Math.* **48** 1005–1026. [MR2871291](https://doi.org/10.1007/s12013-011-9129-1)
- TATUSKO, R. (1990). Cooperation in climate research: An evaluation of the activities conducted under the US–USSR agreement for environmental protection since 1974.
- URIBE, C. A., DVINSKIKH, D., DVURECHENSKY, P., GASNIKOV, A. and NEDIĆ, A. (2018). Distributed computation of Wasserstein barycenters over networks. In *2018 IEEE Conference on Decision and Control (CDC)* 6544–6549. IEEE.
- VAN DE GEER, S. (2006). *Empirical Processes in M-Estimation*. Cambridge Univ. Press, Cambridge.
- VILLANI, C. (2009). *Optimal Transport: Old and New. Grundlehren der Mathematischen Wissenschaften [Fundamental Principles of Mathematical Sciences]* **338**. Springer, Berlin. [MR2459454 https://doi.org/10.1007/978-3-540-71050-9](https://doi.org/10.1007/978-3-540-71050-9)
- WASSERMANN, D., BLOY, L., KANTERAKIS, E., VERMA, R. and DERICHE, R. (2010). Unsupervised white matter fiber clustering and tract probability map generation: Applications of a Gaussian process framework for white matter fibers. *NeuroImage* **51** 228–241.