MEASURING DEPENDENCE BETWEEN RANDOM VECTORS VIA OPTIMAL TRANSPORT

Gilles Mordant, Johan Segers

LIDAM Discussion Paper ISBA
2021 / 24

UCLouvain

LIDAM
ISBA | Louvain Institute of Data Analysis and Modeling in economics and statistics

MEASURING DEPENDENCE BETWEEN RANDOM VECTORS VIA OPTIMAL TRANSPORT

G. MORDANT & J. SEGERS

LIDAM/ISBA, UCLouvain
Voie du Roman Pays 20, B-1348 Louvain-la-Neuve, Belgium

Abstract. To quantify the dependence between two random vectors of possibly different dimensions, we propose to rely on the properties of the 2-Wasserstein distance. We first propose two coefficients that are based on the Wasserstein distance between the actual distribution and a reference distribution with independent components. The coefficients are normalized to take values between 0 and 1, where 1 represents the maximal amount of dependence possible given the two multivariate margins. We then make a quasi-Gaussian assumption that yields two additional coefficients rooted in the same ideas as the first two. These different coefficients are more amenable for distributional results and admit attractive formulas in terms of the joint covariance or correlation matrix. Furthermore, maximal dependence is proved to occur at the covariance matrix with minimal von Neumann entropy given the covariance matrices of the two multivariate margins. This result also helps us revisit the RV coefficient by proposing a sharper normalisation. The two coefficients based on the quasi-Gaussian approach can be estimated easily via the empirical covariance matrix. The estimators are asymptotically normal and their asymptotic variances are explicit functions of the covariance matrix, which can thus be estimated consistently too. The results extend to the Gaussian copula case, in which case the estimators are rank-based. The results are illustrated through theoretical examples, Monte Carlo simulations, and a case study involving electroencephalography data.

1. Introduction

Measuring dependence is a fundamental problem in statistics that has applications in nearly all other domains of science. Because of this importance, it is not surprising that early in their careers, most students learn about the Pearson correlation coefficient, quantifying linear association between two univariate random variables. In modern days, the abundance of data makes it possible to consider groups of variables and the question of measuring dependence between two random vectors appears naturally.

Hotelling (1936) proposed to address the matter by finding the linear combinations of both groups of variables that maximise the correlation coefficient. Canonical correlation analysis was born. Not much attention was devoted to the problem for decades and the

E-mail address: gilles.mordant@uclouvain.be.

Date: April 30, 2021.

Key words and phrases. Dependence, Optimal Transport, Bures–Wasserstein distance, RV coefficient, EEG data.

next development we are aware of is the RV coefficient proposed by Escoufier (1973). For a partitioned $d \times d$ covariance matrix

$$\Sigma = \begin{bmatrix} \Sigma_1 & \Psi \\ \Psi^\top & \Sigma_2 \end{bmatrix}, \tag{1}$$

with $d = p + q$ and with diagonal blocks Σ_1 and Σ_2 of dimensions $p \times p$ and $q \times q$, respectively, the RV coefficient (Escoufier, 1973; Robert and Escoufier, 1976) is

$$\mathrm{RV}(\Sigma) = \frac{\mathrm{tr}(\Psi\Psi^\top)}{\left(\mathrm{tr}(\Sigma_1^2)\,\mathrm{tr}(\Sigma_2^2)\right)^{1/2}}, \tag{2}$$

where $\mathrm{tr}(\,\cdot\,)$ is the trace operator and $(\,\cdot\,)^\top$ denotes matrix transposition. The coefficient is based on the scalar product between certain linear operators associated to the random vectors and is the first extension of the correlation coefficient that is multivariate in nature. Still, for given diagonal blocks Σ_1 and Σ_2 the maximal value attainable is in general smaller than one. In the course of our developments, we will propose another scaling that repairs this minor deficiency (Remark 3.13).

The following milestone is the work by Székely, Rizzo, and Bakirov (2007), where a weighted L_2 distance between characteristic functions is used to construct a dependence measure. Since then, a renewed interest for the question of quantifying dependence between random vectors has grown. The measure proposed by Zhu, Xu, Li, and Zhong (2017) is of the same nature, involving a weighted integral of the squared covariances between indicators associated to linear combinations with varying coefficient vectors.

To test for independence between several random vectors, Quessy (2010) studies a Cramér–von Mises statistic comparing the joint empirical copula with the product of the empirical copulas of the vectors separately. In Medovikov and Prokhorov (2017), the population version of this quantity lies at the basis of a copula-based dependence measure between several random vectors.

Another line of research considered measuring dependence relying on an aggregation of vectors into variables, an approach which can be seen as extending canonical correlation analysis. The multivariate generalisations of Spearman's ρ and Kendall's τ in Grothe, Schnieders, and Segers (2014) fall into this framework. In the same vein, Hofert, Oldford, Prasad, and Zhu (2019) proposed to compute the correlation between collapsing functions of groups of variables.

Recently, Puccetti (2019) proposed a dependence coefficient based on optimal transportation theory. Alike the RV-coefficient, it is based on traces of covariance matrices but the scaling accommodates for those that are attainable given the ones of both vectors of interest. The coefficient cannot be used for vectors with different dimensions and is not invariant with respect to permutations of variables within a group.

Still, as we shall see, the (2-)Wasserstein distance is a particularly convenient metric on the space of probability distributions with finite (second) moments and it can be leveraged to construct new dependence coefficients. The interest of this distance for statistical inference is not new but blossomed recently. We refer to Panaretos and Zemel (2019a) and Panaretos and Zemel (2019b) for background and surveys.

In this paper, we propose new dependence coefficients based on the 2-Wasserstein distance. As the asymptotic theory of the empirical Wasserstein distance is currently

not yet sufficiently developed to derive the results needed for statistical inference for these coefficients, we also propose quasi-Gaussian counterparts in terms of a partitioned covariance or correlation matrix. Our approach thus shares common points with both Escoufier's RV and Puccetti's coefficients. The proper normalisation of the coefficients involves the interesting side-problem of characterising, among all partitioned covariance matrices Σ of the form (1) with fixed diagonal blocks Σ_1 and Σ_2, the $p \times q$ cross-covariance matrix Ψ that yields the strongest dependence.

We then propose plug-in estimators and prove their asymptotic normality by means of the delta method. The asymptotic variances admit analytic formulas and can therefore be estimated by a plug-in approach too, avoiding the need for resampling procedures. The Fréchet differentiability of the maps that send a covariance or correlation matrix to the coefficients means that the asymptotic distributions of plug-in estimators can be studied in a wide variety of settings, including time series, graphical models, and rank-based estimators. The approach is akin to the one of estimating the Wasserstein distance between Gaussian distributions in Rippl, Munk, and Sturm (2016). In passing, our calculations shed new light on the Fréchet differentiability of the Wasserstein distance derived in that article.

Rescaling the univariate margins to the standard Gaussian distribution prior to computing the correlation matrix has two advantages: first, no moment conditions are required and second, the coefficients become invariant under component-wise increasing transformations. The proposed standardisation is particularly natural in the Gaussian copula case, a model assumption which has been gaining popularity since Liu, Lafferty, and Wasserman (2009), for instance for graphical models. We illustrate the coefficients on electroencephalogram (EEG) data modelled in this way in Solea and Li (2020). The estimates relies on the matrix of normal scores rank correlation coefficients, asymptotic expansions of which were established in Klaassen and Wellner (1997).

The outline of this paper is the following. In Section 2, we propose new dependence coefficients between random vectors exploiting the properties of the Wasserstein distance. In Section 3, we introduce a quasi-Gaussian version of the coefficients based on the Bures–Wasserstein distance (Bhatia, Jain, and Lim, 2019) between certain covariance matrices. Plug-in estimators and their limiting distributions are treated in Section 4. We then study the performance of the proposed estimator via Monte Carlo simulations in Section 5 and propose an application to the already mentioned EEG data in Section 6. Section 7 concludes and paves the way for further developments. The Appendices contain proofs and the results of some additional numerical experiments.

2. Wasserstein dependence coefficients

Let $\mathcal{P}(\mathbb{R}^d)$ be the set of Borel probability measures on $\mathbb{R}^d$ and let $\mathcal{P}_2(\mathbb{R}^d) \subset \mathcal{P}(\mathbb{R}^d)$ be the set of such measures with finite second moments. For $(\pi, \pi') \in \mathcal{P}_2(\mathbb{R}^d)^2$, let $\Gamma(\pi, \pi')$ be the set of couplings $\gamma \in \mathcal{P}_2(\mathbb{R}^{2d})$ of π and π', that is, probability measures γ such that $\gamma(B \times \mathbb{R}^d) = \pi(B)$ and $\gamma(\mathbb{R}^d \times B) = \pi'(B)$ for Borel sets $B \subseteq \mathbb{R}^d$. Let W_2 denote the 2-Wasserstein distance on $\mathcal{P}_2(\mathbb{R}^d)$: its square is

$$W_2^2(\pi, \pi') = \inf_{\gamma \in \Gamma(\pi, \pi')} \int_{\mathbb{R}^{2d}} \|v - v'\|^2 \, \mathrm{d}\gamma(v, v'), \qquad \pi, \pi' \in \mathcal{P}_2(\mathbb{R}^d).$$

This defines a metric on $\mathcal{P}_2(\mathbb{R}^d)$, the origins of which go back to Kantorovich; see Panaretos and Zemel (2019b) for a survey and historical notes. The infimum is attained and the corresponding γ is called an optimal coupling between π and π'.

For a random vector (X, Y) of dimension $d = p + q$ and with joint law $\pi \in \mathcal{P}_2(\mathbb{R}^d)$, we seek to quantify the dependence between the subvectors X and Y. Let $\mu \in \mathcal{P}_2(\mathbb{R}^p)$ and $\nu \in \mathcal{P}_2(\mathbb{R}^q)$ denote the distributions of X and Y, respectively. Note that π belongs to $\Gamma(\mu, \nu)$, the set of couplings of μ and ν. The assumption that π has finite second moments is not a real restriction since we can first transform its univariate margins to a suitable distribution, see Remark 2.4.

To quantify the dependence between X and Y, we compare π to $\mu \otimes \nu$, where $\otimes$ denotes product measure—the distribution of an independent coupling. Let $\mathcal{P}_{2,0}(\mathbb{R}^r)$ be the subset of $\mathcal{P}_2(\mathbb{R}^r)$ of all non-degenerate distributions. Choose reference laws $\upsilon_1 \in \mathcal{P}_{2,0}(\mathbb{R}^p)$ and $\upsilon_2 \in \mathcal{P}_{2,0}(\mathbb{R}^q)$ and put

$$\begin{aligned} T_{p,q}(\pi; \upsilon_1, \upsilon_2) &= W_2^2(\pi, \upsilon_1 \otimes \upsilon_2) - W_2^2(\mu \otimes \nu, \upsilon_1 \otimes \upsilon_2) \\ &= W_2^2(\pi, \upsilon_1 \otimes \upsilon_2) - W_2^2(\mu, \upsilon_1) - W_2^2(\nu, \upsilon_2). \end{aligned} \tag{3}$$

Lemma 2.1. *For $\pi, \mu, \nu, \upsilon_1, \upsilon_2$ as above, $T_{p,q}$ in (3) satisfies the following properties:*

(i) $T_{p,q}(\pi; \upsilon_1, \upsilon_2) \geq 0$.
(ii) $T_{p,q}(\mu \otimes \nu; \upsilon_1, \upsilon_2) = 0$.
(iii) If either $\upsilon_1 = \mu$ and $\upsilon_2 = \nu$ or if both υ_1 and υ_2 are absolutely continuous, then $T_{p,q}(\pi; \upsilon_1, \upsilon_2) = 0$ implies $\pi = \mu \otimes \nu$.

The proof of Lemma 2.1 is given in Appendix A. For υ_1 and υ_2 as in Lemma 2.1(iii), we have $T_{p,q}(\pi; \upsilon_1, \upsilon_2) \geq 0$ with equality if and only if $\pi = \mu \otimes \nu$. This fact motivates the use of $T_{p,q}$ to quantify dependence between the subvectors X and Y of a random vector $X = (X, Y)$ with law π. To obtain a coefficient between 0 and 1, we propose to rescale $T_{p,q}(\pi; \upsilon_1, \upsilon_2)$ by the largest possible value over all couplings $\tilde{\pi}$ of μ and ν, provided these are both non-degenerate:

$$\tilde{\mathfrak{D}}(\pi; \upsilon_1, \upsilon_2) = \frac{T_{p,q}(\pi; \upsilon_1, \upsilon_2)}{\sup_{\tilde{\pi} \in \Gamma(\mu,\nu)} T_{p,q}(\tilde{\pi}; \upsilon_1, \upsilon_2)}. \tag{4}$$

The coefficient is indicated with a tilde to indicate the link and difference with the covariance-matrix-based coefficients defined in Section 3. Under the conditions of Lemma 2.1(iii) and as μ and ν are non-degenerate, the supremum in the denominator in (4) is positive. In that case, $\tilde{\mathfrak{D}}(\pi; \upsilon_1, \upsilon_2) \in [0, 1]$, while $\tilde{\mathfrak{D}}(\pi; \upsilon_1, \upsilon_2) = 0$ if and only if $\pi = \mu \otimes \nu$. The supremum in the denominator is attained since $T_{p,q}(\,\cdot\,; \upsilon_1, \upsilon_2)$ is W_2-continuous on $\mathcal{P}_2(\mathbb{R}^d)$ and $\Gamma(\mu, \nu)$ is W_2-compact in $\mathcal{P}_2(\mathbb{R}^d)$, as W_2-convergence implies convergence in distribution and the margins are fixed.

From Eq. (4), we can define two dependence measures that are theoretically particularly appealing. For integer $m \geq 1$, let $\gamma_m = \mathcal{N}_m(0, I_m)$ denote the m-variate centred and isotropic Gaussian distribution, with I_m the $m \times m$ identity matrix.

Definition 2.2 (Wasserstein dependence coefficients)**.** *For positive integer $d = p + q$ and for $\pi \in \Gamma(\mu, \nu)$ with $\mu \in \mathcal{P}_{2,0}(\mathbb{R}^p)$ and $\nu \in \mathcal{P}_{2,0}(\mathbb{R}^q)$, define*

$$\tilde{\mathfrak{D}}_1(\pi; p, q) = \tilde{\mathfrak{D}}(\pi; \gamma_p, \gamma_q) = \frac{W_2^2(\pi, \gamma_d) - W_2^2(\mu, \gamma_p) - W_2^2(\nu, \gamma_q)}{\sup_{\tilde{\pi} \in \Gamma(\mu,\nu)} W_2^2(\tilde{\pi}, \gamma_d) - W_2^2(\mu, \gamma_p) - W_2^2(\nu, \gamma_q)}$$

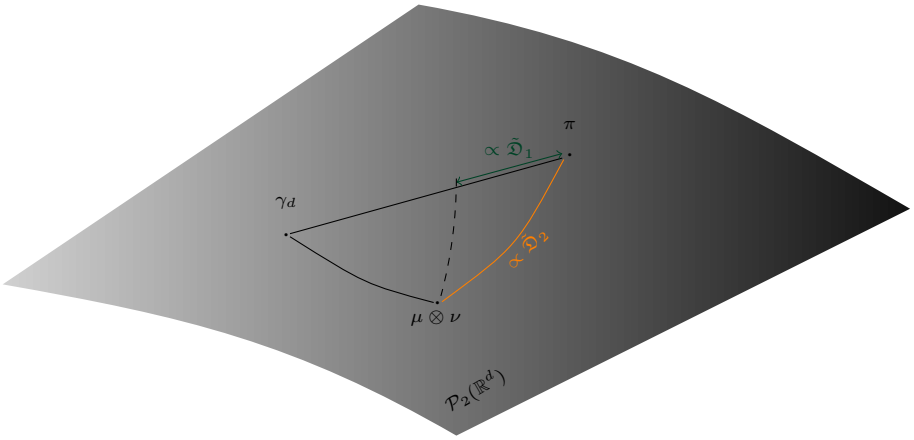

FIGURE 1. Representation of the proposed dependence coefficients

and

$$\tilde{\mathfrak{D}}_2(\pi; p, q) = \tilde{\mathfrak{D}}(\pi; \mu, \nu) = \frac{W_2^2(\pi, \mu \otimes \nu)}{\sup_{\tilde{\pi} \in \Gamma(\mu,\nu)} W_2^2(\tilde{\pi}, \mu \otimes \nu)}.$$

If the dimensions p and q are clear from the context, we just write $\tilde{\mathfrak{D}}_r(\pi)$ for $r \in \{1, 2\}$.

These measures enjoy the following properties (proof in Appendix A). Recall that an orthogonal transformation of Euclidean space is a linear transformation induced by an orthogonal matrix.

Proposition 2.3. *Let $d = p + q$, let $\mu \in \mathcal{P}_{2,0}(\mathbb{R}^p)$ and $\nu \in \mathcal{P}_{2,0}(\mathbb{R}^q)$ and let $\pi \in \Gamma(\mu, \nu)$. The dependence coefficients $\tilde{\mathfrak{D}}_r = \tilde{\mathfrak{D}}_r(\,\cdot\,; p, q)$ for $r \in \{1, 2\}$ satisfy the following properties:*

(i) $\tilde{\mathfrak{D}}_r(\pi) \in [0, 1]$, while $\tilde{\mathfrak{D}}_r(\pi) = 0$ if and only if $\pi = \mu \otimes \nu$.
(ii) There exists $\pi^{(r)} \in \Gamma(\mu, \nu)$ such that $\tilde{\mathfrak{D}}_r(\pi^{(r)}) = 1$.
(iii) $\tilde{\mathfrak{D}}_r$ is invariant w.r.t. orthogonal linear transformations within the first p and the last q coordinates.

Remark 2.4. If the univariate margins of π are continuous, then one can apply the dependence coefficients not to π but rather to a measure sharing the same copula and with margins admitting a finite second moment. The resulting coefficient would then be invariant with respect to permutations within the first p and last q coordinates and also to monotone increasing and decreasing transformations of the d univariate margins.

The two dependence measures are illustrated in Figure 1. Up to scaling, $\tilde{\mathfrak{D}}_2$ is the (squared) distance between π and $\mu \otimes \nu$, whereas $\tilde{\mathfrak{D}}_1$ is the excess squared distance from π to γ_d compared to the one between $\mu \otimes \nu$ and γ_d.

3. A quasi-Gaussian approach

Although theoretically appealing, the actual computation of the two Wasserstein dependence coefficients in Definition 2.2 is involved, not in the least because of the suprema in the denominators. Moreover, statistical inference on the coefficients is

hampered by a lack of a comprehensive large-sample theory for the Wasserstein distance involving empirical measures. We refer to Panaretos and Zemel (2019b) for a recent review of the known results.

Despite these drawbacks, the story does not end here. We instead propose a quasi-Gaussian approach based on covariance matrices. We start in Section 3.1 by defining the modified coefficients. The calculation of the two coefficients relies on an interesting optimisation problem yielding an elegant solution in terms of the minimum-entropy covariance matrix with given diagonal blocks in Section 3.2. The same matrix also realises the maximum value of the RV coefficient for fixed diagonal blocks, motivating the definition of an adjusted RV coefficient with range $[0, 1]$. The coefficients are illustrated for various families of structured covariance matrices in Section 3.3. We conclude in Section 3.4 with some thoughts on the application of the coefficients to distributions with standard Gaussian margins, which we call G-copulas. The proofs of all results in this section are given in Appendix B.

3.1. **Definition and basic properties.** The Wasserstein distance between centred Gaussian distributions is given by the so-called Bures–Wasserstein distance between their covariance matrices. We refer to Bhatia, Jain, and Lim (2019) for an introduction to this distance between positive semi-definite matrices and to Dowson and Landau (1982) and Olkin and Pukelsheim (1982) for a proof that this distance coincides with the Wasserstein distance for two (centred) measures belonging to the same elliptical family. Let $\mathbb{S}^d = \{A \in \mathbb{R}^{d\times d} : A^\top = A\}$ be the set of real symmetric $d \times d$ matrices, $\mathbb{S}^d_\geq \subset \mathbb{S}^d$ the set of positive semi-definite ones and $\mathbb{S}^d_> \subset \mathbb{S}^d_\geq$ the set of positive definite ones.

Definition 3.1. *The squared* Bures–Wasserstein distance *between* $\Sigma, \Xi \in \mathbb{S}^d_\geq$ *is*

$$d^2_W(\Sigma, \Xi) := W_2^2\big(\mathcal{N}_d(0, \Sigma), \mathcal{N}_d(0, \Xi)\big) = \operatorname{tr}(\Sigma) + \operatorname{tr}(\Xi) - 2\operatorname{tr}\big((\Sigma^{1/2}\Xi\Sigma^{1/2})^{1/2}\big). \tag{5}$$

To introduce the quasi-Gaussian version of the Wasserstein dependence coefficients in Definition 2.2, let $d = p + q$ be integer, let $\Sigma_1 \in \mathbb{S}^p_\geq$ and $\Sigma_2 \in \mathbb{S}^q_\geq$, and introduce the set

$$\Gamma(\Sigma_1, \Sigma_2) = \left\{\Sigma \in \mathbb{S}^d_\geq :\ \Sigma = \begin{bmatrix} \Sigma_1 & \Psi \\ \Psi^\top & \Sigma_2 \end{bmatrix} \text{ for some } \Psi \in \mathbb{R}^{p\times q}\right\}. \tag{6}$$

If (X, Y) is a random vector of dimension d such that X and Y have covariance matrices Σ_1 and Σ_2, respectively, then its joint covariance matrix Σ belongs to $\Gamma(\Sigma_1, \Sigma_2)$. Put

$$\Sigma_0 := \begin{bmatrix} \Sigma_1 & 0 \\ 0 & \Sigma_2 \end{bmatrix}, \tag{7}$$

the covariance matrix of an independent coupling of X and Y. To avoid division by zero in the next definition, we need to exclude the zero matrix: let $\mathbb{S}^d_{\geq,0} = \mathbb{S}^d_\geq \setminus \{0\}$.

Definition 3.2 (Quasi-Gaussian Wasserstein dependence coefficients)**.** *For* $\Sigma \in \Gamma(\Sigma_1, \Sigma_2)$ *with* $\Sigma_1 \in \mathbb{S}^p_{\geq,0}$ *and* $\Sigma_2 \in \mathbb{S}^q_{\geq,0}$*, define*

$$\mathfrak{D}_1(\Sigma; p, q) = \frac{d^2_W(\Sigma, I_d) - d^2_W(\Sigma_1, I_p) - d^2_W(\Sigma_2, I_q)}{\sup_{\tilde\Sigma\in\Gamma(\Sigma_1,\Sigma_2)} d^2_W(\tilde\Sigma, I_d) - d^2_W(\Sigma_1, I_p) - d^2_W(\Sigma_2, I_q)},$$

and

$$\mathfrak{D}_2(\Sigma; p, q) = \frac{d_W^2(\Sigma, \Sigma_0)}{\sup_{\tilde{\Sigma} \in \Gamma(\Sigma_1, \Sigma_2)} d_W^2(\tilde{\Sigma}, \Sigma_0)}.$$

If the random vector (X, Y) in dimension $d = p + q$ has law $\pi \in \mathcal{P}_2(\mathbb{R}^d)$ and covariance matrix Σ, then we also put $\mathfrak{D}_r(X, Y) = \mathfrak{D}_r(\pi; p, q) = \mathfrak{D}_r(\Sigma; p, q)$ for $r \in \{1, 2\}$.

These coefficients are to be compared with those in Definition 2.2. The Wasserstein distances in the latter have now been replaced by those between the centred Gaussian distributions with the same covariance matrices. Furthermore, in the denominator, the supremum is now with respect to all Gaussian couplings rather than between all couplings, Gaussian or not. Even when X and Y are themselves Gaussian, it is, to the best of our knowledge, an open question whether the supremum over all Gaussian couplings is equal to the supremum over all couplings.

Proposition 3.3. *Let $d = p + q$ and let $\Sigma \in \Gamma(\Sigma_1, \Sigma_2)$ with $\Sigma_1 \in \mathbb{S}^p_{\geq,0}$ and $\Sigma_2 \in \mathbb{S}^q_{\geq,0}$. The dependence coefficients $\mathfrak{D}_r = \mathfrak{D}_r(\,\cdot\,; p, q)$ for $r \in \{1, 2\}$ satisfy the following properties:*

(i) $\mathfrak{D}_r(\Sigma) \in [0, 1]$, while $\mathfrak{D}_r(\Sigma) = 0$ if and only if $\Sigma = \Sigma_0$ in (7).

(ii) There exists $\Sigma^{(r)} \in \Gamma(\Sigma_1, \Sigma_2)$ such that $\mathfrak{D}_r(\Sigma^{(r)}) = 1$.

(iii) $\mathfrak{D}_r$ is invariant w.r.t. orthogonal transformations within the first p and the last q coordinates: for orthogonal matrices O_1 and O_2 of dimensions $p \times p$ and $q \times q$, respectively, we have

$$\mathfrak{D}_r(O\Sigma O^\top) = \mathfrak{D}_r(\Sigma) \quad \text{with} \quad O = \begin{bmatrix} O_1 & 0 \\ 0 & O_2 \end{bmatrix}.$$

As the coefficients $\mathfrak{D}_r$ in Definition 3.2 are defined in terms of covariance matrices—including correlation matrices—they can be applied whenever such matrices show up and inference on them is feasible. A case we have in mind is when the copula of (X, Y) is Gaussian and Σ is the correlation matrix of the random vector obtained from (X, Y) by transforming the univariate margins to the standard normal distribution (Section 3.4). Plugging in an estimate of the covariance or correlation matrix produces estimates of the coefficients the asymptotic distributions of which can be obtained by the delta method (Section 4). This approach is akin to the one in Rippl, Munk, and Sturm (2016), who propose inference on the Wasserstein distance between Gaussian distributions based on estimated means and covariance matrices.

As one may expect, the simplification to covariance matrices comes at a price: in Proposition 3.3, a vanishing coefficient is no longer a guarantee for independence as it was in Proposition 2.3 but only implies that all cross-covariances are zero. This fact property is shared with the RV coefficient and the one in Puccetti (2019).

Assume all diagonal elements of Σ are positive and let $R = D_\Sigma^{-1/2} \Sigma D_\Sigma^{-1/2}$ be the correlation matrix associated to Σ, where D_Σ is the diagonal matrix having the same diagonal as Σ. Then $\mathfrak{D}_r(\Sigma)$ and $\mathfrak{D}_r(R)$ are different in general. Hence, as in principal component analysis, it may be a good idea to scale variables to have unit variance prior to the use of the coefficients.

Definition 3.2 leaves open the question of the calculation of the suprema in the denominators of $\mathfrak{D}_1$ and $\mathfrak{D}_2$. According to Proposition 3.3, the suprema are attained,

but the matrices where this occurs and the values of the suprema remain unspecified. The problem turns out to have an elegant and explicit solution.

3.2. **Majorisation of vectors of eigenvalues.** To explain the intuition, let R be a $d\times d$ correlation matrix with eigenvalues $\lambda_1 \geq \ldots \geq \lambda_d \geq 0$. Since $\lambda_1 + \cdots + \lambda_d = \mathrm{tr}(R) = d$, the proportion of the total variance explained by the first k principal components is $(\lambda_1 + \cdots + \lambda_k)/d$. The larger this proportion, the better the quality of representation of the d standardised variables on the linear subspace spanned by the first k principal components. Intuitively, the dimension reduction is more successful as the eigenvalues are more spread out. The worst case in this respect occurs when $\mathrm{tr}(R)$ is the identity matrix and all eigenvalues are equal to 1. The idea also applies in general for covariance matrices and underlies many inequalities in mathematics. It goes back to Hardy, Littlewood, and Pólya (1934, 1952) and even earlier to the works of I. Schur. This theory will be key to derive the maxima in $\mathfrak{D}_1$ and $\mathfrak{D}_2$.

We rely on the monograph by Marshall, Olkin, and Arnold (2011), from which the next definition and proposition are taken: see Definition 1.A.1 on page 8 and Proposition 3.C.1 on page 92, as well as the historical remarks on pages 93–95.

Definition 3.4 (Majorization)**.** *For two vectors $x, y \in \mathbb{R}^d$, we say that y majorizes x, notation $x \prec y$, if*

$$\begin{cases} \sum_{i=1}^k x_{[i]} & \leq & \sum_{i=1}^k y_{[i]}, & k = 1, \ldots, d-1, \\ \sum_{i=1}^d x_{[i]} & = & \sum_{i=1}^d y_{[i]}, \end{cases}$$

where $x_{[1]} \geq \ldots \geq x_{[d]}$ denote the elements of x in decreasing order, and similarly for y.

When applied to the vectors of eigenvalues λ and μ of two $d \times d$ covariance matrices Σ and Ξ, respectively, the relation $\lambda \prec \mu$ states that, for any $k = 1, \ldots, d-1$, the reduction to the first k principal components is more successful for Ξ than for Σ in terms of proportion of variance explained. The link between majorisation and the computation of the suprema in the denominators of $\mathfrak{D}_1$ and $\mathfrak{D}_2$ stems from the following property (Marshall, Olkin, and Arnold, 2011, Proposition 3.C.1).

Proposition 3.5 (Majorisation and convexity)**.** *If $I \subseteq \mathbb{R}$ is an interval and if $g : I \to \mathbb{R}$ is convex, then for all $x, y \in I^d$, we have*

$$x \prec y \implies \sum_{i=1}^d g(x_i) \leq \sum_{i=1}^d g(y_j).$$

For fixed diagonal blocks $\Sigma_1 \in \mathbb{S}_{\geq}^p$ and $\Sigma_2 \in \mathbb{S}_{\geq}^q$, does there exist $\Sigma_m \in \Gamma(\Sigma_1, \Sigma_2)$ in (6) whose vector of ordered eigenvalues majorises those of all other covariance matrices of that form? The answer is positive and this matrix turns out to attain the suprema in the definitions of $\mathfrak{D}_1$ and $\mathfrak{D}_2$ in Definition 3.2. The eigendecompositions of Σ_1 and Σ_2 are

$$\Sigma_j = U_j \Lambda_j U_j^\top, \qquad j \in \{1, 2\}, \tag{8}$$

where $\Lambda_1 = \mathrm{diag}(\lambda_{1,1}, \ldots, \lambda_{p,1})$ is the $p \times p$ diagonal matrix containing the p ordered eigenvalues $\lambda_{1,1} \geq \ldots \geq \lambda_{p,1} \geq 0$ of Σ_1, counting multiplicities, and where the columns of the $p \times p$ orthogonal matrix U_1 contain the corresponding eigenvectors. We set similar notation for the elements arising from the eigenvalue decomposition of Σ_2.

Theorem 3.6 (Eigenvalue majorisation given diagonal blocks)**.** *Let $\Sigma_1 \in \mathbb{S}_{\geq}^p$ and $\Sigma_2 \in \mathbb{S}_{\geq}^q$ have eigendecompositions* (8)*. Let $d = p + q$ and define the $d \times d$ matrix*

$$\Sigma_m = \begin{bmatrix} \Sigma_1 & \Psi_m \\ \Psi_m^\top & \Sigma_2 \end{bmatrix} \tag{9}$$

with $p \times q$ off-diagonal block

$$\Psi_m = U_1 \Lambda_2^{1/2} \Pi \Lambda_2^{1/2} U_2^\top, \tag{10}$$

where $\Pi \in \mathbb{R}^{p \times q}$ is the $p \times q$ upper left block of I_d. The eigenvalues of Σ_m are

$$\lambda(\Sigma_m) = (\lambda_{j,1} + \lambda_{j,2})_{j=1}^d \tag{11}$$

where $\lambda_{j,1} = 0$ if $j \geq p + 1$ and $\lambda_{j,2} = 0$ if $j \geq q + 1$. For any $\Sigma \in \Gamma(\Sigma_1, \Sigma_2)$ with eigenvalues $\lambda(\Sigma) = (\lambda_j)_{j=1}^d$, we have

$$\lambda(\Sigma) \prec \lambda(\Sigma_m).$$

Example 3.7 ($d = 2$)**.** If $p = q = 1$ and $\Sigma_j = \sigma_j^2$ for $j \in \{1, 2\}$, then the matrix Σ_m in (9) is

$$\Sigma_m = \begin{bmatrix} \sigma_1^2 & \sigma_1\sigma_2 \\ \sigma_1\sigma_2 & \sigma_2^2 \end{bmatrix}$$

with eigenvalues $\sigma_1^2 + \sigma_2^2$ and 0.

Example 3.8 ($d = 3$)**.** If $p = 1$ with $\Sigma_1 = 1$ and $q = 2$ with $\Sigma_2 = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}$ and $\rho \in [-1, 1]$, then

$$\Sigma_m = \begin{bmatrix} 1 & \sqrt{(1+|\rho|)/2} & \sqrt{(1+|\rho|)/2} \\ \sqrt{(1+|\rho|)/2} & 1 & \rho \\ \sqrt{(1+|\rho|)/2} & \rho & 1 \end{bmatrix},$$

the correlation matrix of (Z_1, X_2, X_3), with $Z_1 = (X_2 + \operatorname{sign}(\rho) X_3)/\sqrt{2}$ the first principal component of $(X_2, X_3) \sim \mathcal{N}_2(0, \Sigma_2)$. The ordered eigenvalues of Σ_2 are $1 + |\rho|$ and $1 - |\rho|$ and those of Σ_m are $2 + |\rho|$, $1 - |\rho|$ and 0.

Among all members of $\Gamma(\Sigma_1, \Sigma_2)$, the matrix Σ_m occupies a special place. According to the following proposition, it maximises the RV coefficient as well as the 2-Wasserstein distance with respect to both $\mathcal{N}_d(0, I_d)$ and $\mathcal{N}_d(0, \Sigma_0)$ for Σ_0 in (7). Given the constraints on the margins, we think of $\mathcal{N}_d(0, \Sigma_m)$ as the Gaussian distribution that is "least random", "most structured", or "farthest away from independence". These claims can be made precise if, as in Remark 3.12, the amount of structure is quantified by the von Neumann entropy.

Proposition 3.9 (Extremal properties of Σ_m)**.** *Let $d = p + q$ be integer and let $\Sigma_1 \in \mathbb{S}_{\geq}^p$ and $\Sigma_2 \in \mathbb{S}_{\geq}^q$. Among all $\Sigma \in \Gamma(\Sigma_1, \Sigma_2)$ in* (6)*, the matrix Σ_m in* (9)*:*

(i) maximizes $d_W(\Sigma, I_d)$;
(ii) maximizes $d_W(\Sigma, \Sigma_0)$ with Σ_0 as in 7;
(iii) maximizes $\operatorname{tr}(\Psi\Psi^\top)$ and therefore maximizes the RV coefficient.

In view of Proposition 3.9, we can now work out the dependence coefficients $\mathfrak{D}_1$ and $\mathfrak{D}_2$ in Definition 3.2. Let $\Sigma_1 \in \mathbb{S}^p_{\geq}$ and $\Sigma_2 \in \mathbb{S}^q_{\geq}$ and let $\Sigma \in \Gamma(\Sigma_1, \Sigma_2)$. Let $\lambda_1 \geq \ldots \geq \lambda_d \geq 0$ denote the eigenvalues of Σ, let $\lambda_{1,1} \geq \ldots \geq \lambda_{p,1} \geq 0$ denote those of Σ_1 and $\lambda_{1,2} \geq \ldots \geq \lambda_{q,2} \geq 0$ those of Σ_2.

Proposition 3.10 (Quasi-Gaussian Wasserstein dependence coefficients: computation)**.** *Let $\Sigma_1, \Sigma_2, \Sigma$ be as above, with Σ_1 and Σ_2 non-zero. For Σ_0 and Σ_m as in* (7) *and* (9)*, respectively, we have*

$$\begin{aligned}\mathfrak{D}_1(\Sigma) &= \frac{\operatorname{tr}(\Sigma_1^{1/2}) + \operatorname{tr}(\Sigma_2^{1/2}) - \operatorname{tr}(\Sigma^{1/2})}{\operatorname{tr}(\Sigma_1^{1/2}) + \operatorname{tr}(\Sigma_2^{1/2}) - \operatorname{tr}(\Sigma_m^{1/2})} \\ &= \frac{\sum_{j=1}^p \lambda_{j,1}^{1/2} + \sum_{j=1}^q \lambda_{j,2}^{1/2} - \sum_{j=1}^d \lambda_j^{1/2}}{\sum_{j=1}^p \lambda_{j,1}^{1/2} + \sum_{j=1}^q \lambda_{j,2}^{1/2} - \sum_{j=1}^{p\vee q} (\lambda_{j,1} + \lambda_{j,2})^{1/2}},\end{aligned}$$

and

$$\begin{aligned}\mathfrak{D}_2(\Sigma) &= \frac{\operatorname{tr}(\Sigma) - \operatorname{tr}\{(\Sigma_0^{1/2}\Sigma\Sigma_0^{1/2})^{1/2}\}}{\operatorname{tr}(\Sigma) - \operatorname{tr}\{(\Sigma_0^{1/2}\Sigma_m\Sigma_0^{1/2})^{1/2}\}} \\ &= \frac{\sum_{j=1}^d \lambda_j - \sum_{j=1}^d \kappa_j^{1/2}}{\sum_{j=1}^d \lambda_j - \sum_{j=1}^{p\vee q} (\lambda_{j,1}^2 + \lambda_{j,2}^2)^{1/2}}\end{aligned}$$

where $\kappa_1 \geq \ldots \geq \kappa_d \geq 0$ denote the eigenvalues of $\Sigma_0^{1/2}\Sigma\Sigma_0^{1/2}$.

The coefficient $\mathfrak{D}_1(\Sigma)$ depends on Σ only through the eigenvalues of Σ_1, Σ_2 and Σ itself. The coefficient $\mathfrak{D}_2(\Sigma)$, instead, requires the eigenvalues of Σ_1, Σ_2 and $\Sigma_0^{1/2}\Sigma\Sigma_0^{1/2}$. We will see in the examples and the case study that the values of $\mathfrak{D}_1(\Sigma)$ and $\mathfrak{D}_2(\Sigma)$ are often rather close. The interpretation of $\mathfrak{D}_2(\Sigma)$ may be more straightforward, comparing Σ directly with Σ_0, but in terms of computations, coefficient $\mathfrak{D}_1(\Sigma)$ is the simpler one.

Remark 3.11 (Perfectly correlated principal components). The matrix Σ_m in Eq. (9) is the covariance matrix of the random vector

$$\begin{bmatrix} U_1\Lambda_2^{1/2}Z_1 \\ U_2\Lambda_2^{1/2}Z_2 \end{bmatrix}$$

where $Z_1 = (Z_{1,1}, \ldots, Z_{1,p})^\top \sim \mathcal{N}_p(0, I_p)$ and $Z_2 = (Z_{2,1}, \ldots, Z_{2,q})^\top \sim \mathcal{N}_q(0, I_q)$ and where $Z_{k,1} = Z_{k,2}$ for $k \in \{1, \ldots, p \wedge q\}$, i.e., Z_1 and Z_2 have the first $p \wedge q$ components in common. If the random vector (X, Y) of dimension $d = p + q$ has covariance matrix Σ_m, then for $k \in \{1, \ldots, p \wedge q\}$, the k-th principal components of X and Y are perfectly correlated.

Remark 3.12 (von Neumann entropy). Among all matrices Σ of the form (6), the matrix Σ_m in Eq. (9) also minimises the von Neumann entropy (Petz, 2001, see Eq. (11))

$$-\operatorname{tr}(\Sigma \log \Sigma) = -\sum_{j=1}^d \lambda_j \log \lambda_j$$

with $\lambda \log \lambda$ to be interpreted as 0 for $\lambda = 0$, and where the sum is over all d eigenvalues of Σ, counting multiplicities. The property follows from Proposition 3.5 and Theorem 3.6 since the function $\lambda \mapsto -\lambda \log \lambda$ is convex. The von Neumann entropy is a generalisation of the concept of entropy that turned useful in quantum physics in which the operators of interest are density matrices. The definition strongly resembles the one of the Shannon entropy in information theory where the eigenvalues in the above display are replaced by the probabilities associated to a finite number of events.

Remark 3.13 (Adjusted RV coefficient). For Ψ_m as in Eq. (10), we have

$$\operatorname{tr}(\Psi_m \Psi_m^\top) = \operatorname{tr}(\Lambda_2 \Pi \Lambda_2) = \sum_{j=1}^{p} \lambda_{j,1} \lambda_{j,2}.$$

Given the diagonal blocks Σ_1 and Σ_2, this is the maximal value of the numerator in the RV coefficient in Eq. (2). We therefore propose to adjust the RV coefficient by

$$\overline{\mathrm{RV}}(\Sigma) = \frac{\mathrm{RV}(\Sigma)}{\mathrm{RV}(\Sigma_m)} = \frac{\operatorname{tr}(\Psi \Psi^\top)}{\operatorname{tr}(\Psi_m \Psi_m^\top)}. \tag{12}$$

We have $0 \leq \mathrm{RV} \leq \overline{\mathrm{RV}} \leq 1$, and in contrast to RV, given Σ_1 and Σ_2, the adjusted version $\overline{\mathrm{RV}}$ can take on all values between 0 and 1.

3.3. **Examples.** We compute the dependence coefficients $\mathfrak{D}_1(\Sigma)$ and $\mathfrak{D}_2(\Sigma)$ for Σ in some parametric families of correlation matrices. For comparison, we also show the RV coefficient and its adjusted version $\overline{\mathrm{RV}}$ in (12). In these low-dimensional examples, the difference between the RV and the adjusted coefficient remains small. The difference however clearly materializes in higher-dimensional examples as in Figure 4 (Top row), for instance.

Example 3.14 (Bivariate correlation matrix)**.** Let $p = q = 1$ and for $\rho \in [-1, 1]$ put

$$\Sigma = \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix}.$$

From Proposition 3.9, we find

$$\mathfrak{D}_1(\Sigma) = \mathfrak{D}_2(\Sigma) = \frac{2 - \sqrt{1+\rho} - \sqrt{1-\rho}}{2 - \sqrt{2}}.$$

The RV coefficient and the adjusted version $\overline{\mathrm{RV}}$ in (12) are both equal to ρ^2 while the coefficient in Puccetti (2019) is equal to ρ itself. In this case, the square of the distance correlation by Székely, Rizzo, and Bakirov (2007) is a function of ρ too. The exact formula for this quantity is given in Theorem 7 in that article and reads

$$\frac{\rho \arcsin(\rho) + \sqrt{1-\rho^2} - \rho \arcsin(\rho/2) - \sqrt{4-\rho^2} + 1}{1 + \pi/3 - \sqrt{3}}.$$

These different coefficients are shown in Figure 2 on the left.

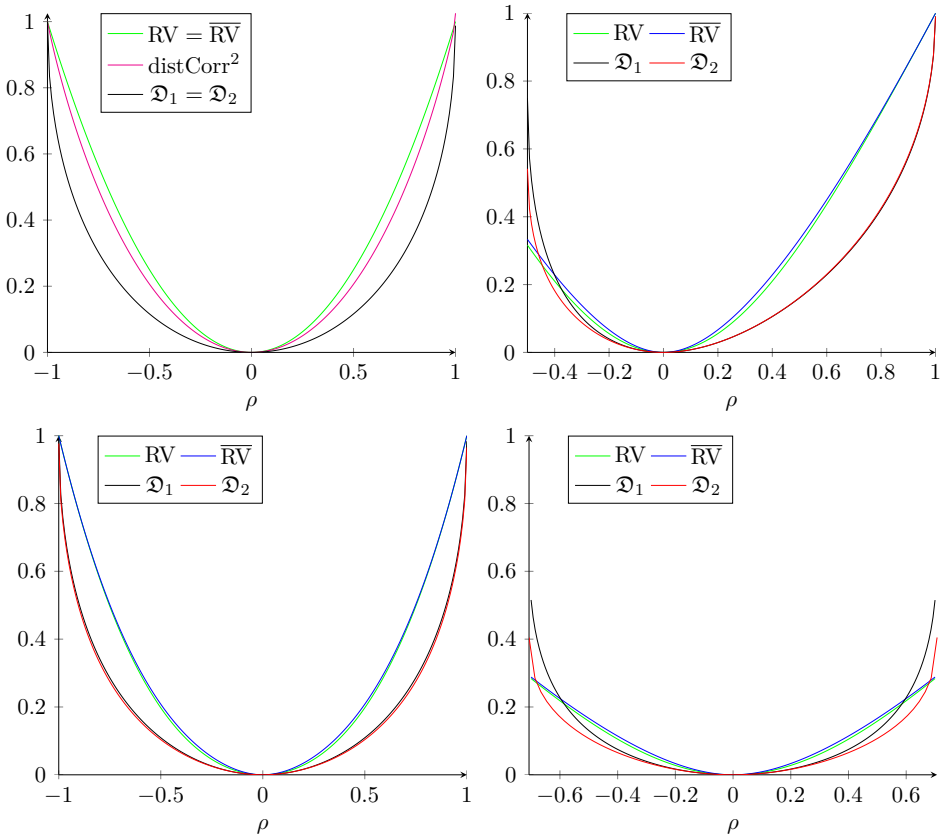

FIGURE 2. *Dependence coefficients in various families of correlation matrices. (Top left) Bivariate correlation matrix. (Top right) Trivariate equicorrelated matrix. (Bottom left) trivariate autoregressive model. (Bottom right) Trivariate moving average model.*

Example 3.15 (Trivariate equicorrelated matrix)**.** Let $p = 1$ and $q = 2$ and for $\rho \in [-1/2, 1]$ put

$$\Sigma = \begin{bmatrix} 1 & \rho & \rho \\ \rho & 1 & \rho \\ \rho & \rho & 1 \end{bmatrix}.$$

The matrix Σ_m was calculated in Example 3.8. Even though $\mathfrak{D}_1(\Sigma) \neq \mathfrak{D}_2(\Sigma)$ in general, both functions are extremely close in this case for ρ positive, with $\sup_{0 \leq \rho \leq 1} |\mathfrak{D}_1(\Sigma) - \mathfrak{D}_2(\Sigma)| < 0.005$. The various coefficients are shown in Figure 2 (Top right). Some closed-form formulas used to produce the graphs exist and are deferred to the end of Appendix B for space considerations.

Example 3.16 (Model comparison)**.** In this example, we measure the dependence between a univariate random variable and a bivariate vector when the joint structure is either moving average or auto-regressive. The result for the various dependence coefficients is shown in the bottom row of Figure 2. The graph on the left pertains to the auto-regressive structure while the graph on the right corresponds to moving averages

structure, that is to the matrices

$$\begin{bmatrix} 1 & \rho & \rho^2 \\ \rho & 1 & \rho \\ \rho^2 & \rho & 1 \end{bmatrix} \quad \text{for } -1 \le \rho \le 1 \quad \text{and} \quad \begin{bmatrix} 1 & \rho & 0 \\ \rho & 1 & \rho \\ 0 & \rho & 1 \end{bmatrix} \quad \text{for } -\frac{1}{\sqrt{2}} \le \rho \le \frac{1}{\sqrt{2}}, \tag{13}$$

respectively. The corresponding formulas are again deferred to the end of Appendix B.

3.4. **G-copulas.** For a random vector (X, Y) in dimension $d = p + q$, the dependence coefficients $\mathfrak{D}_1$ and $\mathfrak{D}_2$ were defined in terms of its joint covariance matrix Σ. As already mentioned, one may first want to rescale the variables and define the coefficients in terms of the joint correlation matrix instead. A more radical standardisation is to transform the univariate margins to a common distribution with finite second moment. This can be achieved by a combination of the probability and quantile transforms, provided the margins are continuous, i.e., do not have atoms. The advantage of such an approach is that the dependence coefficients become invariant with respect to component-wise monotone increasing or decreasing transformations. Also, on the original scale, the distribution is no longer subject to any moment conditions.

In view of the coefficients' origin in the Wasserstein distance between Gaussian distributions, a natural choice for the standardisation target is the standard normal distribution. We call the resulting multivariate distribution a G-copula, as an alternative to classical copulas, whose margins are uniform on the unit interval. The idea is not new: in the context of copula density estimation, Geenens, Charpentier, and Paindaveine (2017) also prefer the standard normal distribution as pivot.

Let Φ denote the standard normal cumulative distribution function (cdf) and let $\Phi^{-1} : [0, 1] \to [-\infty, \infty]$ denote its inverse. A G-copula is simply a multivariate cdf with standard normal margins. By a trivial extension of Sklar's theorem, every multivariate cdf F with univariate margins $F^{(1)}, \dots, F^{(d)}$, admits a G-copula G such that

$$F(z) = G\big(\Phi^{-1} \circ F^{(1)}(z^{(1)}), \dots, \Phi^{-1} \circ F^{(d)}(z^{(d)})\big), \qquad z = (z^{(1)}, \dots, z^{(d)}) \in \mathbb{R}^d.$$

If the margins $F^{(1)}, \dots, F^{(d)}$ are continuous, the G-copula G in the above identity is unique and is equal to the cdf of

$$Z_G = \big(\Phi^{-1} \circ F^{(1)}(Z^{(1)}), \dots, \Phi^{-1} \circ F^{(d)}(Z^{(d)})\big),$$

where the random vector Z has cdf F. The entries of the correlation matrix Σ_G of Z_G are called normal correlation coefficients in Klaassen and Wellner (1997). They are the population versions of the normal scores rank correlation coefficients. The (ordinary) copula of Z is Gaussian if and only if its G-copula is a multivariate Gaussian distribution, i.e., the distribution of Z_G is $\mathcal{N}_d(0, \Sigma_G)$.

Given a random vector $Z = (X, Y)$ of dimension $d = p+q$ with continuous margins, we can now apply the dependence coefficients $\mathfrak{D}_1$ and $\mathfrak{D}_2$ to the random vector $Z_G = (X_G, Y_G)$ with standard normal margins obtained by the above operation. We obtain

$$\mathfrak{D}_{G,r}(X, Y) = \mathfrak{D}_r(X_G, Y_G) = \mathfrak{D}_r(\Sigma_G; p, q),$$

where Σ_G is the correlation matrix of the random vector (X_G, Y_G). Estimating Σ_G by the matrix of normal scores rank correlation coefficients yields a non-parametric rank-based estimator of $\mathfrak{D}_{G,r}(X, Y)$. In Section 4, we study the asymptotic distribution of this estimator in case the copula of (X, Y) is Gaussian.

4. Estimation of quasi-Gaussian Wasserstein dependence coefficients

In this section, we propose plug-in estimators for the Wasserstein-based dependence coefficients (Section 4.1) and establish their limiting distributions, which is paramount for inferential purposes (Section 4.3). Before obtaining the latter results, we establish in Section 4.2 the Fréchet differentiability of the maps $\Sigma \mapsto \mathfrak{D}_r(\Sigma; p, q)$ for $r \in \{1, 2\}$ in Definition 3.2, where Σ must satisfy some conditions. The latter result opens the door to the application of our coefficients in many contexts.

4.1. Estimators.

The dependence coefficients $\mathfrak{D}_r(\Sigma; p, q)$ for $r \in \{1, 2\}$ can be studied in any setting where a covariance or correlation matrix Σ shows up. The coefficient is zero if and only if $\Sigma = \Sigma_0$ in (7). This identity implies independence provided Σ is the covariance or correlation matrix of a Gaussian distribution. The latter may be the distribution of the observations themselves or, as in Section 3.4, it may be their G-copula. Still, the coefficients can be used in non-Gaussian settings too, in the same way as a principal component analysis can be applied to any covariance or correlation matrix.

Recalling that $\mathfrak{D}_r$ for $r \in \{1, 2\}$ is a function from a subset of the $d \times d$ symmetric positive semi-definite matrices to $[0, 1]$, a natural way to estimate the coefficients is to consider a *plug-in* estimator. If $\hat{\Sigma}_n$ is an estimator of the covariance or correlation matrix Σ of interest, we set

$$\hat{\mathfrak{D}}_{n,r} = \mathfrak{D}_r(\hat{\Sigma}_n). \tag{14}$$

An important point to highlight at this stage is the generality of the approach. The matrix $\hat{\Sigma}_n$ could be the empirical covariance or correlation matrix or, in case of a G-copula, the one of normal scores rank correlation coefficients. Constrained covariance matrices could be used for factor models, graphical models etc. In higher dimensions and depending on the context, one could employ a variety of regularization techniques, such as enforcing sparsity of the precision matrix or shrinking the eigenvalues. The impact of the latter will be investigated numerically in Section 5.

4.2. Fréchet differentiability.

First, we will prove the Fréchet differentiability of the maps $\Sigma \mapsto \mathfrak{D}_r(\Sigma; p, q)$ with $r \in \{1, 2\}$ for Σ a positive definite symmetric matrix. To this end, we will need an assumption on the diagonal blocks Σ_1 and Σ_2: we require that Σ_1 has p distinct non-zero eigenvalues and that Σ_2 has q distinct non-zero eigenvalues. Otherwise, the functionals are still compactly (Hadamard) differentiable, but the derivatives are no longer linear and the asymptotic distribution of the plug-in estimator (14) is no longer Gaussian. The phenomenon is caused by the denominator in the definition of the coefficients, which relies on the ordering of the eigenvalues. The issue is visible in Example 3.15 at $\rho = 0$.

As $\mathbb{S}^d$, the space of symmetric real $d \times d$ matrices, is isomorph to a linear subspace of $\mathbb{R}^{d^2}$, any linear map $\mathbb{S}^d \to \mathbb{R}$ can be written as a trace inner product of the form

$$H \mapsto \operatorname{tr}(MH) = \sum_{i=1}^{d} \sum_{j=1}^{d} M_{ij} H_{ij} \tag{15}$$

for some $M \in \mathbb{S}^d$. Fréchet derivatives being linear maps, we will write them in the above form. The main challenge will thus be to identify the matrices M_r in the limits

$$\lim_{t \downarrow 0} t^{-1}\big(\mathfrak{D}_r(\Sigma + tH_t) - \mathfrak{D}_r(\Sigma)\big) = \operatorname{tr}(M_r H) \tag{16}$$

for $r \in \{1, 2\}$, where $H_t, H \in \mathbb{S}^d$ and $H_t \to H$ element-wise as $t \downarrow 0$. We will assume that Σ is positive definite, and then $\Sigma + tH_t$ will be so too for t sufficiently close to zero.

We introduce some notation. Recall that $\mathbb{S}^m_>$ denotes the set of symmetric positive definite real $m \times m$ matrices. Fix positive integer $d = p + q$. Let $\Sigma_1 \in \mathbb{S}^p_>$ and $\Sigma_2 \in \mathbb{S}^q_>$ and let $\Sigma \in \Gamma(\Sigma_1, \Sigma_2)$ as in (6) and Σ_0 as in (7). The eigendecompositions $\Sigma_r = U_r \Lambda_r U_r^\top$ for $r \in \{1, 2\}$ in (8) allow us to define the matrix Σ_m in (9). Let Π_1 be the projection matrix onto the first p coordinates and Π_2 the one onto the last q coordinates, that is,

$$\Pi_1 = \begin{bmatrix} I_p & 0 \end{bmatrix} \in \mathbb{R}^{p \times d}, \qquad \Pi_2 = \begin{bmatrix} 0 & I_q \end{bmatrix} \in \mathbb{R}^{q \times d}. \tag{17}$$

Note that $\Sigma_j = \Pi_j \Sigma \Pi_j^\top$ for $j \in \{1, 2\}$. Assume $q \geq p$ (otherwise, switch the roles of p and q) and partition the second eigenvalue matrix $\Lambda_2 \in \mathbb{S}^q_>$ as

$$\Lambda_2 = \begin{bmatrix} \Lambda_{2,1} & 0 \\ 0 & \Lambda_{2,2} \end{bmatrix} \tag{18}$$

with $\Lambda_{2,1} \in \mathbb{S}^p_>$ containing the first p eigenvalues and $\Lambda_{2,2} \in \mathbb{S}^{q-p}_>$ the remaining $q - p$ ones, the second block being empty if $q = p$. Finally, define

$$\Delta_1 = (\Lambda_1 + \Lambda_{2,1})^{-1/2}, \qquad \Delta_2 = \begin{bmatrix} \Delta_1 & 0 \\ 0 & \Lambda_{2,2}^{-1/2} \end{bmatrix}. \tag{19}$$

We can now state the differentiability of $\mathfrak{D}_1$ and $\mathfrak{D}_2$ with derivatives in the form (16). The meaning of the constants and matrices in the formulas is explained in Remark 4.3.

Theorem 4.1 (Differentiability of $\mathfrak{D}_1$)**.** *Consider the set-up in the previous paragraph. Assume that $\Sigma \in \mathbb{S}^d_>$, that Σ_1 has p distinct eigenvalues and Σ_2 has q distinct eigenvalues. Let $H_t \in \mathbb{S}^d$ for $t > 0$ and $H \in \mathbb{S}^d$ be such that $H_t \to H$ element-wise as $t \downarrow 0$. Then*

$$\lim_{t \to 0} t^{-1}\big(\mathfrak{D}_1(\Sigma + tH_t) - \mathfrak{D}_1(\Sigma)\big) = \operatorname{tr}(M_1 H)$$

with

$$\begin{aligned} M_1 &= \frac{1}{2c_1}\left(-\Sigma^{-1/2} + \big(1 - \mathfrak{D}_1(\Sigma)\big)\Sigma_0^{-1/2} + \mathfrak{D}_1(\Sigma)\Upsilon_1\right), \\ c_1 &= \operatorname{tr}(\Sigma_1^{1/2}) + \operatorname{tr}(\Sigma_2^{1/2}) - \operatorname{tr}(\Sigma_m^{1/2}), \\ \Upsilon_1 &= \begin{bmatrix} U_1 \Delta_1 U_1^\top & 0 \\ 0 & U_2 \Delta_2 U_2^\top \end{bmatrix}. \end{aligned} \tag{20}$$

To state the Fréchet differentiability of $\mathfrak{D}_2$, we need some additional notation. Recall the eigendecompositions (8) of Σ_1 and Σ_2 and recall the partitioning of Λ_2 in (18). Similar to (19), define

$$\Delta_1' = (\Lambda_1^2 + \Lambda_{2,1}^2)^{-1/2}\Lambda_1, \qquad \Delta_2' = \begin{bmatrix} (\Lambda_1^2 + \Lambda_{2,1}^2)^{-1/2}\Lambda_{2,1} & 0 \\ 0 & I_{q-p} \end{bmatrix},$$

the second diagonal block of Δ_2' being empty if $q = p$. Consider the $d \times d$ matrices

$$J = \Sigma_0^{-1/2}(\Sigma_0^{1/2}\Sigma\Sigma_0^{1/2})^{1/2}\Sigma_0^{-1/2} = \begin{bmatrix} J_{11} & J_{12} \\ J_{21} & J_{22} \end{bmatrix}, \tag{21}$$

$$J_0 = \begin{bmatrix} J_{11} & 0 \\ 0 & J_{22} \end{bmatrix}, \tag{22}$$

the dimensions of the two diagonal blocks J_{11} and J_{22} being $p \times p$ and $q \times q$, respectively.

Theorem 4.2 (Differentiability of $\mathfrak{D}_2$)**.** *Under the same assumptions as in Theorem 4.1, we have*

$$\lim_{t\to 0} t^{-1}\big(\mathfrak{D}_2(\Sigma + tH_t) - \mathfrak{D}_2(\Sigma)\big) = \operatorname{tr}(M_2 H)$$

where

$$\begin{aligned} M_2 &= \frac{1}{c_2}\left(-\frac{1}{2}(J_0 + J^{-1}) + (1 - \mathfrak{D}_2(\Sigma))I_d + \mathfrak{D}_2(\Sigma)\Upsilon_2\right), \\ c_2 &= \operatorname{tr}(\Sigma) - \operatorname{tr}\left((\Sigma_0^{1/2}\Sigma_m\Sigma_0^{1/2})^{1/2}\right), \\ \Upsilon_2 &= \begin{bmatrix} U_1\Delta_1'U_1^\top & 0 \\ 0 & U_2\Delta_2'U_2^\top \end{bmatrix}. \end{aligned} \tag{23}$$

Remark 4.3 (Matrices and constants in Theorems 4.1 and 4.2.). The constants c_1 and c_2 are just the denominators of $\mathfrak{D}_1(\Sigma)$ and $\mathfrak{D}_2(\Sigma)$, respectively. The matrices Υ_1 and Υ_2 determine the Fréchet derivatives at Σ of $\operatorname{tr}(\Sigma_m^{1/2})$ and $\operatorname{tr}((\Sigma_0^{1/2}\Sigma_m\Sigma_0^{1/2})^{1/2})$, appearing in the denominators of $\mathfrak{D}_1(\Sigma)$ and $\mathfrak{D}_2(\Sigma)$, see Lemmas C.3 and C.6, respectively. The matrix J is the unique solution in $\mathbb{S}_>^d$ to the equation $J\Sigma_0 J = \Sigma$ and the associated linear operator constitutes the optimal transport with respect to the squared Euclidean distance from $\mathcal{N}_d(0, \Sigma_0)$ to $\mathcal{N}_d(0, \Sigma)$ (Olkin and Pukelsheim, 1982).

Remark 4.4 (Fréchet derivative of Bures–Wasserstein distance). The proof of Theorem 4.2 requires the Fréchet derivative of the squared Bures–Wasserstein distance d_W^2 in (5). The latter is stated in Lemma 2.4 in Rippl, Munk, and Sturm (2016), but the formula is incorrect in case of repeated eigenvalues: the final double sum in their Eq. (21) should extend over all pairs $(i, m) \in \{1, \dots, d\}^2$ such that $i \neq m$, even those with $\lambda_i = \lambda_m$. Their expression is derived from Corollary 2.3 in Gilliam et al. (2009), but the projection matrix P_j in there is the one on the eigenspace of the eigenvalue λ_j, which, in case of repeated eigenvalues, has dimension larger than one. A formula for the Fréchet derivative of d_W^2 in the trace form (15) and not requiring eigendecompositions is given in Lemma C.4.

The matrix estimate used as input of the plug-in estimator in (14) could be a correlation matrix obtained from an estimated covariance matrix by rescaling the d variables by their estimated standard deviations. To find the asymptotic distribution of the resulting plug-in estimator, it is useful to know the Fréchet derivative of the composite map

$$\Sigma \mapsto \varphi(\Sigma) \mapsto \mathfrak{D}_r(\varphi(\Sigma)) \tag{24}$$

for $r \in \{1, 2\}$, where, for $\Sigma \in \mathbb{S}_\geq^d$ with positive diagonal elements, we put

$$\varphi(\Sigma) = D_\Sigma^{-1/2}\Sigma D_\Sigma^{-1/2} \tag{25}$$

with D_A the diagonal matrix having the same dimension and diagonal as the square matrix A. The map φ is scale invariant in the sense that $\varphi(\Delta\Sigma\Delta) = \varphi(\Sigma)$ for any diagonal matrix $\Delta \in \mathbb{S}^d_>$. It will therefore be sufficient to calculate the Fréchet derivative of the map (24) at a $d \times d$ correlation matrix R. Note that $D_R = I_d$ and thus $\varphi(R) = R$ for such a matrix.

Corollary 4.5 (Differentiability of dependence coefficients after rescaling)**.** *Let $R \in \mathbb{S}^d_>$ be a correlation matrix ($D_R = I_d$). Under the assumptions and notation of Theorems 4.1 and 4.2 with Σ replaced by R, we have, for $r \in \{1, 2\}$,*

$$\lim_{t \downarrow 0} t^{-1}\big(\mathfrak{D}_r(\varphi(R + tH_t)) - \mathfrak{D}_r(R)\big) = \mathrm{tr}\big((M_{R,r} - D_{M_{R,r}R})H\big),$$

where $M_{R,r}$ is the matrix M_r with Σ replaced by R.

4.3. **Asymptotic distributions.** Suppose that $\hat{\Sigma}_n$ is an estimator sequence of a covariance matrix Σ such that, for some deterministic sequence $0 < a_n \to \infty$, we have

$$a_n\left(\hat{\Sigma}_n - \Sigma\right) \rightsquigarrow H, \qquad n \to \infty, \tag{26}$$

where H is a random symmetric matrix and the arrow $\rightsquigarrow$ denotes convergence in distribution. The delta method in combination with Theorems 4.1 and 4.2 then yields

$$a_n\left(\mathfrak{D}_r(\hat{\Sigma}_n) - \mathfrak{D}(\Sigma)\right) \rightsquigarrow \mathrm{tr}(M_r H), \qquad n \to \infty, \qquad r \in \{1, 2\}. \tag{27}$$

Next, suppose Σ has correlation matrix $\varphi(\Sigma) = R$ as in (25) and we wish to estimate the dependence coefficient based on the estimated correlation matrix $\varphi(\hat{\Sigma}_n)$. The continuous mapping theorem and (26) imply

$$a_n\left(D_\Sigma^{-1/2}\hat{\Sigma}_n D_\Sigma^{-1/2} - R\right) \rightsquigarrow D_\Sigma^{-1/2} H D_\Sigma^{-1/2}, \qquad n \to \infty.$$

By scale invariance of φ, Corollary 4.5 and the delta method, it follows that, for $r \in \{1, 2\}$,

$$a_n\left(\mathfrak{D}_r(\varphi(\hat{\Sigma}_n)) - \mathfrak{D}_r(R)\right) \rightsquigarrow \mathrm{tr}\big((M_{R,r} - D_{M_{R,r}R})D_\Sigma^{-1/2} H D_\Sigma^{-1/2}\big), \qquad n \to \infty. \tag{28}$$

Often, the joint distribution of the elements of the random matrix H in (26) is Gaussian. By linearity, the weak limits in (27) and (28) are then Gaussian too. This includes for instance the sample covariance matrix of an independent random sample from a distribution with finite fourth moments (Kollo and von Rosen, 2006, Thm 3.1.4) or the matrix of pairwise Spearman's rank correlation coefficients of an independent random sample from a continuous distribution (El Maache and Lepage, 2003, Thm 2.2).

Here, we work out the limit distributions of the plug-in estimators in two settings:

(GD) the sample correlation matrix from an independent random sample from a Gaussian distribution;

(GC) the matrix of normal scores rank correlation coefficients of an independent random sample from a continuous distribution with a Gaussian copula (see Section 3.4).

The common limit distribution in the two cases is centered normal. The asymptotic variance is an explicit and continuous function of the underlying correlation matrix. The latter can therefore be estimated consistently by a plug-in estimator too, permitting the construction of asymptotic confidence intervals.

For setting (GD), let $\xi_1, \dots, \xi_n$ be an independent random sample from the d-variate normal distribution $\mathcal{N}_d(\mu, \Sigma)$ with mean vector $\mu \in \mathbb{R}^d$ and covariance matrix $\Sigma \in \mathbb{S}^d$. We want to estimate the dependence coefficients $\mathfrak{D}_r(R)$ for $r \in \{1, 2\}$ associated to the correlation matrix $R = \varphi(\Sigma)$. The plug-in estimator is $\hat{\mathfrak{D}}_{n,r} = \mathfrak{D}_r(\hat{R}_n)$ where

$$\hat{R}_n = \varphi(\hat{\Sigma}_n) \qquad \text{with} \qquad \hat{\Sigma}_n = \frac{1}{n-1} \sum_{i=1}^{n} (\xi_i - \bar{\xi}_n)(\xi_i - \bar{\xi}_n)^\top \tag{29}$$

is the empirical correlation matrix, based on the empirical covariance matrix $\hat{\Sigma}_n$ and with $\bar{\xi}_n = n^{-1} \sum_{i=1}^{n} \xi_i$ the sample mean vector.

For setting (GC), let $\xi_1, \dots, \xi_n$ be an independent random sample from a d-variate cdf F with continuous univariate margins $F_1, \dots, F_d$ and G-copula equal to the cdf of $\mathcal{N}_d(0, R)$ with correlation matrix R. The plug-in estimator is now $\check{\mathfrak{D}}_{n,r} = \mathfrak{D}_r(\check{R}_n)$ where

$$\check{R}_n = (\check{\rho}_{n,jk})_{j,k=1}^{d} \qquad \text{with} \qquad \check{\rho}_{n,jk} = \frac{1}{n} \sum_{i=1}^{n} \hat{Z}_{ij} \hat{Z}_{ik} \Bigg/ \frac{1}{n} \sum_{i=1}^{n} \big(\Phi^{-1}(\tfrac{i}{n+1})\big)^2, \tag{30}$$

is the matrix of normal scores rank correlation coefficients (Hájek and Šidák, 1967, p. 113), defined in terms of the normal scores

$$\hat{Z}_{ij} = \Phi^{-1}\big(\tfrac{n}{n+1} \hat{F}_{nj}(\xi_{ij})\big)$$

and the marginal empirical cdf $x_j \mapsto \hat{F}_{nj}(x_j) = n^{-1} \sum_{i=1}^{n} \mathbb{1}\{\xi_{ij} \le x_j\}$.

Surprisingly, the estimators $\hat{R}_n$ and $\check{R}_n$ in settings (GD) and (GC), respectively, share the same asymptotic expansions: see Lemma C.8, which repackages Theorem 3.1 in Klaassen and Wellner (1997). This explains why the limit distributions of the plug-in estimators in both settings coincide. The form of the limit variance is a consequence of a particular property of the limit distribution of the empirical covariance matrix of a sample from the multivariate standard Gaussian distribution (Lemma C.7).

Theorem 4.6 (Asymptotic normality of plug-in estimators: Gaussian (copula) case)**.** *Let $R \in \mathbb{S}_>^d$ be a correlation matrix ($D_R = I_d$) such that the conditions of Theorem 4.1 are satisfied with Σ replaced by R. In settings (GD) and (GC) above, we have, for $\mathfrak{D}_{n,r} \in \{\hat{\mathfrak{D}}_{n,r}, \check{\mathfrak{D}}_{n,r}\}$ and $r \in \{1, 2\}$,*

$$\sqrt{n}(\mathfrak{D}_{n,r} - \mathfrak{D}_r(R)) \rightsquigarrow \mathcal{N}(0, \zeta_r^2), \qquad n \to \infty,$$

with asymptotic variance

$$\zeta_r^2 = 2 \operatorname{tr}\Big(\big(R(M_{R,r} - D_{M_{R,r}R})\big)^2\Big) \tag{31}$$

and $M_{R,r}$ the matrix M_r in Theorems 4.1 and 4.2 with Σ replaced by R.

For $r \in \{1, 2\}$, let $\zeta_{n,r}^2$ be the plug-in estimator of ζ_r^2 given by replacing R in (31) by $\hat{R}_n$ and $\check{R}_n$ in settings (GD) and (GC), respectively.

Corollary 4.7 (Asymptotic normality of studentized plug-in estimators)**.** *In the set-up of Theorem 4.6, we have $\zeta_{n,r}^2 \rightsquigarrow \zeta_r^2$ as $n \to \infty$ for $r \in \{1, 2\}$. If $\zeta_r^2 > 0$, then also*

$$\sqrt{n}(\mathfrak{D}_{n,r} - \mathfrak{D}_r(R))/\zeta_{n,r} \rightsquigarrow \mathcal{N}(0, 1), \qquad n \to \infty.$$

Corollary 4.7 permits a standard construction of asymptotic confidence intervals for $\mathfrak{D}_r(R)$. An alternative would be to employ the bootstrap as in Rippl, Munk, and Sturm (2016). We do not develop this here in view of the satisfactory finite-sample performance (Section 5.3) of the confidence intervals based on the normal approximation.

Remark 4.8 (Zero coefficient and testing independence). If $\mathfrak{D}_r(R) = 0$, then necessarily $\zeta_r^2 = 0$ in Theorem 4.6: $\sqrt{n}(\mathfrak{D}_{n,r} - \mathfrak{D}_r(R))$ is non-negative and its limit distribution is centered normal, so the asymptotic variance must be zero. This means that Theorem 4.6 and Corollary 4.7 cannot be used to construct tests for independence. Instead, a higher-order result would be needed, stating weak convergence of $n\mathfrak{D}_{n,r}$ to a non-degenerate limit distribution, as in Rippl, Munk, and Sturm (2016, Theorem 2.3). Since $\mathfrak{D}_r(R) = 0$ does not imply independence anyway, we do not pursue this idea further.

Remark 4.9 ($d = 2$). For bivariate correlation matrices, the dependence coefficient $\mathfrak{D}_1(R) = \mathfrak{D}_2(R)$ is a smooth function of the pairwise correlation ρ (Example 3.14). The estimator $\mathfrak{D}_{n,r}$ is then equal to the corresponding value of the coefficient at the estimated correlation. The limit distribution in Theorem 4.6 is equal to the one given by the delta method in combination with the asymptotic normality of the empirical correlation for the bivariate normal distribution in setting (GD) and the normal scores rank correlation for the bivariate Gaussian copula in setting (GC).

5. Simulation experiments

In this section, we investigate the plug-in estimators for the dependence coefficients by means of various simulation experiments. First, we evaluate the quality of the approximation of their finite-sample distributions by the asymptotically normal one (Section 5.1). We then numerically assess the impact of shrinking the eigenvalues of the empirical covariance matrix to reduce the inherent bias (Section 5.2) and finally we evaluate the actual coverage of confidence intervals based on the normal approximation (Section 5.3).

5.1. **Gaussian goodness-of-fit for finite samples.** The Figure 3 presents P-P plots illustrating the asymptotic normality of the plug-in estimators in Section 4.3. The results are resented for for Gaussian data (GD) with correlation matrix estimated by $\hat{R}_n$ in (29).The standard normal distribution function is on the vertical axis while the actual sampling distribution function of $\sqrt{n}(\mathfrak{D}_{n,r} - \mathfrak{D}_r(R))/\zeta_{n,r}$ based on 3000 independent replications is on the horizontal one. From left to right, the sample sizes are 50, 200, 1000 and 5000, respectively. For brevity we only present the results for $\mathfrak{D}_1$ and for the more useful case when the estimation error is rescaled using a plug-in estimator of its asymptotic variance (Corollary 4.7). Additional simulation results for $\mathfrak{D}_2$ and for the (GC) case are shown in Appendix D.1.

The three rows correspond to the three following settings.

(1) A trivariate autoregressive matrix ($p = 1$, $q = 2$) as in (13) with coefficient $\rho = 0.25$. The true values of $\mathfrak{D}_1$ and $\mathfrak{D}_2$ are 0.026 and 0.025 respectively.
(2) A trivariate autoregressive matrix ($p = 1$, $q = 2$) with coefficient $\rho = 0.8$. The true values of $\mathfrak{D}_1$ and $\mathfrak{D}_2$ are 0.34 and 0.33 respectively.

(3) A five-variate correlation matrix with $p = 2$ and $q = 3$ without any particular structure:

$$\begin{bmatrix} 1.00 & 0.20 & 0.15 & 0.10 & 0.25 \\ 0.20 & 1.00 & 0.05 & 0.30 & 0.35 \\ 0.15 & 0.05 & 1.00 & 0.40 & 0.50 \\ 0.10 & 0.30 & 0.40 & 1.00 & 0.45 \\ 0.25 & 0.35 & 0.50 & 0.45 & 1.00 \end{bmatrix}.$$

The true values of $\mathfrak{D}_1$ and $\mathfrak{D}_2$ are 0.051 and 0.050 respectively.

Observing Figure 3 one can clearly see that in case $n = 50$, the quality of the normal approximation is much better for larger values of the coefficients. For the five-dimensional example, the lack-of-fit at $n = 50$ is rather pronounced, as one could expect given the number of matrix entries to estimate. In particular, the estimator has a large positive bias. In all three settings, the goodness-of-fit improves with the sample size, as expected. We evoke the high-dimensional case, that is, when the number of matrix entries is of the order of magnitude of n, in Section 7.

5.2. **Eigenvalue shrinkage.** The simulations in Section 5.1 reveal the plug-in estimator to have a positive bias for small sample sizes. This is not surprising; it was already noted by C. Stein in the '60s and '70s that the eigenvalues of the empirical covariance matrix tend to be more spread out than their population counterparts. We refer to Dey and Srinivasan (1985) and Donoho, Gavish, and Johnstone (2018) for references about the subject.

In the aforementioned works, new estimators of the covariance matrix were proposed. The idea is to shrink the largest eigenvalues and increase the smaller ones to correct for the discrepancy arising. We follow Dey and Srinivasan (1985). Let S be distributed according to the Wishart $W_d(\Sigma, n-1)$ distribution. The maximum likelihood estimator of the covariance matrix of the $\mathcal{N}_d(\mu, \Sigma)$ distribution with unknown μ and Σ based on an independent random sample of size n has distribution S/n.

Let $\hat{\Sigma} = \hat{U}\hat{\Lambda}\hat{U}^\top$ where $\hat{U}$ is an orthogonal matrix and $\hat{\Lambda}$ is a diagonal matrix with elements $l_1 \geq \ldots \geq l_d$. *Orthogonally invariant estimators* of Σ are those of the form

$$\hat{\Sigma}_\ell = \hat{U}\ell(\hat{\Lambda})\hat{U}^\top$$

where $\ell(\hat{\Lambda})$ is a diagonal matrix with elements $\ell_1(\hat{\Lambda}), \ldots, \ell_d(\hat{\Lambda})$. Many functions ℓ_j have been proposed that correspond to certain loss functions. The maximum likelihood estimator corresponds to $\ell_j^0(\hat{\Lambda}) = n^{-1}l_j$. In Dey and Srinivasan (1985), the following choices are considered:

- $\ell_j^m(\hat{\Lambda}) = d_j l_j$ (Theorem 3.1) referred to as DS1.
- $\ell_j^S(\hat{\Lambda}) = d_j l_j - (l_j \log l_j)\tau(u)/(b_1 + u)$ (Theorem 3.2) where $u = \sum_{j=1}^p (\log l_j)^2$, $b_1 > 5.76(d-2)^2/(n+d-1)^2$ and $\tau(u)$ is a function satisfying, among others, $0 < \tau(u) < 2.4(d-2)/(n+d-1)^2$. In their Section 4, they propose $b_1 = 5.8(d-2)^2/(n+d-1)$ and $\tau(u) = 1.2(d-2)/(n+d-1)^2$. This method is referred to as DS2.

The above shrinkage methods are based upon $\hat{\Sigma}$ sampled from $W_d(\Sigma, n)$, see Dey and Srinivasan (1985). Therefore, we replace n by $n-1$. These are but two choices out of a

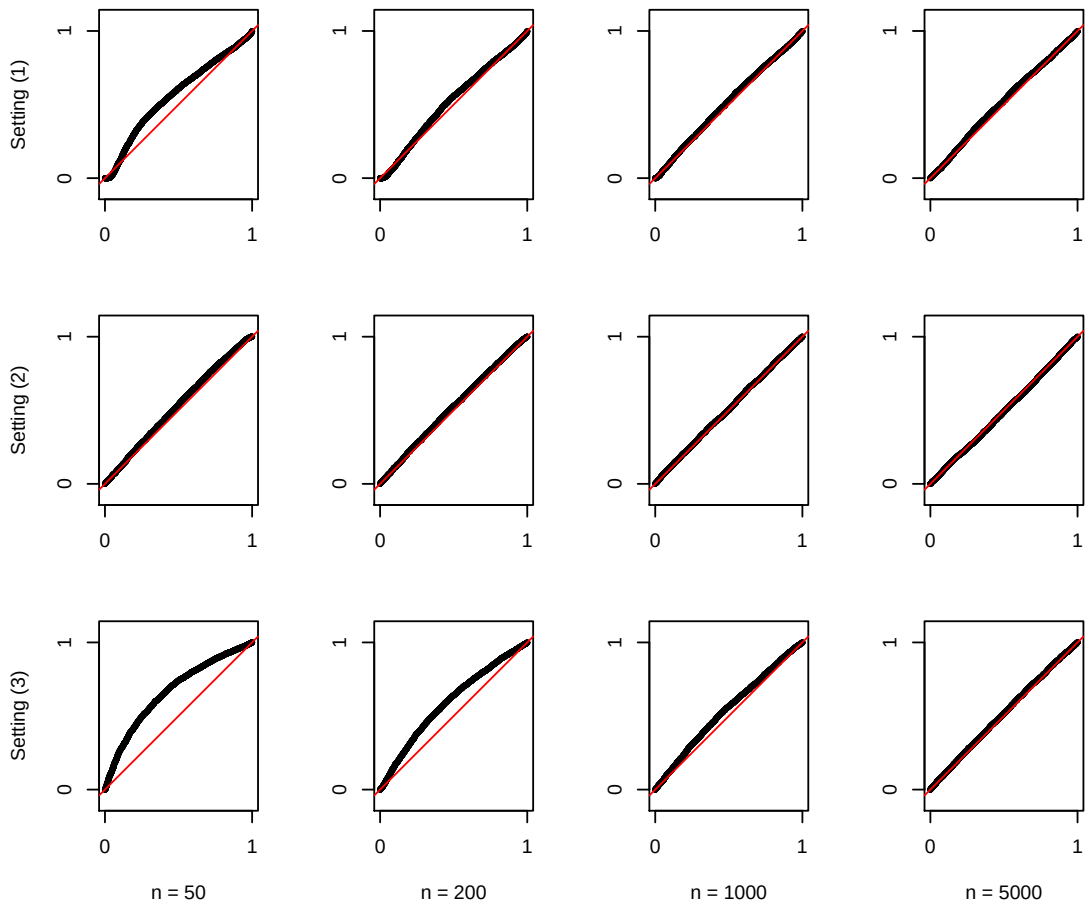

FIGURE 3. *P-P plots for 3000 repetitions of the centred and (empirically) rescaled estimator of $\mathfrak{D}_1$ in the three settings from Section 5.1 for increasing sample sizes (from left to right). The results are presented for an empirical correlation matrix in the case of Gaussian data.*

large number of shrinkage methods that depend on the loss function and the model. We refer to Donoho, Gavish, and Johnstone (2018) for a survey.

In Table 1, we consider settings (1) and (3) from Section 5.1 for sample size $n = 200$. The number of replications is 3000 and the results are obtained for the empirical correlation matrix in the fully Gaussian case, that is, case (GD) in Section 4.3. The entries in the table show the observed mean, median and standard deviation of the quantity $\sqrt{n}(\mathfrak{D}_r(\varphi(\hat{\Sigma}_\ell)) - \mathfrak{D}_r(R))/\zeta_{n,r}$, where the estimator $\hat{\Sigma}_\ell$ of the covariance matrix uses one of the shrinkage functions defined above and where the estimated standard error $\zeta_{n,r}$ is based on plugging in the estimated correlation matrix $\varphi(\hat{\Sigma}_\ell)$, similar to what was done in Corollary 4.7. From the results, we observe that shrinkage moves the median closer to zero in both settings while leaving the standard deviation close to one. There does not seem to be an important difference between DS1 and DS2.

		$\mathfrak{D}_1$			$\mathfrak{D}_2$		
Setting	Method	Mean	Median	SD	Mean	Median	SD
(1)	MLE	−0.026	0.118	1.292	−0.020	0.146	1.282
	DS1	−0.125	0.031	1.320	−0.116	0.058	1.303
	DS2	−0.126	0.031	1.320	−0.117	0.058	1.303
(3)	MLE	0.279	0.335	0.979	0.220	0.261	0.984
	DS1	0.118	0.174	0.981	0.075	0.120	0.986
	DS2	0.117	0.173	0.981	0.074	0.119	0.986

TABLE 1. *Effect of eigenvalue shrinkage methods on the studentised estimator,* $\sqrt{n}(\mathfrak{D}_r(\varphi(\hat{\Sigma}_\ell)) - \mathfrak{D}_r(R))/\zeta_{n,r}$, *at* $n = 200$ *in settings (1) and (3) in Section 5.1.*

		$\mathfrak{D}_1$				$\mathfrak{D}_2$			
Setting	n	True	LB	UB	Cov.	True	LB	UB	Cov.
(1)	50	0.026	0.000	0.104	93.8%	0.025	0.000	0.098	92.8%
	200	0.026	0.001	0.057	93.5%	0.025	0.001	0.056	93.5%
	1000	0.026	0.014	0.039	94.3%	0.025	0.014	0.038	94.4%
	5000	0.026	0.021	0.032	95.8%	0.025	0.020	0.030	95.5%
(3)	50	0.051	0.012	0.129	94.0%	0.050	0.009	0.128	94.4%
	200	0.051	0.028	0.083	94.8%	0.050	0.027	0.083	94.9%
	1000	0.051	0.040	0.064	94.6%	0.050	0.039	0.064	94.8%
	5000	0.051	0.045	0.056	95.4%	0.050	0.045	0.056	94.3%

TABLE 2. *Means of lower and upper bounds and actual coverage of rank-based asymptotic 95% confidence intervals* $[\mathfrak{D}_r(\check{R}_n) \pm z_{0.975} \times \zeta_{n,r}/\sqrt{n}] \cap [0,1]$ *in settings (1) and (3) in Section 5.1 over 3000 independent replications.*

5.3. **Coverage of confidence intervals.** We investigate the actual coverage of the asymptotic $(1-\alpha) \times 100\%$ confidence intervals $[\mathfrak{D}_r(\check{R}_n) \pm z_{1-\alpha/2} \times \zeta_{n,r}/\sqrt{n}] \cap [0,1]$ for various sample sizes, where z_p is the quantile of a standard normal distribution at level p. We consider settings (1) and (3) in Section 5.1 in the Gaussian copula (GC) case, so $\check{R}_n$ and $\zeta_{n,r}$ are as in (30) and Corollary 4.7. The chosen coverage probability is 95%. The results are presented in Tabless 2. For each coefficient $\mathfrak{D}_1$ and $\mathfrak{D}_2$, we give the true value, the mean of the lower and upper bounds over 3000 independent replications, and, finally, the empirical coverage. We did not rely on shrinkage methods in this part.

6. Case study: EEG data

We now turn to an application on real data exhibiting a possible use of the new dependence coefficients. We consider the electroencephalogram (EEG) dataset gathered

by Henri Begleiter[1] and first analysed in Zhang et al. (1995). Data are available for two types of patients: those suffering from alcoholism and a control group. The dataset consists of 120 trials for 122 subjects and is available on the UCI Machine Learning Archive (Dua and Graff, 2020).

An EEG measures the electric activity of the brain and thus helps to understand its functioning. In the dataset we consider, the data are gathered through 64 electrodes placed on the patient's scalp.[2] The electrical activity for each electrode is measured in μV through time. Each patient is exposed to a visual stimulus during a one-second timespan during which 256 measurements are collected. The 120 trials are divided into three types of stimuli tested: a single visual stimulus, two stimuli where the second one matches the first one and two stimuli where the second one does not match the first one. In each trial a different picture or different sets of pictures are used.

This dataset was recently analysed in Solea and Li (2020) and Anuragi and Sisodia (2020). In this first paper, the dependence structure is modelled under a Gaussian copula assumption, which has become classical since the seminal work of Liu, Lafferty, and Wasserman (2009). Even though the Gaussian copula hypothesis may seem restrictive, it turned out quite successful and is well accepted in the field, as stressed in Solea and Li (2020). In the sequel, we also make the assumption that the copula is Gaussian and thus use the rank-based estimator $\mathfrak{D}_r(\check{R}_{n,r})$ with $\check{R}_{n,r}$ the matrix of normal scores rank correlation coefficients in (30).

The graphs in Solea and Li (2020, p. 11) present the results of different estimation procedures for the dependence graph. A visual inspection shows that the connectivity networks estimated by the different methods largely differ from one estimation procedure to another. These discrepancies motivate our analysis of the dependence between the pre-frontal (FP) and the anterio-frontal (AF) electrodes, as the methods seem to estimate different network structures for these particular blocks. The AF region consists of the electrodes AF1, AF2, AF7, AF8 and AFZ while the FP region consists of the FP1, FP2 and FPZ electrodes. In our notation we are thus seeking to quantify dependence between a group of $p = 3$ variables and another one with $q = 5$ variables.

We chose to focus on trial No. 26. This choice is purely random and was made prior to the analysis. The only check that was made concerns the number of patients in the trial. Indeed, even though the experiment was carried out on 122 patients, certain results are missing. For the trial selected, the data for 99 patients were available. Among these 99 patients, 60 were alcoholic. Preliminary Monte-Carlo simulations evaluating the coverage probabilities of estimated confidence intervals—reported in Appendix D.3—suggested the use of the shrinkage estimator DS1 (Section 5.2) of the correlation matrix which is then standardised again via the square roots of the diagonal elements. This finding is purely empirical and theoretical justifications for this or other shrinkage methods in the context of the matrix of normal scores rank correlation coefficients are yet to be developed.

In the top row of Figure 4, we show estimates of various dependence coefficients for the two groups of patients. The coefficients are estimated at one out of five time instants to avoid overloading the graphs. To enable a proper comparison, the RV and $\overline{\text{RV}}$ are

[1]At the Neurodynamics Laboratory at the State University of New York Health Center at Brooklyn.

[2]The position of the electrodes follows the Standard Electrode Position Nomenclature put forward by the American Electroencephalographic Association in 1990.

computed on the same, shrinked matrix as the $\mathfrak{D}_r$ coefficients. The interest of correcting the RV coefficient as in Remark 3.13 is clear. The various coefficients exhibit quite similar profiles over time. Interestingly, the curve of the square of the adjusted RV coefficient (not shown) would be close to $\mathfrak{D}_1$ and $\mathfrak{D}_2$.

In the middle row of Figure 4, we compare the coefficients $\mathfrak{D}_1$ and $\mathfrak{D}_2$ for both types of patients and provide pointwise confidence bands. The latter are formed out of a confidence interval at each time instant, are based on the estimated asymptotic variance and are chosen to have a 95% coverage probability.

Assuming independence between alcoholics and control patients, an asymptotic two-sided $(1-\alpha)$ confidence interval for the difference $\mathfrak{D}_r^{\mathrm{ctr}} - \mathfrak{D}_r^{\mathrm{alc}}$ is

$$\hat{\mathfrak{D}}_r^{\mathrm{ctr}} - \hat{\mathfrak{D}}_r^{\mathrm{alc}} \pm z_{1-\alpha/2}\sqrt{\frac{(\hat{\zeta}_r^{\mathrm{ctr}})^2}{n^{\mathrm{ctr}}} + \frac{(\hat{\zeta}_r^{\mathrm{alc}})^2}{n^{\mathrm{alc}}}}$$

with $n^{\mathrm{ctr}} = 99 - 60 = 39$ and $n^{\mathrm{alc}} = 60$ and with z the standard normal quantile. We present the confidence intervals corresponding to the difference above in the bottom row of Figure 4. From the data one cannot conclude that the two groups of patients have different dependence coefficients between the AF and FP regions.

7. Discussion

In this paper, we investigated the possibility to rely on the properties of the 2-Wasserstein distance to define new dependence coefficients that are easy to interpret. We mostly developed the theory under a Gaussian lens, thus moving from the Wasserstein distance between distributions to the Bures–Wasserstein distance between covariance or correlation matrices. Further, we have shown that the coefficients are particularly natural in this case and that they enjoy desirable properties. They can be estimated easily from an empirical covariance or correlation matrix. The asymptotic distributions of the resulting plug-in estimators can be found by the delta method, with explicit expressions for the asymptotic variances, enabling inference. Some questions remain open and are expected to lead to further research.

The plug-in estimators turned out to have a positive bias, which we proposed to correct by eigenvalue shrinkage. Some more developments towards bias correction would certainly be welcome, for instance in the context of the matrix of normal scores rank correlation coefficients for data drawn from a distribution with a Gaussian copula.

The Fréchet-differentiability of the maps that send a covariance matrix to its dependence coefficients paves the way for further developments in large-sample theory. In a high-dimensional setting, the correlation matrix could be estimated using regularisation techniques or exploiting modelling assumptions. In time series analysis, the focus would be on auto-covariance matrices.

A technical challenge is to obtain the limit distribution of the plug-in estimators in case all cross-covariances are zero so that the dependence coefficients are zero. The rate of convergence may then be conjectured to be $\mathrm{O}_p(n^{-1})$ and the limit laws linear combinations of independent chi-squared random variables. Equally interesting is to quantify the impact of the non-linearity of the (Hadamard) derivatives in case of repeated eigenvalues. A further refinement would be to allow for positive semi-definite correlation matrices instead of positive definite ones.

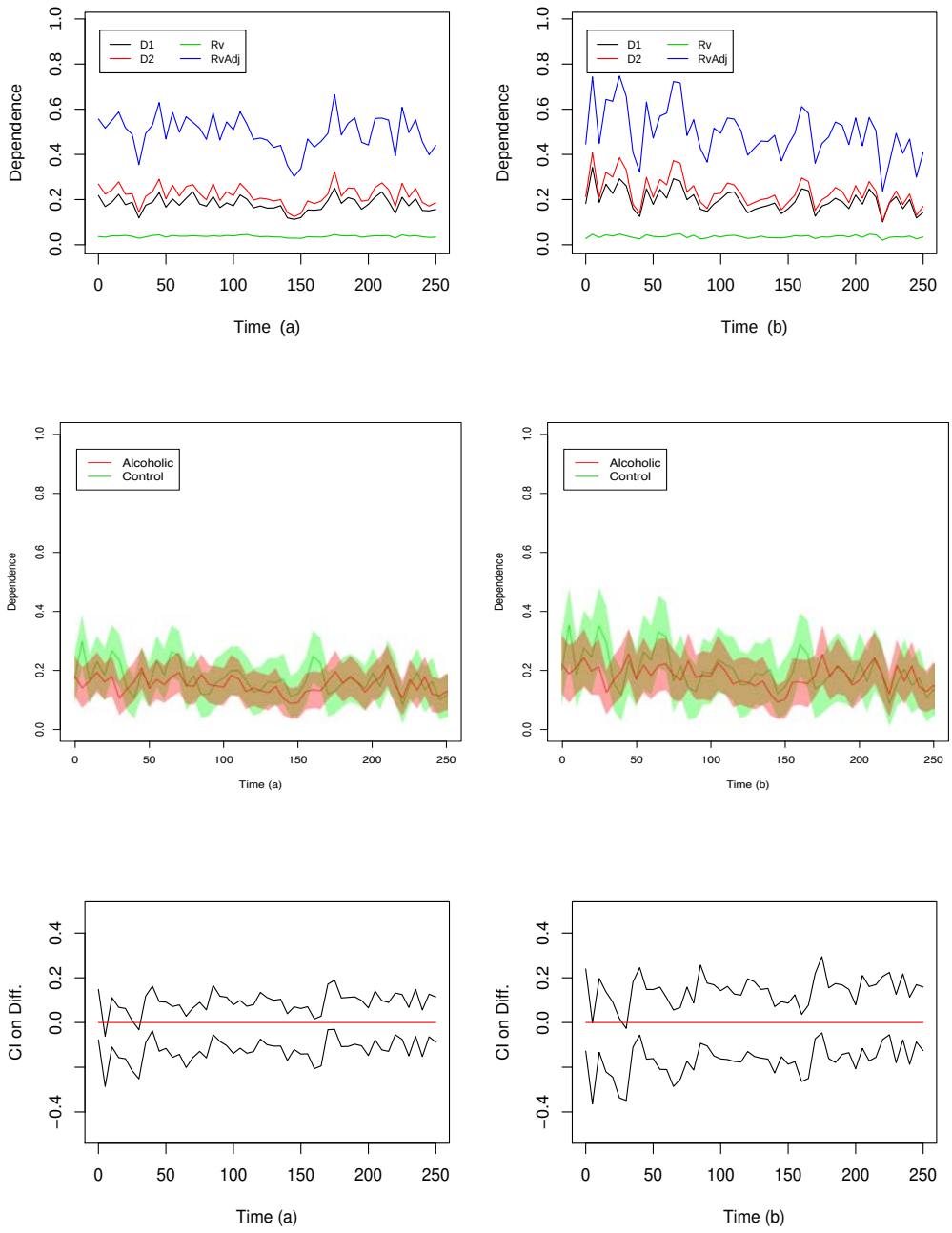

FIGURE 4. *Top: Dependence across time for alcoholics (a) and control patients (b). Middle: Comparison of patients suffering from alcoholism versus control group using dependence coefficients* $\mathfrak{D}_1$ *(a) and* $\mathfrak{D}_2$ *(b). Estimates as solid lines and point-wise 95% confidence bands as coloured shaded areas. Bottom: Asymptotic 95% confidence intervals for the difference* $\mathfrak{D}_r^{\mathrm{ctr}} - \mathfrak{D}_r^{\mathrm{alc}}$ *between the two groups of patients for* $\mathfrak{D}_1$ *(a) and* $\mathfrak{D}_2$ *(b).*

The differentiability questions we referred to are important for resampling. Indeed, the n-out-of-n bootstrap is not consistent when the Fréchet derivative is not linear. A comprehensive and careful analysis of the bootstrap consistency in this case could also be potentially interesting per se.

Finally, one could seek for nonparametric estimators of the distribution-based dependence coefficients $\tilde{\mathfrak{D}}_r$. This will require new probabilistic results to derive their limit laws—or at least guarantee the possibility to approximate their sampling distributions through a numeric scheme—as well as new algorithmic developments to determine the couplings in the maximally dependent case. Identifying the couplings furthest away from a given reference point in Wasserstein space is also an interesting theoretical challenge.

Acknowledgments

The second author gratefully acknowledges funding by FNRS-F.R.S. grant CDR J.0146.19. Enlightening discussions with Professor E. Pircalabelu regarding brain imaging are also gratefully acknowledged.

References

Anuragi, A. and D. S. Sisodia (2020). "Empirical wavelet transform based automated alcoholism detecting using EEG signal features". *Biomedical Signal Processing and Control* 57, p. 101777.

Bhatia, R., T. Jain, and Y. Lim (2019). "On the Bures–Wasserstein distance between positive definite matrices". *Expositiones Mathematicae* 37.2, pp. 165–191.

Dey, G. K. and C. Srinivasan (1985). "Estimation of a covariance matrix under Stein's loss". *The Annals of Statistics* 13.4, pp. 1581–1591.

Donoho, D., M. Gavish, and I. Johnstone (2018). "Optimal shrinkage of eigenvalues in the spiked covariance model". *The Annals of Statistics* 46.4, pp. 1742–1778.

Dowson, D. C. and B. V. Landau (1982). "The Fréchet distance between multivariate normal distributions". *Journal of Multivariate Analysis* 12.3, pp. 450–455.

Dua, D. and C. Graff (2020). *UCI Machine Learning Repository*.

El Maache, H. and Y. Lepage (2003). "Spearman's rho and Kendall's tau for multivariate data sets". *Lecture Notes-Monograph Series*, pp. 113–130.

Escoufier, Y. (1973). "Le traitement des variables vectorielles". *Biometrics* 29, pp. 751–760.

Geenens, G., A. Charpentier, and D. Paindaveine (2017). "Probit transformation for nonparametric kernel estimation of the copula density". *Bernoulli* 23.3, pp. 1848–1873.

Gilliam, D. S., T. Hohage, X. Ji, and F. Ruymgaart (2009). "The Fréchet derivative of an analytic function of a bounded operator with some applications". *International Journal of Mathematics and Mathematical Sciences*, Art. ID 239025, 17.

Grothe, O., J. Schnieders, and J. Segers (2014). "Measuring association and dependence between random vectors". *Journal of Multivariate Analysis* 123, pp. 96–110.

Hájek, J. and Z. Šidák (1967). *Theory of Rank Tests*. Prague: Academia.

Hardy, G. H., J. E. Littlewood, and G. Pólya (1934, 1952). *Inequalities*. 1st, 2nd. Cambridge University Press, London and New York.

Hiriart-Urruty, J.-B. and A. S. Lewis (1999). "The Clarke and Michel-Penot subdifferentials of the eigenvalues of a symmetric matrix". Vol. 13. 1-3. Computational optimization—a tribute to Olvi Mangasarian, Part II, pp. 13–23.

Hofert, M., W. Oldford, A. Prasad, and M. Zhu (2019). "A framework for measuring association of random vectors via collapsed random variables". *Journal of Multivariate Analysis* 172.C, pp. 5–27.

Hotelling, H. (1936). "Relations between two sets of variates". *Biometrika* 28, pp. 321–377.

Klaassen, C. A. J. and J. A. Wellner (1997). "Efficient estimation in the bivariate normal copula model: normal margins are least favourable". *Bernoulli* 3.1, pp. 55–77.

Kollo, T. and D. von Rosen (2006). *Advanced Multivariate Statistics with Matrices*. Vol. 579. Springer Science & Business Media.

Liu, H., J. Lafferty, and L. Wasserman (2009). "The nonparanormal: Semiparametric estimation of high dimensional undirected graphs." *Journal of Machine Learning Research* 10.10.

Marshall, A. W., I. Olkin, and B. C. Arnold (2011). *Inequalities: Theory of Majorization and its Applications*. New York, Springer.

Medovikov, I. and A. Prokhorov (Apr. 2017). "A New Measure of Vector Dependence, with Applications to Financial Risk and Contagion". *Journal of Financial Econometrics* 15.3, pp. 474–503.

Olkin, I. and F. Pukelsheim (1982). "The distance between two random vectors with given dispersion matrices". *Linear Algebra and its Applications* 48, pp. 257–263.

Panaretos, V. and Y. Zemel (2019a). *An Invitation to Statistics in Wasserstein Space*.

– (2019b). "Statistical aspects of Wasserstein distances". *Annual Review of Statistics and Its Application* 6, pp. 405–431.

Petz, D. (2001). "Entropy, von Neumann and the von Neumann entropy". *John von Neumann and the foundations of quantum physics*. Springer, pp. 83–96.

Puccetti, G. (2019). "Measuring Linear Correlation Between Random Vectors". *Available at SSRN 3116066*.

Quessy, J.-F. (2010). "Applications and asymptotic power of marginal-free tests of stochastic vectorial independence". *Journal of Statistical Planning and Inference* 140, pp. 3058–3075.

Rippl, T., A. Munk, and A. Sturm (2016). "Limit laws of the empirical Wasserstein distance: Gaussian distributions". *Journal of Multivariate Analysis* 151, pp. 90–109.

Robert, P. and Y. Escoufier (1976). "A unifying tool for linear multivariate statistical methods: the RV-coefficient". *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 25.3, pp. 257–265.

Solea, E. and B. Li (2020). "Copula Gaussian graphical models for functional data". *Journal of the American Statistical Association*, pp. 1–13.

Székely, G. J., M. L. Rizzo, and N. K. Bakirov (2007). "Measuring and testing dependence by correlation of distances". *The Annals of Statistics* 35.6, pp. 2769–2794.

Thompson, R. C. and S. Therianos (1972). "Inequalities connecting the eigenvalues of a Hermitian matrix with the eigenvalues of complementary principal submatrices". *Bulletin of the Australian Mathematical Society* 6.1, pp. 117–132.

Villani, C. (2008). *Optimal Transport: Old and New*. Vol. 338. Springer Science & Business Media.

Zhang, X. L., H. Begleiter, B. Porjesz, W. Wang, and A. Litke (1995). "Event related potentials during object recognition tasks". *Brain Research Bulletin* 38.6, pp. 531–538.
Zhu, L., K. Xu, R. Li, and W. Zhong (2017). "Projection correlation between two random vectors". *Biometrika* 104.4, pp. 829–843.

Appendix A. Proofs for Section 2

Proof of Lemma 2.1. (i) Let $V = (V_1, V_2)$ be a random vector with law $\upsilon_1 \otimes \upsilon_2$ and let $((X, Y), V)$ be a coupling of (X, Y) and V. Then (X, V_1) and (Y, V_2) are couplings of μ and υ_1 and of ν and υ_2, respectively, and thus

$$\mathbb{E}[\|(X, Y) - V\|^2] = \mathbb{E}[\|X - V_1\|^2] + \mathbb{E}[\|Y - V_2\|^2] \geq W_2^2(\mu, \upsilon_1) + W_2^2(\nu, \upsilon_2). \tag{32}$$

Take the infimum over all couplings $((X, Y), V)$.

(ii) See for instance the beginning of Section 2 in Panaretos and Zemel (2019b).

(iii) If $\mu = \upsilon_1$ and $\nu = \upsilon_2$, then $T_{p,q}(\pi; \upsilon_1, \upsilon_2) = W_2^2(\pi, \mu \otimes \nu)$ and the statement is trivial. Suppose that υ_1 and υ_2 are absolutely continuous. Equality to zero means that there exists an optimal coupling $((X, Y), V)$ of π and $\upsilon_1 \otimes \upsilon_2$ such that the inequality in Eq. (32) is an equality and thus that (X, V_1) and (Y, V_2) are optimal couplings of $\mu \otimes \upsilon_1$ and $\nu \otimes \upsilon_2$ respectively. As υ_1 and υ_2 are absolutely continuous, then, by Brenier's theorem (Villani, 2008, Theorem 2.12), there exist two convex functions $\varphi_1 : \mathbb{R}^p \to \mathbb{R} \cup \{\infty\}$ and $\varphi_2 : \mathbb{R}^q \to \mathbb{R} \cup \{\infty\}$ such that $X = \nabla\varphi_1(V_1)$ and $Y = \nabla\varphi_1(V_1)$ almost surely. Hence, X and Y are independent and their distribution is $\pi = \mu \otimes \nu$. □

Proof of Proposition 2.3. Assertions (i) and (ii) follow in a straightforward way from Lemma 2.1.

Assertion (iii) follows from the invariance of the 2-Wasserstein distance and the multivariate standard Gaussian distribution with respect to orthogonal transformations. For instance, for any orthogonal transformation O of $\mathbb{R}^p$ we have $W_2^2(\mu \circ O^{-1}, \gamma_p) = W_2^2(\mu \circ O^{-1}, \gamma_p \circ O^{-1}) = W_2^2(\mu, \gamma_p)$. □

Appendix B. Proofs for Section 3

Proof of Proposition 3.3. Assertion (i) follows from Assertion (i) in Proposition 2.3 upon identifying d_W^2 with the squared Wasserstein distance between centered Gaussian distributions as in (5). Assertion (ii) is a consequence of continuity of d_W and the fact that the set $\Gamma(\Sigma_1, \Sigma_2)$ is compact. Assertion (iii), finally, follows from the invariance of d_W with respect to orthogonal transformations. □

Proof of Theorem 3.6. We need to show two things: first, the eigenvalues of Σ_m are as in Eq. (11) (which implies that Σ_m is positive semi-definite) and second, the eigenvalues of any other Σ of the form (6) are majorized by those of Σ_m. For ease of writing, we assume that $p \leq q$; otherwise, switch the roles of the two parts in the partition. The matrix Π then becomes

$$\Pi = \begin{bmatrix} I_p & 0_{p\times(q-p)} \end{bmatrix} \in \mathbb{R}^{p\times q}.$$

First, since $\begin{bmatrix} U_1 & 0 \\ 0 & U_2 \end{bmatrix}$ is orthogonal, the eigenvalues of Σ_m are the same as those of

$$\Lambda_m = \begin{bmatrix} \Lambda_2 & \Lambda_2^{1/2}\Pi\Lambda_2^{1/2} \\ \Lambda_2^{1/2}\Pi^\top\Lambda_2^{1/2} & \Lambda_2 \end{bmatrix}.$$

The eigenvalues and eigenvectors of Λ_m can be found explicitly. For integer $1 \le r \le s$, let $e_{r,s}$ be the r-th canonical unit vector in $\mathbb{R}^s$. Then:

- For $j = 1, \ldots, p$, the vector $(\lambda_{j,1}^{1/2} e_{j,p}^\top, \lambda_{j,2}^{1/2} e_{j,q}^\top)^\top$ is an eigenvector of Λ_m with eigenvalue $\lambda_{j,1} + \lambda_{j,2}$.
- For $j = 1, \ldots, p$, the vector $(\lambda_{j,2}^{1/2} e_{j,p}^\top, -\lambda_{j,1}^{1/2} e_{j,q}^\top)^\top$ is an eigenvector of Λ_m with eigenvalue 0.
- For $j = p+1, \ldots, q$, the vector $(0^\top, e_{j,q}^\top)^\top$ is an eigenvector of Λ_m with eigenvalue $\lambda_{j,2}$.

Second, let $\lambda_1 \ge \ldots \ge \lambda_d \ge 0$ be the eigenvalues of Σ. We need to show that

$$\sum_{j=1}^{k} \lambda_j \le \sum_{j=1}^{k} (\lambda_{j,1} + \lambda_{j,2}), \qquad k = 1, \ldots, p,$$

$$\sum_{j=1}^{k} \lambda_j \le p + \sum_{j=1}^{k} \lambda_{j,2}, \qquad k = p+1, \ldots, q.$$

By Theorem 1 in Thompson and Therianos (1972), we have, for any choice of integers

$$1 \le i_1 < \ldots < i_\mu \le p, \qquad 1 \le j_1 < \ldots < j_\nu \le q$$

that

$$\sum_{s=1}^{\mu+\nu} \lambda_{i_s + j_s - s} \le \sum_{s=1}^{\mu} \lambda_{i_s,1} + \sum_{s=1}^{\nu} \lambda_{j_s,2},$$

where $i_s = p - \mu + s$ for $s > \mu$ and $j_s = q - \nu + s$ for $s > \nu$. Now:

- For $k = 1, \ldots, p$, set $\mu = \nu = k$ and $i_s = j_s = s$ to find the first inequality to be proved.
- For $k = p+1, \ldots, q$, set $\mu = p$ with $i_s = s$ for $s = 1, \ldots, p$ and set $\nu = k$ with $j_s = s$ for $s = 1, \ldots, q$ to find the second inequality to be proved. □

Proof of Proposition 3.9. (i) Recall that $\lambda_1 \ge \ldots \ge \lambda_d \ge 0$ are the eigenvalues of Σ. By Eq. (5), we have

$$W_2^2\big(\mathcal{N}_d(0, \Sigma), \mathcal{N}_d(0, I_d)\big) = d + \operatorname{tr}(\Sigma) - 2\operatorname{tr}(\Sigma^{1/2}) = d + \sum_{j=1}^{d} \lambda_j - 2\sum_{j=1}^{d} \lambda_j^{1/2}.$$

Since the function $\lambda \mapsto \lambda - 2\lambda^{1/2}$ is convex on $\lambda \in \mathbb{R}_{\ge}$, the claim of maximality follows from Proposition 3.5 and Theorem 3.6.

(ii) We have

$$\Sigma_0^{1/2}\Sigma\Sigma_0^{1/2} = \begin{bmatrix} \Sigma_1^{1/2} & 0 \\ 0 & \Sigma_2^{1/2} \end{bmatrix} \begin{bmatrix} \Sigma_1 & \Psi \\ \Psi^\top & \Sigma_2 \end{bmatrix} \begin{bmatrix} \Sigma_1^{1/2} & 0 \\ 0 & \Sigma_2^{1/2} \end{bmatrix} = \begin{bmatrix} \Sigma_1^2 & \Sigma_1^{1/2}\Psi\Sigma_2^{1/2} \\ \Sigma_2^{1/2}\Psi^\top\Sigma_1^{1/2} & \Sigma_2^2 \end{bmatrix}.$$

Recall the eigendecomposition (8) of Σ_j. For $r \in \{1, 2\}$ and for $\alpha > 0$, the eigendecomposition of Σ_r^a is $U_r\Lambda_r^\alpha U_r$, i.e., the eigenvectors are the same as those of Σ_r while the eigenvalues are raised to the exponent α. For Ψ_m as in Eq. (10), we get

$$\begin{aligned}\Sigma_1^{1/2}\Psi_m\Sigma_2^{1/2} &= \left(U_1\Lambda_2^{1/2}U_1^\top\right)\left(U_1\Lambda_2^{1/2}\Pi\Lambda_2^{1/2}U_2^\top\right)\left(U_1\Lambda_2^{1/2}U_1^\top\right) \\ &= U_1\Lambda_2\Pi\Lambda_2U_2.\end{aligned}$$

The latter matrix is of the same form as Ψ_m in Eq. (10) but with Λ_r replaced by Λ_r^2. By Theorem 3.6 with Σ_r replaced by Σ_r^2 for $r \in \{1, 2\}$, it follows that of all positive semidefinite $d \times d$ matrices with diagonal blocks Σ_1^2 and Σ_2^2, the eigenvalues are majorised by those of the matrix $\Sigma_0^{1/2}\Sigma_m\Sigma_0^{1/2}$. In view of Eq. (5), we have

$$W_2^2\big(\mathcal{N}_d(0,\Sigma), \mathcal{N}_d(0,\Sigma_0)\big) = 2\operatorname{tr}\Sigma - 2\operatorname{tr}\big\{(\Sigma_0^{1/2}\Sigma\Sigma_0^{1/2})^{1/2}\big\} = 2\operatorname{tr}\Sigma - 2\sum_{j=1}^d \kappa_j^{1/2}$$

with $\kappa_1, \ldots, \kappa_d$ the eigenvalues of $\Sigma_0^{1/2}\Sigma\Sigma_0^{1/2}$, counting multiplicities. The function $\kappa \mapsto -\kappa^{1/2}$ being convex on $\kappa \in \mathbb{R}_\geq$, the maximality follows from Proposition 3.5 and Theorem 3.6.

(iii) For any rectangular matrix A, we have $\operatorname{tr}(AA^\top) = \sum_i \sum_j A_{ij}^2$. It follows that

$$\operatorname{tr}(\Sigma^2) = \operatorname{tr}(\Sigma_1^2) + \operatorname{tr}(\Sigma_2^2) + 2\operatorname{tr}(\Psi\Psi^\top).$$

Given the diagonal blocks Σ_1 and Σ_2, maximising $\operatorname{tr}(\Psi\Psi^\top)$ is thus equivalent to maximising $\operatorname{tr}(\Sigma^2)$. As the function $\lambda \mapsto \lambda^2$ is convex, Proposition 3.5 and Theorem 3.6 imply that $\operatorname{tr}(\Sigma^2) = \sum_{j=1}^d \lambda_j^2$ is maximal for Σ equal to Σ_m. □

Proof of Proposition 3.10. First we calculate $\mathfrak{D}_1(\Sigma)$. By Eq. (5), we have

$$W_2^2\big(\mathcal{N}_d(0,\Sigma), \mathcal{N}_d(0,I_d)\big) = d + \operatorname{tr}(\Sigma) - 2\operatorname{tr}(\Sigma^{1/2}).$$

Apply this result to the three terms in the numerator of $\mathfrak{D}_1(\Sigma)$ and use the content of Theorem 3.6 for the denominator. The claim about $\mathfrak{D}_1(\Sigma)$ follows from straightforward simplifications, using $d = p + q$ and $\operatorname{tr}(\Sigma) = \operatorname{tr}(\Sigma_m) = \operatorname{tr}(\Sigma_0) = \operatorname{tr}(\Sigma_1) + \operatorname{tr}(\Sigma_2)$.

The value of $\mathfrak{D}_2(\Sigma)$ is obtained in a similar way. □

Formulas for dependence coefficients in parametric models. We now present some closed-form formulas for some of the coefficients presented in the examples in Section 3.3. In Example 3.15, as the eigenvalues of Σ are $1 + 2\rho$, $1 - \rho$ and $1 - \rho$, we get, after some simplifications,

$$\mathfrak{D}_1(\Sigma) = \frac{1 + \sqrt{1+\rho} - \sqrt{1+2\rho} - \sqrt{1-\rho}}{1 + \sqrt{1+|\rho|} - \sqrt{2+|\rho|}}.$$

For the second coefficient, a more involved calculation yields

$$\mathfrak{D}_2(\Sigma) = \frac{2+\rho-\sqrt{\lambda_+(\rho)}-\sqrt{\lambda_-(\rho)}}{2+|\rho|-\sqrt{\rho^2+2\,|\rho|+2}},$$

with $\lambda_\pm(\rho) = \frac{1}{2}[\rho^2+2\rho+2 \pm \rho\sqrt{\rho^2+12\rho+12}]$. The RV coefficient and its adjusted version in (12) are

$$\mathrm{RV}(\Sigma) = \frac{2\rho^2}{\sqrt{2(1+\rho^2)}}, \qquad \overline{\mathrm{RV}}(\Sigma) = \frac{2\rho^2}{1+|\rho|}.$$

In Example 3.16, for the trivariate autoregressive matrix, one has

$$\mathrm{RV}(\Sigma,1) = \frac{\rho^4+\rho^2}{\sqrt{2(1+\rho^2)}} \quad \text{and} \quad \overline{\mathrm{RV}}(\Sigma,1) = \frac{\rho^4+\rho^2}{1+|\rho|},$$

while

$$\mathfrak{D}_1(\Sigma,1) = \frac{1+\sqrt{1+\rho}+\sqrt{1-\rho}-\sqrt{1-\rho^2}-\sqrt{\lambda_{1,+}(\rho)}-\sqrt{\lambda_{1,-}(\rho)}}{1+\sqrt{1+|\rho|}-\sqrt{2+|\rho|}}$$

and

$$\mathfrak{D}_2(\Sigma,1) = \frac{3-\sqrt{1-\rho^2}-\sqrt{\lambda_{2,+}(\rho)}-\sqrt{(}\lambda_{2,-}(\rho)}{2+|\rho|-\sqrt{2+2|\rho|+\rho^2}}$$

with $\lambda_{1,\pm}(\rho) = \rho^2/2 \pm \rho\sqrt{\rho^2+8}/2+1$ and $\lambda_{1,\pm}(\rho) = 3\rho^2/2 \pm (\sqrt{5}\rho\sqrt{\rho^2+4})/2+1$. For the trivariate moving average matrix, it holds that

$$\mathrm{RV}(\Sigma,1) = \frac{\rho^2}{\sqrt{2(1+\rho^2)}} \quad \text{and} \quad \overline{\mathrm{RV}}(\Sigma,1) = \frac{\rho^2}{1+|\rho|},$$

while

$$\mathfrak{D}_1(\Sigma,1) = \frac{\sqrt{1+\rho}+\sqrt{1-\rho}-\sqrt{1+\rho\sqrt{2}}-\sqrt{1-\rho\sqrt{2}}}{1+\sqrt{1+|\rho|}-\sqrt{2+|\rho|}}.$$

In this case, the formula for $\mathfrak{D}_2$ is not particularly convenient and the eigendecomposition was obtained numerically.

Appendix C. Proofs for Section 4

Recall that $\mathbb{S}^d$ denotes the set of real symmetric $d \times d$ matrices and $\mathbb{S}^d_> \subset \mathbb{S}^d$ the subset of positive definite such matrices.

Lemma C.1. *Let $B \in \mathbb{S}^d_>$ and let $H_t, H \in \mathbb{S}^d$ for $t > 0$ be such that $H_t \to H$ element-wise as $t \downarrow 0$. Then*

$$\lim_{t\downarrow 0} t^{-1}\big((B+tH_t)^{1/2} - B^{1/2}\big) = X, \tag{33}$$

where $X \in \mathbb{S}^d$ is the solution to the Sylvester equation

$$B^{1/2}X + XB^{1/2} = H.$$

Moreover,

$$\lim_{t\downarrow 0} t^{-1}\left(\mathrm{tr}\big((B+tH_t)^{1/2}\big) - \mathrm{tr}\big(B^{1/2}\big)\right) = \mathrm{tr}(X) = \tfrac{1}{2}\,\mathrm{tr}(B^{-1/2}H). \tag{34}$$

In the sequel, we will also use the notation $\psi : \mathbb{S}^d_> \to \mathbb{S}^d_> : B \mapsto B^{1/2}$ and denote the Fréchet derivative of the latter map at B evaluated in G by $D\psi_B(G)$.

Proof. The existence of the limit (33) follows from the fact that function $z \mapsto z^{1/2}$ is analytic on the positive part of the complex plane and the fact that B has positive eigenvalues. Squaring both sides of the expansion $(B + tH_t)^{1/2} = B^{1/2} + tX + \mathrm{o}(t)$ as $t \downarrow 0$ yields

$$B + tH_t = \big(B^{1/2} + tX + \mathrm{o}(t)\big)^2 = B + t(B^{1/2}X + XB^{1/2}) + \mathrm{o}(t), \qquad t \downarrow 0.$$

Examining the terms linear in t yields the stated Sylvester equation (33). In that equation, premultiply both sides with $B^{-1/2}$ and take the trace to see that

$$\mathrm{tr}(X) + \mathrm{tr}(B^{-1/2}XB^{1/2}) = \mathrm{tr}(B^{-1/2}H).$$

But $\mathrm{tr}(B^{-1/2}XB^{1/2}) = \mathrm{tr}(XB^{1/2}B^{-1/2}) = \mathrm{tr}(X)$ and thus

$$\mathrm{tr}(X) = \tfrac{1}{2}\,\mathrm{tr}(B^{-1/2}H). \qquad \square$$

For $A \in \mathbb{S}^d$, let $L(A) \in \mathbb{S}^d$ be the diagonal matrix whose diagonal is equal to the d eigenvalues (counting multiplicities) of A in decreasing order.

Lemma C.2. *Let $A \in \mathbb{S}^d$ have d distinct (real) eigenvalues and let the orthogonal matrix $U \in \mathbb{R}^{d\times d}$ contain the associated eigenvectors as columns. Let $H_t, H \in \mathbb{S}^d$ for $t > 0$ be such that $H_t \to H$ element-wise as $t \downarrow 0$. Then*

$$\lim_{t\downarrow 0} t^{-1}\big(L(A + tH_t) - L(A)\big) = D_{U^\top HU} =: \dot{L}_A(H).$$

Proof. This is a special case of Theorem 3.3 in Hiriart-Urruty and Lewis (1999). $\square$

Lemma C.3. *Under the conditions of Theorem 4.1, it holds that*

$$\lim_{t\downarrow 0} t^{-1}\Big(\mathrm{tr}\big((\Sigma + tH_t)^{1/2}_m\big) - \mathrm{tr}(\Sigma^{1/2}_m)\Big) = \frac{1}{2}\,\mathrm{tr}(\Upsilon_1 H),$$

with $(\Sigma + tH_t)_m$ the matrix in (9) for Σ replaced by $\Sigma + tH_t$ and with Υ_1 defined in (20).

Proof. The diagonal elements of the diagonal matrix $L(\Sigma_r) = \Lambda_r$ are $\lambda_{1,1} \geq \ldots \geq \lambda_{p,1}$ for $r = 1$ and $\lambda_{1,2} \geq \ldots \geq \lambda_{q,2}$ for $r = 2$. We need to deal with the term

$$\mathrm{tr}(\Sigma^{1/2}_m) = \sum_{j=1}^{q} (\lambda_{j,1} + \lambda_{j,2})^{1/2} =: g(\Lambda_1, \Lambda_2), \tag{35}$$

where $\lambda_{j,1} = 0$ if $j \in \{p+1, \ldots, q\}$ (recall $q \geq p$). Similarly,

$$\mathrm{tr}\big((\Sigma + tH_t)^{1/2}_m\big) = g\big(L(\Sigma_1 + tH_{t,11}), L(\Sigma_2 + tH_{t,22})\big),$$

where $H_{t,11}$ and $H_{t,22}$ are the upper $p \times p$ and lower $q \times q$ diagonal blocks of H_t. In view of Lemma C.2 and the differentiability of g in (35), the chain rule gives

$$\begin{aligned}\lim_{t\downarrow 0} t^{-1}\,&\big(g\big(L(\Sigma_1 + tH_{t,11}), L(\Sigma_2 + tH_{t,22})\big) - g(\Lambda_1, \Lambda_2)\big) \\ &= \sum_{j=1}^{p} \frac{1}{2(\lambda_{j,1} + \lambda_{j,2})^{1/2}} [U_1^\top H_{11} U_1]_{jj} + \sum_{j=1}^{q} \frac{1}{2(\lambda_{j,1} + \lambda_{j,2})^{1/2}} [U_2^\top H_{22} U_2]_{jj},\end{aligned}$$

where H_{11} and H_{22} are the upper $p \times p$ and lower $q \times q$ diagonal blocks of H. The right-hand side can be simplified as follows: with Π_1 and Π_2 as in (17),

$$\begin{aligned}\ldots &\overset{\text{(a)}}{=} \frac{1}{2}\Big(\operatorname{tr}(\Delta_1 U_1^\top H_{11} U_1) + \operatorname{tr}(\Delta_2 U_2^\top H_{22} U_2)\Big) \\ &\overset{\text{(b)}}{=} \frac{1}{2}\Big(\operatorname{tr}(\Pi_1^\top U_1 \Delta_1 U_1^\top \Pi_1 H) + \operatorname{tr}(\Pi_2^\top U_2 \Delta_2 U_2^\top \Pi_2 H)\Big) \\ &\overset{\text{(c)}}{=} \frac{1}{2}\operatorname{tr}\left(\begin{bmatrix} U_1\Delta_1 U_1^\top & 0 \\ 0 & U_2 \Delta_2 U_2^\top \end{bmatrix} H\right) = \frac{1}{2}\operatorname{tr}(\Upsilon_1 H),\end{aligned}$$

using the following arguments:

(a) by the identity $\operatorname{tr}(A \operatorname{diag}(B)) = \operatorname{tr}(\operatorname{diag}(A)B)$ for square matrices A and B;
(b) by the cyclic permutation property of the trace operator together with $H_{rr} = \Pi_r H \Pi_r^\top$ for $r \in \{1, 2\}$;
(c) by the identity $\Pi_1^\top A_1 \Pi_1 + \Pi_2^\top A_2 \Pi_2 = \begin{bmatrix} A_1 & 0 \\ 0 & A_2 \end{bmatrix}$ for matrices A_1 and A_2 of dimensions $p \times p$ and $q \times q$, respectively. □

Proof of Theorem 4.1. Note that for t close enough to zero, $\Sigma + tH_t$ is positive definite since Σ is so and since $\Sigma + tH_t \to \Sigma$ element-wise as $t \downarrow 0$. Consider the function

$$f(\bar{x}, \bar{y}, \bar{z}, \bar{w}) = \frac{\bar{y} + \bar{z} - \bar{x}}{\bar{y} + \bar{z} - \bar{w}}.$$

We have $\mathfrak{D}_1(\Sigma) = f(x, y, z, w)$ and $\mathfrak{D}_1(\Sigma + tH_t) = f(x_t, y_t, z_t, w_t)$ where

$$\begin{aligned} x &= \operatorname{tr}(\Sigma^{1/2}), & y &= \operatorname{tr}(\Sigma_1^{1/2}), \\ z &= \operatorname{tr}(\Sigma_2^{1/2}), & w &= \operatorname{tr}(\Sigma_m^{1/2}), \end{aligned}$$

and similarly

$$\begin{aligned} x_t &= \operatorname{tr}\big((\Sigma + tH_t)^{1/2}\big), & y_t &= \operatorname{tr}\big((\Sigma + tH_t)_1^{1/2}\big), \\ z_t &= \operatorname{tr}\big((\Sigma + tH_t)_2^{1/2}\big), & w_t &= \operatorname{tr}\big((\Sigma + tH_t)_m^{1/2}\big). \end{aligned}$$

Here, $(\Sigma + tH_t)_1$ and $(\Sigma + tH_t)_2$ are the upper $p \times p$ and lower $q \times q$ diagonal blocks of $\Sigma + tH_t$, respectively, while $(\Sigma + tH_t)_m$ is the matrix in (9) with Σ replaced by $\Sigma + tH_t$.

Provided the quantities $(x_t - x)/t$ and so on converge, we have

$$\frac{\mathfrak{D}_1(\Sigma + tH_t) - \mathfrak{D}_1(\Sigma)}{t} = \dot{f}_x \frac{x_t - x}{t} + \dot{f}_y \frac{y_t - y}{t} + \dot{f}_z \frac{z_t - z}{t} + \dot{f}_w \frac{w_t - w}{t} + \mathrm{o}(1), \qquad t \downarrow 0,$$

where $\dot{f}_x$ and so on are the partial derivatives of f evaluated at (x, y, z, w). Using the notation $c_1 = y + z - w$, straightforward computation gives

$$\dot{f}_x = -\frac{1}{c_1}, \qquad \dot{f}_y = \dot{f}_z = \frac{1 - \mathfrak{D}_1(\Sigma)}{c_1}, \qquad \dot{f}_w = \frac{\mathfrak{D}_1(\Sigma)}{c_1}.$$

It follows that, as $t \downarrow 0$ and provided $(x_t - x)/t$ and so on converge,

$$\frac{\mathfrak{D}_1(\Sigma + tH_t) - \mathfrak{D}_1(\Sigma)}{t} = \frac{1}{c_1}\left(-\frac{x_t - x}{t} + (1 - \mathfrak{D}_1(\Sigma))\frac{y_t - y + z_t - z}{t} + \mathfrak{D}_1(\Sigma)\frac{w_t - w}{t}\right) + \mathrm{o}(1). \quad (36)$$

Let H_{11} and H_{22} be the upper $p \times p$ and lower $q \times q$ diagonal blocks of H. By (34),

$$\lim_{t \downarrow 0} \frac{x_t - x}{t} = \frac{1}{2}\operatorname{tr}(\Sigma^{-1/2}H), \quad (37)$$

as well as

$$\begin{aligned}\lim_{t \downarrow 0} \frac{y_t - y + z_t - z}{t} &= \frac{1}{2}\operatorname{tr}(\Sigma_1^{-1/2}H_{11}) + \frac{1}{2}\operatorname{tr}(\Sigma_2^{-1/2}H_{22}) \\ &= \frac{1}{2}\operatorname{tr}\left(\begin{bmatrix}\Sigma_1^{-1/2} & 0 \\ 0 & \Sigma_2^{-1/2}\end{bmatrix} H\right) = \frac{1}{2}\operatorname{tr}(\Sigma_0^{-1/2}H). \quad (38)\end{aligned}$$

Lemma C.3 further yields

$$\lim_{t \downarrow 0} \frac{z_t - z}{t} = \frac{1}{2}\operatorname{tr}(\Upsilon_1 H). \quad (39)$$

Combine equations (36), (37), (38) and (39) to see that

$$\lim_{t \downarrow 0} \frac{\mathfrak{D}_1(\Sigma + tH_t) - \mathfrak{D}_1(\Sigma)}{t} = \frac{1}{2c_1}\left(-\operatorname{tr}(\Sigma^{-1/2}H) + (1 - \mathfrak{D}_1(\Sigma))\operatorname{tr}(\Sigma_0^{-1/2}H) + \mathfrak{D}_1(\Sigma)\operatorname{tr}(\Upsilon_1 H)\right).$$

The claim follows by the linearity of the trace operator followed by isolating H. □

The following lemma provides the Fréchet derivative of the squared 2-Wasserstein distance (5) between Gaussian distributions. As explained in Remark 4.4, it rectifies the formula in Lemma 2.4 in Rippl, Munk, and Sturm (2016).

Lemma C.4 (Differentiability of the Bures–Wasserstein distance)**.** *The Fréchet derivative of the map*

$$\phi : (\mathbb{S}^d_>)^2 \to \mathbb{R} : (A, B) \mapsto 2\operatorname{tr}\big((A^{1/2}BA^{1/2})^{1/2}\big)$$

at $(A, B) \in (\mathbb{S}^d_>)^2$ *evaluated at* $(G, H) \in (\mathbb{S}^d)^2$ *is*

$$\lim_{t \downarrow 0} t^{-1}\big(\phi(A + tG_t, B + tH_t) - \phi(A, B)\big) = \operatorname{tr}(JG) + \operatorname{tr}(J^{-1}H) =: D\phi_{(A,B)}(G, H) \quad (40)$$

where $G_t, H_t \in \mathbb{S}^d$ *for* $t > 0$ *are such that* $G_t \to G$ *and* $H_t \to H$ *element-wise as* $t \downarrow 0$ *and where*

$$\begin{aligned} J &= A^{-1/2}(A^{1/2}BA^{1/2})^{1/2}A^{-1/2} = B^{1/2}(B^{1/2}AB^{1/2})^{-1/2}B^{1/2}, \\ J^{-1} &= A^{1/2}(A^{1/2}BA^{1/2})^{-1/2}A^{1/2} = B^{-1/2}(B^{1/2}AB^{1/2})^{1/2}B^{-1/2}. \end{aligned} \quad (41)$$

As a consequence, the Fréchet derivative of the squared Bures–Wasserstein distance is

$$\lim_{t \downarrow 0} t^{-1}\big(d_W^2(A + tG_t, B + tH_t) - d_W^2(A, B)\big) = \operatorname{tr}\big((I_d - J)G\big) + \operatorname{tr}\big((I_d - J^{-1})H\big). \quad (42)$$

The matrices J and J^{-1} in (41) are the unique solutions in $\mathbb{S}^d_>$ to the matrix equations $JAJ = B$ and $J^{-1}BJ^{-1} = A$. They operationalize the optimal couplings between $\mathcal{N}_d(0, A)$ and $\mathcal{N}_d(0, B)$ with the squared Euclidean distance as cost function (Olkin and Pukelsheim, 1982).

Proof. Equation (42) is an immediate consequence of (40) and the linearity of the trace operator. So it suffices to show (40).

We start by showing the two identities following the definitions of J and J^{-1}. A direct calculation gives

$$\left(A^{1/2}B^{1/2}(B^{1/2}AB^{1/2})^{-1/2}B^{1/2}A^{1/2}\right)^2 = A^{1/2}BA^{1/2}.$$

Since the left-hand side is the square of a symmetric matrix, we find

$$A^{1/2}B^{1/2}(B^{1/2}AB^{1/2})^{-1/2}B^{1/2}A^{1/2} = (A^{1/2}BA^{1/2})^{1/2}. \tag{43}$$

Pre- and post-multiply with $A^{1/2}$ to find

$$B^{1/2}(B^{1/2}AB^{1/2})^{-1/2}B^{1/2} = A^{-1/2}(A^{1/2}BA^{1/2})^{1/2}A^{-1/2},$$

which is the identity following the definition of J. The identity following the definition of J^{-1} follows in the same way, by changing the roles of A and B. Note that, by (43) and the cyclic permatution property of the trace operator,

$$\begin{aligned}\phi(A, B) &= 2\,\mathrm{tr}\big((A^{1/2}BA^{1/2})^{1/2}\big)\\ &= 2\,\mathrm{tr}\big(A^{1/2}B^{1/2}(B^{1/2}AB^{1/2})^{-1/2}B^{1/2}A^{1/2}\big)\\ &= 2\,\mathrm{tr}\big((B^{1/2}AB^{1/2})^{1/2}\big) = \phi(B, A),\end{aligned}$$

confirming the symmetry of ϕ.

By Lemma C.1, we have, as $t \downarrow 0$,

$$\begin{aligned}&(A + tG_t)^{1/2}(B + tH_t)(A + tG_t)^{1/2}\\ &= \big(A^{1/2} + tD\psi_A(G) + \mathrm{o}(t)\big)\big(B + tH + \mathrm{o}(t)\big)\big(A^{1/2} + tD\psi_A(G) + \mathrm{o}(t)\big)\\ &= A^{1/2}BA^{1/2} + t\big(D\psi_A(G)BA^{1/2} + A^{1/2}HA^{1/2} + A^{1/2}BD\psi_A(G)\big) + \mathrm{o}(t).\end{aligned}$$

In Eq. (34), we have calculated the Fréchet derivative of the map $\mathbb{S}^d_> \to \mathbb{R} : C \mapsto 2\,\mathrm{tr}(C^{1/2})$ to be the linear operator $\mathbb{S}^d \to \mathbb{R} : K \mapsto \mathrm{tr}(C^{-1/2}K)$. Therefore,

$$\begin{aligned}&D\phi_{(A,B)}(G, H)\\ &\qquad= \mathrm{tr}\left((A^{1/2}BA^{1/2})^{-1/2}\big(D\psi_A(G)BA^{1/2} + A^{1/2}HA^{1/2} + A^{1/2}BD\psi_A(G)\big)\right).\end{aligned}$$

Isolating the term involving H, we find $\mathrm{tr}(J^{-1}H)$, as required. It remains to deal with the terms involving G. By symmetry of ϕ, we have $D\phi_{(A,B)}(G, H) = D\phi_{(B,A)}(H, G)$. The terms involving G must therefore simplify to become the term involving H but with the roles of A and B reversed: this transformation leads from J^{-1} to J. □

For a $d \times d$ matrix A partitioned into blocks

$$A = \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

of dimensions $p \times p$, $p \times q$, $q \times p$ and $q \times q$, respectively, we put

$$A_0 = \begin{bmatrix} A_{11} & 0 \\ 0 & A_{22} \end{bmatrix}, \tag{44}$$

with zero off-diagonal blocks. This notation is coherent with the one used for Σ_0 in (7) and for J_0 in (22).

Corollary C.5. *The Fréchet derivative of the map*

$$\eta : \mathbb{S}_>^d \to \mathbb{R} : \Sigma \mapsto \operatorname{tr}\left(\left(\Sigma_0^{1/2}\Sigma\Sigma_0^{1/2}\right)^{1/2}\right)$$

is given by

$$\lim_{t\downarrow 0} t^{-1}\big(\eta(\Sigma + tH_t) - \eta(\Sigma)\big) = \tfrac{1}{2}\operatorname{tr}\big((J_0 + J^{-1})H\big)$$

for $H_t, H \in \mathbb{S}^d$ such that $H_t \to H$ element-wise as $t \downarrow 0$, with J and J_0 as in (21) *and* (22)*, respectively.*

Proof. We apply Lemma C.4 with $A = \Sigma_0$, $B = \Sigma$, and, following the convention in (44), $G_t = (H_t)_0$ as well as $G = H_0$ obtained from H_t and H, respectively. The limit is equal to $\frac{1}{2}(\operatorname{tr}(JH_0) + \operatorname{tr}(J^{-1}H))$ with J as in (21). Now $\operatorname{tr}(JH_0) = \operatorname{tr}(J_0H)$ in view of (15). □

It remains to treat the last term in the denominator in the expression for $\mathfrak{D}_2(\Sigma)$ in Proposition 3.10. This is not particularly involved in the light of the earlier developments.

Lemma C.6. *Under the conditions of Theorem 4.2, it holds that*

$$\lim_{t\downarrow 0} t^{-1} \operatorname{tr}\left(\left((\Sigma + tH_t)_0^{1/2}(\Sigma + tH_t)_m(\Sigma + tH_t)_0^{1/2}\right)^{1/2} - \left(\Sigma_0^{1/2}\Sigma_m\Sigma_0^{1/2}\right)^{1/2}\right) = \operatorname{tr}(\Upsilon_2 H),$$

with $(\Sigma + tH_t)_0$ as in (44)*, with $(\Sigma + tH_t)_m$ the matrix in* (9) *for Σ replaced by $\Sigma + tH_t$, and with Υ_2 defined in* (23)*.*

Proof of Lemma C.6. The proof is similar to the one of Lemma C.3, exploiting the eigenvalue map L in Lemma C.2. Recall from Proposition 3.10 that the trace of interest can be written as

$$\operatorname{tr}\left(\left(\Sigma_0^{1/2}\Sigma_m\Sigma_0^{1/2}\right)^{1/2}\right) = \sum_{j=1}^{p\vee q}(\lambda_{j,1}^2 + \lambda_{j,2}^2)^{1/2} =: h(\Lambda_1, \Lambda_2).$$

This expression is similar to the one for $\operatorname{tr}(\Sigma_m^{1/2})$ in (35), so that one can see, using the same arguments and the same notation, that

$$\begin{aligned}
&\lim_{t\downarrow 0} t^{-1}\Big(h\big(L(\Sigma_1 + tH_{11}), L(\Sigma_2 + tH_{22})\big) - h(\Lambda_1, \Lambda_2)\Big) \\
&= \sum_{j=1}^{p} \frac{\lambda_{j,1}}{(\lambda_{j,1}^2 + \lambda_{j,2}^2)^{1/2}}[U_1^\top H_{11} U_1]_{jj} + \sum_{j=1}^{q} \frac{\lambda_{j,2}}{(\lambda_{j,1}^2 + \lambda_{j,2}^2)^{1/2}}[U_2^\top H_{22} U_2]_{jj} \\
&= \operatorname{tr}(\Delta_1' U_1^\top H_{11} U_1) + \operatorname{tr}(\Delta_2' U_2^\top H_{22} U_2) \\
&= \operatorname{tr}(\Upsilon_2 H).
\end{aligned}$$

□

Proof of Theorem 4.2. The proof is similar to the one of Theorem 4.1. Writing $f(\bar{x}, \bar{y}, \bar{z}) = (\bar{z} - \bar{x})/(\bar{z} - \bar{y})$, we have

$$\mathfrak{D}_2(\Sigma) = f(x, y, z), \qquad \mathfrak{D}_2(\Sigma + tH_t) = f(x_t, y_t, z_t)$$

where

$$x = \operatorname{tr}\left((\Sigma_0^{1/2}\Sigma\Sigma_0^{1/2})^{1/2}\right), \qquad y = \operatorname{tr}\left((\Sigma_0^{1/2}\Sigma_m\Sigma_0^{1/2})^{1/2}\right), \qquad z = \operatorname{tr}(\Sigma),$$

and similarly for x_t, y_t, z_t, with Σ replaced by $\Sigma + tH_t$. If we can show that the three expressions $(x_t - x)/t$, $(y_t - y)/t$ and $(z_t - z)/t$ converge as $t \downarrow 0$, the chain rule yields

$$\frac{\mathfrak{D}(\Sigma + tH_t) - \mathfrak{D}_2(\Sigma)}{t} = \dot{f}_x \frac{x_t - x}{t} + \dot{f}_y \frac{y_t - y}{t} + \dot{f}_z \frac{z_t - z}{t} + \mathrm{o}(1), \qquad t \downarrow 0,$$

with partial derivatives

$$\dot{f}_x = -\frac{1}{z - y}, \qquad \dot{f}_y = \frac{z - x}{(z - y)^2} = \frac{\mathfrak{D}_2(\Sigma)}{z - y}, \qquad \dot{f}_z = \frac{1 - \mathfrak{D}_2(\Sigma)}{z - y}.$$

By Corollary C.5 and Lemma C.6, we have, respectively

$$\lim_{t\downarrow 0} \frac{x_t - x}{t} = \frac{1}{2}\operatorname{tr}((J_0 + J^{-1})H), \qquad \lim_{t\downarrow 0} \frac{y_t - y}{t} = \operatorname{tr}(\Upsilon_2 H).$$

Further, $(z_t - z)/t = \operatorname{tr}(H_t) \to \operatorname{tr}(H)$ as $t \downarrow 0$. It follows that

$$\begin{aligned}
&\frac{\mathfrak{D}(\Sigma + tH_t) - \mathfrak{D}_2(\Sigma)}{t} \\
&= \frac{1}{z - y}\left(-\frac{x_t - x}{t} + \mathfrak{D}_2(\Sigma)\frac{y_t - y}{t} + (1 - \mathfrak{D}_2(\Sigma))\frac{z_t - z}{t}\right) + \mathrm{o}(1) \\
&\to \frac{1}{z - y}\left(-\frac{1}{2}\operatorname{tr}((J_0 + J^{-1})H) + \mathfrak{D}_2(\Sigma)\operatorname{tr}(\Upsilon_2 H) + (1 - \mathfrak{D}_2(\Sigma))\operatorname{tr}(H)\right)
\end{aligned}$$

as $t \downarrow 0$. Isolating H yields the stated limit. □

Proof of Corollary 4.5. Write $H_t = (h_{t,jk})_{j,k=1}^d$ and $H = (h_{jk})_{j,k=1}^d$. For any $j \in \{1, \ldots, d\}$, we have

$$[R + tH_t]_{jj}^{-1/2} = (1 + th_{t,jj})^{-1/2} = 1 - \tfrac{1}{2}th_{jj} + \mathrm{o}(t), \qquad t \downarrow 0.$$

Write $R = (\rho_{jk})_{j,k=1}^d$. It follows that, for $j, k \in \{1, \ldots, d\}$,

$$\begin{aligned}
[\varphi(R + tH_t)]_{jk} &= \left(1 - \tfrac{1}{2}th_{jj} + \mathrm{o}(t)\right)^{-1/2}\left(\rho_{jk} + th_{jk} + \mathrm{o}(t)\right)\left(1 - \tfrac{1}{2}th_{kk} + \mathrm{o}(t)\right)^{-1/2} \\
&= \rho_{jk} + t\left(h_{jk} - \tfrac{1}{2}(h_{jj}\rho_{jk} + \rho_{jk}h_{kk})\right) + \mathrm{o}(t), \qquad t \downarrow 0.
\end{aligned}$$

In matrix form, we find

$$\lim_{t\downarrow 0} t^{-1}(\varphi(R + tH_t) - R) = H - \tfrac{1}{2}(D_H R + RD_H) =: \dot{\varphi}_R(H). \tag{45}$$

Note that the operator $\dot{\varphi}_R : \mathbb{S}^d \to \mathbb{S}^d$ is indeed linear. By the chain rule, we have

$$\lim_{t\downarrow 0} t^{-1}\left(\mathfrak{D}_r(\varphi(R + tH_t)) - \mathfrak{D}_r(R)\right) = \operatorname{tr}(M_{R,r}\dot{\varphi}_R(H)).$$

By the cyclic permutation property of the trace operator, the identity $\operatorname{tr}(A\operatorname{diag}(B)) = \operatorname{tr}(\operatorname{diag}(A)B)$ for square matrices A and B, and the fact that R and $M_{R,r}$ are symmetric and thus $RM_{R,r}$ and $M_{R,r}R$ share the same diagonal, we get

$$\begin{aligned}\operatorname{tr}(M_{R,r}\dot{\varphi}_R(H)) &= \operatorname{tr}\left(M_{R,r}(H - \tfrac{1}{2}(D_H R + R D_H))\right)\\ &= \operatorname{tr}(M_{R,r}H) - \tfrac{1}{2}\big(\operatorname{tr}(M_{R,r}D_H R) + \operatorname{tr}(M_{R,r}RD_H)\big)\\ &= \operatorname{tr}(M_{R,r}H) - \operatorname{tr}(D_{M_{R,r}R}H)\\ &= \operatorname{tr}\big((M_{R,r} - D_{M_{R,r}R})H\big).\end{aligned}$$

□

Before turning to the proof of Theorem 4.6 let us state a particularly convenient lemma.

Lemma C.7 (Empirical covariance matrix, standard Gaussian case)**.** *Let $\epsilon_1, \dots, \epsilon_n$ be independent $\mathcal{N}_d(0, I_d)$ random vectors and let*

$$W_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (\epsilon_i \epsilon_i^\top - I_d). \tag{46}$$

Then $W_n \rightsquigarrow W$ as $n \to \infty$, with W a random symmetric matrix such that

$$W_{jk} \sim \begin{cases} \mathcal{N}(0,2), & \text{for } j = k \in \{1, \dots, d\},\\ \mathcal{N}(0,1), & \text{for } 1 \le j < k \le d, \end{cases}$$

all entries being independent (except for the symmetry of W). For $A, B \in \mathbb{S}^d$, we have

$$\mathbb{E}[\operatorname{tr}(AW)\operatorname{tr}(BW)] = 2\operatorname{tr}(AB). \tag{47}$$

Proof of Lemma C.7. The weak convergence $W_n \rightsquigarrow W$ with W as stated is a direct consequence of the multivariate central limit theorem. For symmetric $A \in \mathbb{R}^{d\times d}$, we have, by symmetry of W,

$$\operatorname{tr}(AW) = \sum_{j=1}^{d}\sum_{k=1}^{d} A_{jk}W_{jk} = \sum_{j=1}^{d} A_{jj}W_{jj} + 2\sum_{1\le j<k\le d} A_{jk}W_{jk}.$$

Since the random variables appearing on the last line are independent and have zero mean, it follows that, for symmetric $A, B \in \mathbb{R}^{d\times d}$,

$$\begin{aligned}\mathbb{E}[\operatorname{tr}(AW)\operatorname{tr}(BW)] &= \sum_{j=1}^{d} A_{jj}B_{jj}\mathbb{E}[W_{jj}^2] + 4\sum_{1\le j<k\le d} A_{jk}B_{jk}\mathbb{E}[W_{jk}^2]\\ &= 2\sum_{j=1}^{d} A_{jj}B_{jj} + 4\sum_{1\le j<k\le d} A_{jk}B_{jk}\\ &= 2\sum_{j=1}^{d}\sum_{k=1}^{d} A_{jk}B_{jk}\\ &= 2\operatorname{tr}(AB).\end{aligned}$$

□

Lemma C.8 (Asymptotic expansion of correlation matrix estimates)**.** *Let $R = (\rho_{jk})_{j,k=1}^d$ be a $d \times d$ correlation matrix and let $R_n = (\rho_{n,jk})_{j,k=1}^d$ be either the empirical correlation*

matrix $\hat{R}_n$ *in* (29) *in the Gaussian distribution setting (GD) or the matrix* $\check{R}_n$ *in* (30) *of normal scores rank correlation coefficients in the Gaussian copula setting (GC). In both cases, for* $j, k \in \{1, \ldots, d\}$,

$$\sqrt{n}(\rho_{n,jk} - \rho_{jk}) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left(Z_{ij} Z_{ik} - \frac{1}{2} \rho_{jk} (Z_{ij}^2 + Z_{ik}^2) \right) + \mathrm{o}_p(1), \qquad n \to \infty, \tag{48}$$

or, in matrix form,

$$\sqrt{n}(R_n - R) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \dot{\varphi}_R(Z_i Z_i^\top) + \mathrm{o}_p(1), \qquad n \to \infty \tag{49}$$

with $\dot{\varphi}_R$ *as in* (45) *and with* $Z_1, \ldots, Z_n$ *an independent random sample from* $\mathcal{N}_d(0, R)$.

Proof. The matrix formula (49) is just a repackaging of the element-wise one (48) exploiting (45).

In the Gaussian distribution setting (GD), put $Z_i = D_\Sigma^{-1/2}(\xi_i - \mu)$ for $i \in \{1, \ldots, n\}$. The common distribution of Z_i is $\mathcal{N}_d(0, R)$. Let $\hat{\Sigma}_{n,Z}$ be their empirical covariance matrix, replacing ξ_i by Z_i in (29). We have $\xi_i = \mu + \xi_i D_\Sigma^{1/2}$ and thus

$$\hat{\Sigma}_n = D_\Sigma^{1/2} \hat{\Sigma}_{n,Z} D_\Sigma^{1/2}.$$

As φ reduces variables to unit scale anyway, we have

$$\hat{R}_n = \varphi(\hat{\Sigma}_n) = \varphi(\hat{\Sigma}_{n,Z}).$$

By the multivariate central limit theorem and Slutsky's lemma,

$$\sqrt{n}(\hat{\Sigma}_{n,Z} - R) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (Z_i Z_i^\top - R) + \mathrm{o}_p(1), \qquad n \to \infty.$$

The delta method and the identity $\varphi(R) = R$ yield

$$\sqrt{n}(\hat{R}_n - R) = \dot{\varphi}_R(\sqrt{n}(\hat{\Sigma}_{n,Z} - R)) + \mathrm{o}_p(1), \qquad n \to \infty.$$

The combination of the last two expansions gives (49) in view of linearity of $\dot{\varphi}_R$ and the identity $\dot{\varphi}_R(R) = 0$, as R has unit diagonal.

In the Gaussian copula setting (GC), the expansion (48) is Theorem 3.1 in Klaassen and Wellner (1997). We have $Z_i = (Z_{i1}, \ldots, Z_{id})$ with $Z_{ij} = \Phi^{-1} \circ F_j^{-1}(\xi_{ij})$ for $i \in \{1, \ldots, n\}$ and $j \in \{1, \ldots, d\}$. The common distribution of the random vectors Z_i is $\mathcal{N}_d(0, R)$ by the assumption that the copula of the law of ξ_i is Gaussian with correlation matrix R. □

Remark C.9. The expansion (49) remains valid for the empirical correlation matrix from an independent random sample $\xi_1, \ldots, \xi_n$ from a distribution with finite fourth moments and positive variances, upon defining $Z_i = D_\Sigma^{-1/2}(\xi_i - \mu)$ with μ and Σ the population mean vector and covariance matrix, respectively. The random vectors Z_i have zero means and unit variances but are no longer Gaussian. From the expansion, the asymptotic distribution of the empirical correlation matrix can be found using the multivariate central limit theorem. The asymptotic distribution of $\sqrt{n}(\hat{R}_n - R)$ is a random matrix whose d^2 elements have a centered multivariate normal distribution the covariance matrix of which can be derived from (48). See also Kollo and von Rosen (2006, Theorem 3.1.6).

Proof of Theorem 4.6. We have $\mathfrak{D}_{n,r} = \mathfrak{D}_r(R_n)$ with R_n equal to either $\hat{R}_n$ in (29) in the Gaussian distribution setting (GD) or $\check{R}_n$ in (30) in the Gaussian copula setting (GC). In both cases, we have the expansion (49) and thus

$$\sqrt{n}(R_n - R) = \sqrt{n}\left(\varphi\left(\tfrac{1}{n}\textstyle\sum_{i=1}^n Z_i Z_i^\top\right) - R\right) + \mathrm{o}_p(1), \qquad n \to \infty.$$

Let the eigendecomposition of R be $R = U\Lambda U^\top$, where the diagonal matrix Λ contains the eigenvalues of R on the diagonal and the columns of the orthogonal matrix U contain the associated eigenvectors. Then $Z_i = U\Lambda^{1/2}\epsilon_i$ for $i \in \{1, \ldots, n\}$ where $\epsilon_1, \ldots, \epsilon_n$ is an independent random sample from $\mathcal{N}_d(0, I_d)$. For W_n as in (46), we find

$$\frac{1}{\sqrt{n}}\left(\frac{1}{n}\sum_{i=1}^n Z_i Z_i^\top - R\right) = U\Lambda^{1/2} W_n \Lambda^{1/2} U^\top.$$

Combining the previous expansions with the delta method and Corollary 4.5, we get

$$\begin{aligned}
\sqrt{n}(\mathfrak{D}_{n,r} - \mathfrak{D}_r(R)) &= \operatorname{tr}\left((M_{R,r} - D_{M_{R,r}R})U\Lambda^{1/2}W_n\Lambda^{1/2}U^\top\right) + \mathrm{o}_p(1) \\
&= \operatorname{tr}\left(\Lambda^{1/2}U^\top(M_{R,r} - D_{M_{R,r}R})U\Lambda^{1/2}W_n\right) + \mathrm{o}_p(1) \\
&\rightsquigarrow \operatorname{tr}\left(\Lambda^{1/2}U^\top(M_{R,r} - D_{M_{R,r}R})U\Lambda^{1/2}W\right), \qquad n \to \infty,
\end{aligned}$$

with W the random matrix in Lemma C.7. By the covariance formula (47) in the same lemma, the limit is centered Gaussian with asymptotic variance

$$2\operatorname{tr}\left((\Lambda^{1/2}U^\top(M_{R,r} - D_{M_{R,r}R})U\Lambda^{1/2})^2\right) = 2\operatorname{tr}\left((R(M_{R,r} - D_{M_{R,r}R}))^2\right)$$

for $r \in \{1, 2\}$, using the cyclical property of the trace. □

Proof of Corollary 4.7. Since $\hat{R}_n$ in setting (GD) and $\check{R}_n$ in setting (GC) are consistent estimators of R, it suffices to check that $M_{R,r}$ is a continuous function of R. To do so, we need to inspect the formulas for M_1 and M_2 in Theorems 4.1 and 4.2. The crucial point is that the eigenvalues and eigenvectors of the upper and lower diagonal blocks R_1 (dimension $p \times p$) and R_2 (dimension $q \times q$) depend continuously on R, since by assumption these two blocks have p and q distinct eigenvalues, respectively. □

Appendix D. Additional numerical experiments

D.1. **Goodness-of-fit for finite samples for $\mathfrak{D}_2$.** The results of Section 5.1 were only presented for $\mathfrak{D}_1$ in view of space considerations. Here we conduct the exact same simulations using $\mathfrak{D}_2$. The results are presented in Figure 5.

D.2. **Goodness-of-fit for rank-based estimation of the correlation matrix.** We now repeat the simulations in the same settings as those of Section 5.1 for the Gaussian copula case, that is when the estimated correlation matrix is $\check{R}_n$. The results for $\mathfrak{D}_1$ and $\mathfrak{D}_2$ are shown in Figures 6 and 7, respectively.

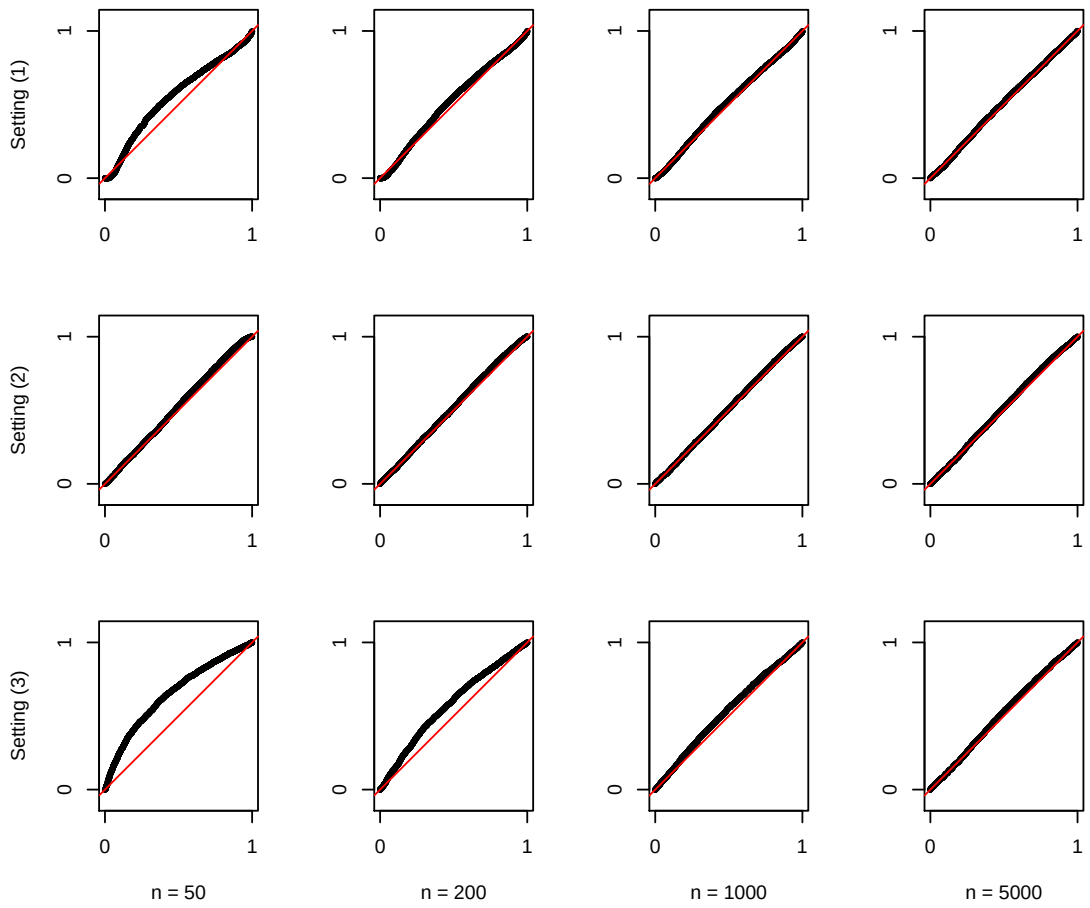

FIGURE 5. *P-P plots for 3000 repetitions of the centred and (empirically) rescaled estimator of $\mathfrak{D}_2$ in the three settings from Section 5.1 for increasing sample sizes (from left to right). The results are presented for an empirical correlation matrix in the case of Gaussian data.*

D.3. **Shrinkage evaluation for EEG data.** For the EEG case study in Section 6, we conducted a preliminary assessment to evaluate whether the shrinkage methods in Section 5.2 produce confidence intervals performing as they should. The sample size and parameter values were taken to match those of the data. The empirical coverage of the confidence intervals was estimated based on 2000 replications. The results are presented in Figure 8. In plots (a) and (c), the advantage of shrinking the eigenvalues is clearly visible for coefficient $\mathfrak{D}_1$.

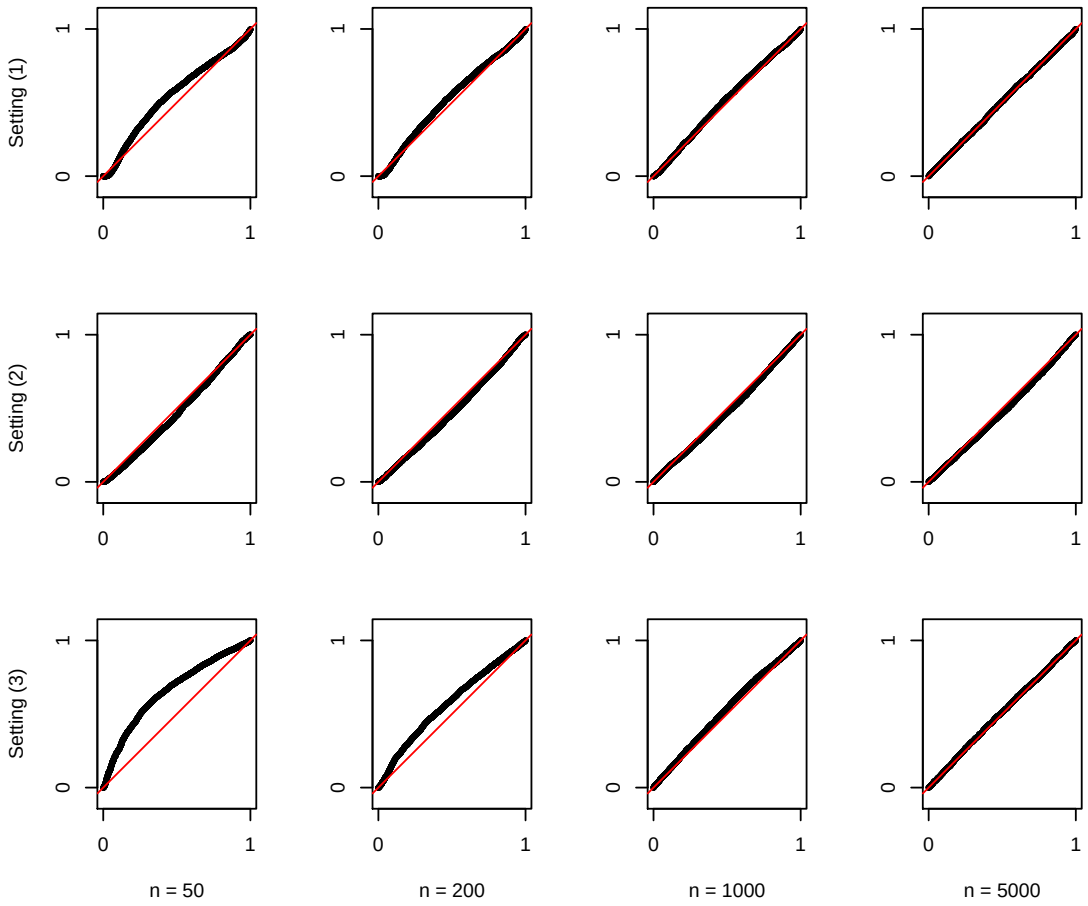

FIGURE 6. *P-P plots for 3000 repetitions of the centred and (empirically) rescaled estimator of $\mathfrak{D}_2$ in the three settings from Section 5.1 for increasing sample sizes (from left to right). The results are presented for a Gaussian copula relying on $\check{R}_n$.*

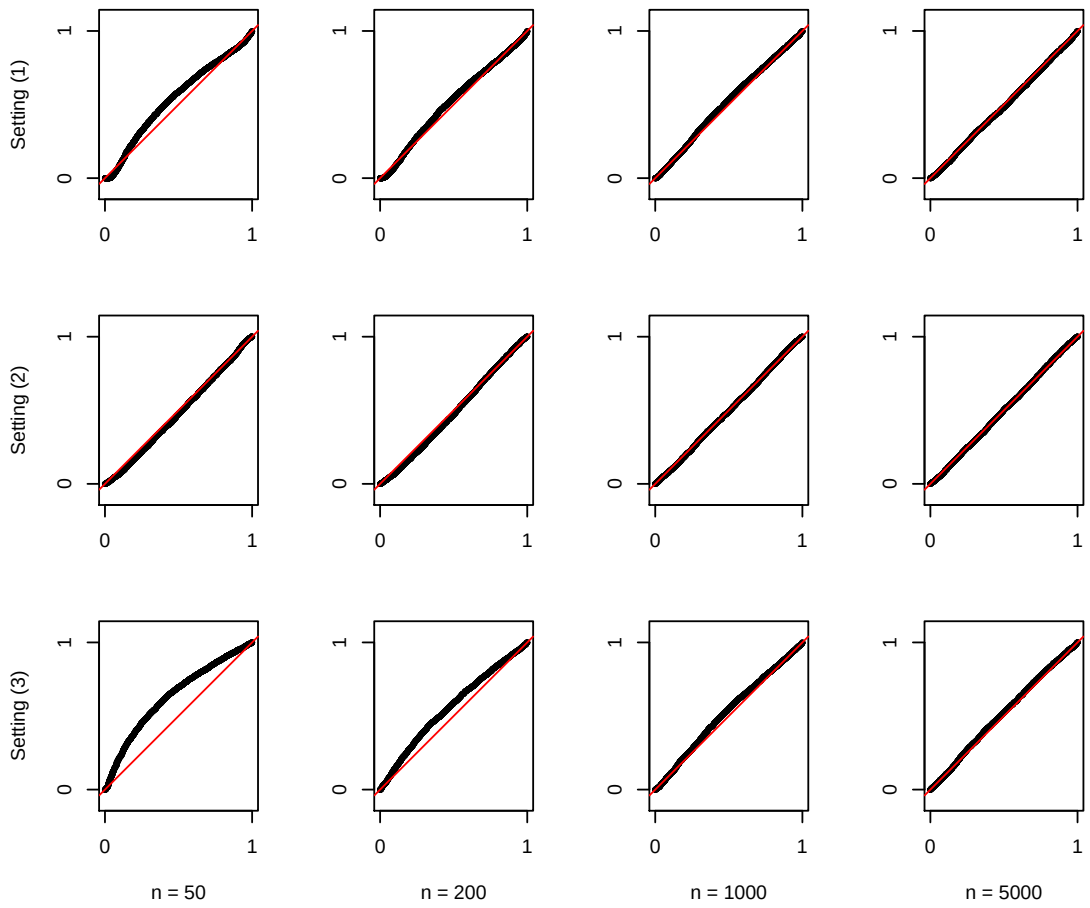

FIGURE 7. *P-P plots for 3000 repetitions of the centred and (empirically) rescaled estimator of $\mathfrak{D}_2$ for the three settings from Section 5.1 for increasing sample sizes (from left to right). The results are presented for a Gaussian copula relying on $\check{R}_n$.*

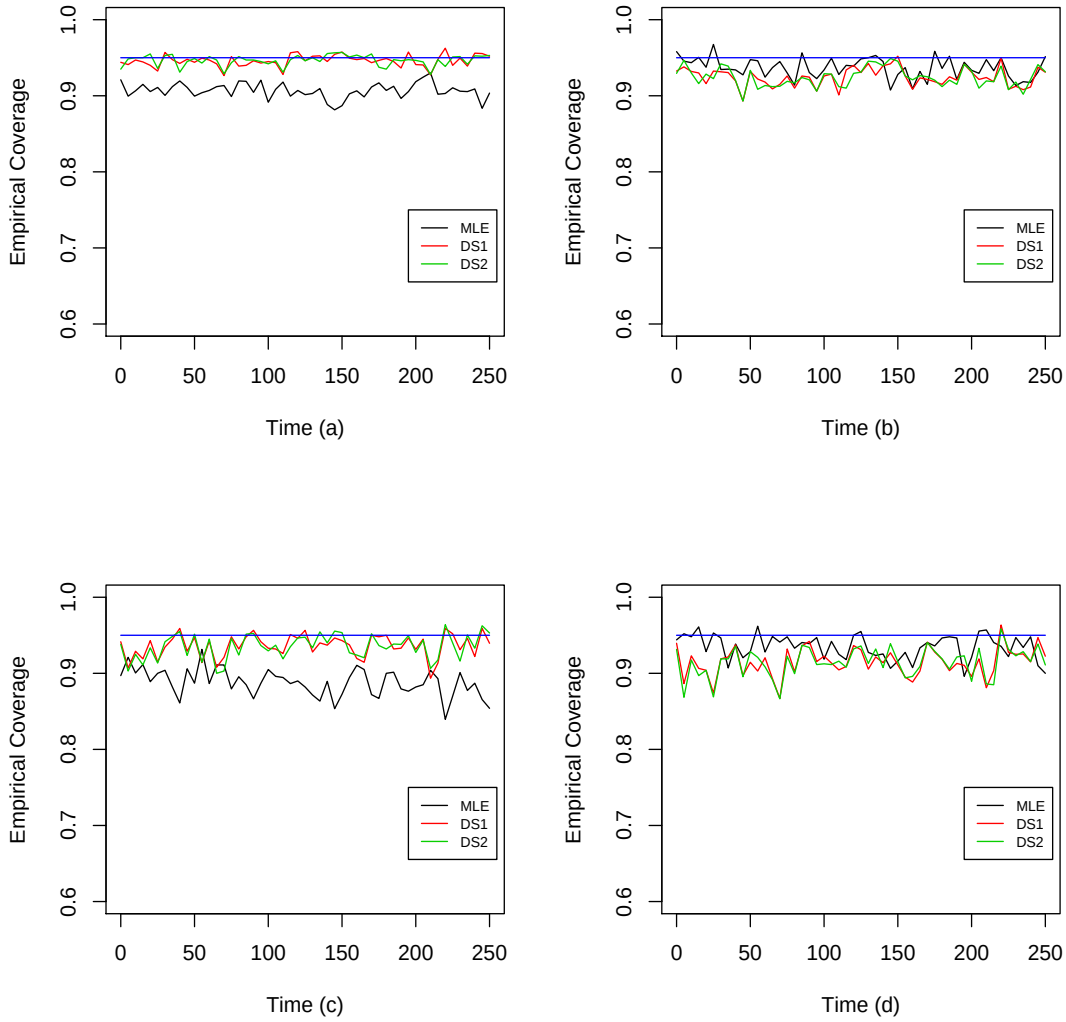

FIGURE 8. *Empirical coverage of 95% confidence intervals estimated from 2000 replications in the Gaussian copula setting with sample size and parameters derived from the case study in Section 6. MLE refers to no shrinkage while DS1 and DS2 refer to the two shrinkage methods in Section 5.2. (a) Alcoholic group with* $\mathfrak{D}_1$. *(b) Alcoholic group with* $\mathfrak{D}_2$. *(c) Control group with* $\mathfrak{D}_1$. *(d) Control group with* $\mathfrak{D}_2$.