MATHEMATICAL ENGINEERING
TECHNICAL REPORTS

Properties of Divergence for Semiparametric Copula Models

Tomonari SEI and Kazuya MATSUMOTO

METR 2019-11 June 2019

DEPARTMENT OF MATHEMATICAL INFORMATICS
GRADUATE SCHOOL OF INFORMATION SCIENCE AND TECHNOLOGY
THE UNIVERSITY OF TOKYO
BUNKYO-KU, TOKYO 113-8656, JAPAN

WWW page: https://www.keisu.t.u-tokyo.ac.jp/research/techrep/

Properties of divergence for semiparametric copula models[*]

Tomonari Sei[†] and Kazuya Matsumoto[†]

June 17, 2019

Abstract

A semiparametric copula model is a statistical model where the copula part is assumed to be parametric and the marginal distribution is arbitrary. In this paper, properties of divergence for the model is investigated. In particular, a relation between the rank divergence induced from the marginal distribution of the multivariate rank statistic and the profile divergence defined by infimum of the Kullback–Leibler divergence with respect to the nuisance parameter is established. Formulas for piecewise uniform and Gaussian copulas are also obtained.

Keywords: Composite transformation model, Copula, Divergence, Holonomic gradient method, Information geometry, Optimal transport.

1 Introduction

A d-dimensional probability density function $c(x)$ $(x = (x_1, \ldots, x_d) \in [0,1]^d)$ is called a copula density if all the one-dimensional marginal density is uniform over $[0,1]$. By Sklar's theorem, any probability density function $p(x) = p(x_1, \ldots, x_d)$ on $\mathbb{R}^d$ is uniquely represented as

$$p(x_1, \ldots, x_d) = c(F_1(x_1), \ldots, F_d(x_d)) \prod_{i=1}^{d} F_i'(x_i), \tag{1}$$

where c is a copula density, F_i is the marginal distribution function of p and F_i' is the derivative of F_i. A statistical model where the copula part c is parametric and the marginal distribution F_i is nonparametric is called a semiparametric copula model. More precise definition is given in Section 3.

There are a number of researches on the estimation problem of semiparametric copula models. Klaassen and Wellner (1997) showed that the normal scores rank correlation coefficient for the two-dimensional Gaussian copula model is

[*] This paper is an English version of a paper written in Japanese.

[†] Graduate School of Information Science and Technology, The University of Tokyo.

asymptotically efficient. Genest and Werker (2002) pointed out that the pseudo-maximum likelihood estimator is not asymptotically efficient for models other than the Gaussian copula. Chen et al. (2006) constructed an asymptotically efficient estimator by using the sieve method for marginal estimation. Tsukahara (2005) provided a class of estimators depending only on the rank statistic and derived its asymptotic properties. Construction of an asymptotically efficient estimator depending only on the rank statistic is an important open problem. For submodels of the Gaussian copula, Hoff et al. (2014) characterized the information bound and Segers et al. (2014) constructed a rank-based asymptotically efficient estimator.

In this paper, we investigate properties of divergence measures for semiparametric copula models. Based on the Kullback–Leibler divergence, we define the rank divergence and profile divergence, which are population characteristics of rank likelihood (Hoff, 2007) and profile likelihood, respectively. Both divergences are not explicitly calculated in general. However, we can derive some explicit formulas for piecewise-uniform copulas. As a result, under the regularity conditions, the rank divergence converges to the profile divergence (Theorem 2). For Gaussian copula models, calculation of the profile divergence is reduced to a finite-dimensional optimization problem, and the rank divergence is represented by orthant probability of the multivariate normal distributions. The latter is numerically evaluated by the holonomic gradient method. These results will be a fundamental step of obtaining the asymptotic properties of estimators.

The paper is organized as follows. In Section 2, we give a simple example that motivates to study the divergence of semiparametric copula models. In Section 3, we define the divergences and provide fundamental theorems on them. In Section 4 and 5, we show explicit results for piecewise uniform and Gaussian copulas, respectively. Finally, we discuss future problems in Section 6.

2 A toy example

Consider the following two-dimensional piecewise uniform distribution. Divide the square region $[0,1]^2$ into four small squares and define a copula density by

$$c(x_1,x_2) = \begin{cases} 1.8 & \text{if } (x_1,x_2) \in [0,\frac{1}{2})^2 \cup [\frac{1}{2},1]^2, \\ 0.2 & \text{otherwise} \end{cases}$$

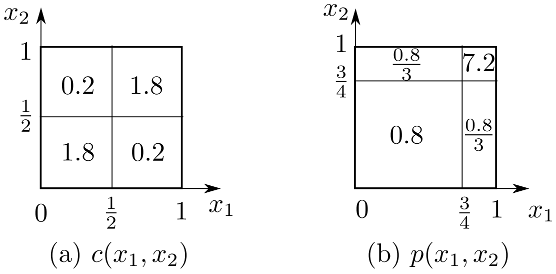
which is constant on each region (Fig. 1 (a)). Choose a one-dimensional distribution

$$F_1(\xi) = F_2(\xi) = \begin{cases} \frac{2}{3}\xi & \text{if } \xi \in [0,\frac{3}{4}), \\ \frac{1}{2} + 2(\xi - \frac{3}{4}) & \text{if } \xi \in [\frac{3}{4},1] \end{cases} \qquad (2)$$

and define a density function $p(x_1, x_2)$ by Eq. (1). Specifically,

$$p(x_1, x_2) = \begin{cases} 0.8 & \text{if } (x_1, x_2) \in [0, \frac{3}{4})^2, \\ 0.8/3 & \text{if } (x_1, x_2) \in ([0, \frac{3}{4}) \times [\frac{3}{4}, 1]) \cup ([\frac{3}{4}, 1] \times [0, \frac{3}{4})) \\ 7.2 & \text{if } (x_1, x_2) \in [\frac{3}{4}, 1]^2. \end{cases}$$

See Fig. 1 (b). Note that only the difference between $c(x_1, x_2)$ and $p(x_1, x_2)$ is the marginal distribution, and the dependence is the same.



(a) $c(x_1, x_2)$ (b) $p(x_1, x_2)$

Figure 1: Two density functions with different marginal and the same copula part. The number specified in each region denotes the value of density functions. The density function $p(x_1, x_2)$ has a smaller Kullback–Leibler divergence from the uniform density than $c(x_1, x_2)$.

However, the Kullback–Leibler divergence between u and c, where $u(x_1, x_2) = 1$ denotes the uniform density, is

$$\begin{aligned} \mathrm{KL}(u, c) &= \int_{[0,1]^2} u(x) \log \frac{u(x)}{c(x)} \, \mathrm{d}x \\ &= \frac{1}{2} \log \frac{1}{1.8} + \frac{1}{2} \log \frac{1}{0.2} \\ &\approx 0.511 \end{aligned}$$

and the divergence between u and p is

$$\begin{aligned} \mathrm{KL}(u, p) &= \left(\frac{3}{4}\right)^2 \log \frac{1}{0.8} + 2 \left(\frac{3}{4}\right) \left(\frac{1}{4}\right) \log \frac{1}{0.8/3} + \left(\frac{1}{4}\right)^2 \log \frac{1}{7.2} \\ &\approx 0.498. \end{aligned}$$

Therefore, p is closer than c from u.

Likewise, the divergence between densities changes with the marginal distributions. The minimum is called the profile divergence in this paper. On the other hand, as discussed in the following section, information of semiparametric copula models is aggregated to the rank statistic. Hence we call the divergence defined by the rank statistic the rank divergence. Our purpose is to elucidate their relation.

3 Rank divergence and profile divergence

In this section, we first explain that the semiparametric copula models are invariant under coordinate-wise transformations (Hoff, 2007; Hoff et al., 2014). Thus the model is regarded as a composite transformation model (Appendix A). From this point of view, two divergences called the rank divergence and the profile divergence are defined. Their relationship is clarified.

3.1 Semiparametric copula models

Let us precisely define the semiparametric copula model. Denote the set of all positive probability density functions on $[0,1]^d$ by $\mathcal{P}$, where $p(x)$ is said to be positive if $p(x) > 0$ almost surely. The set of all coordinate-wise transformations $T(x) = (T_1(x_1), \ldots, T_d(x_d))$ such that $T_i : [0,1] \to [0,1]$ for each i is a monotone increasing, bijective and absolutely continuous function is denoted by $\mathcal{T}$. For example, the pair of marginal distributions (F_1, F_2) in Eq. (2) is an element of $\mathcal{T}$. The set $\mathcal{T}$ forms a group with respect to function composition. For a density $p \in \mathcal{P}$ and a transformation $T \in \mathcal{T}$, the push-forward density $T_* p \in \mathcal{P}$ is defined by

$$(T_*p)(x_1, \ldots, x_d) = p(T_1^{-1}(x_1), \ldots, T_d^{-1}(x_d)) \prod_{i=1}^{d} (T_i^{-1})'(x_i). \tag{3}$$

This is the density function of a transformed random vector $T(X)$ if X is distributed according to p. The map $(T, p) \mapsto T_* p$ defines an action of $\mathcal{T}$ to $\mathcal{P}$. Denote the orbit (equivalence class) with respect to the action by $[p] = \{T_* p \mid T \in \mathcal{T}\}$. Sklar's theorem implies that each orbit contains a unique copula density. In other words, copulas and orbits have one-to-one correspondence.

Although the density functions are restricted to those on $[0,1]^d$, it is possible to deal with density functions on $\mathbb{R}^d$. Indeed, we can fix any function from $\mathbb{R}^d$ to $(0,1)^d$, then transformations from $\mathbb{R}^d$ to $\mathbb{R}^d$ are obtained as a result.

Under these notations, the semiparametric copula model is defined by

$$\mathcal{M} = \{T_* c_\theta \mid \theta \in \Theta,\ T \in \mathcal{T}\}, \tag{4}$$

where $\{c_\theta \mid \theta \in \Theta\}$ is a parametric family of copula densities. See Nelsen (2006) for examples of parametric copulas. We will deal with the piecewise uniform copulas and Gaussian copulas in Sections 4 and 5, respectively.

The model $\mathcal{M}$ is a composite transformation model with respect to the action of $\mathcal{T}$. The parameter of interest is θ, that is, the orbit. We try to define the divergence between two orbits $[p]$ and $[q]$.

3.2 Rank divergence

Let $p \in \mathcal{P}$ be the true density and $X = (x_{ti})_{1 \le t \le n, 1 \le i \le d}$ be a random sample generated from p, where n is the sample size. Since we consider continuous distributions, the values $\{x_{ti}\}_{t=1}^{n}$ are assumed to be different from each other for each i. The following lemma is well known for one-dimensional case (e.g. Eaton (1983)). The multi-dimensional case is similarly proved.

Lemma 1 (Hoff (2007))**.** The maximal invariant of the semiparametric copula model is the multivariate rank statistic

$$r_{ti} = \sharp\{s \in \{1, \ldots, n\} \mid x_{si} \leq x_{ti}\}, \quad 1 \leq t \leq n, \quad 1 \leq i \leq d,$$

where $\sharp A$ denotes the cardinality of a set A. We also use a matrix notation $R = (r_{ti})$.

We call the multivariate rank statistic simply the rank statistic. Denote the marginal distribution of the rank statistic R by $\bar{p}_n(R)$ and call it the rank likelihood. Since R can take only finite number of values (precisely $(n!)^d$), $\bar{p}_n$ is a discrete distribution.

The rank likelihood is described by a high-dimensional integral. Indeed, let $\{R(X) = R\}$ be the set of values $X \in \mathbb{R}^{n\times d}$ that are consistent with R. Then the rank likelihood of p is

$$\bar{p}_n(R) = \int_{\{R(X)=R\}} \prod_{t=1}^{n} p(x_{t1}, \ldots, x_{td}) dX. \tag{5}$$

As will be stated in Section 4, we can write down the rank likelihood without integrals when p is piecewise uniform.

The rank divergence is defined as follows.

Definition 1. For given density functions $p, q \in \mathcal{P}$ and the sample size n, the rank divergence is defined by

$$\begin{aligned} D_n([p], [q]) &= \frac{1}{n}\mathrm{KL}(\bar{p}_n, \bar{q}_n) \qquad (6) \\ &= \frac{1}{n}\sum_{R} \bar{p}_n(R) \log \frac{\bar{p}_n(R)}{\bar{q}_n(R)}. \end{aligned}$$

The function $D_n([p], [q])$ is well-defined since $\bar{p}_n(R)$ does not depend on the representative of $[p]$. The reason why the right hand side of Eq. (6) is divided by n is that the statistic R has information of order $\mathrm{O}(n)$. The additivity and monotonicity of the Kullback–Leibler divergence imply

$$D_n([p], [q]) \leq \mathrm{KL}(p, q). \tag{7}$$

The rank divergence is not positive in general, that is, $D_n([p], [q])$ may be zero even if $[p] \neq [q]$. For example, if a two-dimensional copula density p has a symmetry $p(x_1, x_2) = p(1 - x_1, x_2)$, then $\bar{p}_2(R)$ is the uniform distribution (on four points). Thus if both p and q are symmetric, then $D_2([p], [q]) = 0$. On the other hand, there is not $[p] \neq [q]$ such that $D_n([p], [q]) = 0$ for all n (under regularity conditions). This fact is confirmed by Theorem 1 and Theorem 2 later.

3.3 Profile divergence

It is natural to consider the following divergence as an analogue of composite transformation models (Appendix A).

Definition 2. For given density functions $p, q \in \mathcal{P}$, the profile divergence is defined by

$$\begin{aligned}\tilde{D}([p],[q]) &= \inf_{T,U \in \mathcal{T}} \mathrm{KL}(T_* p, U_* q) \qquad (8)\\ &= \inf_{T \in \mathcal{T}} \mathrm{KL}(T_* p, q).\end{aligned}$$

The second equality follows from invariance of the Kullback–Leibler divergence $\mathrm{KL}(T_* p, T_* q) = \mathrm{KL}(p, q)$.

Even if p and q are copula densities, $\tilde{D}([p],[q]) < \mathrm{KL}(p,q)$ in general. The example given in Section 2 is such an example.

We derive a condition when $\tilde{D}([p],[q]) = \mathrm{KL}(p,q)$ holds. For a density $q \in \mathcal{P}$ and a map $T \in \mathcal{T}$, define the pull-back density $T^* q$ by

$$(T^* q)(x) = q(T_1(x_1), \ldots, T_d(x_d)) \prod_{i=1}^{d} T_i'(x_i). \qquad (9)$$

This is the inverse of the push-forward operation defined by Eq. (3). Note that $\mathrm{KL}(T_* p, q) = \mathrm{KL}(p, T^* q)$.

Lemma 2. *Let p and q be continuously differentiable probability density functions on $(0,1)^d$, which are not necessarily copula densities. Then the equality $\tilde{D}([p],[q]) = \mathrm{KL}(p,q)$ holds only if*

$$\partial_i \log p_i(x_i) = \mathrm{E}_p[\partial_i \log q(x) | x_i], \quad i = 1, \ldots, d. \qquad (10)$$

Here p_i is the marginal density of p, ∂_i is the partial derivative by x_i and $\mathrm{E}_p[\cdot|\cdot]$ is the conditional expectation with respect to p. Furthermore, if q is log-concave, then Eq. (10) is also a sufficient condition.

Proof. The proof is based on the variational method. Let $T_i(x_i) = x_i + \delta T_i(x_i)$, where $\delta T_i(x_i)$ is a smooth function with a compact support in $(0,1)$. Expand $\mathrm{KL}(p, T^* q)$ with respect to δT_i up to the first order term to obtain

$$\begin{aligned}\mathrm{KL}(p, T^* q) &= \int p(x) \log \frac{p(x)}{q(T(x)) \prod_i T_i'(x_i)} \mathrm{d}x \qquad (11)\\ &\simeq \mathrm{KL}(p,q) - \sum_i \int p(x)(\partial_i \log q(x)) \delta T_i(x_i) \mathrm{d}x - \sum_i \int p(x) \delta T_i'(x_i) \mathrm{d}x\\ &= \mathrm{KL}(p,q) + \sum_i \int \{-p(x) \partial_i \log q(x) + \partial_i p(x)\} \delta T_i(x_i) \mathrm{d}x \qquad (12)\\ &= \mathrm{KL}(p,q) + \sum_i \int p_i(x_i) \{\mathrm{E}_p[-\partial_i \log q(x) | x_i] + \partial_i \log p_i(x_i)\} \delta T_i(x_i) \mathrm{d}x_i.\end{aligned}$$

Here the equality in (12) follows from the integral-by-parts and the boundary condition $\delta T_i(0) = \delta T_i(1) = 0$. Hence we have the stationary condition (10). If q is log-concave, the functional $T \mapsto \mathrm{KL}(p, T^* q)$ on $\mathcal{T}$ is convex due to the form of Eq. (11). Thus the stationary condition implies optimality. □

In the last part of the proof, we used convexity of the set $\mathcal{T}$ and convexity of the functional $T \mapsto \mathrm{KL}(p, T^* q)$. They are called displacement convexity in the context of the optimal transport theory. Lemma 2 holds even if the support of p and q is not $[0,1]^d$.

If q is the uniform density on $[0,1]^d$, then Eq. (10) in Lemma 2 is equivalent to $\partial_i p_i = 0$, which means p is a copula density. In this case, $\mathrm{KL}(p,q) = \int p(x) \log p(x) \mathrm{d}x$ is equal to the negative entropy. Hence we obtain the following consequence.

Lemma 3. The density $p \in \mathcal{P}$ is a copula density function if and only if p maximizes the entropy over the orbit $[p]$.

3.4 Main theorems

We provide two theorems on the rank divergence and profile divergence. Note that the two quantities satisfy $D_n([p],[q]) \leq \tilde{D}([p],[q])$ in general from Eq. (7).

First we have the following theorem on positivity of the profile divergence.

Theorem 1. Let p and q be positive copula density functions and assume that q is bounded from above and upper semi-continuous. Then there exists $T \in \mathcal{T}$ such that $\tilde{D}([p],[q]) = \mathrm{KL}(p, T^* q)$. In particular, $\tilde{D}([p],[q]) > 0$ if $[p] \neq [q]$.

The proof is given in Appendix B.1. The optimal transport theory is relevant, where the minimization problem $\tilde{D}([p],[q]) = \inf_{T \in \mathcal{T}} \mathrm{KL}(T_* p, q)$ is interpreted as energy minimization problem with respect to the transport map T.

The boundedness of q in Theorem 1 is assumed for the sake of proof and not necessary. In fact, the Gaussian copula does not satisfy the condition, but the positivity of the profile divergence is directly proved (Section 5). To weaken the condition is a future work.

The following result is an analogue of a known fact in the composite transformation model (Appendix A).

Theorem 2. Let p and q be positive copula density functions and assume that both are bounded from below and above and continuous. If the true density is p, then we have

$$\lim_{n \to \infty} \frac{1}{n} \log \frac{\bar{p}_n(R)}{\bar{q}_n(R)} = \tilde{D}([p],[q])$$

with probability one. Furthermore, the rank divergence converges to the profile divergence:

$$\lim_{n \to \infty} D_n([p],[q]) = \tilde{D}([p],[q]).$$

The proof is given in Appendix B.2. From the theorem, the asymptotic information of semiparametric copula models is contained in the profile divergence.

4 Piecewise uniform copula

The rank divergence and profile divergence do not have explicit formulas in general. However, the piecewise uniform distributions and Gaussian distributions are exceptions. This section deals with the piecewise uniform distributions. The discussion is restricted to two-dimensional cases but the multi-dimensional case is similar. Note that piecewise uniform distributions are also called the chessboard distributions (Ghosh and Henderson, 2001).

4.1 Profile divergence

Let I and J be positive integers. Divide $[0,1]^2$ into $I \times J$ small rectangles and call them $A_{ij} = \left[\frac{i-1}{I}, \frac{i}{I}\right) \times \left[\frac{j-1}{J}, \frac{j}{J}\right)$ $(1 \leq i \leq I,\, 1 \leq j \leq J)$. A density function p is called piecewise uniform if

$$p(x_1, x_2) = p_{ij} \quad \text{if} \quad (x_1, x_2) \in A_{ij}, \quad 1 \leq i \leq I, \quad 1 \leq j \leq J.$$

Here $\{p_{ij}\}$ is a set of positive numbers such that $\sum_{i=1}^{I} \sum_{j=1}^{J} p_{ij}/IJ = 1$. A piecewise uniform density is a copula density if and only if

$$\sum_{i=1}^{I} \frac{p_{ij}}{I} = 1 \quad (1 \leq j \leq J), \quad \sum_{j=1}^{J} \frac{p_{ij}}{J} = 1 \quad (1 \leq i \leq I).$$

Therefore the set of piecewise copula densities is a $(I-1)(J-1)$-dimensional parametric model. Although it looks like a contingency table model, the divergence structure is different from it as already discussed in Section 2.

The following lemma shows that calculation of the profile divergence $\tilde{D}([p],[q])$ is reduced to a finite-dimensional optimization problem as long as q is piecewise uniform.

Lemma 4. Let p be an arbitrary copula density and q be a piecewise uniform copula density. Denote the value of $q(x)$ over A_{ij} by q_{ij}. Then the map $T \in \mathcal{T}$ that attains $\tilde{D}([p],[q]) = \mathrm{KL}(p, T^* q)$ is a piecewise linear transformation that satisfies

$$\xi_i = T_1^{-1}(i/I), \quad \eta_j = T_2^{-1}(j/J) \quad (1 \leq i \leq I-1, \quad 1 \leq j \leq J-1). \tag{13}$$

Furthremore, ξ_i and η_j are the solution of the following minimization problem:

$$\begin{aligned}\text{Minimize}\quad & \sum_i\sum_j\left(\int_{\xi_{i-1}}^{\xi_i}\int_{\eta_{j-1}}^{\eta_j} p(x)\mathrm{d}x\right)\log\frac{1}{q_{ij}} \qquad (14)\\ & +\sum_i(\xi_i-\xi_{i-1})\log(\xi_i-\xi_{i-1})+\sum_j(\eta_j-\eta_{j-1})\log(\eta_j-\eta_{j-1})\\ \text{subject to}\quad & 0=\xi_0<\xi_1<\cdots<\xi_I=1,\quad 0=\eta_0<\eta_1<\cdots<\eta_J=1.\end{aligned}$$

The objective function is equal to $\mathrm{KL}(p,T^*q)$ up to a constant term.

Proof. Fix $\{\xi_i\}$ and $\{\eta_j\}$. We prove that the minimizer of $\mathrm{KL}(p,T^*q)$ under the condition (13) is piecewise uniform. For such a T, the value of $q(T(x))$ does not depend on T because of the piecewise uniformity of q. Therefore we have

$$\begin{aligned}\mathrm{KL}(p,T^*q) &= \int p(x)\log\frac{p(x)}{q(T(x))T_1'(x_1)T_2'(x_2)}dx \qquad (15)\\ &= (\text{const.})-\int_0^1\log T_1'(x_1)dx_1-\int_0^1\log T_2'(x_2)dx_2,\end{aligned}$$

where the condition $\int p(x)\mathrm{d}x_2=\int p(x)\mathrm{d}x_1=1$ of copula densities is used. In general, the minimizer of $-\int_0^1\log t'(x)\mathrm{d}x$ under the boundary condition $t(0)<t(1)$ is a linear function. Indeed, concavity of the logarithm implies that $-\int_0^1\log t'(x)\mathrm{d}x\geq-\log\int_0^1 t'(x)\mathrm{d}x=-\log(t(1)-t(0))$, and the equality holds if and only if $t''(x)=0$. From these, T_1 and T_2 are piecewise linear. The objective function in Eq. (14) is obtained from Eq. (15). Note that $T_1'(x_1)=1/(I(\xi_i-\xi_{i-1}))$ for $x_1\in[\xi_{i-1},\xi_i]$. □

The optimization problem (14) is not convex in general and the solution is not unique. However, there exists a solution due to Theorem 1.

From now on, we investigate the simplest case $I=J=2$. Consider the following copula density:

$$c_\theta(x)=\begin{cases}1+\theta & \text{if } x\in[0,1/2]^2\cup[1/2,1]^2,\\ 1-\theta & \text{otherwise,}\end{cases}$$

where $-1<\theta<1$. The copula density used in Section 2 was of the form. If $\theta=0$, c_θ is equal to the uniform density.

In fact, the following "bifurcation phenomenon" holds. Here the definition of ξ_1 and η_1 is the same as above.

Lemma 5. Let $\theta>0$. Then the profile divergence from the uniform density u to c_θ is

$$\tilde{D}([u],[c_\theta])=\begin{cases}\mathrm{KL}(u,c_\theta) & \text{if } 0<\theta\leq\tanh(1),\\ \mathrm{KL}(u,T^*c_\theta) & \text{if } \tanh(1)<\theta<1,\end{cases}$$

where $T = (T_1, T_2)$ is a piecewise linear transformation with knots $\xi_1 = \eta_1 = \xi$ or $1 - \xi$. The quantity ξ is the unique solution of the following equation:

$$\xi = \frac{1}{2}\left(1 + \frac{\log \frac{\xi}{1-\xi}}{\log \frac{1+\theta}{1-\theta}}\right), \quad \frac{1}{2} < \xi < 1. \tag{16}$$

Proof. Denote the objective function in Eq. (14) by $f(\xi_1, \eta_1)$. The stationary condition of f is

$$\eta_1 = \frac{1}{2}\left(1 + \frac{\log \frac{\xi_1}{1-\xi_1}}{\log \frac{1+\theta}{1-\theta}}\right), \quad \xi_1 = \frac{1}{2}\left(1 + \frac{\log \frac{\eta_1}{1-\eta_1}}{\log \frac{1+\theta}{1-\theta}}\right).$$

This equation has a unique solution $(\xi_1, \eta_1) = (1/2, 1/2)$ if $|\log \frac{1+\theta}{1-\theta}| \leq 2$, or equivalently $|\theta| \leq \tanh(1)$, and two symmetric solutions (ξ, ξ) and $(1-\xi, 1-\xi)$ together with $(1/2, 1/2)$ if $|\theta| > \tanh(1)$. Here ξ is the solution of Eq. (16). The Hessian matrix of f is

$$\begin{pmatrix} \frac{1}{\xi_1(1-\xi_1)} & 2\log\frac{1-\theta}{1+\theta} \\ 2\log\frac{1-\theta}{1+\theta} & \frac{1}{\eta_1(1-\eta_1)} \end{pmatrix}.$$

In particular, if $\theta > \tanh(1)$, then $\xi_1 = \eta_1 = 1/2$ is not a minimal point, and $\xi_1 = \eta_1 = \xi$ and $\xi_1 = \eta_1 = 1 - \xi$ are minimal. □

Solve (16) with respect to θ to obtain

$$\theta = \frac{(\frac{\xi}{1-\xi})^{1/(2\xi-1)} - 1}{(\frac{\xi}{1-\xi})^{1/(2\xi-1)} + 1}, \quad \frac{1}{2} < \xi < 1.$$

For example, if $\xi = \frac{3}{4}$, then $\theta = 0.8$, which corresponds to the example given in Section 2.

From Lemma 5, we see that a bifurcation phenomenon occurs at $\theta = \tanh(1)$. See Figure 2. It is also shown that $\tilde{D}([u], [c_\theta])$ converges to $\log 2$ as $\theta \to 1$. In particular, it is remarkable that the profile divergence is bounded.

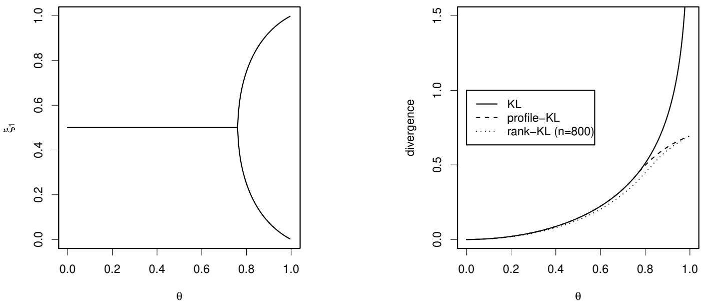
4.2 Rank likelihood of piecewise uniform copulas

We determine the rank likelihood for piecewise uniform copulas. Consider a piecewise uniform copula density

$$p_\theta(x_1, x_2) = \theta_{ij} \quad \text{if } (x_1, x_2) \in A_{ij}, \tag{17}$$

where A_{ij} is a small region defined in the preceding subsection and θ_{ij} is a positive number that satisfies $\sum_i \theta_{ij}/I = 1$ and $\sum_j \theta_{ij}/J = 1$. The joint density function of a random sample $X = \{(x_{t1}, x_{t2})\}_{t=1}^n$ is

$$(IJ)^{-n} \prod_{i=1}^{I} \prod_{j=1}^{J} \theta_{ij}^{n_{ij}}, \tag{18}$$



(a) Optimal $\xi_1 (= \eta_1)$

(b) Divergence

Figure 2: A bifurcation phenomenon. (a) The optimal map T changes at $\theta = \tanh(1)$ (b) Kullback–Leibler divergence $\mathrm{KL}(u, c_\theta)$, rank divergence $D_n([u], [c_\theta])$ $(n = 800)$ and profile divergence $\tilde{D}([u], [c_\theta])$ are plotted as a function of θ.

where n_{ij} denotes the number of observations that belong to A_{ij} and is the sufficient statistic of θ. However, since X itself is not observed and only the rank statistic R is observed, (n_{ij}) is a latent variable.

Denote the marginal frequency of (n_{ij}) by $\sigma_i = n_{i+} = \sum_j n_{ij}$ and $\tau_j = n_{+j} = \sum_i n_{ij}$. Once $\sigma = (\sigma_i)$ and $\tau = (\tau_j)$ are given, (n_{ij}) is determined from R. We denote the relation by $n_{ij} = n_{ij}(R, \sigma, \tau)$.

Theorem 3. For a piecewise copula density p, the rank likelihood is given by

$$\bar{p}_n(R) = (IJ)^{-n} \sum_{\sigma} \sum_{\tau} \frac{1}{\prod_i \sigma_i! \prod_j \tau_j!} \prod_i \prod_j \theta_{ij}^{n_{ij}(R,\sigma,\tau)}, \tag{19}$$

where σ and τ range over the whole set of marginal frequency vectors with the total frequency n.

The proof is given in Appendix B.3. The rank divergence plotted in Figure 2 (b) is computed by Theorem 3 together with the Monte Carlo method.

5 Gaussian copulas

The Gaussian copula is a copula induced from the multivariate Gaussian distribution. It is shown that the profile divergence between Gaussian copulas has a simple form. The rank divergence is reduced to the problem of calculating orthant probability.

As noted in Section 3, the divergence between two densities on $\mathbb{R}^d$ is defined via the densities on $(0,1)^d$. However, as a result, it is enough to consider transformations on $\mathbb{R}^d$ and not necessary to go through $(0,1)^d$.

5.1 Profile divergence

Calculation of the profile divergence of Gaussian copulas is reduced to a finite-dimensional convex optimization problem. Denote the density function of Gaussian density with the mean vector 0 and the covariance matrix Σ by ϕ_Σ. For a vector u, $\mathrm{diag}(u)$ is the diagonal matrix with the diagonal part u.

Lemma 6. Let P and Q be d-dimensional symmetric positive definite matrices, and set $p = \phi_P$ and $q = \phi_Q$. Then a map $T \in \mathcal{T}$ satisfying $\tilde{D}([p],[q]) = \mathrm{KL}(p, T^*q)$ is given by a linear map $T(x) = \mathrm{diag}(u)x$, where $u = (u_i)$ is the unique solution of the following convex programming:

$$\begin{aligned} &\text{Minimize} \quad -\sum_i \log u_i + \frac{1}{2}\mathrm{tr}(Q^{-1}\mathrm{diag}(u)P\,\mathrm{diag}(u)) \\ &\text{subject to} \quad u_1, \ldots, u_d > 0. \end{aligned} \tag{20}$$

The objective function is equal to $\mathrm{KL}(p, T^*q)$ up to a constant term.

Proof. The density q is log-concave because it is Gaussian. Hence it is sufficient to prove that $T(x) = \mathrm{diag}(u)x$ satisfies the stationary condition

$$\partial_i \log p_i(x_i) = \mathrm{E}_p[(\partial_i \log(T^*q)(x)|x_i]$$

of Lemma 2. The left hand side is equal to $\partial_i \log p_i(x_i) = -P_{ii}^{-1}x_i$. The right hand side is

$$\begin{aligned} \mathrm{E}_p[(\partial_i \log(T^*q))(x)|x_i] &= -\sum_j u_i(Q^{-1})_{ij}u_j \mathrm{E}_p[x_j|x_i] \\ &= -\sum_j u_i(Q^{-1})_{ij}u_j P_{ji}P_{ii}^{-1}x_i \\ &= -P_{ii}^{-1}x_i. \end{aligned}$$

The last equality follows from the stationary condition of Eq. (20). The fact that Eq. (20) has a unique solution is shown in Marshall and Olkin (1968). □

If $d = 2$, the profile divergence between p and q is

$$\tilde{D}([p],[q]) = \log \frac{1 - \rho_p\rho_q}{\sqrt{(1-\rho_p^2)(1-\rho_q^2)}}$$

where ρ_p and ρ_q are the correlation coefficient of p and q, respectively. This is a symmetric divergence. By using Fisher's Z transform $\rho = \tanh z$, we have

$$\tilde{D}([p],[q]) = \log\cosh(z_p - z_q),$$

which is invariant with respect to location shift of z's.

We confirm that, for Gaussian copulas, the metric induced from the profile divergence coincides with the efficient information derived in Segers et al. (2014).

Here the metric induced from the profile divergence for a statistical model $\{p_\theta \mid \theta \in \mathbb{R}^m\}$ is defined by

$$\tilde{g}_{ij} = -\left.\frac{\partial^2 \tilde{D}([p_\theta],[p_\phi])}{\partial\theta_i \partial\phi_j}\right|_{\theta=\phi}, \quad i,j \in \{1,\ldots,m\}$$

(see Eguchi (1983); Amari (1985)). The following theorem is proved in Appendix B.4.

Theorem 4. For a parametric model of covariance matrices $\{P = P_\theta \mid \theta \in \mathbb{R}^m\}$, the metric is

$$\tilde{g}_{ij} = \frac{1}{2}\operatorname{tr}\left\{P^{-1}(\partial_i P - \Pi(\partial_i P))P^{-1}(\partial_j P - \Pi(\partial_j P))\right\}, \tag{21}$$

where ∂_i is an abbreviation of $\partial/\partial\theta_i$ and the projection $\Pi : \mathbb{R}^{d\times d} \to \mathbb{R}^{d\times d}$ is defined by

$$\Pi(A) = P\operatorname{diag}(b(A)) + \operatorname{diag}(b(A))P, \tag{22}$$

$$b(A) = (P^{-1}\circ P + I)^{-1}(P^{-1}\circ A)1_d, \quad 1_d = (1,\ldots,1)^\top \in \mathbb{R}^d. \tag{23}$$

The symbol $A \circ B$ is the element-wise product of matrices A and B (Hadamard product).

5.2 Rank likelihood of Gaussian copulas

For Gaussian copulas, the rank divergence (5) is represented by orthant probability of Gaussian measures since the integration region $\{R(X) = R\}$ is the intersection of half spaces

$$x_{t(s,i),i} < x_{t(s+1,i),i}, \quad s \in [n-1], \quad i \in [d],$$

where $t(s,i)$ denotes the data number t such that the rank of i-th coordinate is $r_{ti} = s$. We also defined $[n] = \{1,\ldots,n\}$. The result is summarized in the following theorem. The proof is given in Appendix B.5.

Theorem 5. The rank likelihood of the d-dimensional Gaussian distribution with the covariance matrix Σ is

$$\bar{p}_n(R|\Sigma) = \frac{1}{n^{d/2}|\Sigma|^{(n-1)/2}|B|^{1/2}}\int_{\mathbb{R}_+^{(n-1)d}} \phi(w|0,B^{-1})dw,$$

where $\mathbb{R}_+ = (0,\infty)$, ϕ is the Gaussian density and B is a $(n-1)d$-dimensional symmetric positive definite matrix defined by

$$B_{(r-1)d+i,(s-1)d+j} = (\Sigma^{-1})_{ij}\left(\sum_{t=1}^n I_{\{r_{ti}\le r, r_{tj}\le s\}} - \frac{rs}{n}\right), \quad r,s \in [n-1], \quad i,j \in [d].$$

Table 1: The rank likelihood for $n = 2$ and $r_{t1} = t$ $(\forall t)$. The values are multiplied by $n!$.

θ	.00	.10	.20	.30	.40	.50	.60	.70	.80	.90	.95
$(r_{t2}) = (1,2)$	.5000	.5319	.5641	.5970	.6310	.6667	.7048	.7468	.7952	.8564	.8989
$(2,1)$	.5000	.4681	.4359	.4030	.3690	.3333	.2952	.2532	.2048	.1436	.1011

Table 2: The rank likelihood for $n = 3$ and $r_{t1} = t$ $(\forall t)$. The values are multiplied by $n!$.

θ	.00	.10	.20	.30	.40	.50	.60	.70	.80	.90	.95
$(r_{t2}) = (1,2,3)$	.1667	.1918	.2196	.2509	.2866	.3280	.3773	.4381	.5179	.6359	.7302
$(1,3,2)$	.1667	.1780	.1880	.1964	.2028	.2067	.2068	.2017	.1875	.1549	.1212
$(2,1,3)$	.1667	.1780	.1880	.1964	.2028	.2067	.2068	.2017	.1875	.1549	.1212
$(2,3,1)$	.1667	.1542	.1407	.1262	.1109	.0947	.0777	.0597	.0408	.0210	.0106
$(3,1,2)$	.1667	.1542	.1407	.1262	.1109	.0947	.0777	.0597	.0408	.0210	.0106
$(3,2,1)$	.1667	.1439	.1230	.1037	.0859	.0693	.0537	.0391	.0253	.0123	.0061

It is known that orthant probability of multivariate Gaussian distributions is computed in high accuracy by the holonomic gradient method (Koyama and Takemura, 2015). An R package is available (Koyama et al., 2014). Therefore the rank likelihood and rank divergence is computed via Theorem 5. The computational complexity of the holonomic gradient method is expressed by the holonomic rank. For m-dimensional Gaussian probability, the holonomic rank is known to be 2^m and the computation cost rapidly increases with the sample size.

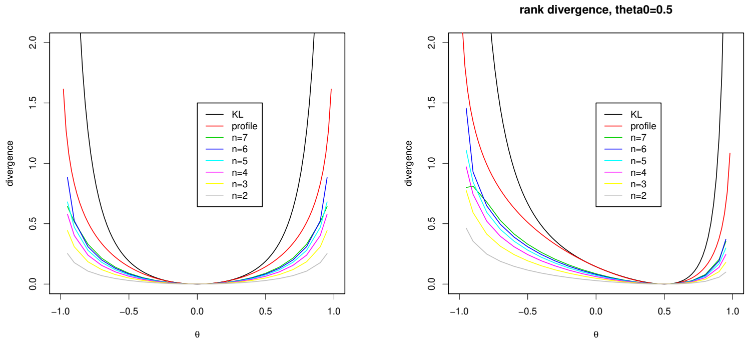
Table 1 and Table 2 show the rank likelihood of the two-dimensional Gaussian copula with the covariance matrix

$$\Sigma = \begin{pmatrix} 1 & \theta \\ \theta & 1 \end{pmatrix}, \quad \theta \in \{0.0, 0.1, \ldots, 0.9, 0.95\},$$

for $n = 2$ and $n = 3$, respectively. The first column of R is fixed to $(1, \ldots, n)$. Instead, the probability $\bar{p}_n(R)$ is multiplied by $n!$. As a result, the sum of each column in the table is one.

Figure 3 shows the graph of the rank divergence $D_n([p_{\theta_0}], [p_\theta])$ as a function of θ, where p_θ is the two-dimensional Gaussian copula and $\theta_0 = 0$ or $\theta_0 = 0.5$. The sample size ranges from $n = 2$ to $n = 7$. In the same figure, the profile divergence $\tilde{D}([p_{\theta_0}], [p_\theta])$ and the Kullback–Leibler divergence $\mathrm{KL}(p_{\theta_0}, p_\theta)$ are also plotted. A theoretical relation $\mathrm{KL} \geq \tilde{D} \geq D_n$ is also observed in the figure.

If θ is fixed, the rank divergence tends to increase as n increases from 2 to 7. However, the curve for $n = 7$ crosses over those for $n \leq 6$. The reason may be the computational accuracy for $n = 7$. Future investigation is necessary.



(a) $\theta_0 = 0$ の場合

(b) $\theta_0 = 0.5$ の場合

Figure 3: The rank divergence $D_n(\theta_0, \theta)$ for Gaussian copulas.

6 Discussion

In this paper, we defined rank divergence and profile divergence. The former converges to the latter (Theorem 2). However, the copula density function was assumed to be bounded in order to prove the theorem. This assumption is too strong and not applicable to practical copula models. A future work is to weaken this assumption.

In Theorem 4, the efficient information of the Gaussian copula was obtained by the profile divergence. The fact will be generalized to wider class of copula models, but is not demonstrated yet. The efficient information is represented by a system of Sturm–Liouville equations and not obtained explicitly (Bickel et al., 1993).

We assumed from the beginning that the marginal distributions $F_1, \ldots, F_d$ are completely unknown. However, marginal distributions will be of interest in practice and it may be natural to assume prior information. The estimator by Chen et al. (2006) is interpreted as such an example. The divergence for these cases should be derived.

There are various divergence measures between probability distributions other than the Kullback–Leibler one. It is possible to formally define the rank divergence and profile divergence on the basis of them. However, it should be careful if invariance and additivity of divergence do not hold.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 26108003 and 17K00044. The first author appreciates Kenichi Hashimoto for providing a relevant topic of the current research.

Appendices

A Composite transformation models

Composite transformation models are invariant statistical models with respect to some transformations. More specifically, consider a family of probability distributions $\{P_{\theta,\nu} \mid \theta \in \Theta, \nu \in N\}$ on a sample space $\mathcal{X}$. Let θ be a parameter of interest and ν be a nuisance parameter. Suppose that there exists a group G acting both $\mathcal{X}$ and N and it satisfies

$$x \sim P_{\theta,\nu} \quad \Rightarrow \quad gx \sim P_{\theta,g\nu}.$$

We further assume that the action G to N is transitive. Then the family $\{P_{\theta,\nu}\}$ is called a composite transformation model (Barndorff-Nielsen and Jupp, 1988). It is also called an invariant probability model (Eaton, 1983). The space $\mathcal{X}$ is partitioned into mutually disjoint orbits by the action of G. A statistic is called maximal invariant if it has a one-to-one correspondence with the orbits. The distribution of a maximal invariant depends only on θ.

Example 1. Consider a random sample $x = (x_1, \ldots, x_n)$ according to the normal distribution $\mathrm{N}(\mu, \sigma^2)$. If we scale the data as $x_i \mapsto ax_i$ $(a > 0)$, then the parameter (μ, σ) is transformed to $(a\mu, a\sigma)$. Thus $\mathrm{N}(\mu, \sigma^2)$ is a composite transformation model with respect to the scale transformation, where the parameter of interest is $\theta = \mu/\sigma$ and the nuisance parameter is $\nu = \sigma$. An orbit containing an observation $(x_1, \ldots, x_n) \in \mathbb{R}^n$ is given by $\{(ax_1, \ldots, ax_n) \mid a > 0\}$. The maximal invariant is, for example, $w(x) = (x_1/\hat{\sigma}, \ldots, x_n/\hat{\sigma})$, where $\hat{\sigma}^2$ denotes the sample variance. The distribution of $w(x)$ depends only on θ. Note that the maximal invariant on the minimum sufficient statistic $(\bar{x}, \hat{\sigma})$ is $\bar{x}/\hat{\sigma}$, which is a constant multiple of Student's t statistic (Cox and Hinkley, 1974, Example 5.16).

Now consider a random sample $x = (x_1, \ldots, x_n)$ of size n from the distribution $P_{\theta,\nu}$ with the density function $p_{\theta,\nu}$. It is natural to make inference based on the maximal invariant $w = w(x)$. Denote the marginal density function of w by $\bar{p}_\theta(w)$ and consider the Kullback–Leibler divergence between two density functions $\bar{p}_{\theta_1}$ and $\bar{p}_{\theta_2}$:

$$D_n(\theta_1, \theta_2) = \frac{1}{n} \int \bar{p}_{\theta_1}(w) \log \frac{\bar{p}_{\theta_1}(w)}{\bar{p}_{\theta_2}(w)} \mathrm{d}w. \tag{24}$$

We call it the marginal divergence. The reason why the right hand side is divided by n is that w has information of order $\mathrm{O}(n)$. On the other hand, a divergence function independent of the nuisance parameter is defined by

$$\tilde{D}(\theta_1, \theta_2) := \inf_{\nu_2 \in N} \mathrm{KL}(p_{\theta_1,\nu_1}, p_{\theta_2,\nu_2}) \quad (\forall \nu_1), \tag{25}$$

where the right hand side does not depend on ν_1 because of the invariance of the Kullback–Leibler divergence. We call $\tilde{D}$ the profile divergence, which is also called the profile discrimination information in Section 4 of Barndorff-Nielsen and Jupp (1988). The following lemma is obtained by Laplace approximation of the marginal distribution. The fact is essentially stated in Section 6 of Barndorff-Nielsen and Jupp (1988).

Lemma 7. If the true parameter is θ, then we have

$$\lim_{n\to\infty} \frac{1}{n} \log \frac{\bar{p}_\theta(w)}{\bar{p}_\phi(w)} = \lim_{n\to\infty} D_n(\theta, \phi) = \tilde{D}(\theta, \phi)$$

under some regularity conditions.

Example 2 (cont.)**.** For the last example, the marginal divergence is the Kullback–Leibler divergence between two non-central t distributions and not explicitly expressed. However, the profile divergence is obtained explicitly:

$$\begin{aligned}\tilde{D}(\theta_1, \theta_2) &= \inf_{\sigma_2 > 0} \left(-\frac{1}{2} \log \frac{\sigma_1^2}{\sigma_2^2} - \frac{1}{2} + \frac{(\theta_1\sigma_1 - \theta_2\sigma_2)^2 + \sigma_1^2}{2\sigma_2^2} \right) \\ &= -\log \left(\frac{\theta_1\theta_2 + \sqrt{\theta_1^2\theta_2^2 + 4(\theta_1^2+1)}}{2(\theta_1^2+1)} \right) - \frac{\theta_1\theta_2}{2} \left(\frac{\theta_1\theta_2 + \sqrt{\theta_1^2\theta_2^2 + 4(\theta_1^2+1)}}{2(\theta_1^2+1)} \right) + \frac{\theta_2^2}{2}.\end{aligned}$$

In particular, if $\theta_1 = 0$, then $\tilde{D}(0, \theta_2) = \theta_2^2/2$, and if $\theta_2 = 0$, then $\tilde{D}(\theta_1, 0) = (1/2)\log(\theta_1^2 + 1)$. Hence $\tilde{D}$ is quite asymmetric. The fact reflects to the asymmetry between size and power of the t test. In this way, we can study properties of the composite transformation models by focusing on the profile divergence.

B Proofs

B.1 Proof of Theorem 1

From the definition of profile divergence, we can choose a sequence of density functions $\{p_m\}_{m=1}^\infty \subset [p]$ such that $\mathrm{KL}(p_m, q)$ converges to $\tilde{D}([p],[q])$. Since $\{p_m\}$ is a tight sequence, we assume that it weakly converges to a distribution (which may not be absolutely continuous) without loss of generality. Then, letting $q(x) \leq M < \infty$ since $q(x)$ is bounded, we obtain

$$\begin{aligned}\int p_m(x) \log p_m(x)\, \mathrm{dx} &= \mathrm{KL}(p_m, q) + \int p_m(x) \log q(x) \mathrm{d}x \\ &\leq \mathrm{KL}(p_m, q) + \log M.\end{aligned}$$

Hence we have

$$\liminf_{m\to\infty} \int p_m(x) \log p_m(x) \mathrm{d}x \leq \tilde{D}([p],[q]) + \log M < \infty.$$

Then from Corollary 3.5 of McCann (1997), the weak limit of $\{p_m\}$ is absolutely continuous. Denote the limit density by p_∞. Then Lemma 3.4 of McCann (1997) implies

$$\liminf_{m\to\infty} \int p_m(x) \log p_m(x) \mathrm{d}x \geq \int p_\infty(x) \log p_\infty(x) \mathrm{d}x.$$

Since $q(x)$ is upper semi-continuous, we have

$$\liminf_{m\to\infty} \int p_m(x) \log(1/q(x)) \mathrm{d}x \geq \int p_\infty(x) \log(1/q(x)) \mathrm{d}x$$

(e.g. van der Vaart (2000)). Therefore we obtain

$$\tilde{D}([p],[q]) = \lim_{m\to\infty} \int p_m(x) \log \frac{p_m(x)}{q(x)} \mathrm{d}x \geq \int p_\infty(x) \log \frac{p_\infty(x)}{q(x)} \mathrm{d}x > 0,$$

where the last inequality follows since p_∞ is different from q. Indeed, if $p_\infty = q$, then p_∞ is also a copula density, but $p_\infty = p$ since p_∞ is the weak limit of $p_m \in [p]$. This contradicts to the assumption $p \neq q$.

B.2 Proof of Theorem 2

We only prove the two-dimensional case. First consider the two-dimensional piecewise uniform densities $p = p_\theta$ and $q = p_\phi$ of the form (17). Assume that the true density is p and derive the asymptotic form of $\bar{q}_n(R)$. Theorem 3 implies

$$\bar{q}_n(R) = (IJ)^{-n} \sum_\sigma \sum_\tau \frac{1}{\prod_i \sigma_i! \prod_j \tau_j!} \prod_i \prod_j \phi_{ij}^{n_{ij}(R,\sigma,\tau)},$$

where $\sigma = (\sigma_i)$ and $\tau = (\tau_j)$ are marginal frequencies. Letting $\hat{\pi}_{ij} = n_{ij}/n$, we obtain

$$\log \bar{q}_n(R) = -n \log(IJ) + \log \left(\sum_\sigma \sum_\tau \frac{\prod_i \prod_j \phi_{ij}^{n\hat{\pi}_{ij}}}{\prod_i (n\hat{\pi}_{i+})! \prod_j (n\hat{\pi}_{+j})!} \right).$$

For any non-negative integers m, an inequality $m \log m - m \leq \log m! \leq (m+1)\log(m+1) - m$ holds (see e.g. Feller (1968, Section II.9)). The number of distinct values of σ and τ are $\binom{n+I-1}{n}$ and $\binom{n+J-1}{n}$, respectively, which are of polynomial order. Hence Laplace's approximation can be applied as

$$\frac{1}{n} \log((n!)^2 \bar{q}_n(R)) = \sup_{\sigma,\tau} \Big(\sum_i \sum_j \hat{\pi}_{ij} \log \phi_{ij} - \sum_i \hat{\pi}_{i+} \log(I\hat{\pi}_{i+}) - \sum_j \hat{\pi}_{+j} \log(J\hat{\pi}_{+j}) \Big) + \mathrm{o}(1),$$

where o(1) denotes a term converging to 0 as $n \to \infty$ uniformly in R. Now using an empirical measure $\hat{P} = n^{-1}\sum_{t=1}^{n}\delta_{(x_{t1},x_{t2})}$, we write $\hat{\pi}_{ij} = \hat{\pi}_{ij}^{T} := \hat{P}(T^{-1}(A_{ij}))$ with a map $T \in \mathcal{T}$. Here A_{ij} is the small rectangular region defined in Section 4. Furthermore, T is linear on each region $T^{-1}(A_{ij})$. Then determining T is equivalent to determining σ and τ. Then we have

$$\frac{1}{n}\log((n!)^2\bar{q}_n(R)) = \sup_{T}\Big(\sum_i\sum_j \hat{\pi}_{ij}^{T}\log\phi_{ij} - \sum_i \hat{\pi}_{i+}^{T}\log(I\hat{\pi}_{i+}) - \sum_j \hat{\pi}_{+j}^{T}\log(J\hat{\pi}_{+j})\Big) + \mathrm{o}(1).$$

From Glivenko–Cantelli's theorem (e.g. van der Vaart (2000)), the probability $\hat{\pi}_{ij}^{T}$ converges to $\pi_{ij}^{T} := P(T^{-1}(A_{ij}))$ uniformly in T with probability one. Here P is the true distribution, that is, $P(dx) = p(x)dx$. Now, with probability one,

$$\lim_{n\to\infty}\frac{1}{n}\log((n!)^2\bar{q}_n(R)) = \sup_{T}\Big(\sum_i\sum_j \pi_{ij}^{T}\log\phi_{ij} - \sum_i \pi_{i+}^{T}\log(I\pi_{i+}^{T}) - \sum_j \pi_{+j}^{T}\log(J\pi_{+j}^{T})\Big).$$

On the other hand, since T is linear on $T^{-1}(A_{ij})$, the derivative is $T_1'(x_1) = 1/(I\pi_{i+}^{T})$ if $x \in T^{-1}(A_{ij})$, and so on. Therefore

$$\begin{aligned}\mathrm{KL}(p, T^{*}q) &= \int p(x)\log p(x)dx + \sum_i\sum_j\int_{T^{-1}(A_{ij})} p(x)\log\frac{1}{\phi_{ij}T_1'(x_1)T_2'(x_2)}dx \\ &= \int p(x)\log p(x)dx - \sum_i\sum_j \pi_{ij}^{T}\log\phi_{ij} + \sum_i \pi_{i+}^{T}\log(I\pi_{i+}^{T}) + \sum_j \pi_{+j}^{T}\log(J\pi_{+j}^{T}).\end{aligned}$$

We proved that

$$\lim_{n\to\infty}\frac{1}{n}\log\Big((n!)^2\bar{q}_n(R)\Big) = \int p(x)\log p(x)dx - \inf_{T}\mathrm{KL}(p, T^{*}q).$$

The same thing holds for $\bar{p}_n(R)$. Finally we have

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{n}\log\Big(\frac{\bar{p}_n(R)}{\bar{q}_n(R)}\Big) &= -\inf_{T}\mathrm{KL}(p, T^{*}p) + \inf_{T}\mathrm{KL}(p, T^{*}q) \\ &= \inf_{T}\mathrm{KL}(p, T^{*}q) \\ &= \tilde{D}([p],[q]),\end{aligned}$$

where the last equality comes from Lemma 4.

If p and q are not piecewise uniform, then the density is approximated by a piecewise one. More specifically, for any $\varepsilon > 0$, there exists a piecewise uniform p_ε such that $(1-\varepsilon)p_\varepsilon \le p \le (1+\varepsilon)p_\varepsilon$. Define q_ε in a similar way. The rank likelihood is evaluated as

$$\frac{1}{n}\log\frac{\bar{p}_n(R)}{\bar{q}_n(R)} \le \frac{1}{n}\log\frac{(1+\varepsilon)^n\bar{p}_{\varepsilon,n}(R)}{(1-\varepsilon)^n\bar{q}_{\varepsilon,n}(R)} \to \log\frac{1+\varepsilon}{1-\varepsilon} + \tilde{D}([p_\varepsilon],[q_\varepsilon]) \quad (n\to\infty).$$

The lower bound is similarly evaluated. The limit $\varepsilon \to 0$ gives the result.

The convergence of the rank divergence follows from the bounded convergence theorem. Indeed, since p is bounded from below and above by the assumption, we obtain $C_0^n \le \bar{p}_n(R) \le C_1^n$ using $0 < C_0 \le p \le C_1 < \infty$. The expression on $\bar{q}_n(R)$ is similar. Therefore $n^{-1}\log(\bar{p}_n(R)/\bar{q}_n(R))$ is bounded.

B.3 Proof of Theorem 3

We abbreviate $\bar{p}_n(R)$ as $p(R)$. From Eq. (18), the marginal distribution of $N := (n_{ij})$ is

$$p(N) = \frac{n!}{\prod_i \prod_j n_{ij}!} (IJ)^{-n} \prod_i \prod_j \theta_{ij}^{n_{ij}}.$$

We derive the conditional distribution $p(R|N)$ of R given N. Since N is a sufficient statistic on θ, $p(R|N)$ does not depend on θ. Therefore we can assume that X is distributed according to the uniform distribution without loss of generality. Since σ and τ are functions of N, we obtain

$$\begin{aligned} p(R|N) &= p(R|N,\sigma,\tau) \\ &= \frac{p(R,N|\sigma,\tau)}{p(N|\sigma,\tau)} \\ &= \begin{cases} \dfrac{p(R|\sigma,\tau)}{p(N|\sigma,\tau)} & \text{if } N = N(R,\sigma,\tau), \\ 0 & \text{otherwise.} \end{cases} \end{aligned}$$

Since X is a random sample from the uniform distribution, $p(N|\sigma,\tau)$ is the hypergeometric distribution and $p(R|\sigma,\tau) = p(R) = 1/(n!)^2$. In summary, the marginal distribution of R is

$$\begin{aligned} p(R) &= \sum_N p(R|N)p(N) \\ &= \sum_\sigma \sum_\tau p(R|\sigma,\tau) \frac{1}{p(N|\sigma,\tau)} p(N) \Bigg|_{N=N(R,\sigma,\tau)} \\ &= \sum_\sigma \sum_\tau \frac{1}{(n!)^2} \frac{n! \prod_i \prod_j n_{ij}!}{\prod_i \sigma_i! \prod_j \tau_j!} \frac{n!}{\prod_i \prod_j n_{ij}!} (IJ)^{-n} \prod_i \prod_j \theta_{ij}^{n_{ij}} \Bigg|_{N=N(R,\sigma,\tau)}, \end{aligned}$$

which yields Eq. (19).

B.4 Proof of Theorem 4

It is widely known that the Fisher information matrix of the multivariate normal distribution is

$$g_{ij} = \frac{1}{2} \operatorname{tr}(P^{-1}(\partial_i P) P^{-1}(\partial_j P)).$$

In general, the metric induced from the profile divergence in the sense of Eq. (25) is given by

$$\tilde{g}_{ij} = \mathrm{E}[\{\partial_i \ell - \Pi(\partial_i \ell)\}\{\partial_j \ell - \Pi(\partial_j \ell)\}]$$

(Barndorff-Nielsen and Jupp, 1988, Theorem 7.2), where $\ell = \log p_{\theta,\nu}$ is the log likelihood and Π is the orthogonal projection onto the nuisance tangent space spanned by $\partial \ell / \partial \nu$. Since the correspondence $\partial_i P \mapsto \partial_i \ell$ of tangent vectors is a linear isomorphism, we have Eq. (21). Now determine the orthogonal projection Π. The nuisance parameter u appears in $\mathrm{diag}(u)P\,\mathrm{diag}(u)$ and therefore the tangent vector at $u = 1_d$ is

$$\partial_{u_k}(\mathrm{diag}(u)P\,\mathrm{diag}(u)) = \mathrm{diag}(e_k)P + P\,\mathrm{diag}(e_k), \quad k = 1, \ldots, d,$$

where e_k is the k-th unit vector. Hence Π is of the form (22). Let $\Pi(A) = PB + BP$ where $B = \mathrm{diag}(b)$. Then, for any diagonal matrix $C = \mathrm{diag}(c)$, we have

$$\begin{aligned} 0 &= \frac{1}{2}\mathrm{tr}\{P^{-1}(PC + CP)P^{-1}(A - (PB + BP))\} \\ &= \sum_k c_k(P^{-1}A)_{kk} - \sum_k c_k b_k - \sum_k \sum_l c_k b_l (P^{-1})_{kl} P_{kl}, \end{aligned}$$

which implies $(P^{-1} \circ A)1_d - (I + P^{-1} \circ P)b = 0$ and Eq. (23) follows.

B.5 Proof of Theorem 5

Let $S = \Sigma^{-1}$. Consider a transformation $v_i = x_{t(n,i),i}$ and $w_{ri} = x_{t(r+1,i),i} - x_{t(r,i),i}$. Then the integration region is $v_i \in \mathbb{R}$ and $w_{ri} > 0$. The inverse transformation is $x_{ti} = v_i - \sum_{r=1}^{n-1} I_{\{r_{ti} \leq r\}} w_{ri}$, where I is the definition function. The exponential part of the joint density function of X is

$$\begin{aligned} \sum_{t=1}^{n}\sum_{i=1}^{d}\sum_{j=1}^{d} S_{ij}x_{ti}x_{tj} &= \sum_{t=1}^{n}\sum_{i=1}^{d}\sum_{j=1}^{d} S_{ij}\left(v_i - \sum_{r=1}^{n} I_{\{r_{ti} \leq r\}} w_{ri}\right)\left(v_j - \sum_{s=1}^{n} I_{\{r_{tj} \leq s\}} w_{sj}\right) \\ &= \sum_i \sum_j \left(nS_{ij}v_iv_j - 2nS_{ij}m_iv_j + \sum_r \sum_s S_{ij} I_{\{r_{ti} \leq r\}} I_{\{r_{tj} \leq s\}} w_{ri}w_{sj}\right), \end{aligned}$$

where

$$m_i = \frac{1}{n}\sum_{t=1}^{n}\sum_{r=1}^{n-1} I_{\{r_{ti} \leq r\}} w_{ri} = \frac{1}{n}\sum_{r=1}^{n-1} r w_{ri}.$$

Thus the exponential part of the joint density function of $(v_i), (w_{ri})$ is

$$n\sum_i \sum_j S_{ij}(v_i - m_i)(v_j - m_j) + \sum_i \sum_j \sum_r \sum_s B_{(r-1)d+i,(s-1)d+j} w_{ri}w_{sj}.$$

Integration with respect to v_i yields the result.

References

Amari, S. (1985). *Differential-Geometrical Methods in Statistics*, Springer, New York.

Barndorff-Nielsen, O. E. and Jupp, P. E. (1988). Differential geometry, profile likelihood, L-sufficiency and composite transformation models, *The Annals of Statistics*, **16** (3), 1009–1043.

Bickel, P. J., Klaassen, C. A. J., Ritov, Y. and Wellner, J. A. (1993). *Efficient and adaptive estimation for semiparametric models*, Johns Hopkins University Press, Baltimore.

Chen, X., Fan, Y. and Tsyrennikov, V. (2006). Efficient estimation of semiparametric multivariate copula models, *Journal of the American Statistical Association*, **101** (475), 1228–1240.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*, Chapman and Hall, London.

Eaton, M. L. (1983). *Multivariate Statistics – A Vector Space Approach*, Wiley, New York.

Eguchi, S. (1983). Second order efficiency of minimum contrast estimators in a curved exponential family, *The Annals of Statistics*, **11** (3), 793–803.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications*, 3rd ed., John Wiley & Sons, New York.

Genest, C. and Werker, B. J. M. (2002). Conditions for the asymptotic semiparametric efficiency of an omnibus estimator of dependence parameters in copula models, *Distributions with Given Marginals and Statistical Modelling* (eds. C. M. Cuadras and J. A. R. Lallena), 103–112, Kluwer Academic, Dordrecht.

Ghosh, S. and Henderson, S. G. (2001). Chessboard distributions and random vectors with specified marginals and covariance matrix, *Operations Research*, **50** (5), 820–834.

Hoff, P. D. (2007). Extending the rank likelihood for semiparametric copula estimation, *The Annals of Applied Statistics*, **1** (1), 265–283.

Hoff, P. D., Niu, X. and Wellner, J. A. (2014). Information bounds for Gaussian copulas, *Bernoulli*, **20** (2), 604–622.

Klaassen, C. A. J. and Wellner, J. A. (1997). Efficient estimation in the bivariate normal copula model: normal margins are least favourable, *Bernoulli*, **3** (1), 55–77.

Koyama, T. and Takemura, A. (2015). Calculation of orthant probabilities by the holonomic gradient method, *Japan Journal of Industrial and Applied Mathematics*, **32**, 187–204.

Koyama, T., Nakayama, H., Ohara, K., Sei, T. and Takayama, N. (2014). Software packages for holonomic gradient method, *Mathematical Software – ICMS 2014. ICMS 2014. Lecture Notes in Computer Science* (eds. H. Hong and C. Yap), 8592, 706–712, Springer, Berlin.

Marshall, A. W. and Olkin, I. (1968). Scaling of matrices to achieve specified row and column sums, *Numerische Mathematik*, **12**, 83–90.

McCann, R. J. (1997). A convexity principle for interacting gases, *Advances in Mathematics*, **128**, 153–179.

Nelsen, R. B. (2006). *An Introduction to Copulas*, 2nd ed., Springer, New York.

Segers, J., van den Akker, R. and Werker, B. J. M. (2014). Semiparametric Gaussian copula models: geometry and efficient rank-based estimation, *The Annals of Statistics*, **42** (5), 1911–1940.

Tsukahara, H. (2005). Semiparametric estimation in copula models, *The Canadian Journal of Statistics*, **33** (3), 357–375.

van der Vaart, A. W. (2000). *Asymptotic Statistics*, Cambridge University Press, Cambridge.