Coordinate-wise transformation of probability distributions to achieve a Stein-type identity

Tomonari Sei¹

Received: 20 December 2019 / Revised: 1 March 2021 / Accepted: 17 June 2021 /
Published online: 28 June 2021
© The Author(s) 2021

Abstract

It is shown that for any given multi-dimensional probability distribution with regularity conditions, there exists a unique coordinate-wise transformation such that the transformed distribution satisfies a Stein-type identity. A sufficient condition for the existence is referred to as copositivity of distributions. The proof is based on an energy minimization problem over a totally geodesic subset of the Wasserstein space. The result is considered as an alternative to Sklar's theorem regarding copulas, and is also interpreted as a generalization of a diagonal scaling theorem. The Stein-type identity is applied to a rating problem of multivariate data. A numerical procedure for piece-wise uniform densities is provided. Some open problems are also discussed.

Keywords Copositive distribution · Copula · Energy minimization · Optimal transportation · Stein-type distribution · Wasserstein space

Mathematics Subject Classification 60E05 · 62E10

1 Introduction

One of the important concepts in information geometry is the maximum entropy principle. The unique probability distribution that maximizes entropy under fixed moments is characterized by an exponential family ([15]). The exponential and mixture families induce a dually flat structure in the space of probability distributions (e.g. [2]).

We establish a version of the maximum entropy principle over the family of distributions generated by coordinate-wise transformations. The coordinate-wise transformations naturally arise in copula theory to make distributions have uniform marginals (e.g. [27]) and are considered as a subset of optimal transport maps.

✉ Tomonari Sei
sei@mist.i.u-tokyo.ac.jp

¹ Graduate School of Information Science and Technology, The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, Japan

Before going into details, we first consider a linear analogue of the problem. Marshall and Olkin [25] proved the following diagonal scaling theorem on matrices.

Theorem 1 ([25]) *Let $S = (S_{ij}) \in \mathbb{R}^{d \times d}$ be a positive semi-definite matrix and assume that S is strictly copositive in the sense that*

$$\inf_{w_1, \dots, w_d > 0} \frac{\sum_i \sum_j w_i S_{ij} w_j}{\sum_i w_i^2} > 0. \quad (1)$$

Then, there exists a unique positive-definite diagonal matrix D such that

$$\sum_{j=1}^d (DSD)_{ij} = 1 \quad (2)$$

for each $i \in \{1, \dots, d\}$.

Note that (1) is satisfied if S is positive definite. The equation (2) is, as pointed out by [20], the stationary condition of a convex function

$$\psi(w) = \sum_i (-\log w_i) + \frac{1}{2} \sum_i \sum_j w_i S_{ij} w_j, \quad (3)$$

where $w = (w_1, \dots, w_d)$ is the diagonal component of D .

Theorem 1 is interpreted in a probabilistic framework as follows. Let μ be a normal distribution on $\mathbb{R}^d$ with mean zero and covariance matrix $S_{ij} = \int x_i x_j d\mu$. Let ν be the push-forward measure of μ by a linear transformation $x \mapsto Dx$ in $\mathbb{R}^d$. Since the covariance matrix of ν is DSD , the equation (2) is rewritten as

$$\sum_{j=1}^d \int x_i x_j d\nu = 1 \quad (4)$$

for each $i \in \{1, \dots, d\}$. In other words, each coordinate x_i and the sum $\sum_j x_j$ have unit covariance under the law ν .

In the present paper, we provide a nonlinear analogue of the theorem. We admit a nonlinear monotone coordinate-wise transformation to achieve a stronger condition than (4). The condition will be referred to as the Stein-type identity (see Sect. 2 for the precise definition). Under some mild conditions on μ , it is shown that there exists such a unique transformation and it minimizes a free energy functional like (3), which has an entropy term. The space we use in the proof is the Wasserstein space, a distance space induced from optimal transportation. A key observation is that our functional is displacement convex in the sense of [26]. Refer to [30, 38] for comprehensive studies of optimal transportation and its applications.

Under the Stein-type identity, the sum of variables has positive correlation with each variable. This property is applied to a rating problem of multivariate data in Sect. 3.

As is well known, Sklar’s theorem (see, e.g., [27]) states that any multi-dimensional distribution is transformed by the probability integral transformation into a distribution with uniform marginals. The transformed distribution is called a copula. A linear analogue of Sklar’s theorem is that for any covariance matrix S there exists a unique positive-definite diagonal matrix D such that every diagonal element of DSD is unity. This is nothing but the correlation matrix corresponding to S .

There are some papers relevant to our study. A relation between copula and diagonal scaling is investigated in [4] from different perspectives. Their scaling operation does not correspond to transformation of random variables. Optimal transportation between two distributions sharing the same copula is considered in [1], where the various cost functions are the center of discussion. Optimal transportation is used to determine multi-dimensional quantiles in [7] and [13]. Although our motivation is also to define a kind of quantile functions of multivariate data, the construction is different from theirs (see Sect. 3). A particular class of optimal transport maps called moment maps has a deep connection to another Stein-type identity as investigated in [10].

The remainder of the present paper is organized as follows. In Sect. 2, we define the Stein-type distributions and transformations. In Sect. 3, we briefly explain its application to a rating problem of multivariate data. In Sect. 4, we describe the existence and uniqueness theorem as well as a variational characterization theorem. In Sect. 5, we prove the main results using the theory of optimal transportation. In Sect. 6, a numerical method to find the transformation for piecewise uniform distributions is proposed. Finally, we discuss open problems in Sect. 7.

2 Definition of Stein-type distributions and transformations

We define a class of distributions that satisfy a stronger condition than (4). Let $\mathcal{P}^2 = \mathcal{P}^2(\mathbb{R}^d)$ be the set of probability distributions μ on $\mathbb{R}^d$ with mean zero and finite second moments such that each marginal distribution μ_i of μ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}$. Note that μ itself is not assumed to be absolutely continuous. The mean-zero condition is imposed only for simplicity. We say that a function $f : \mathbb{R} \rightarrow \mathbb{R}$ is absolutely continuous if there exists a locally integrable function f' such that $f(x) = f(0) + \int_0^x f'(y)dy$ in Lebesgue’s sense.

Definition 1 A distribution $\mu \in \mathcal{P}^2$ is said to be *Stein-type* if it satisfies

$$\int f(x_i) \left(\sum_{j=1}^d x_j \right) d\mu = \int f'(x_i) d\mu, \quad i = 1, \dots, d, \quad (5)$$

for any absolutely continuous function $f : \mathbb{R} \rightarrow \mathbb{R}$ with bounded derivative f' .

Note that the equation (4) is a special case of (5) with $f(x_i) = x_i$.

We refer to the equation (5) as the *Stein-type identity*. Indeed, if $d = 1$, it reduces to the Stein identity $\int f(x_1)x_1 d\mu = \int f'(x_1) d\mu$, which implies that μ is the standard normal distribution (see [35] and [6]). The Stein identity is used to evaluate distance between a given distribution and the normal distribution. Although the Stein-type

identity we defined is a generalization of the Stein identity, the author does not aware of its applications to the distance evaluation. Instead, we develop a different application. More specifically, if a random vector $(X_1, \dots, X_d)$ has a Stein-type distribution, then the sum $\sum_j X_j$ is positively correlated with any monotone transformation of X_i due to (5). This property is applied to a rating problem in Sect. 3.

If μ is completely independent in the sense that μ is the direct product of its marginals μ_i , then the Stein-type distribution has to be the d -dimensional standard normal distribution. Hereafter, we focus on dependent cases.

For Gaussian random variables, we obtain the following lemma. We denote the expectation by E .

Lemma 1 (Theorem 5 of [33]) *Let μ denote the d -dimensional normal distribution with mean zero and covariance matrix $S = (S_{ij})$. Then, μ is Stein-type if and only if $\sum_j S_{ij} = 1$ for each i .*

Proof Let $X = (X_i)$ be a random vector distributed according to μ . Then, the conditional expectation of X_j given X_i is $E[X_j|X_i] = S_{ij}X_i/S_{ii}$. The left hand side of (5) is

$$\begin{aligned} E \left[f(X_i) \sum_j E[X_j|X_i] \right] &= \frac{\sum_j S_{ij}}{S_{ii}} E[f(X_i)X_i] \\ &= \left(\sum_j S_{ij} \right) E[f'(X_i)], \end{aligned}$$

where the last equality follows from the Stein identity for one-dimensional case. Hence (5) holds if and only if $\sum_j S_{ij} = 1$. $\square$

The following example gives a rich class of Stein-type distributions.

Example 1 Let W be a random variable with the standard normal distribution and let U be any random variable independent of W such that $E[U] = 0$ and $E[U^2] < \infty$. The condition $E[U] = 0$ is assumed to make the following distribution belong to $\mathcal{P}^2$ and not essential here. Consider two variables

$$X_1 = \frac{W + U}{\sqrt{2}} \quad \text{and} \quad X_2 = \frac{W - U}{\sqrt{2}}.$$

Then the distribution of (X_1, X_2) is Stein-type. Indeed, we obtain

$$E \left[f \left(\frac{W \pm U}{\sqrt{2}} \right) \sqrt{2} W \right] = E \left[f' \left(\frac{W \pm U}{\sqrt{2}} \right) \right]$$

for any f by the one-dimensional Stein identity with respect to W conditional on U . The variable $W = (X_1 + X_2)/\sqrt{2}$ has a meaning of “an overall score” of the two variables X_1 and X_2 . The identity implies that W is positively correlated with any increasing function $f(X_i)$ of X_i .

This example is generalized to the case $d \geq 3$. Define a random vector $(X_1, \dots, X_d)$ by $X_i = (W + U_i)/\sqrt{d}$, where W has the standard normal distribution independent of $U_1, \dots, U_{d-1}$ and $U_d = -\sum_{j=1}^{d-1} U_j$. Then, the distribution of $(X_1, \dots, X_d)$ is Stein-type as long as $E[U_i] = 0$ and $E[U_i^2] < \infty$. As in the two-dimensional case, W has positive correlation with any increasing function $f(X_i)$ of X_i . In Sect. 3, we will show by an example that the positive-correlation property does not hold in general for copulas.

The example does not cover the entire class of Stein-type distributions. Other examples are given in Sect. 6 and Appendix B.

For each $\mu \in \mathcal{P}^2$, let $\mathcal{T}_{\text{cw}}(\mu)$ be the set of coordinate-wise transformations

$$T(x) = (T_1(x_1), \dots, T_d(x_d)), \quad x = (x_1, \dots, x_d) \in \mathbb{R}^d,$$

such that each $T_i : \mathbb{R} \rightarrow \mathbb{R} \cup \{-\infty, \infty\}$ is non-decreasing and $T_{\#}\mu$ belongs to $\mathcal{P}^2$. Here, $T_{\#}\mu$ is the push-forward measure defined by $(T_{\#}\mu)(A) = \mu(T^{-1}(A))$ for any measurable set A . The set $\mathcal{T}_{\text{cw}}(\mu)$ depends only on the marginal distributions of μ . Two maps T and U in $\mathcal{T}_{\text{cw}}(\mu)$ are identified if $\mu(T = U) = 1$. Note that T_i may have discontinuous points if the support of the i -th marginal $(T_{\#}\mu)_i$ is not connected.

We consider a problem to find a map $T \in \mathcal{T}_{\text{cw}}(\mu)$ such that $T_{\#}\mu$ is Stein-type.

Definition 2 A map $T \in \mathcal{T}_{\text{cw}}(\mu)$ is called a *Stein-type transformation* of μ if $T_{\#}\mu$ is a Stein-type distribution.

For example, if μ is the product measure of one-dimensional continuous distributions μ_i , then the map T defined by $T_i(x_i) = \Phi^{-1}(\mu_i((-\infty, x_i]))$ is the Stein-type transformation of μ , where Φ is the cumulative distribution function of the standard normal distribution. The map is nothing but the Brenier map between the two product measures.

The following lemma is immediate.

Lemma 2 Let μ be the normal distribution with a covariance matrix $S = (S_{ij})$. Then, μ has a Stein-type transformation if S is strictly copositive in the sense of (1).

Proof Let D be the diagonal matrix with entries $w_1, \dots, w_d$ satisfying (4). Set $T(x) = Dx$. Then, $T_{\#}\mu$ is Stein-type due to Lemma 1. $\square$

3 Application to a rating problem

In this section, we briefly describe an application of Stein-type transformations. We first explain a linear rating method of multivariate data according to [33]. Let S be a covariance matrix of an $\mathbb{R}^d$ -valued random vector $x = (x_i)$. Suppose that each variable x_i has a meaning of “score”. For example, x_1, x_2 and x_3 are student scores of mathematics, physics and history, and so forth. We want to determine positive weights w_i such that $\sum_j w_j x_j$ reflects the scores $x_1, \dots, x_d$. A candidate of such a weight w_i is the i -th diagonal element of D in (2). Indeed, under (2), the variable x_i and

the overall score $\sum_j w_j x_j$ have positive covariance for each i . This property is not attained in general for other methods of weighting. The quantity $\sum_j w_j x_j$ is called the *objective general index* (OGI) in [33], which reflects the d scores in this sense.

In a similar manner, we can define a nonlinear version of the objective general index via Stein-type transformations. Let μ be a probability distribution on $\mathbb{R}^d$ and x be a random vector distributed according to μ . Again each coordinate x_i is assumed to have a meaning of score. Then the nonlinear general index is defined by

$$g(x) = \sum_j T_j(x_j),$$

where $T = (T_i)$ is the Stein-type transformation of μ . The quantity $g(x)$ satisfies $E[g(X)f(X_i)] \geq 0$ for any increasing function $f(x_i)$ of x_i due to the Stein-type identity (5). In particular, by taking the step function $f(x_i) = I_{[a, \infty)}(x_i)$ and noting that $E[g(X)] = 0$, we obtain

$$E[g(X) \mid X_i < a] \leq E[g(X) \mid X_i \geq a] \quad (6)$$

for each $i \in \{1, \dots, d\}$ and $a \in \mathbb{R}$. The property means that the conditional average of the overall score of students taking higher score on x_i is larger than that of students taking lower score. In this sense, $g(x)$ reflects the d scores.

General indices $g(x)$ satisfying the inequality (6) are not unique. Indeed, for two-dimensional distributions, the probability integral transformation $T_i(x_i) = \mu_i((-\infty, x_i])$ provides (6); see Appendix D for a proof. However, for higher dimensional distributions, it is not trivial to find such a transformation. The Stein-type transformation solves the problem. We confirm this point by an example.

Example 2 Suppose that x_1, x_2 and x_3 are random variables that have a probability distribution μ on $[-1, 1]^3$ with the density function

$$f(x_1, x_2, x_3) = \begin{cases} 3/8 & \text{if } (x_1, x_2, x_3) \in ([-1, 0] \times [0, 1]^2) \cup ([0, 1] \times [-1, 0]^2), \\ 1/24 & \text{otherwise.} \end{cases}$$

The three marginal distributions are uniform over $[-1, 1]$, i.e., μ is a copula. We can see

$$E \left[\sum_j X_j \mid X_1 < 0 \right] = \frac{1}{6} > E \left[\sum_j X_j \mid X_1 > 0 \right] = -\frac{1}{6}.$$

Hence the property in (6) does not hold if we adopt $g(x) = \sum_i x_i$. As will be demonstrated in Sect. 6, the unique Stein-type transformation of μ is

$$T_i(x_i) = \begin{cases} -c_i + \Phi^{-1}((1 + x_i)\Phi(c_i)) & \text{if } x_i < 0, \\ c_i - \Phi^{-1}((1 - x_i)\Phi(c_i)) & \text{if } x_i > 0 \end{cases}$$

with $c_1 = 1.2490$ and $c_2 = c_3 = 0.3445$, where Φ denotes the standard normal distribution function and c_1, c_2 are numerically obtained. The nonlinear objective general index $g(x) = \sum_i T_i(x_i)$ satisfies the relation (6) as a result.

4 Main results

For given $\mu \in \mathcal{P}^2$, denote the set of coordinate-wise transformed distributions of μ by

$$\mathcal{F}_\mu = \{T_\# \mu \mid T \in \mathcal{T}_{\text{cw}}(\mu)\} \subset \mathcal{P}^2.$$

We refer to $\mathcal{F}_\mu$ as a *fiber*. The following lemma is a direct consequence of the one-dimensional optimal transportation. See Appendix A.

Lemma 3 *For given $\mu \in \mathcal{P}^2$ and $\nu \in \mathcal{F}_\mu$, the map $T \in \mathcal{T}_{\text{cw}}(\mu)$ satisfying $\nu = T_\# \mu$ is uniquely determined μ -almost everywhere. Furthermore, the relation $\nu \in \mathcal{F}_\mu$ between two measures μ and ν is an equivalence relation. In particular, $\mathcal{P}^2$ is partitioned into mutually disjoint fibers.*

Now, we state our three main theorems. All proofs are presented in Sect. 5.

The first main theorem characterizes Stein-type distributions in terms of the variational principle. Define an energy functional $\mathcal{E}(\mu)$ of μ by

$$\mathcal{E}(\mu) = \sum_{i=1}^d \int p_i(x_i) \log p_i(x_i) dx_i + \int \left(\sum_{j=1}^d x_j \right)^2 d\mu, \quad (7)$$

where $p_i = d\mu_i/dx_i$ is the marginal density function. The first term of $\mathcal{E}(\mu)$ represents the negative entropy of the marginal distributions, and the second term is the variance of the diagonal part $\sum_j x_j$. The functional has displacement convexity over each fiber in the sense of [26]. See Sect. 5 for details. If we replace the entropy term with the joint entropy, the functional becomes displacement convex over the whole space as proved by McCann [26].

Theorem 2 *A measure $\mu \in \mathcal{P}^2$ is Stein-type if and only if $\mathcal{E}(\mu)$ is finite and μ minimizes $\mathcal{E}$ over the fiber $\mathcal{F}_\mu$.*

The second main theorem is on the uniqueness of Stein-type transformations. A distribution μ on $\mathbb{R}^d$ is said to have a *regular support* if the support of μ is equal to the direct product of the supports of the marginal distributions μ_i . This property is invariant under coordinate-wise transformations. Note that the regular support condition does not imply absolute continuity of μ with respect to $\prod_{i=1}^d \mu_i$.

Theorem 3 (Uniqueness) *Assume that $\mu \in \mathcal{P}^2$ has a regular support. Then, a Stein-type transformation of μ is unique if it exists.*

We conjecture that the uniqueness follows without the regular support condition. See Sect. 7 for more details.

The third main theorem is on existence. A measure $\mu \in \mathcal{P}^2$ is said to be *copositive* if

$$\beta(\mu) := \inf_{T \in \mathcal{T}_{\text{cw}}(\mu)} \frac{\int (\sum_i T_i(x_i))^2 d\mu}{\sum_i \int T_i(x_i)^2 d\mu} > 0. \quad (8)$$

For example, if μ is pairwise independent, then $\int (\sum_i T_i(x_i))^2 d\mu = \sum_i \int T_i(x_i)^2 d\mu$ for any $T \in \mathcal{T}_{\text{cw}}(\mu)$, and therefore $\beta(\mu) = 1$. It is not difficult to see that $\beta(\mu) \leq 1$ for any μ . If μ is associated in the sense of [9, 11], and [21], then $\int T_i T_j d\mu \geq 0$ for each pair of i and j , and therefore $\beta(\mu) \geq 1$. On the other hand, if $d = 2$ and $\mu(\{x \mid x_1 + x_2 = 0\}) = 1$, then $\beta(\mu) = 0$ because $\int (x_1 + x_2)^2 d\mu = 0$. Sufficient conditions for copositivity are presented in Appendix E.

Theorem 4 (Existence) *Let $\mu \in \mathcal{P}^2$ be copositive. Then, there exists a Stein-type transformation of μ .*

We now present a few remarks before proceeding to the following section.

The uniqueness and existence results in Theorem 3 and Theorem 4 are consequences of the variational characterization in Theorem 2, as will be shown in Sect. 5. For $d = 1$, the functional $\mathcal{E}(\mu)$ is the Kullback-Leibler divergence from μ to the standard normal density up to a constant term. For $d \geq 2$, however, $\mathcal{E}$ is not even bounded from below. Indeed, for each $t > 0$, let μ^t be the multivariate normal distribution with mean zero and covariance matrix $\Sigma_t = P + t(I - P)$, where I is the identity matrix and

$$P = \frac{1}{d} \begin{pmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{pmatrix} \in \mathbb{R}^{d \times d}.$$

Then, each marginal distribution of μ^t is normal with variance $\sigma_t^2 = (1/d) + t(1 - 1/d)$. We can show that $\mathcal{E}(\mu^t) = -(d/2) \log(2\pi\sigma_t^2)$, which tends to $-\infty$ as $t \rightarrow \infty$. Therefore, it is not trivial if there is a minimizer of $\mathcal{E}$ over the fiber. Nevertheless, the existence and uniqueness theorems are obtained.

If μ has the joint density function $p(x)$ with respect to the Lebesgue measure, then the negative joint entropy is defined by

$$\mathcal{U}_d(\mu) = \int p(x) \log p(x) dx.$$

In most cases, we can replace the marginal entropy term in $\mathcal{E}(\mu)$ with the joint entropy because the difference $\mathcal{U}_d(\mu) - \sum_{i=1}^d \mathcal{U}_1(\mu_i)$, which is referred to as the multi-information function or the measure of multivariate dependence, is invariant in each fiber (e.g., [16] and [36]). However, in some pathological cases, the difference diverges. Therefore, it is more appropriate to adopt the marginal entropy.

According to Sklar's theorem (e.g. [27]), any d -dimensional distribution μ is transformed by the probability integral transformation $T_i(x_i) = \int_{-\infty}^{x_i} d\mu_i$ into the distribution $T_{\#}\mu$ with uniform marginals unless some μ_i has an atom. The resultant distribution $T_{\#}\mu$ is a copula. The Stein-type distribution we defined is considered as an alternative representation of the copula. Copulas are also characterized by an energy minimization problem. Here, the potential term in (7) is replaced with $\int V(x)d\mu$, where $V(x) = \infty$ if $x \notin [0, 1]^d$ and 0 otherwise. In parallel, we have to remove the condition $\int x_i d\mu_i = 0$ from the definition of $\mathcal{P}^2$. Maximum entropy copulas under a given diagonal section are discussed in [5], where, in contrast to the present paper, the marginals are fixed to be uniform.

5 Proofs based on the theory of optimal transportation

In this section, we prove the three main theorems stated in Sect. 4. The proof is based on the theory of optimal transportation. Necessary facts about one-dimensional optimal transportation are summarized in Appendix A.

5.1 Regularity of Stein-type distributions

The Stein-type identity forces regularity of marginal density functions. We first characterize this by an integral equation.

Theorem 5 *Let $\mu \in \mathcal{P}^2$. Denote the marginal density functions of μ with respect to the Lebesgue measure by $p_i(x_i) = d\mu_i/dx_i$. Then, μ is a Stein-type distribution if and only if it satisfies a set of integral equations*

$$p_i(a) = \int_a^\infty p_i(x_i)m_i(x_i)dx_i, \quad a \in \mathbb{R}, \quad i = 1, \dots, d, \quad (9)$$

where $m_i(x_i)$ denotes the conditional expectation of $\sum_{j=1}^d x_j$ given x_i with respect to μ .

Proof First, note that $m_i(x_i)$ is finite μ_i -almost everywhere because μ belongs to $\mathcal{P}^2$.

Assume μ is Stein-type. For $-\infty < a < b < \infty$, let $h_{ab}(x) = (b - a)^{-1} \int_{-\infty}^x I_{(a,b)}(\xi)d\xi$, where $I_{(a,b)}$ is the indicator function of (a, b) . The Stein-type identity for h_{ab} implies

$$\begin{aligned} \int h'_{ab}(x_i)d\mu_i &= \int h_{ab}(x_i) \sum_j x_j d\mu \\ &= \int h_{ab}(x_i)m_i(x_i)p_i(x_i)dx_i. \end{aligned}$$

Letting $b \rightarrow a$, we obtain (9).

Conversely, assume (9). The right-hand side of (9) converges to zero as $a \rightarrow \pm\infty$ because $\int x_j d\mu_j = 0$ for all j . Then, for any bounded and absolutely continuous

function f with bounded derivative f' , we obtain the Stein-type identity for f :

$$\begin{aligned}\int f'(a)p_i(a)da &= \int_{-\infty}^{\infty} f'(a) \left(\int_a^{\infty} p_i(x_i)m_i(x_i)dx_i \right) da \\ &= \int_{-\infty}^{\infty} f(x_i)p_i(x_i)m_i(x_i)dx_i \\ &= \int f(x_i) \sum_j x_j d\mu,\end{aligned}$$

where the second equality follows from the integral-by-parts formula. If f is not bounded, then let $f_M(x) = f(0) + \int_0^x f'(u)1_{\{|u| \leq M\}}du$ and take $M \rightarrow \infty$. $\square$

As a corollary, the regularity of the marginal density functions is established.

Corollary 1 *Let μ be Stein-type. Then, its marginal density functions $p_i(x_i)$ are bounded, absolutely continuous, and converge to zero as $x_i \rightarrow \pm\infty$.*

Proof From the formula (9), it is obvious that p_i is absolutely continuous and bounded by $\int \sum_i |x_i| d\mu_i < \infty$. We also have $p_i(x_i) \rightarrow 0$ as $x_i \rightarrow \pm\infty$ because the right-hand side of (9) vanishes as $a \rightarrow \pm\infty$. $\square$

Although the marginal density function of any Stein-type distribution is absolutely continuous, it can have non-differentiable points as shown in an example in Sect. 6. The continuous differentiability of $p_i(x_i)$ follows from the regularity of the pair-wise copula of μ along formula (9). We do not pursue this line of investigation here. On the other hand, we conjecture that the marginal density of any Stein-type distribution is positive everywhere. See Sect. 7 for more details.

The following corollary will be used later.

Corollary 2 *Let μ be Stein-type. Then, its negative marginal entropy $\int p_i(x_i) \log p_i(x_i) dx_i$ is finite.*

Proof Since the marginal density $p_i(x_i)$ is bounded, we have $\int p_i(x_i) \log p_i(x_i) dx_i < \infty$. To prove $\int p_i(x_i) \log p_i(x_i) dx_i > -\infty$, we use the non-negativity of the Kullback-Leibler information from p_i to the standard normal density $\phi(x_i) = e^{-x_i^2/2}/\sqrt{2\pi}$. Indeed,

$$\begin{aligned}\int p_i(x_i) \log p_i(x_i) dx_i &\geq \int p_i(x_i) \log \phi(x_i) dx_i \\ &= \int \{-(1/2) \log(2\pi) - x_i^2/2\} d\mu_i \\ &> -\infty\end{aligned}$$

because $\int x_i^2 d\mu_i < \infty$. $\square$

Other properties of Stein-type distributions are given in Appendix C.

5.2 Variational problem over a fiber of Wasserstein space

Let $\mathcal{F}$ be a fiber of $\mathcal{P}^2$ (see Sect. 4 for the definition) and choose two measures μ and ν in $\mathcal{F}$, where ν is written as $\nu = T_{\#}\mu$ with some $T \in \mathcal{T}_{\text{cw}}(\mu)$ by definition. Define the geodesic, which is also referred to as the displacement interpolation [26], from μ to ν by

$$[\mu, \nu]_t = [(1-t)\text{Id} + tT]_{\#}\mu, \quad t \in [0, 1],$$

where Id denotes the identity map. Based on the one-dimensional optimal transportation, it follows that $[\mu, \nu]_t \in \mathcal{F}$ and $[\mu, \nu]_t = [\nu, \mu]_{1-t}$ for each t . A functional on $\mathcal{F}$ is said to be *displacement convex* if it is convex along a geodesic. Refer to [26] and [3] for further details on displacement convexity.

Although a geodesic between any pair of distributions in $\mathcal{P}^2$ is similarly defined, we need only geodesics in a common fiber. It is known that a geodesic actually attains the minimum length of a path between two measures with respect to the L^2 -Wasserstein distance (see e.g. [3] and [38]). Here the L^2 -Wasserstein distance is the infimum of $(\int \|x - y\|^2 d\gamma(x, y))^{1/2}$ over the set of joint distributions γ on $\mathbb{R}^{2d}$ with the marginals μ and ν . Each fiber $\mathcal{F}$ is totally geodesic in the sense of [37].

Recall that μ is said to have a regular support if its support is the direct product of the supports of marginal distributions.

Lemma 4 *Let $\mathcal{F}$ be a fiber and choose any two distributions μ and ν in $\mathcal{F}$, where $\mu \neq \nu$. Then, $\mathcal{E}([\mu, \nu]_t)$ is convex in t , that is, $\mathcal{E}$ is displacement convex over $\mathcal{F}$. Furthermore, $\mathcal{E}([\mu, \nu]_t)$ is strictly convex if one of the following conditions is satisfied:*

- (i) μ (and therefore ν) has a regular support, or
- (ii) the supports of μ_i and ν_i are connected, respectively, for each i .

Proof Let $\nu = T_{\#}\mu$, with $T \in \mathcal{T}_{\text{cw}}(\mu)$. Let $p_i = d\mu_i/dx_i$ be the marginal density of μ . By the change-of-variable formula (Lemma 7 in Appendix A), we obtain

$$\mathcal{E}([\mu, \nu]_t) = \sum_i \int p_i(x_i) \log \frac{p_i(x_i)}{(1-t) + tT'_i(x_i)} dx_i + \frac{1}{2} \int \left(\sum_i ((1-t)x_i + tT_i(x_i)) \right)^2 d\mu, \quad (10)$$

where $T'_i(x_i)$ is the derivative of T_i if it exists, and $T'_i(x_i) = 0$ otherwise. Both terms in (10) are convex in t .

Assume (i) and that $\mathcal{E}([\mu, \nu]_t)$ is not strictly convex. Then, there is an interval over which $\mathcal{E}([\mu, \nu]_t)$ is linear. It is deduced from (10) that $\sum_{i=1}^d (T_i(x_i) - x_i) = 0$, μ -almost everywhere. Let I be the set of indices i such that $\mu_i(T_i(x_i) \neq x_i) > 0$. Then, I is not empty because $T \neq \text{Id}$. For each $i \in I$, the probability $\mu_i(T_i(x_i) - x_i > 0)$ is positive because $\int (T_i(x_i) - x_i) d\mu_i = 0$. Then, by the regular support condition, we have $\mu(\sum_{i \in I} (T_i(x_i) - x_i) > 0)$ is positive. However, this contradicts $\sum_{i=1}^d (T_i(x_i) - x_i) = 0$. Thus, $\mathcal{E}([\mu, \nu]_t)$ should be strictly convex under (i).

Next, assume (ii). Then, T_i has no discontinuous points. Assume $\mathcal{E}([\mu, \nu]_t)$ is not strictly convex. Then, it follows from (10) that $T'_i(x_i) = 1$ and, therefore, $T_i(x_i) = x_i$ by the connectedness of the support together with the condition $\int T_i d\mu_i = 0$. However, this contradicts $\mu \neq \nu$. Thus, $\mathcal{E}([\mu, \nu]_t)$ is strictly convex. $\square$

Example 3 The strict convexity of $\mathcal{E}([\mu, \nu]_t)$ can fail if neither condition (i) nor condition (ii) in Lemma 4 is satisfied. For example, let $d = 2$ and assume that μ is uniformly distributed over the region $([-1, 0] \times [0, 1]) \cup ([0, 1] \times [-1, 0])$. Define the map T by $T_i(x_i) = x_i + 1$ if $x_i > 0$, and $T_i(x_i) = x_i - 1$ otherwise for each i . Let $\nu = T_\# \mu$. Then, $\mathcal{E}([\mu, \nu]_t)$ is constant along $t \in [0, 1]$ because $T'_i(x_i) = 1$ and $\sum_i T_i(x_i) = \sum_i x_i$, μ -almost everywhere. In this case, μ_i is supported on $[-1, 1]$, whereas ν_i is supported on $[-2, -1] \cup [1, 2]$.

5.3 Proof of Theorem 2

Let μ be a Stein-type distribution. Corollary 2 implies that μ belongs to $\text{dom} \mathcal{E}$. From the convexity (Lemma 4), it is sufficient to show that

$$\left. \frac{d}{dt_+} \mathcal{E}([\mu, \nu]_t) \right|_{t=0} \geq 0$$

for any $\nu = T_\# \mu \in \mathcal{F}$, where d/dt_+ denotes the right derivative. It follows from formula (10) that

$$\left. \frac{d}{dt_+} \mathcal{E}([\mu, \nu]_t) \right|_{t=0} = - \sum_i \int p_i(x_i) (T'_i(x_i) - 1) dx_i + \sum_i \sum_j \int (T_i(x_i) - x_i) x_j d\mu. \quad (11)$$

If T_i is absolutely continuous, the right-hand side vanishes by the Stein-type identity, where the boundedness of the derivatives T'_i can be assumed by a standard approximation argument, as in the proof of Theorem 5. If T_i is not absolutely continuous, T_i can be decomposed into an absolutely continuous part and a discontinuous part as $T_i = T_i^{\text{ac}} + T_i^{\text{d}}$. See Appendix A. The contribution of T_i^{ac} in (11) vanishes due to the Stein-type identity. It is sufficient to prove that $\sum_j \int T_i^{\text{d}}(x_i) x_j d\mu \geq 0$ for each i because $(T_i^{\text{d}})' = 0$ by definition. We can take a sequence $\{f_{i,n}\}_{n=1}^\infty$ of non-decreasing differentiable functions with a bounded derivative such that $f_{i,n}(x_i)$ converges to $T_i^{\text{d}}(x_i)$ μ -almost everywhere. More specifically, a step function $I_{[\xi, \infty)}(x_i)$ at each $\xi \in \mathbb{R}$ is approximated by a logistic function $1/(1 + \exp(-n(x_i - \xi)))$. Then, by Lebesgue's dominated convergence theorem and the Stein-type identity, we obtain

$$\begin{aligned} \sum_j \int T_i^{\text{d}}(x_i) x_j d\mu &= \lim_{n \rightarrow \infty} \sum_j \int f_{i,n}(x_i) x_j d\mu \\ &= \lim_{n \rightarrow \infty} \int f'_{i,n}(x_i) d\mu \\ &\geq 0. \end{aligned}$$

Conversely, assume that $\mathcal{E}(T_{\sharp}^*\mu)$ is minimized at $T = \text{Id}$. Fix $1 \leq i \leq d$, and let f be an absolutely continuous function with bounded derivative f' . For sufficiently small $\varepsilon > 0$, two maps $T(x) = x \pm \varepsilon f(x_i)e_i$, where e_i is the i -th unit vector, belong to $\mathcal{T}_{\text{cw}}(\mu)$. Thus, the right derivative (11) has to be zero, and μ satisfies the Stein-type identity.

5.4 Proof of Theorem 3

Assume that μ has a regular support and admits a Stein-type transformation T . Then, Theorem 2 implies that $T_{\sharp}^*\mu$ minimizes $\mathcal{E}$ over the fiber $\mathcal{F}_{\mu}$. However, it is deduced from Lemma 4 that $\mathcal{E}$ is strictly convex over $\mathcal{F}_{\mu}$. Thus, the minimizer is unique.

5.5 Proof of Theorem 4

Assume that μ is copositive. Denote the functional $\mathcal{E}$ restricted to the fiber $\mathcal{F}_{\mu}$ by $\mathcal{E}_{\mu}$. From Theorem 2, it is sufficient to show that $\mathcal{E}_{\mu}$ has a minimum point. We first show that $\mathcal{E}_{\mu}$ is bounded from below and that the level set $\{v \mid \mathcal{E}_{\mu}(v) \leq c\}$ for each $c \in \mathbb{R}$ is tight. For any $v \in \mathcal{F}_{\mu}$, the copositivity condition implies

$$\mathcal{E}_{\mu}(v) \geq \sum_{i=1}^d \left\{ \int q_i(x_i) \log q_i(x_i) dx_i + \frac{\beta}{2} \int x_i^2 dv_i \right\},$$

where $q_i = dv_i/dx_i$ and $\beta = \beta(v) = \beta(\mu) > 0$. We obtain

$$\begin{aligned} \int q_i(x_i) \log q_i(x_i) dx_i &= \int q_i(x_i) \log \frac{q_i(x_i)}{\sqrt{\beta/(4\pi)} e^{-\beta x_i^2/4}} dx_i - \frac{\beta}{4} \int x_i^2 dv_i + \frac{1}{2} \log \frac{\beta}{4\pi} \\ &\geq -\frac{\beta}{4} \int x_i^2 dv_i + \frac{1}{2} \log \frac{\beta}{4\pi}, \end{aligned}$$

where the last inequality follows from the nonnegativity of the Kullback-Leibler divergence. Then, $\mathcal{E}_{\mu}$ is bounded from below as

$$\mathcal{E}_{\mu}(v) \geq C + \frac{\beta}{4} \sum_{i=1}^d \int x_i^2 dv_i,$$

where C is a constant independent of v . This inequality also implies that the level set $\{v \mid \mathcal{E}_{\mu}(v) \leq c\}$ is tight.

Now there exists a weakly converging sequence v_k such that $\mathcal{E}_{\mu}(v_k)$ converges to $\inf \mathcal{E}_{\mu}(v)$. Let v_* be the weak limit. Then, Corollary 3.5 of [26] shows that $v_* \in \mathcal{P}^2$ and $\mathcal{E}_{\mu}(v_*) \leq \lim_k \mathcal{E}_{\mu}(v_k)$. The distribution v_* gives a minimum point of $\mathcal{E}_{\mu}$. This completes the proof.

6 Piecewise uniform densities

In this section, it is shown that if μ has piecewise uniform density function, then the Stein-type transformation of μ is obtained by finite-dimensional optimization. Here, we do not impose the zero mean condition $\int x_i d\mu = 0$ on μ . We can always translate it to have zero mean if necessary.

We say that a probability density function $c(u)$ on $[0, 1]^d$ is piecewise uniform if its two-dimensional marginal densities c_{ij} ($1 \leq i < j \leq d$) are written as

$$c_{ij}(u_i, u_j) = n^2 \pi_{ab}^{ij} \quad \text{if } (u_i, u_j) \in \left(\frac{a-1}{n}, \frac{a}{n}\right] \times \left(\frac{b-1}{n}, \frac{b}{n}\right], \quad a, b \in \{1, \dots, n\}, \quad (12)$$

for some n , where π_{ab}^{ij} is a positive number such that

$$\sum_{a=1}^n \sum_{b=1}^n \pi_{ab}^{ij} = 1.$$

Let $\pi_a^i = \sum_{b=1}^n \pi_{ab}^{ij}$. Although c is not a copula density unless $\pi_a^i = 1/n$ for all i and a , it is transformed by a piecewise linear transform into a copula density. Then Corollary 3 in Appendix E, together with Theorem 4, guarantees the existence of a Stein-type transformation as long as the support of $c(u)$ is $[0, 1]^d$.

By solving Equation (9), we obtain an expression of the Stein-type transformation of c as follows. Denote the cumulative distribution function and density function of the standard normal distribution by Φ and ϕ , respectively.

Lemma 5 *Suppose that $c(u)$ satisfies (12) and its support is $[0, 1]^d$. Let p be the unique Stein-type density transformed from c . Then, there exist real constants $\alpha_{1i}, \dots, \alpha_{ni}$ and $\xi_{1i} < \dots < \xi_{n-1,i}$ such that*

$$p_i(x_i) = \pi_a^i \frac{\phi(x_i - \alpha_{ai})}{Z_{ai}} \quad \text{for } \xi_{a-1,i} < x_i \leq \xi_{ai} \quad (13)$$

where $\xi_{0i} = -\infty$, $\xi_{ni} = \infty$, and $Z_{ai} = \Phi(\xi_{ai} - \alpha_{ai}) - \Phi(\xi_{a-1,i} - \alpha_{ai})$. The Stein-type transformation is

$$x_i = T_i(u_i) = \alpha_{ai} + \Phi^{-1} \left(\Phi(\xi_{a-1,i} - \alpha_{ai}) + n(u_i - \frac{a-1}{n})Z_{ai} \right), \quad u_i \in \left(\frac{a-1}{n}, \frac{a}{n}\right], \quad (14)$$

and the two-dimensional marginal density is

$$p_{ij}(x_i, x_j) = \pi_{ab}^{ij} \frac{\phi(x_i - \alpha_{ai})}{Z_{ai}} \frac{\phi(x_j - \alpha_{bj})}{Z_{bj}}, \quad (x_i, x_j) \in (\xi_{a-1,i}, \xi_{ai}] \times (\xi_{b-1,j}, \xi_{bj}]. \quad (15)$$

Furthermore, the following identity is satisfied:

$$\alpha_{ai} = - \sum_{j \neq i} \sum_{b=1}^n \frac{\pi_{ab}^{ij}}{\pi_a^i} \int_{\xi_{b-1,j}}^{\xi_{bj}} \frac{x_j \phi(x_j - \alpha_{bj})}{Z_{bj}} dx_j. \quad (16)$$

Proof Equation (9) implies that $\partial_i p_i(x_i) = -(x_i + \sum_{j \neq i} E[X_j | x_i]) p_i(x_i)$, where $\partial_i = \partial / \partial x_i$. Since the conditional expectation $E[X_j | x_i]$ has to be piecewise constant, $p_i(x_i)$ is piecewise Gaussian up to a normalizing constant. Since the mass of each piece is preserved under a coordinate-wise transformation, we obtain the form (13). Then, the unique monotone transformation (14) is derived from $c_i(u_i) du_i = p_i(x_i) dx_i$. Equation (15) results from the transformation of $c_{ij}(u_i, u_j)$. Finally, Equation (16) is obtained from $\partial_i \log p_i(x_i) = -(x_i + \sum_{j \neq i} E[X_j | x_i])$. $\square$

The parameters α_{ai} and ξ_{ai} are determined by the continuity of (13) at $x_i = \xi_{ai}$ and the identity (16). However, instead of solving the simultaneous equations directly, we adopt an optimization approach.

Assume the density of a distribution μ obeys the parametric form given by Equation (13). Then, the energy function $\mathcal{E}(\mu)$ defined in Sect. 5 is a function of α and ξ , which is denoted by $F(\alpha, \xi)$ and is obtained as follows:

$$\begin{aligned} F(\alpha, \xi) &= \sum_i \int p_i(x_i) \log p_i(x_i) + \frac{1}{2} \sum_i \int x_i^2 p_i(x_i) dx_i + \sum_{i < j} \int x_i x_j p_{ij}(x_i, x_j) dx_i dx_j \\ &= \sum_i \sum_a \pi_a^i \int_{\xi_{a-1,i}}^{\xi_{ai}} \frac{\phi(x_i - \alpha_{ai})}{Z_{ai}} \left(\log \frac{\pi_a^i}{\sqrt{2\pi}} - \frac{(x_i - \alpha_{ai})^2}{2} - \log Z_{ai} + \frac{x_i^2}{2} \right) \\ &\quad + \sum_{i < j} \sum_a \sum_b \frac{\pi_{ab}^{ij}}{Z_{ai} Z_{bj}} \int_{\xi_{a-1,i}}^{\xi_{ai}} \int_{\xi_{b-1,j}}^{\xi_{bj}} x_i x_j \phi(x_i - \alpha_{ai}) \phi(x_j - \alpha_{bj}) dx_i dx_j \\ &= \sum_i \sum_a \pi_a^i \log \frac{\pi_a^i}{\sqrt{2\pi}} + \sum_i \sum_a \pi_a^i \left(-\frac{\alpha_{ai}^2}{2} + \alpha_{ai} M_{ai} - \log Z_{ai} \right) \\ &\quad + \sum_{i < j} \sum_a \sum_b \pi_{ab}^{ij} M_{ai} M_{bj}, \end{aligned}$$

where

$$\begin{aligned} M_{ai} &= \frac{1}{Z_{ai}} \int_{\xi_{a-1,i}}^{\xi_{ai}} x_i \phi(x_i - \alpha_{ai}) dx_i \\ &= \alpha_{ai} + \frac{1}{Z_{ai}} (-\phi(\xi_{ai} - \alpha_{ai}) + \phi(\xi_{a-1,i} - \alpha_{ai})). \end{aligned}$$

Since Z_{ai} and M_{ai} are functions of three parameters α_{ai} , ξ_{ai} , and $\xi_{a-1,i}$, we denote the corresponding partial derivative by D_1 , D_2 , and D_3 . The derivatives of F are

$$\begin{aligned} \frac{\partial F}{\partial \alpha_{ai}} &= \pi_a^i \left(-\alpha_{ai} + M_{ai} + \alpha_{ai} D_1 M_{ai} - \frac{D_1 Z_{ai}}{Z_{ai}} \right) + \sum_{j \neq i} \sum_b \pi_{ab}^{ij} (D_1 M_{ai}) M_{bj} \\ &= \pi_a^i \left(\alpha_{ai} + \sum_{j \neq i} \sum_b \frac{\pi_{ab}^{ij}}{\pi_a^i} M_{bj} \right) (D_1 M_{ai}), \end{aligned} \quad (17)$$

$$\begin{aligned} \frac{\partial F}{\partial \xi_{ai}} &= \pi_a^i \left(\alpha_{ai} D_2 M_{ai} - \frac{D_2 Z_{ai}}{Z_{ai}} \right) + \pi_{a+1}^i \left(\alpha_{a+1,i} D_3 M_{a+1,i} - \frac{D_3 Z_{a+1,i}}{Z_{a+1,i}} \right) \\ &\quad + \sum_{j \neq i} \sum_b \left\{ \pi_{ab}^{ij} (D_2 M_{ai}) + \pi_{a+1,b}^{ij} (D_3 M_{a+1,i}) \right\} M_{bj} \\ &= \pi_a^i \left(\alpha_{ai} + \sum_{j \neq i} \sum_b \frac{\pi_{ab}^{ij}}{\pi_a^i} M_{bj} \right) D_2 M_{ai} + \pi_{a+1}^i \left(\alpha_{a+1,i} + \sum_{j \neq i} \sum_b \frac{\pi_{a+1,b}^{ij}}{\pi_{a+1}^i} M_{bj} \right) D_3 M_{a+1,i} \\ &\quad - \pi_a^i \frac{D_2 Z_{ai}}{Z_{ai}} - \pi_{a+1}^i \frac{D_3 Z_{a+1,i}}{Z_{a+1,i}}. \end{aligned} \quad (18)$$

By using these formulas, we obtain the following theorem.

Theorem 6 A stationary point of F together with formula (13) provides the global minimum point of the energy functional $\mathcal{E}(\mu)$ over the fiber. In other words, F has a unique stationary point that corresponds to the Stein-type density.

Proof Since $M_{ai} = \int x_i \phi(x_i - \alpha_{ai}) dx_i / Z_{ai}$ is the expectation parameter of an exponential family $\phi(x_i - \alpha_{ai}) / Z_{ai}$, it is an increasing function of α_{ai} (e.g., [23]). Therefore, $D_1 M_{ai} > 0$. Thus, the stationary condition $\partial F / \partial \alpha_{ai} = 0$ is equivalent to

$$\alpha_{ai} + \sum_{j \neq i} \sum_b \frac{\pi_{ab}^{ij}}{\pi_a^i} M_{bj} = 0,$$

which is equivalent to (16) and solves the integral equation (9) except at boundary points ξ_{ai} . Furthermore, substituting this relation into (18), we obtain

$$\begin{aligned} \frac{\partial F}{\partial \xi_{ai}} &= -\pi_a^i \frac{D_2 Z_{ai}}{Z_{ai}} - \pi_{a+1}^i \frac{D_3 Z_{a+1,i}}{Z_{a+1,i}} \\ &= -p_i(\xi_{ai}-) + p_i(\xi_{ai}+). \end{aligned}$$

Therefore, $\partial F / \partial \xi_{ai} = 0$ is equivalent to the continuity of p_i at ξ_{ai} . Then, the density p is the Stein-type density, which is unique due to Theorem 3. $\square$

The minimization problem of $F(\alpha, \xi)$ over $\alpha_{ai} \in \mathbb{R}$ and $\xi_{1i} < \dots < \xi_{n-1,i}$ is performed using a standard optimization package (e.g., the function `optim` in R [29]) when the coordinate $\tau_{ai} = \xi_{ai} - \xi_{a-1,i}$, rather than ξ_{ai} , is used for $2 \leq a \leq n-1$.

Fig. 1 Coordinate-wise transformation for the two-dimensional discretized Clayton copula with $\theta = 2$ and $n = 10$ is shown. The joint density function (top-left) is transformed into a Stein-type density function (top-right) by the inverse function of a cumulative distribution (bottom-left), the density of which is piece-wise Gaussian (bottom-right). The dashed curve in the bottom figures represents the standard normal distribution

Example 4 We numerically obtain the Stein-type densities of discretized copulas. The result is shown in Fig. 1. The copula used here is the Clayton copula

$$C_{\theta}(x_1, x_2) = \left[\max(x_1^{-\theta} + x_2^{-\theta} - 1, 0) \right]^{-1/\theta}.$$

The discretized copula density of $n \times n$ cells is given by (12) with

$$\pi_{ab}^{12} = C_{\theta}\left(\frac{a}{n}, \frac{b}{n}\right) - C_{\theta}\left(\frac{a-1}{n}, \frac{b}{n}\right) - C_{\theta}\left(\frac{a}{n}, \frac{b-1}{n}\right) + C_{\theta}\left(\frac{a-1}{n}, \frac{b-1}{n}\right).$$

7 Discussion

In the present paper, we showed that a class of multi-dimensional distributions has a unique representation via the Stein-type identity. Now, we describe areas for future study and some open problems.

In Sect. 5.1, we derived some properties of Stein-type distributions. The author could not find any counter-example against the following conjecture.

Conjecture 1 The marginal density function of any Stein-type distribution is positive everywhere.

A partial answer to Conjecture 1 is given in the following lemma.

Lemma 6 Let μ be a Stein-type distribution. If the copula of μ has pair-wise marginal densities c_{ij} such that

$$D = \sup_{i,j:i \neq j} \sup_{u_i \in [0,1]} \int_0^1 c_{ij}(u_i, u_j)^2 du_j < \infty,$$

then each marginal density p_i of μ is positive everywhere. In particular, if the copula density of μ is bounded, then the same consequence follows.

Proof The density $p_i(x_i)$ satisfies $\partial_i p_i(x_i) + p_i(x_i)m_i(x_i) = 0$ with $m_i(x_i) = E[\sum_j X_j | x_i]$ by Theorem 5. The conditional expectation satisfies

$$|E[X_j | x_i]| = \left| \int x_j c_{ij}(F_i(x_i), F_j(x_j)) p_j(x_j) dx_j \right| \leq (DE[X_j^2])^{1/2},$$

where $F(x_i) = \int_{-\infty}^{x_i} p_i(\xi) d\xi$. Let $D_* = \sum_{j \neq i} (DE[X_j^2])^{1/2}$. Then, we obtain an inequality

$$-(x_i + D_*)p_i(x_i) \leq \partial_i p_i(x_i) \leq -(x_i - D_*)p_i(x_i)$$

Let $a \in \mathbb{R}$ be a point at which $p_i(a) > 0$. Then, Gronwall's lemma shows that $p_i(x_i) \geq p_i(a)e^{-(x_i+D_*)^2/2+(a+D_*)^2/2} > 0$ for $x_i > a$, and similarly $p_i(x_i) > 0$ for $x_i < a$. $\square$

If Conjecture 1 is positively solved, then the following conjecture, which is based on Theorem 3, is also positive according to Lemma 4 (ii).

Conjecture 2 A Stein-type transformation is unique if it exists.

We state a relevant conjecture that is the converse of Theorem 4.

Conjecture 3 A distribution is copositive if it has a Stein-type transformation.

In Sect. 5, we showed that a Stein-type distribution is characterized by the stationary point of an energy functional $\mathcal{E}$ over a fiber $\mathcal{F}$. From the perspective of optimal transportation, we can construct the gradient flow of the energy functional with respect to the L^2 -Wasserstein space ([19,28] and [38]). The formal equation is as follows

$$\partial_t p_i = \partial_i(\partial_i p_i + m_i p_i), \quad t \geq 0, \quad i = 1, \dots, d, \quad (19)$$

where $m_i(x_i) = \mathbb{E}[\sum_j X_j | x_i]$. Although this appears to be an independent system of one-dimensional Fokker-Planck equations, the equations interact with each other via $m_i(x_i)$. The physical meaning of the equation is not clear. From Theorem 5, it follows that each Stein-type density is a stationary point of (19). The time evolution will be theoretically of interest.

In Appendix E, we presented sufficient conditions for copositivity of distributions. In particular, a Gaussian distribution is copositive if its covariance matrix is not degenerated. Conversely, if a Gaussian distribution is copositive, then the covariance matrix must, by definition, be strictly copositive (see Equation (1)). The following conjecture naturally arises but is not proven. This is positively solved if Conjecture 3 is correct, due to Lemma 2.

Conjecture 4 A Gaussian distribution is copositive if the covariance matrix is strictly copositive.

As stated in Sect. E.4, tail-dependent copulas do not satisfy the sufficient condition in Theorem 7. The copositivity of tail-dependent copulas remains unclear.

In the present paper, we did not consider statistical models that explain a given data set. A statistical model involving a Stein-type distribution is essentially equivalent to a copula model because such models correspond to each other through coordinate-wise transformations, whereas the marginal distributions are not of much interest in copula modelling. The class given in Example 1 provides a flexible model because the distribution of U_i 's in the construction can be selected arbitrarily.

Finally, it is expected that there is a coordinate-wise transformation to satisfy

$$\mathbb{E}[f(X_i)g(X_1 + \dots + X_d)] > 0, \quad i = 1, \dots, d,$$

for any monotone increasing functions f and g . If g is fixed, we can make a similar discussion as the present paper (see [34]).

A One-dimensional optimal transportation

Necessary information about one-dimensional optimal transportation is summarized. Refer to [30] and [38] for further details.

Let $\mathcal{P}^2(\mathbb{R})$ be the set of absolutely continuous probability distributions μ on $\mathbb{R}$ such that $\int x d\mu = 0$ and $\int x^2 d\mu < \infty$. For given $\mu \in \mathcal{P}^2(\mathbb{R})$, let $\mathcal{T}(\mu)$ be the set of non-decreasing functions $T : \mathbb{R} \rightarrow \mathbb{R} \cup \{-\infty, \infty\}$ such that $T_{\#}\mu \in \mathcal{P}^2(\mathbb{R})$.

For given μ and ν in $\mathcal{P}^2(\mathbb{R})$, there exists $T \in \mathcal{T}(\mu)$ such that $\nu = T_{\#}\mu$. The map is uniquely determined μ -almost everywhere. More explicitly, T is given by

$T = G^- \circ F$, where $F(x) = \int_{-\infty}^x d\mu$, $G(x) = \int_{-\infty}^x dv$, and $G^-(u) = \inf\{x \in \mathbb{R} \mid G(x) > u\}$. The map T is called the optimal transportation from μ to ν because this map minimizes the functional $\int (T(x) - x)^2 d\mu$ over $\{T \mid T_{\#}\mu = \nu\}$. Since μ and ν are absolutely continuous, T is decomposed into an absolutely continuous part, T^{ac} , and a discontinuous part, T^{d} , without a singular continuous part. This is because G^- constructed above has the same property. The decomposition is unique up to a μ -negligible set.

The following lemmas are used in Sect. 5 and Sect. E. These lemmas were originally proven for multi-dimensional measures but here we simplify them for the one-dimensional case.

Lemma 7 (Theorem 4.4 of [26]) *For given μ and ν in $\mathcal{P}^2(\mathbb{R})$, let T be a unique monotone map such that $\nu = T_{\#}\mu$. Let p and q be density functions of μ and ν , respectively. Let $X \subset \mathbb{R}$ denote the set of points where the derivative T' is defined and positive. Then, $\mu(X) = 1$. Furthermore,*

$$\int A(q(y))dy = \int_X A\left(\frac{p(x)}{T'(x)}\right) T'(x)dx$$

for any measurable function A on $[0, \infty)$ with $A(0) = 0$.

Lemma 8 (Proposition 1.3 of [26]) *Let $\mu \in \mathcal{P}^2$ and $T \in \mathcal{T}(\mu)$. Then, $(1-t)\text{Id} + tT \in \mathcal{T}(\mu)$ for each $t \in [0, 1]$.*

Lemma 9 (Proposition 4.2 of [26]) *Let $\mu \in \mathcal{P}^2$. If $T : \mathbb{R} \rightarrow \mathbb{R}$ is a non-decreasing function written as $T = T^{\text{ac}} + T^{\text{d}}$ and the derivative $(T^{\text{ac}})'$ of the absolutely continuous part is strictly positive μ -almost everywhere, then $T_{\#}\mu$ is absolutely continuous.*

B Explicit expression of Stein-type distributions

We formally derive an explicit expression of the Stein-type distributions.

Assume that $\mu \in \mathcal{P}^2$ has a smooth density function p with decay at infinity. Then, μ is Stein-type if and only if there exists a function $r(x)$ such that

$$\sum_{j=1}^p \left(\frac{\partial p(x)}{\partial x_j} + x_j p(x) \right) = r(x), \quad \int_{\mathbb{R}^{d-1}} r(x) dx_{-i} = 0 \quad \text{for all } i, \quad (20)$$

where dx_{-i} means $dx_1 \cdots dx_{i-1} dx_{i+1} \cdots dx_d$. In fact, formula (20) is rewritten as $\partial p_i / \partial x_i + p_i(x_i) m_i(x_i) = 0$, where $m_i(x_i)$ is the conditional expectation of $\sum_j x_j$ given x_i , and this equation is equivalent to (9).

Equation (20) is explicitly solved if $r(x)$ is given. Let Q be a fixed orthogonal matrix such that $(Q^{\top}x)_1 = \sum_j x_j / \sqrt{d}$, where $Q^{\top}$ denotes the matrix transpose of Q . Then, (20) is written as

$$\frac{\partial p(Qw)}{\partial w_1} + w_1 p(Qw) = \frac{r(Qw)}{\sqrt{d}}, \quad w = Q^{\top}x.$$

The general solution is

$$p(Qw) = \frac{1}{\sqrt{2\pi}} e^{-w_1^2/2} \left(q(w_2, \dots, w_d) + \int_0^{w_1} \frac{\sqrt{2\pi}}{\sqrt{d}} e^{v_1^2/2} r(Q(v_1, w_2, \dots, w_d)) dv_1 \right),$$

where q is any probability density function on $\mathbb{R}^{d-1}$.

In particular, if $r(x) = 0$, we obtain a simple formula

$$p(x) = \frac{1}{\sqrt{2\pi}} e^{-w_1^2/2} q(w_2, \dots, w_d), \quad w = Q^\top x. \quad (21)$$

Example 1 in Sect. 4 is this solution. The class of densities (21) is characterized by a stronger condition than the Stein-type identity, i.e.,

$$\int f(x) \sum_j x_j d\mu = \int \sum_i \frac{\partial f(x)}{\partial x_i} d\mu$$

for any function $f(x) = f(x_1, \dots, x_d)$.

C Properties of Stein-type distributions

We provide some properties of Stein-type distributions that are not used in the main stream of the paper.

We first point out that Stein-type distributions have finite Fisher information. The Fisher information of a density function q on $\mathbb{R}$ is defined by

$$I(q) = \int \left(\frac{q'(x)}{q(x)} \right)^2 q(x) dx,$$

where q is assumed to be absolutely continuous, and $q'(x)/q(x)$ is set to 0 if q is not differentiable or not positive at x . See [18] for properties implied by finite Fisher information. Note that the Fisher information we defined is that of location family $\{q(x - \theta) \mid \theta \in \mathbb{R}\}$ in statistics (e.g., [23]).

Lemma 10 *For any Stein-type distribution μ , the Fisher information $I(p_i)$ of each marginal density p_i is bounded by the dimension d . In particular, p_i has bounded variation.*

Proof From (9), the score function $p'_i(x_i)/p_i(x_i)$ is equal to $-m_i(x_i)$. Since $m_i(x_i)$ is the conditional expectation of $\sum_j x_j$ given x_i , we obtain

$$I(p_i) = \int m_i(x_i)^2 d\mu \leq \int \left(\sum_j x_j \right)^2 d\mu = d,$$

where the last equality follows from the Stein-type identity with $f(x_i) = x_i$. By the Cauchy-Schwarz inequality, we also have $\int |p'_i(x_i)| dx_i \leq \sqrt{I(p_i)}$. Then, p_i has bounded variation. $\square$

Let $\mathcal{S}$ be the set of Stein-type distributions on $\mathbb{R}^d$. We prove that $\mathcal{S}$ is closed under mixture, normalized convolution, and weak limit.

Lemma 11 (Mixture) *If μ and ν are two distributions in $\mathcal{S}$, then $(1-t)\mu + t\nu$ belongs to $\mathcal{S}$ for any $t \in [0, 1]$.*

Proof This follows from the linearity of the Stein-type identity (5) with respect to μ . $\square$

Lemma 12 (Normalized convolution) *Let $X = (X_1, \dots, X_d)$ and $Y = (Y_1, \dots, Y_d)$ be independent random vectors with Stein-type distributions. Let a and b be real numbers with $a^2 + b^2 = 1$. Then, $aX + bY$ has a Stein-type distribution.*

Proof The Stein-type identity with respect to X implies that

$$\mathbb{E} \left[f(aX_i + bY_i) \left(\sum_j X_j \right) \right] = a \mathbb{E} [f'(aX_i + bY_i)]$$

for each i , because X and Y are independent. By changing the roles of X and Y , we have

$$\mathbb{E} \left[f(aX_i + bY_i) \left(\sum_j Y_j \right) \right] = b \mathbb{E} [f'(aX_i + bY_i)]$$

Their average is

$$\mathbb{E} \left[f(aX_i + bY_i) \left(\sum_j (aX_j + bY_j) \right) \right] = (a^2 + b^2) \mathbb{E} [f'(aX_i + bY_i)].$$

Thus, the Stein-type identity for $aX + bY$ holds if and only if $a^2 + b^2 = 1$. $\square$

The set $\mathcal{S}$ is also closed under weak limit in the following sense. Denote the Euclidean norm on $\mathbb{R}^d$ by $\|x\|$ for $x \in \mathbb{R}^d$.

Lemma 13 (Weak convergence) *Let $\mu^{(n)}$ be a sequence in $\mathcal{S}$. If $\mu^{(n)}$ converges to μ in law and $\int \|x\|^2 d\mu^{(n)}$ converges to $\int \|x\|^2 d\mu < \infty$, then μ belongs to $\mathcal{S}$.*

Proof These conditions imply that $\int \varphi d\mu^{(n)} \rightarrow \int \varphi d\mu$ for any continuous function φ such that $|\varphi(x)| \leq C(1 + \|x\|^2)$ for some $C > 0$. (Refer to Theorem 7.12 of [38].) Letting $\varphi(x)$ be $f(x_i) \sum_j x_j$ and $f'(x_i)$, respectively, we obtain the Stein-type identity for μ . Absolute continuity of μ_i is shown in the same manner as in the proof of Theorem 5. $\square$

The condition regarding moment convergence in Lemma 13 is necessary. Indeed, we can construct a sequence $(W, U^{(n)})$ of Stein-type random variables in the same manner as in Example 1 of Sect. 4 such that $U^{(n)}$ converges in law to a random variable U with $E[U^2] = \infty$.

By Lemma 12 and Lemma 13 together with the central limit theorem, if we have independent and identically distributed samples $X^1, \dots, X^n$ according to a Stein-type distribution μ , then the limit distribution of $(X^1 + \dots + X^n)/\sqrt{n}$ is a Stein-type normal distribution that is characterized by Lemma 1.

Note that the set of copulas satisfies the same consequence as Lemma 11 and Lemma 13. If we modify the definition of the copulas in such a way that the marginal distribution is standard normal, then the same consequence as Lemma 12 also follows.

D A property of two-dimensional copulas

We prove that the simple sum $g(x) = x_1 + x_2$ satisfies the inequality (6) for any two-dimensional continuous copula density function $c(x_1, x_2)$.

Lemma 14 *Let $c(x_1, x_2)$ be a two-dimensional continuous copula density. Then*

$$\int_0^1 x_2 \left(\int_a^1 c(x_1, x_2) dx_1 \right) dx_2 \geq \frac{1}{2}(1-a)^2$$

holds for any $a \in [0, 1]$.

Proof Let $h_a(x_2) = \int_a^1 c(x_1, x_2) dx_1$ and $H_a(x_2) = \int_0^{x_2} h_a(\eta) d\eta$. Then we have $h_a(x_2) \leq h_0(x_2) = 1$ and $H_a(1) = 1 - a$ by the definition of copula. We also obtain $H_a(x_2) \leq \min(x_2, 1 - a)$. The integral-by-parts formula yields

$$\begin{aligned} \int_0^1 x_2 h_a(x_2) dx_2 &= H_a(1) - \int_0^1 H_a(x_2) dx_2 \\ &\geq 1 - a - \int_0^{1-a} x_2 dx_2 - \int_{1-a}^1 (1-a) dx_2 \\ &= \frac{1}{2}(1-a)^2. \end{aligned}$$

This proves the lemma. □

From the lemma, we have

$$\begin{aligned} \int_0^1 \left(\int_a^1 (x_1 + x_2) c(x_1, x_2) dx_1 \right) dx_2 &= \int_a^1 x_1 dx_1 + \int_0^1 x_2 \left(\int_a^1 c(x_1, x_2) dx_1 \right) dx_2 \\ &\geq \frac{1}{2}(1-a^2) + \frac{1}{2}(1-a)^2 \\ &= 1 - a. \end{aligned}$$

Since $\int_0^1 \int_0^1 (x_1 + x_2) c(x_1, x_2) dx_1 dx_2 = 1$, we obtain $\int_0^1 (\int_0^a (x_1 + x_2) c(x_1, x_2) dx_1) dx_2 \leq a$. Finally, we deduce that

$$\frac{1}{a} \int_0^1 \left(\int_0^a (x_1 + x_2) c(x_1, x_2) dx_1 \right) dx_2 \leq 1 \leq \frac{1}{1-a} \int_0^1 \left(\int_a^1 (x_1 + x_2) c(x_1, x_2) dx_1 \right) dx_2.$$

E Sufficient conditions for copositivity

We present the sufficient conditions for copositivity (8) of a given distribution μ . In Sect. E.1, we first take into account the measures with a non-zero mean as well as coordinate-wise transformations that are constant over an interval. We then present a lower bound of the quantity $\beta(\mu)$ in (8). Subsequent subsections are devoted to finding sufficient conditions for copositivity. Refer to [34] for other sufficient conditions based on positive super-modular dependence.

E.1 Extension of the definition and a lower bound

Let $\mathcal{P}_*^2$ be the set of measures on $\mathbb{R}^d$ such that each marginal μ_i is absolutely continuous and $\int x_i^2 d\mu_i < \infty$ without assuming $\int x_i d\mu_i = 0$. The set $\mathcal{T}_{\text{cw}*}(\mu)$ for $\mu \in \mathcal{P}_*^2$ is defined by the set of coordinate-wise non-decreasing map $T : \mathbb{R}^d \rightarrow \mathbb{R}^d$ such that $\int T_i d\mu_i = 0$ and $\int T_i^2 d\mu_i < \infty$ for each i .

The following lemma is useful to consider copositivity. Denote the inner product and norm of $L^2(\mu)$ by $\langle f, g \rangle = \int f(x)g(x)d\mu$ and $\|f\| = \langle f, f \rangle^{1/2}$, respectively.

Lemma 15 *If $\mu \in \mathcal{P}_*^2$, then*

$$\beta(\mu) = \inf_{0 \neq T \in \mathcal{T}_{\text{cw}*}(\mu)} \frac{\|\sum_i T_i\|^2}{\sum_i \|T_i\|^2}. \quad (22)$$

Proof Denote the right-hand side of (22) by $\delta(\mu)$. Then, it is obvious that $\beta(\mu) \geq \delta(\mu)$ since $\mathcal{T}_{\text{cw}}(\mu) \subset \mathcal{T}_{\text{cw}*}(\mu)$. In order to prove the converse inequality, choose $0 \neq T \in \mathcal{T}_{\text{cw}*}(\mu)$ such that $\|\sum_i T_i\|^2 / (\sum_i \|T_i\|^2) \leq \delta(\mu) + \varepsilon$ for given ε . It follows from Lemma 9 in Appendix A that a map T^η defined by $T^\eta(x) = T(x) + \eta x$ belongs to $\mathcal{T}_{\text{cw}}(\mu)$ for each $\eta > 0$. Then, we have

$$\frac{\|\sum_i T_i\|^2}{\sum_i \|T_i\|^2} = \lim_{\eta \rightarrow 0} \frac{\|\sum_i T_i^\eta\|^2}{\sum_i \|T_i^\eta\|^2},$$

implying $\beta(\mu) \leq \delta(\mu) + \varepsilon$. $\square$

We extend the definition of $\beta(\mu)$ for any $\mu \in \mathcal{P}_*^2$ by (22). In this section, μ is a measure in $\mathcal{P}_*^2$ unless otherwise stated.

Let $L_0^2(\mu_i)$ be the set of functions $T_i : \mathbb{R} \rightarrow \mathbb{R}$ such that $\int T_i d\mu_i = 0$ and $\int T_i^2 d\mu_i < \infty$. The set $\mathcal{T}_{\text{cw}*}(\mu)$ is a subset of $\prod_{i=1}^d L_0^2(\mu_i)$. The space $L_0^2(\mu_i)$ is

considered to be a subspace of $L^2(\mu)$. More precisely, $T_i \in L^2_0(\mu_i)$ is identified with the function $x \mapsto T_i(x_i)$ in $L^2(\mu)$.

By relaxing the set $\mathcal{T}_{\text{cw}*}(\mu)$ in (22), we obtain a lower bound of $\beta(\mu)$ as

$$\beta_L(\mu) := \inf_{0 \neq T \in \prod_{i=1}^d L^2_0(\mu_i)} \frac{\|\sum_i T_i\|^2}{\sum_i \|T_i\|^2} \leq \beta(\mu).$$

Therefore, μ is copositive if $\beta_L(\mu) > 0$.

It is shown that $\beta(\mu)$ and $\beta_L(\mu)$ are invariant under coordinate-wise transformations. Thus, $\beta(\mu)$ and $\beta_L(\mu)$ depend only on the copula of μ . Furthermore, they depend only on the set of two-dimensional marginal copulas of μ .

If $d = 2$, then the quantity $\beta_L(\mu)$ is related to the Hirschfeld-Gebelein-Rényi maximal correlation coefficient (refer to [12,31] and [24])

$$\gamma(\mu) = \sup_{0 \neq T_1 \in L^2_0(\mu_1), 0 \neq T_2 \in L^2_0(\mu_2)} \frac{\langle T_1, T_2 \rangle}{\|T_1\| \|T_2\|}.$$

Lemma 16 *Let $d = 2$. Then, $\beta_L(\mu) = 1 - \gamma(\mu)$. In particular, μ is copositive if $\gamma(\mu) < 1$.*

Proof Let $\gamma = \gamma(\mu)$. For any $T_1 \in L^2_0(\mu_1)$ and $T_2 \in L^2_0(\mu_2)$, we have

$$\begin{aligned} \|T_1 + T_2\|^2 &= \|T_1\|^2 + \|T_2\|^2 + 2\langle T_1, T_2 \rangle \\ &\geq \|T_1\|^2 + \|T_2\|^2 - 2\gamma \|T_1\| \|T_2\| \\ &= \gamma(\|T_1\| - \|T_2\|)^2 + (1 - \gamma)(\|T_1\|^2 + \|T_2\|^2) \\ &\geq (1 - \gamma)(\|T_1\|^2 + \|T_2\|^2). \end{aligned}$$

Thus, we have $\beta_L(\mu) \geq 1 - \gamma$. In order to prove the converse inequality, take sequences T_{1n} and T_{2n} satisfying $\|T_{1n}\| = \|T_{2n}\| = 1$ and $\lim_{n \rightarrow \infty} \langle T_{1n}, T_{2n} \rangle = \gamma$. Then,

$$\beta_L(\mu) \leq \lim_{n \rightarrow \infty} \frac{\|T_{1n} - T_{2n}\|^2}{\|T_{1n}\|^2 + \|T_{2n}\|^2} = 1 - \gamma.$$

□

In the literature, two subspaces H_1 and H_2 of a Hilbert space with the property

$$\sup_{h_1 \in H_1, h_2 \in H_2} \frac{\langle h_1, h_2 \rangle}{\|h_1\| \|h_2\|} < 1$$

are said to satisfy the strengthened Cauchy-Schwarz inequality [8]. In our setting, μ is copositive if $L^2_0(\mu_1)$ and $L^2_0(\mu_2)$ satisfy the strengthened Cauchy-Schwarz inequality.

E.2 Gaussian case

We obtain an explicit expression of $\beta_L(\mu)$ if μ is a multivariate normal distribution.

Lemma 17 *Let μ be the multivariate normal distribution with mean vector 0 and covariance matrix S . Then, $\beta_L(\mu)$ is the minimum eigenvalue of the correlation matrix of S . In particular, μ is copositive if S is non-singular.*

Proof The case of $d = 2$ has been proven by [22].

Assume that all the marginal densities of μ are the standard normal $\phi(x) = (2\pi)^{-1/2}e^{-x^2/2}$ without loss of generality. Then, the covariance matrix coincides with the correlation matrix $R = (\rho_{ij})$. We prove that $\beta_L(\mu) = \lambda_{\min}(R)$, where the minimum eigenvalue of a positive definite matrix A is denoted by $\lambda_{\min}(A)$. Note that $\lambda_{\min}(R) \leq 1$ because $\text{tr}(R) = d$.

Denote the Hermite polynomial of order k by $\eta_k(x) = (-1)^k \phi(x)^{-1} (d^k/dx^k) \phi(x)$ for $x \in \mathbb{R}$. Any function $T_i \in L_0^2(\mu_i)$ is expanded as

$$T_i(x_i) = \sum_{k \geq 1} c_{ik} \eta_k(x_i), \quad \sum_{k \geq 1} k! c_{ik}^2 < \infty.$$

Since $\int \eta_k(x_i) \eta_l(x_j) d\mu = \delta_{kl} (k! \rho_{ij}^k)$, we obtain

$$\int \left(\sum_i T_i \right)^2 d\mu = \sum_{k \geq 1} k! \sum_i \sum_j c_{ik} c_{jk} \rho_{ij}^k$$

and

$$\sum_i \int T_i^2 d\mu_i = \sum_{k \geq 1} k! \sum_i c_{ik}^2.$$

For any $k \geq 1$, we can show that

$$\sum_i \sum_j c_{ik} c_{jk} \rho_{ij}^k \geq \lambda_{\min}(R) \sum_i c_{ik}^2.$$

Indeed, set $A_{ij} = \rho_{ij}$ and $B_{ij} = c_{ik} c_{jk} \rho_{ij}^{k-1}$ in an inequality $\text{tr}(AB) \geq \lambda_{\min}(A) \text{tr}(B)$ for any positive definite matrices A and B . Thus, we have

$$\int \left(\sum_i T_i \right)^2 d\mu \geq \lambda_{\min}(R) \sum_i \int T_i^2 d\mu_i,$$

which implies $\beta_L(\mu) \geq \lambda_{\min}(R)$.

Conversely, let $(v_1, \dots, v_d)$ be the eigenvector corresponding to $\lambda_{\min}(R)$ and $T_i(x_i) = v_i x_i$. Then, we have $\int (\sum_i T_i)^2 d\mu = \lambda_{\min}(R) \sum_i \int T_i^2 d\mu_i$. Thus, $\beta_L(\mu) \leq \lambda_{\min}(R)$. $\square$

We conjecture that $\beta(\mu)$ coincides with (1) if μ is Gaussian and S is its covariance matrix. See Sect. 7.

E.3 Rényi's condition of positive copula densities

The following theorem, which has been proven by [31] for $d = 2$, provides a checkable condition for copositivity.

Theorem 7 ([31] for $d = 2$) *Assume that μ has a regular support (see Sect. 4 for the definition) and for each pair $i \neq j$, the two-dimensional marginal copula density function c_{ij} of μ is square integrable. Then, $\beta_L(\mu) > 0$. In particular, μ is copositive.*

Proof We first prove that if $T \in \prod_{i=1}^d L_0^2(\mu_i)$ satisfies an equation $\sum_i T_i = 0$, then $T = 0$. Assume $\sum_i T_i = 0$. Let $I \subset \{1, \dots, d\}$ be the set of indices i such that $\mu(T_i \neq 0) > 0$. Next, by contradiction, assume I is not empty. Let $A_i = \{x_i \mid T_i > 0\}$ for $i \in I$. Since $\int T_i d\mu_i = 0$, we have $\mu_i(A_i) > 0$. However, based on the assumption about the support, we obtain $\mu(\cap_{i \in I} A_i) > 0$, which implies that $\mu(\sum_i T_i > 0) > 0$ and contradiction. Thus, I is empty, and $T = 0$.

Now, we prove that $\beta_L(\mu) > 0$ using elementary concepts of functional analysis (refer to [39]). Assume that μ_i is uniform over $[0, 1]$, i.e., μ is a copula distribution. Let $H = \prod_{i=1}^d L_0^2(\mu_i)$ be a Hilbert space of $\mathbb{R}^d$ -valued functions and define the inner product of H as $\langle T, U \rangle_H = \sum_i \int T_i U_i dx_i$. Let c_{ij} be the pairwise copula density and define an operator $C : H \rightarrow H$ by

$$(CT)_i = \sum_{j \neq i} \int_0^1 c_{ij} T_j dx_j, \quad i = 1, \dots, d.$$

Based on the assumption that $\iint c_{ij}^2 dx_i dx_j < \infty$, we deduce that C is a Hilbert-Schmidt operator. It is easy to see that C is self-adjoint. Now, we can write

$$\int (\sum_i T_i)^2 d\mu = \langle T, (I + C)T \rangle_H, \quad (23)$$

where I is the identity operator. If $(I + C)T = 0$, then (23) implies $\sum_i T_i = 0$ and, therefore, $T = 0$. Thus, $I + C$ is injective. Since the operator $I + C$ is an injective Fredholm operator, it is surjective. By the continuous inverse theorem, we deduce that the inverse operator $(I + C)^{-1}$ is bounded. Therefore, we have

$$\langle T, (I + C)T \rangle_H \geq \frac{1}{\|(I + C)^{-1}\|} \langle T, T \rangle_H,$$

which means $\beta_L(\mu) \geq \|(I + C)^{-1}\|^{-1} > 0$. $\square$

Corollary 3 *If μ has a positive and bounded copula density function, then μ is copositive.*

In Sect. 6, we deal with positive and piecewise uniform copula density functions.

Note that the support of μ is not determined from the support of two-dimensional marginal distributions. See the following example. Refer to [32] for related topics.

Example 5 Let $\mu \in \mathcal{P}^2(\mathbb{R}^4)$ be the uniform measure supported on the region

$$(+, +, -, -) \cup (+, -, +, -) \cup (+, -, -, +) \cup (-, +, +, -) \cup (-, +, -, +) \cup (-, -, +, +),$$

where $(+, +, -, -)$ denotes the set $[0, 1] \times [0, 1] \times [-1, 0] \times [-1, 0]$, and so on. Then μ is not copositive although each two-dimensional marginal distribution is supported on $[-1, 1]^2$. In order to demonstrate this point, let $T_i(x_i) = \text{sign}(x_i)$ for each i . Then $\int T_i d\mu_i = 0$ and $\int T_i^2 d\mu_i > 0$ but $\int (\sum_i T_i)^2 d\mu = 0$. Hence, $\beta(\mu) = 0$.

E.4 Tail dependence

Many useful copulas in application exhibit tail dependence (e.g. [14, 17, 27]). The following lemma shows that, unfortunately, Theorem 7 is not helpful for this class of copulas.

Lemma 18 Let $d = 2$ and assume that the copula density $c(u_1, u_2)$ has lower-tail dependency

$$\lambda = \lim_{\delta \rightarrow 0} \frac{\int_0^\delta \int_0^\delta c(u_1, u_2) du_1 du_2}{\delta} > 0.$$

Then, $\iint c(u_1, u_2)^2 du_1 du_2 = \infty$. Similar results hold for other types of tail dependency.

Proof The Cauchy-Schwarz inequality implies that

$$\int_0^\delta \int_0^\delta c(u_1, u_2)^2 du_1 du_2 \geq \frac{1}{\delta^2} \left(\int_0^\delta \int_0^\delta c(u_1, u_2) du_1 du_2 \right)^2.$$

If c is square-integrable, then the left-hand side should converge to 0 as $\delta \rightarrow 0$, which is impossible. Thus, c is not square-integrable. $\square$

We conjecture that many copulas with tail dependence are copositive. On the other hand, there is a non-copositive measure with positive copula density, as follows.

Example 6 (*Tail counter-comonotonic copula*) It is known that there is a positive copula density function with the property

$$\lim_{\delta \rightarrow 0} P(X_2 < \delta | X_1 < \delta) = 1,$$

which is equivalent to $\lambda = 1$ in Lemma 18. Such a copula is referred to as a lower tail comonotonic copula (see Sect. 2.21 of [17]). Let μ be the induced measure of $Y_1 = X_1$ and $Y_2 = 1 - X_2$. Then, μ is not copositive. Indeed, define a map $T \in \mathcal{T}_{\text{cw}*}(\mu)$ by

$$T_1(y_1) = \delta I_{(\delta,1)}(y_1) - (1-\delta)I_{(0,\delta)}(y_1), \quad T_2(y_2) = (1-\delta)I_{(1-\delta,1)}(y_2) - \delta I_{(0,1-\delta)}(y_2),$$

where I_A denotes the indicator function of a set A . Then, $\|T_1\| = \|T_2\| = \sqrt{\delta(1-\delta)}$ and

$$\langle T_1, T_2 \rangle = -\delta(1-\delta)^2 P(Y_2 > 1-\delta | Y_1 < \delta) + O(\delta^2), \quad \delta \rightarrow 0.$$

Therefore,

$$\lim_{\delta \rightarrow 0} \frac{\langle T_1, T_2 \rangle}{\|T_1\| \|T_2\|} = -1.$$

In a similar manner to Lemma 16, we deduce that $\beta(\mu) = 0$.

Acknowledgements The author is grateful to the co-editor and two anonymous referees for their careful reading and insightful suggestions. He also thanks to Masaaki Fukasawa and Xiao Li for their helpful comments. The study was supported by JSPS KAKENHI Grant Numbers JP26108003 and JP26540013.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Alfonsi, A., Jourdain, B.: A remark on the optimal transport between two probability measures sharing the same copula. *Stat. Probab. Let.* **84**, 131–134 (2014)
2. Amari, S., Nagaoka, H.: *Methods of Information Geometry*, American Mathematical Society (2000)
3. Ambrosio, L., Gigli, N., Savaré, G.: *Gradient Flows – in Metric Spaces and in the Space of Probability Measures*, Birkhäuser (2005)
4. Borwein, J.M., Lewis, A.S., Nussbaum, R.D.: Entropy minimization, DAD problems, and doubly stochastic kernels. *J. Funct. Anal.* **123**, 264–307 (1994)
5. Butucea, C., Delmas, J., Dutfoy, A., Fischer, R.: Maximum entropy copula with given diagonal section. *J. Multivar. Anal.* **137**, 61–81 (2015)
6. Chen, L.H.Y., Goldstein, L., Shao, Q.: *Normal Approximation by Stein's Method*, Springer (2011)
7. Chernozhukov, V., Galichon, A., Hallin, M., Henry, M.: Monge-Kantorovich depth, quantiles, ranks and signs. *Ann. Stat.* **45**(1), 223–256 (2017)
8. De Rossi, A., Rodino, L.: Strengthened Cauchy-Schwarz inequality for biorthogonal wavelets in Sobolev spaces. *J. Math. Anal. Appl.* **299**, 49–60 (2004)
9. Fallat, S., Lauritzen, S., Sadeghi, K., Uhler, C., Wermuth, N., Zwiernik, P.: Total positivity in Markov structures. *Ann. Stat.* **45**(3), 1152–1184 (2017)
10. Fathi, M.: Stein kernels and moment maps. *Ann. Probab.* **47**(4), 2172–2185 (2019)
11. Fortuin, C.M., Kasteleyn, P.W., Ginibre, J.: Correlation inequalities on some partially ordered sets. *Comm. Math. Phys.* **22**, 89–103 (1971)

12. Gebelein, H.: Das statistische Problem der Korrelation als Variations- und Eigenwertproblem und sein Zusammenhang mit der Ausgleichsrechnung. *Z. Angew. Math. Mech.* **21**(6), 364–379 (1941)
13. Hallin, M.: On distribution and quantile functions, ranks and signs in $\mathbb{R}^d$: a measure transportation approach, preprint (2017)
14. Hua, L.: Multivariate Extremal Dependence and Risk Measures, Ph. D. Thesis in the University of British Columbia (2012)
15. Jaynes, E.T.: Information theory and statistical mechanics. *Phys. Rev.* **106**(4), 620–630 (1957)
16. Joe, H.: Relative entropy measures of multivariate dependence. *J. Am. Stat. Assoc.* **84**, 157–164 (1989)
17. Joe, H.: *Dependence Modeling with Copulas*. CRC Press, Boca Raton (2014)
18. Johnson, O., Barron, A.: Fisher information inequalities and the central limit theorem. *Probab. Theory Relat. Fields* **129**, 391–409 (2004)
19. Jordan, R., Kinderlehrer, D., Otto, F.: The variational formulation of the Fokker-Planck equation. *SIAM J. Math. Anal.* **29**(1), 1–17 (1998)
20. Kalantari, B.: A theorem of the alternative for multihomogeneous functions and its relationship to diagonal scaling of matrices. *Linear Algebra Appl.* **236**, 1–24 (1996)
21. Karlin, S., Rinott, Y.: Classes of orderings of measures and related correlation inequalities. I. multivariate totally positive distributions. *J. Multivar. Anal.* **10**, 467–498 (1980)
22. Lancaster, H.O.: Properties of the bivariate normal distribution considered in the form of a contingency table. *Biometrika* **44**, 289–292 (1957)
23. Lehmann, E.L., Casella, G.: *Theory of Point Estimation*, Springer (1998)
24. Lopez-Paz, D., Hennig, P., Schölkopf, B.: The randomized dependence coefficient. *Adv. Neural Inf. Process. Syst.* **16**, 1–9 (2013)
25. Marshall, A.W., Olkin, I.: Scaling of matrices to achieve specified row and column sums. *Numer. Math.* **12**, 83–90 (1968)
26. McCann, R.J.: A convexity principle for interacting gases. *Adv. Math.* **128**, 153–179 (1997)
27. Nelsen, R. B. (2006). *An Introduction to Copulas*, 2nd ed., Springer
28. Otto, F.: The geometry of dissipative evolution equations: the porous medium equation. *Comm. Partial Diff. Eq.* **26**, 101–174 (2001)
29. R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundations for Statistical Computing, Vienna, Austria (2010). (<http://www.R-project.org/>)
30. Rachev, S.T., Rüschendorf, L.: *Mass Transportation Problems I: Theory*. Springer-Verlag, New York (1998)
31. Rényi, : On measures of dependence, *Acta Math. Acad. Sci. Hungar.* **10**, 441–451 (1959)
32. Rüschendorf, L.: *Mathematical Risk Analysis*. Springer, New York (2013)
33. Sei, T.: An objective general index for multivariate ordered data. *J. Multivar. Anal.* **147**, 247–264 (2016)
34. Sei, T., (2017). Coordinate-wise transformation and Stein-type densities. In: Nielsen F., Barbaresco F. (eds) *Geometric Science of Information. GSI*, : Lecture Notes in Computer Science, vol. 10589. Springer, Cham (2017)
35. Stein, C.: A bound for the error in the normal approximation to the distribution of a sum of dependent random variables, *Proc. Sixth Berkeley Symp. on Math. Stat. Prob.* **2**, 583–602 (1972)
36. Studený, M.: *Probabilistic Conditional Independence Structures*. Springer, New York (2005)
37. Takatsu, A.: Wasserstein geometry of Gaussian measures. *Osaka J. Math.* **48**(4), 1005–1026 (2011)
38. Villani, C.: *Topics in Optimal Transportation*. American Mathematical Society, Providence (2003)
39. Zeidler, E.: *Applied Functional Analysis - Main Principles and their Applications*, Applied Mathematical Sciences, 109. Springer, New York (1995)