Minimal Gelbrich Distance to Uncorrelation

Matthieu Borelle, Teodoro Alamo, Cristina Stoica Maniu, Sylvain Bertrand, Eduardo Camacho

Minimal Gelbrich distance to uncorrelation

Matthieu Borelle, Teodoro Alamo, *Member, IEEE*, Cristina Stoica, *Senior Member, IEEE*, Sylvain Bertrand, Eduardo F. Camacho, *Life Fellow, IEEE*

***Abstract*— This paper reports new properties of the Wasserstein/Gelbrich distance and associated ambiguity sets to analyze the correlation between two scalar random variables. A simple closed expression is derived for the Gelbrich distance between two bidimensional random distributions. Moreover, the minimum disturbance in the Gelbrich metric required to reach uncorrelation between two random variables is obtained. This allows us to determine the robustness of the Pearson coefficient within an ambiguity set. A numerical example showcases the potential use of the obtained results in the field of variable selection.**

***Index Terms*— Wasserstein distance, Gelbrich distance, Pearson coefficient, ambiguity set, distributionally robust optimization.**

I. Introduction

VARIABLE selection is a fundamental task in statistical analysis / statistical inference, aiming to identify a subset of variables that are most relevant for a given problem. Traditionally, correlation measures, such as the Pearson coefficient [1] have been widely employed to quantify the strength and direction of the relationship between variables. However, relying solely on correlation measures could overlook important aspects of the data, such as its robustness to the lack of knowledge about the data-generating probability distribution, or its potential time-varying nature.

In this context, the concept of *ambiguity set* [2] plays a fundamental role. An ambiguity set is a collection of probability distributions representing the uncertainty about the true data-generating distribution. It is a key component in models where the underlying distribution is not precisely known, but there is a requirement to make robust decisions that perform well under various possible scenarios. The ambiguity set is often defined as a ball in the space of probability distributions that contains all distributions close to a nominal or a priori most likely distribution with respect to a given probability metric [2]. The motivation of the current paper derives from the possible degradation of the Pearson coefficient in an ambiguity set.

M. Borelle and C. Stoica are with Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des signaux et systèmes, 91190, Gif-sur-Yvette, France (e-mail: {matthieu.borelle; cristina.stoica}@l2s.centralesupelec.fr).

M. Borelle and S. Bertrand are with Université Paris-Saclay, ONERA, Traitement de l'information et systèmes, 91123, Palaiseau, France (e-mail: {matthieu.borelle; sylvain.bertrand}@onera.fr).

T. Alamo and E.F. Camacho are with Department of Ingeniería de Sistemas y Automática, Universidad de Sevilla, Camino de los Descubrimientos, 41092 Sevilla, Spain (e-mail: talamo@us.es, eduardo@esi.us.es).

In the realm of distributionally robust optimization and estimation, different notions of distance between distributions are available. For example, Prokhorov [3] and Wasserstein distances [2] are used to assess how different two probability measures are (see also [4] for many other possibilities). Derived from the optimal transport problem [5], by quantifying the minimum work needed to transport the mass from one distribution to another, the Wasserstein distance captures both global and local features of the distributions, offering a more nuanced understanding of their dissimilarity. Related with the Wasserstein metric is the Gelbrich distance [6], [7], which has a simple expression and is used to lower bound the Wasserstein distance, especially in the Gaussian case. Since both of them define a metric, they differ from other popular measures [8], such as the Kullback-Leibler divergence. The versatility and applicability of the Wasserstein and Gelbrich distances make them a valuable tool fostering advancements in data analysis, machine learning (e.g., in the context of generative adversarial networks [9]), estimation and control (e.g., Kalman filtering [10], [11], model predictive control [12]), game theory [13] and allowing to solve distributionally robust decision problem in various fields: finance [14], energy production [15], imaging [16], and beyond.

The current work proposes a new perspective by incorporating the Gelbrich distance into the analysis of the correlation between two scalar random variables. The proposed results capture the correlation between variables and their resilience to distributional deviations (i.e., distributional robustness of the correlation coefficient). Closed expressions for the minimum Gelbrich distance required to reach uncorrelation between two scalar random variables are further contributions. The proposed methodology enables the identification of variables that exhibit both strong correlation and robustness to various data conditions. A numerical example on system identification illustrates how the results of the paper can be applied in the context of variable selection.

This paper is organized as follows. Section II familiarizes the reader with the Wasserstein distance. Section III introduces the Gelbrich distance used as a lower bound for the Wasserstein distance. The minimal Gelbrich distance to reach uncorrelation is developed in Section IV. In Section V, a numerical example illustrates the proposed results. Concluding remarks are provided in Section VI.

Notation

Denote the set of symmetric positive semidefinite matrices in $\mathbb{R}^{d\times d}$ by $\mathbb{S}^d_+ = \{\, S \in \mathbb{S}^d \;:\; S \succeq 0 \,\}$. The square root of

$S \in \mathbb{S}_+^d$ is denoted by $S^{\frac{1}{2}}$, which is the positive semidefinite matrix satisfying $(S^{\frac{1}{2}})^2 = S$. The notation $\mathrm{Tr}(A)$ designates the trace of the matrix A.

II. Wasserstein distance

The Wasserstein distance offers a unique perspective on measuring the dissimilarity or discrepancy between probability distributions. Derived from the optimal transport problem [5], Wasserstein distances provide a quantitative measure of the required "effort" to transform one distribution into another. This concept draws an intuitive analogy to the transportation of mass, where the distances capture the minimal cost of moving masses from one distribution to another. The key idea behind Wasserstein distances lies in comparing not just the locations of the masses or the shapes of the distributions, but rather their entire structures. It identifies the most cost-effective way to transport the mass from one distribution to another (see Fig. 1). Recent results on the Wasserstein distance can be found, for instance, in [17], [18], [2].

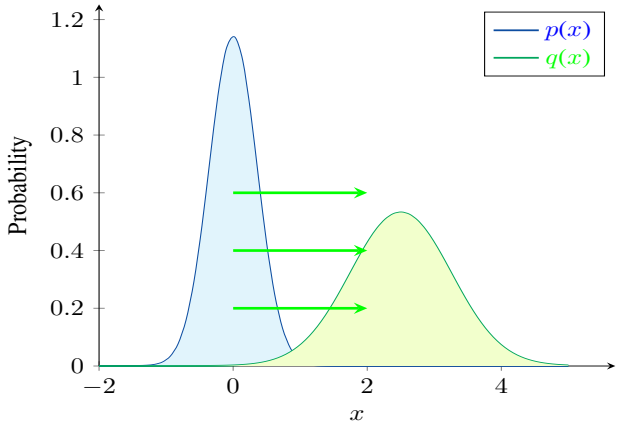

Fig. 1. Illustration of the transport of a normal probability distribution $\boldsymbol{p(x)}$ to another one $\boldsymbol{q(x)}$

Definition 1 (Wasserstein p-distance [19], [7]): Consider a metric space (M, c), e.g., $M = \mathbb{R}^d$ and $c(x, y) = \|x - y\|_2$, $\forall x, y \in M$. The Wasserstein p-distance between two probability measures μ and ν, with finite p-moments on M, is given by

$$W_p(\mu,\nu) = \left(\inf_{\pi \in \Pi(\mu,\nu)} \left\{ \int_{M\times M} c(x,y)^p \, \pi(dx,dy)) \right\} \right)^{\frac{1}{p}} \tag{1}$$

where $\Pi(\mu, \nu)$ is the set of all probability distributions π defined on $M \times M$ with marginals μ and ν, often called couplings or transport plans.

Even though the Wasserstein p-distance between two probability measures seems to be difficult to compute, the Wasserstein 2-distance between Gaussian distributions studied in [19] has a much simpler expression that will be explored in the next section.

III. Gelbrich distance

This section briefly describes the Gelbrich distance, which is further used as a lower bound for the Wasserstein distance.

Definition 2 (Gelbrich distance [6], [7]): Suppose that the tuples $(\mu_\Sigma, \Sigma) \in \mathbb{R}^n \times \mathbb{S}_+^n$ and $(\mu_S, S) \in \mathbb{R}^n \times \mathbb{S}_+^n$ represent the mean and covariance matrices of random distributions $\mathbb{Q}_\Sigma$ and $\mathbb{Q}_S$. Then, the Gelbrich distance $G_D(\mathbb{Q}_\Sigma, \mathbb{Q}_S)$ between $\mathbb{Q}_\Sigma$ and $\mathbb{Q}_S$ is defined as

$$\begin{aligned} G_D(\mathbb{Q}_\Sigma, \mathbb{Q}_S) = \Big(& \|\mu_\Sigma - \mu_S\|_2^2 \\ & + \mathrm{Tr}\left(S + \Sigma - 2(\Sigma^{\frac{1}{2}} S \Sigma^{\frac{1}{2}})^{\frac{1}{2}} \right) \Big)^{\frac{1}{2}}. \end{aligned} \tag{2}$$

This metric allows us to define the notion of ambiguity set. Denote by $\mathcal{P}_d$ the set of all the probability distributions of dimension d. Define the ambiguity set $\mathcal{A}_\alpha$ centred at $\mathbb{Q}_\Sigma \in \mathcal{P}_d$ and its radius $\alpha \geq 0$ as

$$\mathcal{A}_\alpha = \{\mathbb{Q}_S \in \mathcal{P}_d \mid G_D(\mathbb{Q}_\Sigma, \mathbb{Q}_S) \leq \alpha\}. \tag{3}$$

As stated in [7], the squared Gelbrich distance (2) is both convex and continuous on the parameters μ_Σ, μ_S, Σ and S. The next property states that the Gelbrich distance is a lower bound of the Wasserstein distance [6], [7].

Property 1: Suppose that (μ_Σ, Σ) and (μ_S, S) represent the mean and covariance matrices of distributions $\mathbb{Q}_\Sigma$ and $\mathbb{Q}_S$. Then, the following expression holds

$$W_2(\mathbb{Q}_\Sigma, \mathbb{Q}_S) \geq G_D(\mathbb{Q}_\Sigma, \mathbb{Q}_S), \tag{4}$$

with $W_2(\mathbb{Q}_\Sigma, \mathbb{Q}_S)$ the Wasserstein 2-distance between distributions $\mathbb{Q}_\Sigma$ and $\mathbb{Q}_S$.

Notice that the inequality (4) can be replaced with the equality $W_2(\mathbb{Q}_\Sigma, \mathbb{Q}_S) = G_D(\mathbb{Q}_\Sigma, \mathbb{Q}_S)$ if both $\mathbb{Q}_\Sigma$ and $\mathbb{Q}_S$ are Gaussian distributions [7].

The Gelbrich distance relies on the computation of $\Sigma^{\frac{1}{2}}$ and $\left(\Sigma^{\frac{1}{2}} S \Sigma^{\frac{1}{2}}\right)^{\frac{1}{2}}$. The next property shows that, in the particular case of two-dimensional distributions, the Gelbrich distance can be more easily calculated.

Property 2: Suppose that (μ_Σ, Σ) and (μ_S, S) represent the mean and covariance matrices of the *two-dimensional distributions* $\mathbb{Q}_\Sigma$ and $\mathbb{Q}_S$. Then, the Gelbrich distance is

$$\begin{aligned} G_D(\mathbb{Q}_\Sigma, \mathbb{Q}_S) \quad = \quad & \Big(\|\mu_\Sigma - \mu_S\|_2^2 + \mathrm{Tr}(S + \Sigma) \\ & - 2\sqrt{\mathrm{Tr}(S\Sigma) + 2\sqrt{\det(S\Sigma)}} \Big)^{\frac{1}{2}}. \end{aligned}$$

Proof: Using the definition of the Gelbrich distance (2), it suffices to show that for every pair of matrices $\Sigma \in \mathbb{S}_+^2$ and $S \in \mathbb{S}_+^2$, the following equality holds

$$\mathrm{Tr}\left(\left(\Sigma^{\frac{1}{2}} S \Sigma^{\frac{1}{2}} \right)^{\frac{1}{2}} \right) = \sqrt{\mathrm{Tr}(S\Sigma) + 2\sqrt{\det(S\Sigma)}}.$$

Consider a generic matrix $A \in \mathbb{R}^{2\times 2}$. Its characteristic equation $\det(\lambda \mathrm{I} - A) = 0$ is equivalent to

$$\lambda^2 - \mathrm{Tr}(A)\lambda + \det(A) = 0. \tag{5}$$

Recalling the Cayley-Hamilton theorem [20], a square matrix satisfies its characteristic equation, i.e.,

$$A^2 - \mathrm{Tr}(A)A + \det(A)\mathrm{I} = 0, \quad \forall A \in \mathbb{R}^{2\times 2}. \tag{6}$$

The positive semidefinite matrix $H = \Sigma^{\frac{1}{2}} S \Sigma^{\frac{1}{2}}$ has a unique symmetric semidefinite square root (see e.g., Theorem 7.2.6 in [20]) denoted by $H^{\frac{1}{2}}$. Using $A = H^{\frac{1}{2}}$ in (6), the following expression is obtained

$$H - \mathrm{Tr}(H^{\frac{1}{2}}) H^{\frac{1}{2}} + \det(H^{\frac{1}{2}})\mathrm{I} = 0. \tag{7}$$

Applying the trace operator to expression (7) leads to

$$\mathrm{Tr}(H) - \left(\mathrm{Tr}(H^{\frac{1}{2}})\right)^2 + 2\det(H^{\frac{1}{2}}) = 0.$$

Then, using the commutativity of the trace operator, the following result can be inferred

$$\begin{aligned}\left(\mathrm{Tr}(H^{\frac{1}{2}})\right)^2 &= \mathrm{Tr}(H) + 2\det(H^{\frac{1}{2}}) \\ &= \mathrm{Tr}\left(\Sigma^{\frac{1}{2}} S \Sigma^{\frac{1}{2}}\right) + 2\sqrt{\det(H)} \\ &= \mathrm{Tr}\left(S\Sigma\right) + 2\sqrt{\det(\Sigma^{\frac{1}{2}} S \Sigma^{\frac{1}{2}})} \\ &= \mathrm{Tr}\left(S\Sigma\right) + 2\sqrt{\det(S\Sigma)}.\end{aligned}$$

■

Remark 1: In comparison with (2), where it is required to compute two times the square root of a two by two matrix, Property 2 provides a simple expression written in terms of the determinant and trace of a two by two matrix. This expression for the Gelbrich distance, which is a novel result to the best knowledge of the authors, paves the way for the results of the next section.

IV. Minimum Gelbrich distance to uncorrelation

The Pearson coefficient (also known as the correlation coefficient) is a measure of the linear correlation between two sets of data [1]. Consider the random scalar variables x and y. Denote by μ_x, μ_y their mean and by Σ the covariance of the two-dimensional random vector $\begin{bmatrix} x \\ y \end{bmatrix}$, i.e., $\mu_x = \mathrm{E}\{x\}$, $\mu_y = \mathrm{E}\{y\}$ and

$$\Sigma = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{xy} & \Sigma_{yy} \end{bmatrix} = \mathrm{E}\left\{ \begin{bmatrix} x-\mu_x \\ y-\mu_y \end{bmatrix} \begin{bmatrix} x-\mu_x \\ y-\mu_y \end{bmatrix}^\top \right\}.$$

One can compute the Pearson correlation coefficient $\bar{\rho}$ between x and y as

$$\bar{\rho} = \frac{\mathrm{E}\{(x-\mu_x)(y-\mu_y)\}}{\sqrt{\mathrm{E}\{(x-\mu_x)^2\}\mathrm{E}\{(y-\mu_y)^2\}}} = \frac{\Sigma_{xy}}{\sqrt{\Sigma_{xx}\Sigma_{yy}}}.$$

The Pearson coefficient takes values between -1 and 1. A value close to -1 or 1 indicates a strong linear relationship, while a value close to 0 suggests a weak or no linear relationship (no linear correlation). The sign indicates the direction of the relationship, while the magnitude represents the strength [1].

The Pearson coefficient is often used to determine if a given covariate $x \in \mathbb{R}$ is useful to reduce the error $y \in \mathbb{R}$ provided by a given estimator. If the error y is highly correlated with x (absolute value of the Pearson coefficient sufficiently close to 1), then adapting the estimator by incorporating x to the list of covariates would translate into an improvement of the estimator. If the correlation is close to zero, then the incorporation of x will increase the estimator complexity without significantly improving the estimations. However, the previous analysis is valid only when a precise knowledge of the joint probability distribution is available. If one has to take into consideration that the probability distribution belongs to an ambiguity set, then the decision process is more difficult. One should consider not only the nominal Pearson coefficient (corresponding to the centre of the ambiguity set) but also the possible variations of it in the ambiguity set.

This section proposes a robustness analysis of a given nominal Pearson coefficient in terms of its distance to uncorrelation. In particular, we analyze the minimal perturbation to the bivariate distribution (measured by means of the Gelbrich distance) required to force that the two variables of the distribution are no longer correlated.

Let us denote by $\mathcal{U}$ the set of two-dimensional probabilistic distributions for which there is no correlation term. This means that the two-dimensional distribution $\mathbb{Q}_S$ belongs to $\mathcal{U}$ if the covariance matrix $S \in \mathbb{S}_+^2$ corresponding to $\mathbb{Q}_S$ is diagonal, i.e.,

$$S \in \mathbb{S}_{\mathcal{U}} = \left\{ \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} : \sigma_1 \geq 0, \sigma_2 \geq 0 \right\}.$$

The problem under consideration is determining in an analytical way the minimum Gelbrich distance $D_{\mathcal{U}}(\mathbb{Q}_\Sigma)$ of a two-dimensional probability distribution $\mathbb{Q}_\Sigma$ to the set of uncorrelated distributions $\mathcal{U}$. To this aim, we consider the optimization problem

$$D_{\mathcal{U}}(\mathbb{Q}_\Sigma) = \min_{\mathbb{Q}_S \in \mathcal{U}} G_D(\mathbb{Q}_\Sigma, \mathbb{Q}_S).$$

From Property 1, we derive that $D_{\mathcal{U}}(Q_\Sigma)$ constitutes a lower bound on the Wasserstein 2-distance to uncorrelation

$$\min_{\mathbb{Q}_S \in \mathcal{U}} W_2(\mathbb{Q}_\Sigma, \mathbb{Q}_S) \geq \min_{\mathbb{Q}_S \in \mathcal{U}} G_D(\mathbb{Q}_\Sigma, \mathbb{Q}_S) = D_{\mathcal{U}}(\mathbb{Q}_\Sigma).$$

As commented before, the Wasserstein 2-distance equals the Gelbrich distance in the case of Gaussian distributions. Thus, the previous lower bound is exact under a Gaussian assumption on $\mathbb{Q}_\Sigma$ and $\mathbb{Q}_S$.

The following theorem provides a closed-form expression for the minimum Gelbrich distance to uncorrelation.

Theorem 1: Suppose that $\mathbb{Q}_\Sigma$ is a two-dimensional distribution with covariance $\Sigma = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{xy} & \Sigma_{yy} \end{bmatrix}$. Then, the minimum Gelbrich distance $D_{\mathcal{U}}(\mathbb{Q}_\Sigma)$ of $\mathbb{Q}_\Sigma$ to uncorrelation is derived from

$$D_{\mathcal{U}}^2(\mathbb{Q}_\Sigma) = \frac{2\Sigma_{xy}^2}{\mathrm{Tr}(\Sigma) + \sqrt{(\Sigma_{xx}-\Sigma_{yy})^2 + 4\det(\Sigma)}}.$$

Moreover, the following inequalities hold

$$\frac{\Sigma_{xy}^2}{\max\{\Sigma_{xx}, \Sigma_{yy}\}} \geq D_{\mathcal{U}}^2(\mathbb{Q}_\Sigma) \geq \frac{\Sigma_{xy}^2}{\mathrm{Tr}(\Sigma)}.$$

Proof: Denote by μ_Σ the mean of distribution $\mathbb{Q}_\Sigma$. By definition, and taking into consideration the closed-form expression for the Gelbrich distance from Property 2, the

following result holds

$$D^2_{\mathcal{U}} = \min_{\mu_S\in\mathbb{R}^2, S\in\mathbb{S}_{\mathcal{U}}} \Big(\|\mu_\Sigma - \mu_S\|_2^2 + \mathrm{Tr}(S+\Sigma) - 2\sqrt{\mathrm{Tr}(S\Sigma) + 2\sqrt{\det(S\Sigma)}}\Big).$$

Clearly, the minimum is attained at $\mu_S = \mu_\Sigma$. Thus, the previous expression leads to

$$D^2_{\mathcal{U}} = \min_{S\in\mathbb{S}_{\mathcal{U}}} \left(\mathrm{Tr}(S+\Sigma) - 2\sqrt{\mathrm{Tr}(S\Sigma) + 2\sqrt{\det(S\Sigma)}}\right). \quad (8)$$

Given $S = \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \in \mathbb{S}_{\mathcal{U}}$, consider the terms $\mathrm{Tr}(S\Sigma)$ and $\sqrt{\det(S\Sigma)}$ that can be further rewritten

$$\begin{aligned} \mathrm{Tr}(S\Sigma) &= \mathrm{Tr}\left(\begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix} \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{xy} & \Sigma_{yy} \end{bmatrix}\right) \\ &= \sigma_1^2\Sigma_{xx} + \sigma_2^2\Sigma_{yy}, \\ \sqrt{\det(S\Sigma)} &= \sqrt{\det(\Sigma)\det(S)} = \sqrt{\det(\Sigma)}\sigma_1\sigma_2. \end{aligned}$$

With these terms, the expression (8) becomes

$$D^2_{\mathcal{U}} = \min_{\sigma_1\geq 0, \sigma_2\geq 0} \Big(\mathrm{Tr}(\Sigma) + \sigma_1^2 + \sigma_2^2 - 2\sqrt{\sigma_1^2\Sigma_{xx} + \sigma_2^2\Sigma_{yy} + 2\sqrt{\det(\Sigma)}\sigma_1\sigma_2}\Big).$$

Notice that the constraints $\sigma_1 \geq 0$, $\sigma_2 \geq 0$ are not active at the solution of the optimization problem. This is due to the fact that the signs of σ_1 and σ_2 affect only the term $2\sqrt{\det(\Sigma)}\sigma_1\sigma_2$, and the larger this term is, the smaller the functional to be minimized. Therefore, we conclude that the same value for $D^2_{\mathcal{U}}$ is obtained if one removes the constraints $\sigma_1 \geq 0$, $\sigma_2 \geq 0$. Thus, this leads to

$$\begin{aligned} D^2_{\mathcal{U}} &= \min_{\sigma_1\in\mathbb{R}, \sigma_2\in\mathbb{R}} \Big(\mathrm{Tr}(\Sigma) + \sigma_1^2 + \sigma_2^2 - 2\sqrt{\sigma_1^2\Sigma_{xx} + \sigma_2^2\Sigma_{yy} + 2\sqrt{\det(\Sigma)}\sigma_1\sigma_2}\Big) \\ &= \min_{\sigma_1\in\mathbb{R}, \sigma_2\in\mathbb{R}} \Big(\mathrm{Tr}(\Sigma) + \sigma_1^2 + \sigma_2^2 - 2\sqrt{\begin{bmatrix} \sigma_1 \\ \sigma_2 \end{bmatrix}^\top \begin{bmatrix} \Sigma_{xx} & \sqrt{\det(\Sigma)} \\ \sqrt{\det(\Sigma)} & \Sigma_{yy} \end{bmatrix} \begin{bmatrix} \sigma_1 \\ \sigma_2 \end{bmatrix}}\Big). \end{aligned}$$

For a fixed value of $\tau^2 = \sigma_1^2 + \sigma_2^2$, the values of σ_1 and σ_2 that minimize the functional are the ones corresponding to the eigenvector of the largest eigenvalue λ_{max} of

$$H = \begin{bmatrix} \Sigma_{xx} & \sqrt{\det(\Sigma)} \\ \sqrt{\det(\Sigma)} & \Sigma_{yy} \end{bmatrix}.$$

With this result, it can be further derived

$$D^2_{\mathcal{U}} = \min_{\sigma_1\in\mathbb{R}, \sigma_2\in\mathbb{R}} \left(\mathrm{Tr}(\Sigma) + \sigma_1^2 + \sigma_2^2 - 2\sqrt{\lambda_{max}(\sigma_1^2+\sigma_2^2)}\right).$$

Due to the Perron-Frobenious theorem [20], which states that the largest eigenvalue of a matrix with all its components non-negative is non-negative, we have that $\lambda_{max} \geq 0$. Then, the following result is derived

$$\begin{aligned} D^2_{\mathcal{U}} &= \min_{\sigma_1\in\mathbb{R}, \sigma_2\in\mathbb{R}} \Big(\mathrm{Tr}(\Sigma) + \sigma_1^2 + \sigma_2^2 - 2\sqrt{\lambda_{max}}\sqrt{\sigma_1^2+\sigma_2^2}\Big) \\ &= \min_{\tau\geq 0} \left(\mathrm{Tr}(\Sigma) + \tau^2 - 2\sqrt{\lambda_{max}}\tau\right) \end{aligned}$$

The optimal value τ^* for $\tau = \sqrt{\sigma_1^2+\sigma_2^2}$ satisfies $2\tau^* - 2\sqrt{\lambda_{max}} = 0$. Thus, $\tau^* = \sqrt{\lambda_{max}}$ and $D^2_{\mathcal{U}}$ verifies

$$\begin{aligned} D^2_{\mathcal{U}} &= \min_{\tau\geq 0} \left(\mathrm{Tr}(\Sigma) + \tau^2 - 2\sqrt{\lambda_{max}}\tau\right) \\ &= \mathrm{Tr}(\Sigma) - \lambda_{max}. \end{aligned} \quad (9)$$

A closed expression for λ_{max} is obtained in what follows. From (6), the characteristic polynomial of matrix H verifies

$$\begin{aligned} 0 &= \det(\lambda \mathrm{I} - H) \\ &= \lambda^2 - \mathrm{Tr}(H)\lambda + \det(H) \\ &= \lambda^2 - \mathrm{Tr}(H)\lambda + \Sigma_{xx}\Sigma_{yy} - \det(\Sigma) \\ &= \lambda^2 - \mathrm{Tr}(\Sigma)\lambda + \Sigma_{xy}^2. \end{aligned}$$

Thus, the largest eigenvalue of matrix H is computed as the largest root of the previous second order equation

$$\begin{aligned} \lambda_{\max} &= \frac{\mathrm{Tr}(\Sigma) + \sqrt{(\mathrm{Tr}(\Sigma))^2 - 4\Sigma_{xy}^2}}{2} \\ &= \frac{\mathrm{Tr}(\Sigma) + \sqrt{\Sigma_{xx}^2 + \Sigma_{yy}^2 + 2\Sigma_{xx}\Sigma_{yy} - 4\Sigma_{xy}^2}}{2} \\ &= \frac{\mathrm{Tr}(\Sigma) + \sqrt{(\Sigma_{xx}-\Sigma_{yy})^2 + 4\det(\Sigma)}}{2}. \end{aligned} \quad (10)$$

From (9) and (10), we finally obtain

$$\begin{aligned} D^2_{\mathcal{U}} &= \mathrm{Tr}(\Sigma) - \lambda_{max} \\ &= \mathrm{Tr}(\Sigma) - \frac{\mathrm{Tr}(\Sigma) + \sqrt{(\Sigma_{xx}-\Sigma_{yy})^2 + 4\det(\Sigma)}}{2} \\ &= \frac{\mathrm{Tr}(\Sigma) - \sqrt{(\Sigma_{xx}-\Sigma_{yy})^2 + 4\det(\Sigma)}}{2}. \end{aligned}$$

Then, multiplying the numerator and denominator by $\mathrm{Tr}(\Sigma) + \sqrt{(\Sigma_{xx}-\Sigma_{yy})^2 + 4\det(\Sigma)}$, it leads to

$$\begin{aligned} D^2_{\mathcal{U}} &= \frac{(\Sigma_{xx}+\Sigma_{yy})^2 - (\Sigma_{xx}-\Sigma_{yy})^2 - 4\det(\Sigma)}{2(\mathrm{Tr}(\Sigma) + \sqrt{(\Sigma_{xx}-\Sigma_{yy})^2 + 4\det(\Sigma)})} \\ &= \frac{2\Sigma_{xy}^2}{\mathrm{Tr}(\Sigma) + \sqrt{(\Sigma_{xx}-\Sigma_{yy})^2 + 4\det(\Sigma)}}, \end{aligned} \quad (11)$$

which proves the main claim of the theorem.

From (11), and using $\det(\Sigma) \geq 0$, we further obtain

$$\begin{aligned} D^2_{\mathcal{U}} &\leq \frac{2\Sigma_{xy}^2}{\mathrm{Tr}(\Sigma) + \sqrt{(\Sigma_{xx}-\Sigma_{yy})^2}} \\ &= \frac{2\Sigma_{xy}^2}{\mathrm{Tr}(\Sigma) + \max\{\Sigma_{xx}-\Sigma_{yy}, \Sigma_{yy}-\Sigma_{xx}\}} \\ &= \frac{\Sigma_{xy}^2}{\max\{\Sigma_{xx}, \Sigma_{yy}\}}. \end{aligned}$$

The following result can be also derived from (11)

$$\begin{aligned} D_{\mathcal{U}}^2 &= \frac{2\Sigma_{xy}^2}{\mathrm{Tr}(\Sigma)+\sqrt{(\Sigma_{xx}-\Sigma_{yy})^2+4\det(\Sigma)}} \\ &= \frac{2\Sigma_{xy}^2}{\mathrm{Tr}(\Sigma)+\sqrt{(\mathrm{Tr}(\Sigma))^2-4\Sigma_{xy}^2}} \\ &\geq \frac{2\Sigma_{xy}^2}{\mathrm{Tr}(\Sigma)+\sqrt{(\mathrm{Tr}(\Sigma))^2}} = \frac{\Sigma_{xy}^2}{\mathrm{Tr}(\Sigma)}, \end{aligned}$$

which concludes the proof. ■

Remark 2: Robust estimation and *measuring dependency* problems (which are closely related) can benefit from the proposed results. The minimax theorem (e.g., [10]) can be used to minimize the worst-case mean square estimation error across all distributions in a given Wasserstein/Gelbrich ambiguity set.

The main contribution of this section is a straightforward formula that provides the minimal Gelbrich distance from one bi-variate distribution to uncorrelation. The next section provides an application of this analytical expression in the context of robust system identification.

V. Illustrative Example

By means of a system identification example, this section illustrates the novel insights that can be obtained according to the results of this paper.

Suppose a set of noisy data $\{z_k\}_{k=1}^N$ and $\{u_k\}_{k=1}^N$ corresponding to the output and input of a given dynamical system. Consider an estimator $\hat{z}_k$ for z_k obtained by means of a linear regression of the covariates z_{k-1}, z_{k-2} and u_{k-1}

$$\hat{z}_k = -\hat{a}_1 z_{k-1} - \hat{a}_2 z_{k-2} + \hat{b}_1 u_{k-1}. \tag{12}$$

Assume that the coefficients $\hat{a}_1$, $\hat{a}_2$ and $\hat{b}_1$ have already been identified on the training set by any standard identification approach (such as the least square method [21]). In this context, a relevant task is to determine if it is worthwhile to include an additional term in the nominal estimator, e.g., one depending on z_{k-3}. Denote by estimator 1 the nominal estimator (12) and by estimator 2 the following estimator with a supplementary term (obtained with the same identification approach as the nominal estimator)

$$\tilde{z}_k = -\tilde{a}_1 z_{k-1} - \tilde{a}_2 z_{k-2} - \tilde{a}_3 z_{k-3} + \tilde{b}_1 u_{k-1}. \tag{13}$$

Estimator 2 is expected to perform better than estimator 1 under the assumption that covariate z_{k-3} is a relevant ingredient for the prediction of z_k. However, assessing if the incorporation of a given covariate (like z_{k-3}) would translate into a consistently improved performance of the estimator is not simple. To address this question, we first compute the estimation error e_k on the training set, associated with the nominal estimator (12)

$$e_k = z_k - \hat{z}_k = z_k + \hat{a}_1 z_{k-1} + \hat{a}_2 z_{k-2} - \hat{b}_1 u_{k-1},\ k = 3, \ldots, N.$$

Then, we compute the mean of $\{e_k\}_{k=3}^N$ and $\{z_{k-3}\}_{k=4}^N$, and the covariance matrix between $\{e_k\}_{k=3}^N$ and $\{z_{k-3}\}_{k=4}^N$. This leads to the first two moments of the nominal bi-variate distribution between the error e_k and z_{k-3} (denoted by μ_x, μ_y, and Σ in the notation of the paper). The reader can notice that the obtained moments are just an estimation because of the limitations in the number of samples N, the noise in the measurements, etc. Moreover, in a time-varying system, the first moments relating e_k and z_k might vary with time. We suppose now that a bound on the Gelbrich distance α to the real parameters μ_S and Σ_S is available. We notice that the existing methodologies to obtain α are often of probabilistic nature since they are usually based on concentration inequalities (see, e.g., [17], [22], [23], [24]).

Using Theorem 1 we compute the minimal Gelbrich distance $D_{\mathcal{U}}$ of the nominal pair (μ_Σ, Σ) to uncorrelation. When comparing $D_{\mathcal{U}}$ to α, two cases are observed:

- $\alpha \geq D_{\mathcal{U}}$ (case 1): In this case, the ambiguity set contains a bi-variate distribution with no correlation between z_{k-3} and e_k. Thus, we conclude that it might not be advisable to add z_{k-3} as a covariate in the estimator of z_k.
- $\alpha < D_{\mathcal{U}}$ (case 2): In this case, it would be reasonable to add the covariate z_{k-3} in the estimator because the nominal Pearson coefficient does not vanish in the obtained ambiguity set.

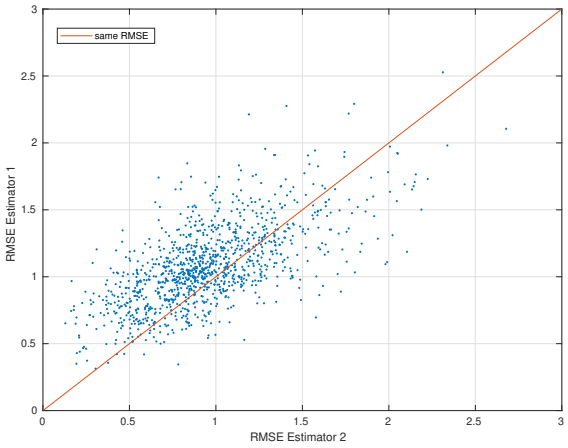

Fig. 2. Comparison RMSE obtained by the estimators for the pairs (μ_S, S) that are at Gelbrich distance larger than $D_{\mathcal{U}}$.

To illustrate how the Gelbrich distance to the nominal values μ_Σ and Σ degrades the performance of estimator 2, which incorporates the additional covariate z_{k-3}, we have randomly generated M pairs (μ_S, S) and computed for each of them its Gelbrich distance to (μ_Σ, Σ) and the corresponding Root Mean Square Error (RMSE) when using estimators 1 and 2. On one hand, our simulations show that the performance of estimator 2 usually outperforms estimator 1 when the Gelbrich distance is smaller than $D_{\mathcal{U}}$, the Gelbrich distance to uncorrelation of (μ, Σ). On the other hand, when the Gelbrich distance is larger than $D_{\mathcal{U}}$, then the advantage of using estimator 2 degrades and using estimator 1 becomes a reasonable option since it provides similar performance with a reduced number of parameters. The obtained results are displayed in the following figures.

In Fig. 2 and Fig. 3, the blue markers represent the RMSE obtained by the two estimators for each pair (μ_S, S). The red

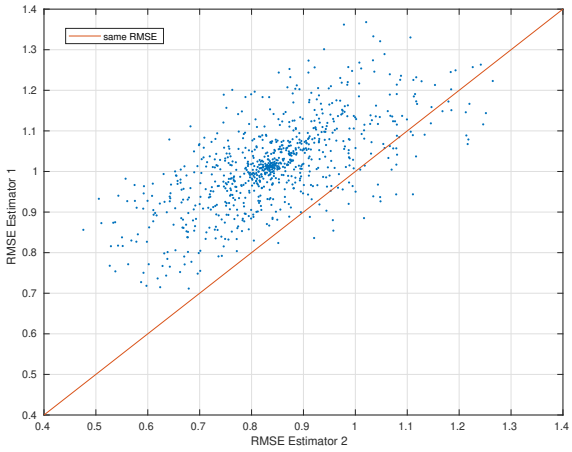

Fig. 3. Comparison RMSE obtained by the estimators for the pairs $(\boldsymbol{\mu_S}, \boldsymbol{S})$ that are at Gelbrich distance no larger than $\boldsymbol{D_{\mathcal{U}}}$.

line indicates when the two RMSE are equal, i.e., it permits to highlight the delimitation. From Fig. 2, when the Gelbrich distance of (μ_S, S) to the nominal pair (μ_Σ, Σ) is greater than the minimum distance to uncorrelation, we can notice that the RMSE of both estimators are similar and no consistent reduction on the RMSE is observed. In Fig. 3, when the Gelbrich distance is no larger than the minimum distance to uncorrelation, the estimator 2 usually has a smaller RMSE than the estimator 1. Summing up, deciding if incorporating a given covariate in an estimator if $\alpha < D_{\mathcal{U}}$ seems, from our simulation experience, an adequate criterion.

VI. Conclusion and Perspectives

This paper has introduced novel insights into the Wasserstein/Gelbrich distance and its ambiguity sets, offering perspective on analyzing correlations between scalar random variables to allow for robust decision making. Through closed expressions, it determines the minimum Gelbrich perturbation needed to reach uncorrelation between two random scalar variables. Numerical simulations validate these concepts, highlighting their relevance in the realm of estimation and control. This work not only enhances our understanding of this distance metric but also underscores its practical value across various fields, e.g., data analysis [9], robust estimation and sensor fusion [10], [11]. Current work focuses on extending the results for new estimation and control applications.

VII. Acknowledgment

The authors acknowledge support from the STIC doctoral school / Université Paris-Saclay, ONERA, Agence de l'innovation de défense (Grant 2022-65-0011-AID_ONERA), Agence Nationale de Recherche (ANR)-France (Grant ANR-21-CE48-0003), and the European Research Council under the advanced grant OCONTSOLAR Grant agreement ID: 789051. T. Alamo acknowledges support from grant PID2022-142946NA-I00 funded by MCIN/AEI/ 10.13039/501100011033 and by ERDF, A way of making Europe.

References

[1] J. Lee Rodgers and W. A. Nicewander, "Thirteen ways to look at the correlation coefficient," *The American Statistician*, vol. 42, no. 1, pp. 59–66, 1988.

[2] D. Kuhn, P. M. Esfahani, V. A. Nguyen, and S. Shafieezadeh-Abadeh, "Wasserstein distributionally robust optimization: Theory and applications in machine learning," in *Operations Research & Management Science in the Age of Analytics*, pp. 130–166, INFORMS, 2019.

[3] Y. V. Prokhorov, "Convergence of random processes and limit theorems in probability theory," *Theory of Probability & Its Applications*, vol. 1, no. 2, pp. 157–214, 1956.

[4] S.-H. Cha, "Comprehensive survey on distance/similarity measures between probability density functions," *International Journal of Mathematical models and Methods in Applied Sciences*, vol. 1, no. 4, pp. 300–307, 2007.

[5] C. Villani, *Topics in optimal transportation*. Graduate studies in mathematics, Providence, Rhode Island: American mathematical society, C 2003.

[6] M. Gelbrich, "On a formula for the L2 Wasserstein metric between measures on Euclidean and Hilbert spaces," *Mathematische Nachrichten*, vol. 147, no. 1, pp. 185–203, 1990.

[7] V. A. Nguyen, S. Shafieezadeh-Abadeh, D. Kuhn, and P. Mohajerin Esfahani, "Bridging Bayesian and minimax mean square error estimation via Wasserstein distributionally robust optimization," *Mathematics of Operations Research*, vol. 48, no. 1, pp. 1–37, 2023.

[8] Y. Cai and L.-H. Lim, "Distances between probability distributions of different dimensions," *IEEE Transactions On Information Theory*, vol. 68, no. 6, pp. 4020–4031, 2022.

[9] M. Arjovsky, S. Chintala, and L. Bottou, "Wasserstein generative adversarial networks," in *Proceedings of the 34th International Conference on Machine Learning* (D. Precup and Y. W. Teh, eds.), vol. 70, pp. 214–223, PMLR, 2017.

[10] S. Shafieezadeh Abadeh, V. A. Nguyen, D. Kuhn, and P. M. Mohajerin Esfahani, "Wasserstein distributionally robust Kalman filtering," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[11] D.-J. Xin, L.-F. Shi, and X. Yu, "Distributed Kalman filter with faulty/reliable sensors based on Wasserstein average consensus," *IEEE Transactions on Circuits and Systems II: Express Briefs*, vol. 69, no. 4, pp. 2371–2375, 2022.

[12] L. Aolaritei, M. Fochesato, J. Lygeros, and F. Dörfler, "Wasserstein Tube MPC with exact uncertainty propagation," 2023.

[13] F. Fabiani and B. Franci, "On distributionally robust generalized Nash games defined over the Wasserstein ball," *Journal of Optimization Theory and Applications*, pp. 1–12, 2023.

[14] T. Rachev, S. Stoyanov, and F. Fabozzi, *A Probability Metrics Approach to Financial Risk Measures*. Wiley-Blackwell, 2011.

[15] B. K. Poolla, A. R. Hota, S. Bolognani, D. S. Callaway, and A. Cherukuri, "Wasserstein distributionally robust look-ahead economic dispatch," *IEEE Transactions on Power Systems*, vol. 36, no. 3, pp. 2010–2022, 2021.

[16] J. Lee, N. P. Bertrand, and C. J. Rozell, "Unbalanced optimal transport regularization for imaging problems," *IEEE Transactions on Computational Imaging*, vol. 6, pp. 1219–1232, 2020.

[17] H. Rahimian and S. Mehrotra, "Frameworks and results in distributionally robust optimization," *Open Journal of Mathematical Optimization*, vol. 3, pp. 1–85, 2022.

[18] J. C. Duchi and H. Namkoong, "Learning models with uniform performance via distributionally robust optimization," *The Annals of Statistics*, vol. 49, no. 3, pp. 1378–1406, 2021.

[19] C. R. Givens and R. M. Shortt, "A class of Wasserstein metrics for probability distributions.," *Michigan Mathematical Journal*, vol. 31, no. 2, pp. 231–240, 1984.

[20] R. A. Horn and C. R. Johnson, *Matrix analysis*. Cambridge University Press, second ed., 2013.

[21] L. Ljung, *System identification: Theory for the user*. Pearson Education, 1998.

[22] N. Fournier and A. Guillin, "On the rate of convergence in Wasserstein distance of the empirical measure," *Probability Theory and Related Fields*, vol. 162, 2013.

[23] R. Ji and M. Lejeune, "Data-driven optimization of reward-risk ratio measures," *INFORMS Journal on Computing*, vol. 33, 2020.

[24] V. A. Nguyen, S. Shafiee, D. Filipović, and D. Kuhn, "Mean-covariance robust risk measurement," *arXiv, 2112.09959*, pp. 1–68, 2023.