Copulas and Information Geometry: Strange Bedfellows?

Robert Zimmerman
Department of Statistical Sciences
University of Toronto
Toronto, ON
`robert.zimmerman@mail.utoronto.ca`

Abstract

Over the past several decades, copulas have been widely applied in probability and statistics to model complex structures governing the dependence between components of random vectors. Despite the nearly concurrent increased interest and theoretical developments in information geometry, it appears that little work has been done in the intersection of these two fields — in fact, we identified only two papers written in the last decade which focus on this intersection. In this report, we critically analyze the discussions in these papers of information-geometric connections to copulas. We find that the connections are relatively superficial and conclude that a paper exploring deep theoretical connections has yet to be written.

1 Introduction

A d-dimensional *copula* is a function $C : [0,1]^d \to [0,1]$ which satisfies three properties:

$$\begin{aligned} C(1,\dots,1,u_i,1,\dots,1) &= u, \quad 0 \leqslant u_i \leqslant 1 \\ C(u_1,\dots,u_{i-1},0,u_{i+1},\dots,u_d) &= 0, \quad 0 \leqslant u_i \leqslant 1 \\ \sum_{i_1=1}^{2} \cdots \sum_{i_d=1}^{2} (-1)^{i_1+\cdots i_d} C\left(u_1^{(i_1)},\dots,u_d^{(i_d)}\right) &\geqslant 0, \quad 0 \leqslant u_j^{(1)} \leqslant u_j^{(2)} \leqslant 1. \end{aligned}$$

In statistics and probability, copulas are operationalized by *Sklar's theorem* [Sklar, 1959]. This foundational result states roughly that any copula is the distribution function of some random vector $(U_1,\dots,U_d)$ with $U_i \sim \text{Unif}[0,1]$ for each $i \in \{1,\dots,d\}$, and conversely that any random vector $(X_1,\dots,X_d)$ with $X_i \sim F_i$ for each $i \in \{1,\dots,d\}$ induces a copula through the relation $C(\boldsymbol{u}) := \mathbb{P}(F_1(X_1) \leqslant u_1,\dots,F_d(X_d) \leqslant u_d)$. The latter correspondence is unique when the marginal distributions $F_1,\dots,F_d$ are continuous; otherwise, C is uniquely determined on $\times_{h=1}^{d} \operatorname{Ran}(F_i)$ [Nelsen, 2007].

Copulas allow the study of a random vector's dependence structure separately from its marginals. A copula is often used when the elliptical dependence structure of the multivariate Gaussian distribution does not capture the (often quite complex) dependencies observed in multivariate modelling. A classic example of this phenomenon is a time series of stock returns, in which dependence often becomes prominent only at the extremes [McNeil et al., 2015].

Many — if not most — copulas used in practice are members of parametric families, through which some parameter θ usually characterizes the "strength" of the dependence between the marginals, often interpolating between perfect negative dependence (when $d = 2$) and independence, or else independence and perfect positive dependence. Parametric families of copulas that interpolate all three are called *comprehensive*. The *Gauss copula*, for example, is defined as $C_R(\boldsymbol{u}) = \Phi_{(\mathbf{0},R)}\left(\Phi^{-1}(u_1),\dots,\Phi^{-1}(u_d)\right)$ — where $\Phi(\cdot)$ is the cdf of the (univariate) $\mathcal{N}(0,1)$

distribution, $\Phi_{(\mathbf{0},R)}$ is the cdf of the (multivariate) $\mathcal{N}_d\left(\mathbf{0}, R\right)$ distribution and R is a correlation matrix. It is perhaps the most widely used copula outside of finance and extreme value modelling, because it is tractable and easy to interpret. Also popular are the *Archimedean copulas*; these comprise families of the form $\{C_\theta(\boldsymbol{u}) = \varphi_\theta^{-1}\left(\varphi_\theta(u_1) + \cdots + \varphi_\theta(u_d)\right) : \theta \in \Theta\}$, where φ_θ is a family-specific *Archimedean generator* satisfying several regularity conditions. Countless modifications and extensions of these copula families have appeared over the years; spin-offs of the Archimedean copulas alone include nested Archimedean copulas [Hofert and Pham, 2013], outer power copulas [Górecki et al., 2021], Archimax copulas [Klement et al., 2005], and many more. Meanwhile, entirely new families of copulas are constantly being developed.

Separately, the field of *information geometry* seeks to analyze parametric families of absolutely continuous distributions through the lens of differential geometry by viewing a parametric family of densities $\{f_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta\}$ as a *statistical manifold* M (that is, a Riemannian manifold equipped a suitable Riemannian metric, such as the Fisher-Rao metric — see Subsection 3.2) [Ay et al., 2017]. Among other innovations, this viewpoint allows one to construct various notions of distances (or, more precisely, *statistical divergences*) between distributions within the same family. Some extensions to non-parametric statistics have been developed as well [Pistone and Sempi, 1995, Zhang, 2013, Pistone, 2013]).

Briefly, a *divergence* $\mathcal{D}(\cdot \,||\, \cdot)$ is a function defined on a statistical manifold M satisfying the following three properties [Amari, 2016]:

1. $\mathcal{D}(\boldsymbol{\theta}_1 \,||\, \boldsymbol{\theta}_2) \geqslant 0$ for all $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in M$;
2. $\mathcal{D}(\boldsymbol{\theta}_1 \,||\, \boldsymbol{\theta}_2) = 0$ if and only if $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$; and
3. $\mathcal{D}$ admits a Taylor expansion of the form

   $$\mathcal{D}(\boldsymbol{\theta} \,||\, \boldsymbol{\theta} + \delta\boldsymbol{\theta}) = \frac{1}{2}\sum_{i,j=1}^{d} g_{ij}(\boldsymbol{\theta})\,\mathrm{d}\theta_i\,\mathrm{d}\theta_j + O\left(|\delta\boldsymbol{\theta}|^3\right)$$

   for some positive definite matrix $G(\boldsymbol{\theta}) = [g_{i,j}(\boldsymbol{\theta})]_{i,j}$ and an infinitesimal displacement $\boldsymbol{\theta} + \delta\boldsymbol{\theta}$ from $\boldsymbol{\theta}$.

Two important families of divergences are the Bregman divergences and the f-divergences. Given a strictly convex differentiable function ψ, a *Bregman divergence* takes the form

$$\mathcal{D}_\psi(\boldsymbol{\theta}_1 \,||\, \boldsymbol{\theta}_2) = \psi(\boldsymbol{\theta}_1) - \psi(\boldsymbol{\theta}_2) - \langle \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2, \nabla\psi(\boldsymbol{\theta}_2)\rangle,$$

which can be interpreted geometrically as the vertical distance between $\psi(\boldsymbol{\theta}_1)$ and the hyperplane tangent to ψ at the point $\boldsymbol{\theta}_2$. Given a strictly convex function $f : \mathbb{R}_+ \to \mathbb{R} \cup \{\infty\}$ which satisfies $f(1) = 0$, an *f-divergence* takes the form

$$\mathcal{D}_f(\mathbb{P}_{\boldsymbol{\theta}_1} \,||\, \mathbb{P}_{\boldsymbol{\theta}_2}) = \int_{\mathcal{X}} f\left(\frac{\mathrm{d}\mathbb{P}_{\boldsymbol{\theta}_1}}{\mathrm{d}\mathbb{P}_{\boldsymbol{\theta}_2}}\right)\mathrm{d}\mathbb{P}_{\boldsymbol{\theta}_2}$$

where $\mathbb{P}_{\boldsymbol{\theta}_1} \ll \mathbb{P}_{\boldsymbol{\theta}_2}$. While f-divergences are defined on a spaces of measures, any absolutely continuous probability measure $\mathbb{P}_{\boldsymbol{\theta}}$ can obviously be identified with $\boldsymbol{\theta} \in M$. The famous *Kullback-Leibler (KL) divergence*, perhaps the most important example of a statistical divergence, can be obtained both as a Bregman divergence with $\psi(\boldsymbol{\theta}) = \sum_{d=1}^{D} \theta_d \cdot \log(\theta_d)$ and as an f-divergence with $f(x) = x \cdot \log(x)$); it is the only statistical divergence which is simultaneously a Bregman divergence and an f-divergence on the space of probability measures [Amari, 2009].

Theoretically, any parametric family of absolutely continuous copula densities $\{c_\theta : \theta \in \Theta\}$ with $\Theta \subseteq \mathbb{R}^p$ — as is virtually always the case — should qualify for an information-geometric analysis. In this report, we examine two papers which deal with the intersection of copulas and information geometry; both focus on well-known statistical methodologies (clustering and variational inference, respectively), and both use copulas and information-geometric notions to devise improvements over more basic schemes. In these papers, the primary connection to information-geometric is the quantification of distances between copulas (or multivariate distributions) via statistical divergences.

The remainder of this report is organized as follows. In Sections 2 and 3, we provide some background for the statistical methodology featured in each paper, discuss the main techniques and results in an information geometric context, and critically evaluate the papers. In Section 4, we end with a discussion and briefly speculate on why there has apparently been so little work featuring both copulas and information geometric concepts.

2 Copula Variational Bayes

Our first paper is *Copula Variational Bayes inference via information geometry* [sic] [Tran, 2018], which as of this writing is still in preprint on arXiv.

2.1 Background

Variational Bayes, sometimes known as *variational inference*,[1] is a popular method for conducting approximate statistical inference of complicated systems. Although (as its name suggests) the technique is most often used in Bayesian applications, the basic approach need not be confined to these. Briefly, a complicated d-dimensional density $f_{\boldsymbol{\theta}} = f_{\boldsymbol{\theta}}(\boldsymbol{\theta})$ — which is impractical or impossible to compute or sample from directly — is approximated in a sensible way by a more analytically convenient density. In Bayesian statistics, for example, posterior densities given by

$$f_{\boldsymbol{\theta}}(\boldsymbol{\theta}) := \frac{\pi(\boldsymbol{\theta}) \cdot f(\boldsymbol{x} \mid \boldsymbol{\theta})}{\int_{\Theta} \pi(\boldsymbol{\theta}) \cdot f(\boldsymbol{x} \mid \boldsymbol{\theta}) \, \mathrm{d}\boldsymbol{\theta}}$$

are typically unavailable in closed form due to the intractable normalizing constant in the denominator. In other cases, the complex dependence structure between the components of $\boldsymbol{\theta}$ may pose its own additional challenges.

In the variational Bayes setup, an *approximating family* of tractable d-dimensional distributions $\mathcal{C}$ is chosen first, and then a unique approximating distribution is selected as

$$\widetilde{f}_{\boldsymbol{\theta}} = \underset{g \in \mathcal{C}}{\arg\min}\, \mathcal{D}_{\mathrm{KL}}\left(g \,||\, f_{\boldsymbol{\theta}}\right) \tag{1}$$

The term "variational" refers to the calculus of variations, since the optimization problem defined by Equation 1 is solved using techniques borrowed from the calculus of variations [Tran, 2018]. More general notions of variational inference were initially used for neural networks [Peterson, 1987, Hinton and Van Camp, 1993] before being applied to parametric graphical models by Jordan et al. [1999]; its application to the Bayesian paradigm then followed naturally.

The classic variational Bayes framework chooses $\mathcal{C}$ to be a class of factorial densities — that is, densities of the form $\widetilde{f}_{\boldsymbol{\theta}} = \prod_{h=1}^{d} \widetilde{f}_h$, so that the marginals in the approximating family are independent. However, the independence requirement can be quite restrictive, especially when the true density $f_{\boldsymbol{\theta}}$ is believed to induce complicated dependence structures among the θ_h's. The present paper — as well as several others published before and since [Tran et al., 2015, Smith et al., 2020, Gunawan et al., 2021, Chi et al., 2022] — considers loosening this restriction to allow $\mathcal{C}$ to be a class of multivariate densities with a fixed copula but varying marginals. This idea is motivated by the fact that one can, at least in principle, compute variational approximations on the marginal distributions in such a way that the calculations does not "interfere" with with the copula that binds the marginals together. Tran develops an algorithm for accomplishing this computation using principles from information geometry, as we discuss below.

2.2 Main Information Geometric Results

Tran introduces Bregman divergences at length, starting with its definition and its basic properties. Among the latter, the so-called *three-point property* is highlighted: for any $\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\gamma} \in \mathbb{R}^d$, a Bregman divergence $\mathcal{D}_{\psi}(\cdot \,||\, \cdot)$ satisfies the following characterizing property [Ay et al., 2017]:

$$\mathcal{D}_{\psi}\left(\boldsymbol{\alpha} \,||\, \boldsymbol{\beta}\right) + \mathcal{D}_{\psi}\left(\boldsymbol{\beta} \,||\, \boldsymbol{\gamma}\right) - \mathcal{D}_{\psi}\left(\boldsymbol{\alpha} \,||\, \boldsymbol{\gamma}\right) = \left\langle \boldsymbol{\beta} - \boldsymbol{\alpha}, \nabla\psi(\boldsymbol{\beta}) - \nabla\psi(\boldsymbol{\gamma}) \right\rangle .$$

Tran then combines two more fundamental results of information geometry — the *projection theorem* and the *generalized Pythagorean theorem* [Amari, 2016] — into one theorem, which provides the foundational basis for the copula variational Bayes algorithm:

Theorem 2.1 (Bregman Pythagorean inequality). *Let $\mathcal{X} \subseteq \mathbb{R}^d$ be closed and convex and let $\boldsymbol{\gamma} \in \mathbb{R}^d$. Then the Bregman projection of $\boldsymbol{\gamma}$ onto $\mathcal{X}$, defined by $\boldsymbol{\beta}_{\mathcal{X}} := \arg\min_{\boldsymbol{\alpha} \in \mathcal{X}} \mathcal{D}_{\psi}(\boldsymbol{\alpha} \,||\, \boldsymbol{\gamma})$, is unique. Moreover, for any $\boldsymbol{\alpha} \in \mathcal{X}$, we have*

$$\mathcal{D}_{\psi}(\boldsymbol{\alpha} \,||\, \boldsymbol{\beta}_{\mathcal{X}}) + \mathcal{D}_{\psi}(\boldsymbol{\beta}_{\mathcal{X}} \,||\, \boldsymbol{\gamma}) \leqslant \mathcal{D}_{\psi}(\boldsymbol{\alpha} \,||\, \boldsymbol{\gamma}), \tag{2}$$

where equality holds if and only if $\boldsymbol{\beta}_{\mathcal{X}} - \boldsymbol{\alpha}$ and $\nabla\psi(\boldsymbol{\beta}_{\mathcal{X}}) - \nabla\psi(\boldsymbol{\gamma})$ are orthogonal.

[1] Some authors consider the two terms as synonyms, while others consider variational Bayes to be a special case of variational inference.

The first part of Theorem 2.1 — the projection theorem, which asserts the uniqueness of the Bregman projection $\boldsymbol{\beta}_{\mathcal{X}}$, also known as the *information projection* [Nielsen, 2018] — has been proved by Amari [2016], Ay et al. [2017], and many others, in varying levels of generality. The inequality in Equation 2 follows from showing that $\langle \boldsymbol{\beta}_{\mathcal{X}} - \boldsymbol{\alpha}, \nabla\psi(\boldsymbol{\beta}_{\mathcal{X}}) - \nabla\psi(\boldsymbol{\gamma})\rangle \leqslant 0$. In greater generality, the condition equivalent for equality to hold may be restated in information-geometric terms as a requirement that the dual geodesic connecting $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}_{\mathcal{X}}$ be orthogonal to the geodesic connecting $\boldsymbol{\beta}_{\mathcal{X}}$ and $\boldsymbol{\gamma}$ [Amari, 2016].

Following a lengthy introduction to copulas, Tran narrows his focus to the KL-divergence and proves the following theorem:

Theorem 2.2. *Let $f_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = c(\boldsymbol{u}(\boldsymbol{\theta})) \cdot \prod_{h=1}^{d} f_h(\theta_h)$ and $\widetilde{f}_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \widetilde{c}(\boldsymbol{u}(\boldsymbol{\theta})) \cdot \prod_{h=1}^{d} \widetilde{f}_h(\theta_h)$ be two densities on $\mathbb{R}^d$. Then*

$$\mathcal{D}_{\mathrm{KL}}\left(\widetilde{f}_{\boldsymbol{\theta}} \,||\, f_{\boldsymbol{\theta}}\right) = \mathcal{D}_{\mathrm{KL}}\left(\widetilde{c}(\boldsymbol{u}) \,||\, c(F(\widetilde{F}^{\leftarrow}(\boldsymbol{u})))\right) + \sum_{h=1}^{d} \mathcal{D}_{\mathrm{KL}}\left(\widetilde{f}_h \,||\, f_h\right) \tag{3}$$

$$\geqslant \sum_{h=1}^{d} \mathcal{D}_{\mathrm{KL}}\left(\widetilde{f}_h \,||\, f_h\right) \tag{4}$$

where $c(F(\widetilde{F}^{\leftarrow}(\boldsymbol{u}))) = \widetilde{c}\left(F(\widetilde{F}_1^{\leftarrow}(u_1)), \ldots, F(\widetilde{F}_d^{\leftarrow}(u_d))\right)$.

Equation 3 above is a substantial generalization of the fact [Ma and Sun, 2011] that the KL-divergence from any copula density $c(\boldsymbol{u})$ to the independence copula — that is, the *copula entropy* of $c(\boldsymbol{u})$ — is equal to the mutual information contained in the random vector $\boldsymbol{U} \sim c$. The inequality Equation 4 follows immediately from the non-negativity of Bregman divergences.

A reasonable approximation to $f_{\boldsymbol{\theta}}$ would be some approximating density $\widetilde{f}_{\boldsymbol{\theta}}$ whose marginals $\widetilde{f}_1, \ldots, \widetilde{f}_d$ minimize Equation 4; however, finding such marginals is typically infeasible because the true marginals themselves $f_h(\theta_h) = \int f_{\boldsymbol{\theta}}(\boldsymbol{\theta})\, \mathrm{d}\boldsymbol{\theta}_{\setminus h}$ are unavailable. In contrast, it is conceivably much easier to find an approximating density $\widetilde{f}_{\boldsymbol{\theta}}$ that directly minimizes Equation 3, since a sensibly chosen optimization scheme may not require access to the true marginals. In addition, such an approximation would be expected to produce reasonably good marginal approximations, simply because Equation 3 is an upper bound for Equation 4. This is the motivation behind Tran's *copula variational Bayes* algorithm.

The main idea of the algorithm is to consider a family of distributions $\widetilde{f}_{\boldsymbol{\theta}} = \widetilde{f}^{*}_{\setminus h|h} \cdot \widetilde{f}_h$ for which the conditional density $\widetilde{f}^{*}_{\setminus h|h}$ is fixed, and then use Theorem 2.2 to determine the optimal marginal $\widetilde{f}_h$ which minimizes the KL-divergence from $\widetilde{f}_{\boldsymbol{\theta}}$ to $f_{\boldsymbol{\theta}}$, repeating the process for each k. Tran refers to each such step as a *conditionally variational approximation*, and encodes it in the following theorem:

Theorem 2.3. *Let $\widetilde{f}_{\boldsymbol{\theta}} = \widetilde{f}^{*}_{\setminus h|h} \cdot \widetilde{f}_h$ and $\widetilde{f}^{*}_{\boldsymbol{\theta}} = \widetilde{f}^{*}_{\setminus h|h} \cdot \widetilde{f}^{*}_h$ be two distributions with the same fixed conditional density $\widetilde{f}^{*}_{\setminus h|h}$. Then*

$$\mathcal{D}_{\mathrm{KL}}\left(\widetilde{f}_{\boldsymbol{\theta}} \,||\, f_{\boldsymbol{\theta}}\right) = \mathcal{D}_{\mathrm{KL}}\left(\widetilde{f}_{\boldsymbol{\theta}} \,||\, \widetilde{f}^{*}_{\boldsymbol{\theta}}\right) + \mathcal{D}_{\mathrm{KL}}\left(\widetilde{f}^{*}_{\boldsymbol{\theta}} \,||\, f_{\boldsymbol{\theta}}\right) \tag{5}$$

$$\geqslant \mathcal{D}_{\mathrm{KL}}\left(\widetilde{f}^{*}_{\boldsymbol{\theta}} \,||\, f_{\boldsymbol{\theta}}\right) \tag{6}$$

The optimal marginal distribution $\widetilde{f}_h = \widetilde{f}^{*}_h$ which minimizes Equation 5 is given by

$$\widetilde{f}^{*}_h(\theta_h) = \frac{1}{\zeta_h} \cdot \frac{f_h(\theta_h)}{\exp\left(\mathcal{D}_{\mathrm{KL}}\left(\widetilde{f}^{*}_{\setminus h|h} \,||\, f_{\setminus h|h}\right)\right)} \tag{7}$$

$$= \frac{1}{\zeta_h} \cdot \exp\left(\mathbb{E}_{\widetilde{f}^{*}_{\setminus h|h}}\left[\log\left(\frac{f_{\boldsymbol{\theta}}(\boldsymbol{\theta})}{\widetilde{f}^{*}_{\setminus h|h}(\boldsymbol{\theta}_{\setminus h} \mid \theta_h)}\right)\right]\right) \tag{8}$$

where ζ_h is the appropriate normalizing constant that makes Equation 7 a density, which makes Equation 6 equal to $-\log(\zeta_h)$.

The proof proceeds roughly as follows. First, Tran notes that since $\widetilde{f}_{\boldsymbol{\theta}}$ is linear in $\widetilde{f}_h$, it is convex in $\widetilde{f}_h$, and thus Theorem 2.1 applies, yielding Equation 5; the inequality in Equation 6 follows trivially from the positivity of Bregman divergences. A calculation using Theorem 2.2 exploits the fact that $\widetilde{f}_{\boldsymbol{\theta}}$ and $\widetilde{f}^*_{\boldsymbol{\theta}}$ differ only in the marginals $\widetilde{f}_h$ and $\widetilde{f}^*_h$, and shows that $\mathcal{D}_{\mathrm{KL}}(\widetilde{f}_{\boldsymbol{\theta}} \mid\mid \widetilde{f}^*_{\boldsymbol{\theta}}) = 0$ is equivalent to Equation 7; a further calculation establishes that under this choice we obtain $\mathcal{D}_{\mathrm{KL}}(\widetilde{f}^*_{\boldsymbol{\theta}} \mid\mid f_{\boldsymbol{\theta}}) = -\log(\zeta_h)$. Equation 8 then follows from rewriting the KL-divergence as an expectation and absorbing into it the "constant" f_h.

Tran notes that if instead we fix the family $\widetilde{f}_{\boldsymbol{\theta}} = \widetilde{f}_{\setminus h|h} \cdot \widetilde{f}^*_h$ where $\widetilde{f}^*_h$ is fixed and $\widetilde{f}_{\setminus h|h}$ is allowed to vary, then $\widetilde{f}_{\boldsymbol{\theta}}$ is convex over $\widetilde{f}_{\setminus h|h}$ and the optimal conditional distribution $\widetilde{f}_{\setminus h|h} = \widetilde{f}^*_{\setminus h|h}$ in Equation 5 is given by the true conditional distribution $\widetilde{f}^*_{\setminus h|h} = f_{\setminus h|h}$. However, he also observes that for most practical purposes this fact is unusable because $f_{\setminus h|h}$ is typically unknown. In contrast, the density given by Equation 8 can be computed, at least up to the normalizing constant ζ_h (see Subsection 2.3).

The conditionally variational approximation yields one optimally approximated marginal $\widetilde{f}^*_h$ given the fixed conditional $\widetilde{f}^*_{\setminus h|h}$, which then yields the updated joint density $\widetilde{f}^*_{\boldsymbol{\theta}} = \widetilde{f}^*_{\setminus h|h} \cdot \widetilde{f}^*_h$. To turn this step into an algorithm, Tran observes that we have the alternative decomposition $\widetilde{f}^*_{\boldsymbol{\theta}} = \widetilde{f}^*_{h|\setminus h} \cdot \widetilde{f}^*_{\setminus h}$, where

$$\widetilde{f}^*_{h|\setminus h} = \frac{\widetilde{f}^*_{\setminus h|h} \cdot \widetilde{f}^*_h}{\int \widetilde{f}^*_{\setminus h|h} \cdot \widetilde{f}^*_h \, \mathrm{d}\theta_h}$$

is the "reverse conditional". According to Tran, this decomposition yields the following algorithm: starting with an initial fixed conditional distribution $\widetilde{f}^{(0)}_{\setminus h|h}$, at iteration ν we set $\widetilde{f}^{(\nu)} := \widetilde{f}^{(\nu-1)}_{\setminus h|h} \cdot \widetilde{f}^{(\nu)}_h$, where $\widetilde{f}^{(\nu)}_h$ is optimized in accordance with Theorem 2.3. We then calculate the "reverse conditional"

$$\widetilde{f}^{(\nu)}_{h|\setminus h} = \frac{\widetilde{f}^{(\nu-1)}_{\setminus h|h} \cdot \widetilde{f}^{(\nu)}_h}{\int \widetilde{f}^{(\nu-1)}_{\setminus h|h} \cdot \widetilde{f}^{(\nu)}_h \, \mathrm{d}\theta_h} \tag{9}$$

and proceed iteratively until we obtain "convergence" of the sequence of approximated densities $\widetilde{f}^{(1)}_{\boldsymbol{\theta}}, \widetilde{f}^{(2)}_{\boldsymbol{\theta}}, \ldots$ in the following sense, at which point the algorithm terminates:

Theorem 2.4. $\mathcal{D}_{\mathrm{KL}}(\widetilde{f}^{(\nu)} \mid\mid f_{\boldsymbol{\theta}}) = -\log\left(\zeta^{(\nu)}_h\right)$ *converges monotonically to a local minimum.*

Tran states that the computations required in the algorithm become tractable if one first assumes that the true joint distribution $f_{\boldsymbol{\theta}}$ is a member of a "conditional exponential family"; that is, $f_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \propto \exp\left(\left\langle \boldsymbol{g}_h(\theta_h), \boldsymbol{g}_{\setminus h}(\boldsymbol{\theta}_{\setminus h})\right\rangle\right)$ for some vector-valued functions $\boldsymbol{g}_h : \mathbb{R} \to \mathbb{R}^q$ and $\boldsymbol{g}_{\setminus h} : \mathbb{R}^{d-1} \to \mathbb{R}^q$. The h'th marginal density is then given by

$$f_h(\theta_h) \propto \int \exp\left(\left\langle \boldsymbol{g}_h(\theta_h), \boldsymbol{g}_{\setminus h}(\boldsymbol{\theta}_{\setminus h})\right\rangle\right) \mathrm{d}\boldsymbol{\theta}_{\setminus h}, \tag{10}$$

which is generally intractable. However, by fixing a conditional density $f^*_{\setminus h|h}$ in the same class, the optimization specified in Equation 8 is available in closed form (up to the normalizing constant), leading to an optimal approximating marginal that can be computed:

Theorem 2.5. *Let* $\widetilde{f}_{\boldsymbol{\theta}} = \widetilde{f}^*_{\setminus h|h} \cdot \widetilde{f}_h$ *be a distribution such that* $\widetilde{f}^*_{\setminus h|h}(\boldsymbol{\theta}) \propto \exp\left(\left\langle \boldsymbol{h}_h(\theta_h), \boldsymbol{h}_{\setminus h}(\boldsymbol{\theta}_{\setminus h})\right\rangle\right)$*. If the true distribution takes the form* $f_{\boldsymbol{\theta}}(\boldsymbol{\theta}) \propto \exp\left(\left\langle \boldsymbol{g}_h(\theta_h), \boldsymbol{g}_{\setminus h}(\boldsymbol{\theta}_{\setminus h})\right\rangle\right)$*, then the optimal marginal distribution in Equation 7 takes the form*

$$\widetilde{f}^*_h(\theta_h) \propto \exp\left(\left\langle \boldsymbol{g}_h(\theta_h) - \boldsymbol{h}_h(\theta_h), \mathbb{E}_{\boldsymbol{\theta}_{\setminus h} \sim \widetilde{f}^*_{\setminus h|h}}\left[\boldsymbol{g}_{\setminus h}(\boldsymbol{\theta}_{\setminus h}) - \boldsymbol{h}_{\setminus h}(\boldsymbol{\theta}_{\setminus h})\right]\right\rangle\right). \tag{11}$$

Tran suggests that Equation 11 is easier to compute than Equation 10, for now the integral is inside the exponential function.

Finally, Tran proceeds to show directly that several commonly-used instances of mean-field approximations — variational Bayes, the EM algorithm, the ICM algorithm, and the k-means algorithm — can all be viewed as special cases of his algorithm.

2.3 Critical Assessment

To begin with, Tran spends far more time than necessary reviewing concepts that could simply be referenced in established publications. While it is traditional for papers involving copulas to state Sklar's theorem, it is far less common to devote an entire section to a review of standard facts about copulas, as the author does here. Similarly, an entire section is spent on a review of information geometry (or more specifically, Bregman divergences), in which Bregman divergences are defined twice: first for functions on $\mathbb{R}^d$, and then for functions on $\mathcal{L}_p$. The author's motivation here is clarity and ease of interpretation, since the first definition allows one to easily visualize Theorem 2.1; however, we believe the repetition is unnecessary given that this paper was (presumably) not intended to be a pedagogical reference for information geometry. We do acknowledge, however, that this section of the paper includes two beautiful figures clearly illustrating Bregman divergences and Theorem 2.1.

The level of information geometry in the paper is not very deep, being fully confined to the use of Bregman divergences and the basic properties thereof. While Bregman divergences are certainly fundamental to information geometry, the field includes a vast number of other important, but mostly unrelated concepts. Moreover, we think that the author has focused unduly on the general case of Bregman divergences themselves; although the paper includes multiple pages of introductory discussion about them, the author actually uses only one particular case — the KL-divergence — in his algorithm. Moreover, the algorithm cannot directly be extended to other Bregman divergences, as the fundamental Theorem 2.2 holds only for KL-divergences. Every result about Bregman divergences mentioned in the paper could just as well have been stated for KL-divergences with no loss of relevance. Going even further, the paper could have been titled more suitably as "Copula Variational Bayes inference via KL-divergence"; ironically, such a title might have increased the paper's exposure because practicioners are likely much more familiar with KL-divergences than with the term "information geometry" itself. Additionally, the use of the generalized Pythagorean theorem in Theorem 2.3 seems vacuous; it is unclear why we should not simply minimize the left-hand side of Equation 5).

The use of copulas is also rather superficial. Tran's approach essentially ignores the copula c in $f_{\boldsymbol{\theta}}(\boldsymbol{\theta}) = c(\boldsymbol{u}(\boldsymbol{\theta})) \cdot \prod_{h=1}^{d} f_h(\theta_h)$ and focuses only on the marginals $f_1, \ldots, f_d$. He notes [Tran, 2018, Remark 22] that an alternative approach finds $\widetilde{f}_{\boldsymbol{\theta}}$ such that $\mathcal{D}_{\text{KL}}(\widetilde{c}(\boldsymbol{u}) \mid\mid c(F(\widetilde{F}^{\leftarrow}(\boldsymbol{u}))))$ approximates $\mathcal{D}_{\text{KL}}(c(\boldsymbol{u}) \mid\mid \widetilde{c}(\boldsymbol{u}))$, which is equivalent to finding each of the exact marginals f_h; however, this would involve "copula's explicit analysis" [sic] and is left for future work — which, based on our literature review, has not yet materialized. Presumably this is because the aforementioned explicit analysis is intractable, as we elaborate on in Section 4.

The paper's biggest flaw, however, is in the algorithm itself — specifically, the need to need to derive the "reverse conditional" Equation 9 in order to proceed from one step of the algorithm to the next, and the need to calculate the normalizing constant ζ_h in Theorem 2.3. Both of these steps appear to be quite intractable in general, even when the true distributions are members of the "conditional exponential family" described in Theorem 2.5. While Tran notes that the shift of the integral in Equation 10 into the exponential function in Equation 11 makes the calculation easier, exactly *how* much easier depends on the complexity of the functions $\boldsymbol{g}_{\setminus h}$ and $\boldsymbol{h}_{\setminus h}$; while the choice of the latter is up to the user, the choice of the former is not. Because of this, the utility of the algorithm is very limited outside of standard cases such as Gaussian distributions (Tran's own example). In any case, it is also unclear how the calculation of Equation 9 actually leads to the next fixed conditional $f^{(\nu)}_{\setminus h|h}$. Even in the two theoretical case studies provided by the author — one involving the approximation of a zero-mean bivariate Gaussian distribution, and the other a finite mixture of bivariate Gaussians — we could not understand why the very complex calculations involved (particularly in the second case study) led to a viable algorithm. Nowhere does Tran write out the algorithm explicitly, nor does he state explicitly that the user must repeat the marginal approximation step for each h'th marginal $\widetilde{f}_h$ (although this latter point is implied by the case studies).

After developing his algorithm, Tran shows directly that several common mean-field approximations can be viewed as special cases. However, these demonstrations are redundant, because the EM algorithm, the ICM algorithm, and the k-means algorithm are *already* known to be special cases of the standard variational Bayes algorithm that Tran's algorithm generalizes (see, for example, [Neal and Hinton, 1998]), a fact noted only vaguely in the paper. Moreover, the k-means algorithm is itself known to be a special case of the ICM algorithm [Frey and Jojic, 2005].

Finally, the paper is rife with typos. At one point Tran refers to the Bregman projection as the "Bayesian projection"; elsewhere, a proof concludes with both a "Q.E.D." and a square. The choice of template is quite unsuitable for a lengthy paper with four or more section levels. Moreover, while the underlying math is not tremendously complicated, Tran's very cumbersome notational choices, combined with math-heavy prose and an unnatural writing style, make the paper quite difficult to follow; with many notations and even phrases repeated nearly verbatim throughout the paper, readers must constantly retrace their steps to know which section they are currently reading.

In summary, Tran [2018] shows some serious deficiencies which may explain why the paper remains in preprint form four years after its posting to arXiv.

3 Paper 2: Clustering Multivariate Time Series

The second paper we review is *Optimal transport vs. Fisher-Rao distance between copulas for clustering multivariate time series* [Marti et al., 2016], which was published as a conference proceeding in the 2016 IEEE/SP Workshop on Statistical Signal Processing (SSP). At only four pages plus references, this is a substantially smaller work than Tran [2018].

3.1 Background

Clustering is one of the oldest statistical paradigms. Briefly, one desires to categorize a set of objects into clusters whose members have more in common with each other than with members of other clusters. In parametric statistics, random effects models are often used for this purpose. In non-parametric statistics, likely the most famous example is *k-means clustering*, in which each of N given objects is assigned to one of *k clusters* in such a manner that the distance from that object to the *centroid* of the cluster (usually its mean) is minimized. The notion of "distance" here is crucial; as noted by the authors, *any* nonparametric clustering algorithm relies on some notion of distance (or divergence; we use the terms interchangeably) between objects.

Typically, the objects to be clustered are observations in a given dataset, but one can also cluster more general "objects" such as sets of multivariate time series. Marti et al. [2016] observe that one can cluster multivariate time series using two broad approaches: one can either compare their entire distributions, or discriminate based on the dependence inherent within the multivariate observations. The latter approach requires a notion of distance between copulas (see Section 1); the present paper tackles that problem using information geometric concepts. In particular, the paper aims to compare the distances between copulas measured with several information-geometric metrics (generally distances based on the Fisher-Rao metric) as well as with the 2-Wasserstein metric.[2]

The authors use a copula-based clustering methodology based on Marti et al. [2016], another very short paper written by three of the same four authors in the same year. In that paper, the authors compare the empirical copula to another pre-specified copula in a fully non-parametric fashion by first computing the 1-Wasserstein distance (also known as the "earth mover's distance") between the two, and then applying *any* clustering algorithm that takes a dissimilarity matrix [Murphy, 2012] as input. [Marti et al., 2016] However, because this approach suffers from scaling issues in high dimensions, the authors of the present paper consider a parametric approach instead, in which parametric copulas are chosen and the distances between them computed using information-geometric divergences.

[2] Generally, the p-Wasserstein distance between two probability measures $\mathbb{P}$ and $\mathbb{Q}$ is given by

$$W_p(\mathbb{P} \mid\mid \mathbb{Q}) = \left(\inf_{\gamma \in \Gamma(\mathbb{P},\mathbb{Q})} \int_{M \times M} d(x,y)^p \, \mathrm{d}\gamma(x,y) \right)^{1/p},$$

where $\Gamma(\mathbb{P},\mathbb{Q})$ denotes the set of all couplings of $\mathbb{P}$ and $\mathbb{Q}$ (i.e., joint distributions whose marginals coincide with $\mathbb{P}$ and $\mathbb{Q}$) and d is a metric on M. These distances are widely used in the field of optimal transport.

3.2 Main Information Geometric Results

The authors begin by introducing the distances they will compare. The first notion of distance is the Fisher-Rao metric $\mathrm{d}s^2(\theta) = \sum_{i,j}^{d} g_{ij}(\theta)\,\mathrm{d}\theta_i\,\mathrm{d}\theta_j$ on the statistical manifold M induced by a parametric model where

$$g_{ij}(\theta) = \mathbb{E}_\theta\left[\frac{\partial}{\partial\theta_i}\log\left(f_\theta(X)\right)\frac{\partial}{\partial\theta_j}\log\left(f_\theta(X)\right)\right]$$

is the Fisher information and the induced distance is $D(\theta_1, \theta_2) = \int_{\theta_1}^{\theta_2} \mathrm{d}s$ [Marti et al., 2016].

This distance is usually intractable in practice, so one typically resorts to other measures of distance; in contrast to Tran [2018], the authors choose to focus on the class of *f-divergences*

$$\mathcal{D}_f\left(\mathbb{P} \mid\mid \mathbb{Q}\right) = \int f\left(\frac{\mathrm{d}\mathbb{P}}{\mathrm{d}\mathbb{Q}}\right)\mathrm{d}\mathbb{Q} = \int_{\mathcal{X}} f\left(\frac{p(x)}{q(x)}\right) q(x)\,\mathrm{d}x \tag{12}$$

for convex differentiable functions f satisfying $f(1) = 0$, since these divergences are also parametrization-invariant in the sense that if $h : \mathcal{X} \to \mathcal{Y}$ is a diffeomorphism and $p'(y) = p'(h(x)) := p(x) \cdot |\mathcal{J}(x)|^{-1}$ (and $q'(y)$ is defined similarly), then $\mathcal{D}_f\left(\mathbb{P} \mid\mid \mathbb{Q}\right) = \mathcal{D}_f\left(\mathbb{P}' \mid\mid \mathbb{Q}'\right)$ [Qiao and Minematsu, 2010]. Moreover, f-divergences also provide a second-order approximation to the Fisher-Rao metric in a certain technical sense (see [Amari and Cichocki, 2010, Theorem 5]).

The authors compare the 2-Wasserstein distance (which is *not* based on the Fisher-Rao metric[3]) and various information-geometric divergences — the Fisher-Rao distance, the KL-divergence, the Jeffreys distance, the Hellinger distance, and the Bhattacharyya distance — between three bivariate Gauss copulas (see Section 1) identified by their correlation matrices:

$$R_A = \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix}, \quad R_B = \begin{pmatrix} 1 & 0.99 \\ 0.99 & 1 \end{pmatrix}, \quad \text{and} \quad R_C = \begin{pmatrix} 1 & 0.9999 \\ 0.9999 & 1 \end{pmatrix}.$$

For such copulas, it is well-known that $C_R(\boldsymbol{u}) \to M(\boldsymbol{u}) := \min\{u_1, \ldots, u_d\}$ as $R \to \mathbf{1}_{d\times d}$ (where the latter convergence is with respect to any matrix norm). The copula $M(\boldsymbol{u})$ on the right-hand side is the *comonotonicity copula* (or the *upper Fréchet-Hoeffding bound*) and represents perfect positive dependence between the components of $\boldsymbol{U} \sim M$. Thus, C_{R_B} and C_{R_C} are intuitively "close" to each other (in an imprecise sense), as they are both "closer" to the comonotonicity copula than C_{R_A}, which induces only mild positive dependence.

Surprisingly, C_{R_A} and C_{R_B} are closer to each other with respect to the information-geometric divergences than are C_{R_B} and C_{R_C}; the latter pair, however, are closer with respect to the 2-Wasserstein distance. The authors explain this apparently unintuitive result in several related ways. From a purely computational perspective, the analytical forms of these divergences (which the authors provide in a table) show that the Fisher-Rao and f-divergences between any C_{R_1} and C_{R_2} are superlinearly increasing functions of $|R_1^{-1}|$ and $|R_2^{-1}|$, and thus become poorly-behaved as either of these correlation matrices approaches $\mathbf{1}_{d\times d}$. In contrast, the Wasserstein distance is stable in these limits.

From an information-geometric perspective, the authors note that we see this result *because* C_{R_B} and C_{R_C} are close to the comonotonicity copula. Specifically, the comonotonicity copula is not absolutely continuous. Since it lacks a density, it does not correspond to a point on the statistical manifold M. On the other hand, the Wasserstein metrics $W_p(\mathbb{P} \mid\mid \mathbb{Q})$ are defined only in terms of the distributions $\mathbb{P}$ and $\mathbb{Q}$ themselves, rather than their densities; thus, the fact that $M(\boldsymbol{u})$ is not absolutely continuous poses no issue.

The authors also point out that when equipped with the Fisher-Rao metric, the space of symmetric positive definite matrices (which clearly includes the current statistical manifold as a submanifold) is a Riemannian manifold of negative sectional curvature [Said et al., 2017]. In contrast, the Wasserstein geometry of the space is of nonnegative curvature and is flat [Takatsu, 2011]. As a consequence, if we denote a generic bivariate correlation matrix as

$$R_\rho = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix},$$

[3] We note here that Chizat et al. [2018] defines a metric tensor which interpolates between the Hellinger distance (a special case of f-divergence) and the squared 2-Wasserstein distance in an attempt to relieve the restriction of optimal transport being defined only between measures having the same mass [Chizat et al., 2018].

the 2-Wasserstein distance $W_2(C_{R_{\rho_1}} \mid\mid C_{R_{\rho_2}})$ increases in a nearly linear fashion away from the line $y = x$ in $[0,1]^2$. We can observe this increase both from a surface plot of $(\rho_1, \rho_2) \mapsto W_2(C_{R_{\rho_1}} \mid\mid C_{R_{\rho_2}})$ provided by the authors, and to a lesser extent, from the closed form expression

$$W_2(C_{R_{\rho_1}} \mid\mid C_{R_{\rho_2}}) = \sqrt{\operatorname{tr}\left(R_{\rho_1} + R_{\rho_2} - 2\sqrt{R_{\rho_1}^{1/2} R_{\rho_2} R_{\rho_1}^{1/2}}\right)}$$

as derived by Barbaresco [2011]. In contrast, the curvature of the Fisher-Rao geometry results in a very high sensitivity of $D(\rho_1, \rho_2)$ to small changes in ρ_1 and ρ_2 when these parameters are already close to 1. The authors note that in small-sample data, the estimation error of the parameters could easily "exceed" the sensitivity of $D(\rho_1, \rho_2)$, rendering this measure useless as a means of discrimination. On the other hand, the authors observe that all of the information-geometric divergences are locally a quadratic form of the Fisher information $I(\rho)$ [Amari and Cichocki, 2010], and $1/I(\rho)$ provides a lower bound on the variance of any unbiased estimator $\widehat{\rho}$ of ρ via the Cramér-Rao lower bound.

3.3 Critical Assessment

This interesting paper serves its purpose well. However, at only four pages (plus references), it is very short and could have been written within a larger paper on differences between Wasserstein and Fisher-Rao based distances, or merged with the authors' companion paper of the same length on time series clustering using the 1-Wasserstein distance [Marti et al., 2016]. The present work contains almost no information directly related to time series, and seems to be more of a supporting work to the other paper. On the other hand, both this paper and Marti et al. [2016] were published as IEEE conference proceedings, and the authors may have been limited by space constraints.

Some elements of the paper could have been expanded. It is understandable that the authors would choose to focus on Gauss copulas, as these are likely the only non-trivial copulas which have 2-Wasserstein, Fisher-Rao, and related f-divergences available in closed form, thus allowing for surface plots of $W_2(C_{\rho_1} \mid\mid C_{\rho_2})$ and $D(\rho_1, \rho_2)$ as well as exact computations for specific ρ_1 and ρ_2. That said, closed form expressions are provided only for divergences between generic zero-mean multivariate Gaussian distributions with arbitrary covariance matrices. It would have been preferable for the authors to actually derive these forms for the bivariate correlation matrices used in the paper, in order to show more clearly the effect of changes on inputs of ρ_1 and ρ_2. These calculations are tedious but not tremendously challenging; we have carried them out ourselves in Table 1. For consistency, we have also recomputed the specific distances for $\mathcal{D}(C_{R_A} \mid\mid C_{R_B})$ and $\mathcal{D}(C_{R_B} \mid\mid C_{R_C})$ to three decimal places.[4] The expressions provided in Table 1 show much more clearly the effect of changes in ρ_1 and ρ_2 when both are close to 1. For additional clarity, Figure 1 includes reproductions of the authors' contour plots (or "heatmaps") for the Fisher-Rao and 2-Wasserstein distances, as well as plots that we generated for the remaining distances in Table 1.

We also note that the authors fail to mention the fact that the Fisher-Rao distance and the KL-divergence are not symmetric in their arguments; we are left wondering whether the same counter-intuitive divergence measures result from comparing $\mathcal{D}_f(C_{R_B} \mid\mid C_{R_A})$ with $\mathcal{D}_f(C_{R_C} \mid\mid C_{R_B})$ — especially given that the 2-Wasserstein distance, being a metric, *is* symmetric in its arguments.

There seems to be a self-contradiction in the authors' comparison of the 2-Wasserstein and Fisher-Rao metric by the curvatures that they induce on the statistical manifold (Section 3.3 of the paper). In Subsection 3.2, the authors claim that the Fisher-Rao metric is highly sensitive to changes in parameters around the upper-right corners of $[0,1]^2$, while the 2-Wasserstein distance increases roughly linearly away from the main diagonal; however, the heatmaps of $D(\rho_1, \rho_2)$ and $(\rho_1, \rho_2) \mapsto W_2(C_{R_{\rho_1}} \mid\mid C_{R_{\rho_2}})$ provided by the authors seem to show the opposite; in those, the Fisher-Rao distance appears to be the more stable of the two. Moreover, in the next paragraph they state that "Fisher-Rao and related divergences do not suffer from this drawback" due to the connection to the Crámer-Rao lower bound. This apparent self-contradiction disappears if we assume a typographical error in which the names of the distances have been switched. However, later on in the discussion section of the paper, the authors remark that "if the dependence is strong between the time series, the use of Fisher-Rao geodesic distance and related divergences may not be appropriate. [...] To measure distance [sic] between copulas, we think that the Wasserstein geometry is more appropriate since it

[4] Strangely, our values for the Hellinger distances vary somewhat from those provided by the authors, which we have not been able to reproduce. We surmise that the most likely reason is an implementation error.

Distance	$\mathcal{D}(C_{\rho_1} \Vert C_{\rho_2})$	$\mathcal{D}(C_{R_A} \Vert C_{R_B})$	$\mathcal{D}(C_{R_B} \Vert C_{R_C})$
Fisher-Rao	$\sqrt{\frac{1}{2}\left(\log\left(\frac{(\rho_1+1)(\rho_2-1)}{\rho_1^2-1}\right)^2+\log\left(\frac{(\rho_2+1)(\rho_1-1)}{\rho_1^2-1}\right)^2\right)}$	2.773	3.256
KL	$\frac{(\rho_1-\rho_2)\rho_2}{\rho_2^2-1}+\frac{1}{2}\log\left(\frac{\rho_2^2-1}{\rho_1^2-1}\right)$	22.562	47.197
Jeffreys	$\frac{(\rho_1-\rho_2)^2(1+\rho_1\rho_2)}{(\rho_1^2-1)(\rho_2^2-1)}$	24.050	49.005
Hellinger	$\sqrt{1-\frac{2(1-\rho_1^2)^{1/4}(1-\rho_2^2)^{1/4}}{\sqrt{4-2\rho_1\rho_2-\rho_1^2-\rho_2^2}}}$	0.690	0.745
Bhattacharyya	$\frac{1}{2}\log\left(\frac{1-\frac{1}{2}\rho_1\rho_2-\frac{1}{4}\rho_1^2-\frac{1}{4}\rho_2^2}{\sqrt{(\rho_1^2-1)(\rho_2^2-1)}}\right)$	0.646	0.810
W_2	$\sqrt{4-2\sqrt{(\rho_1-1)(\rho_2-1)}-2\sqrt{(\rho_1+1)(\rho_2+1)}}$	0.635	0.090

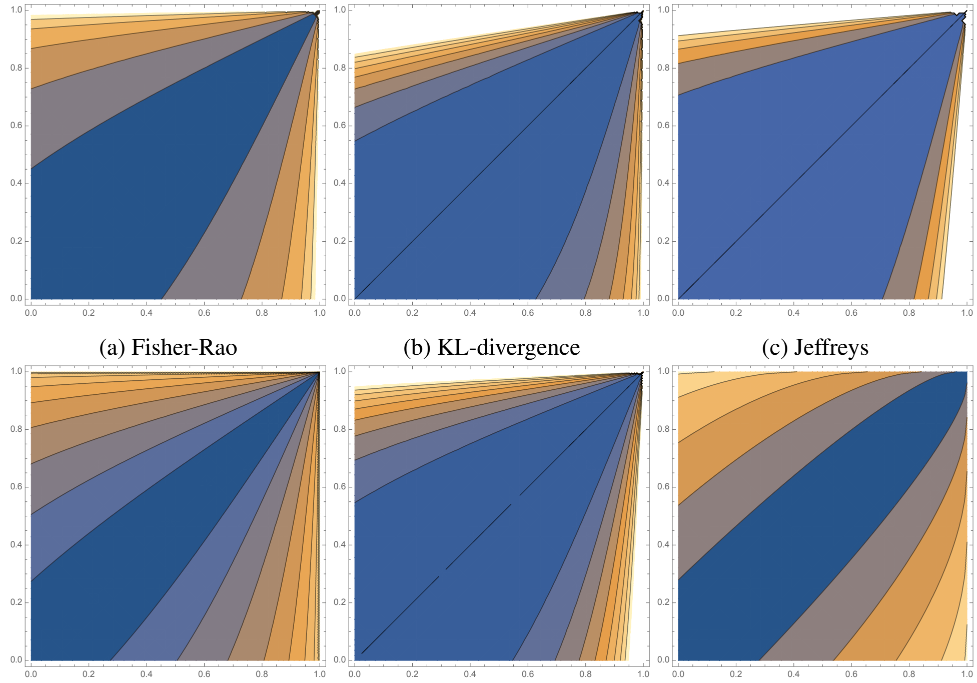
Table 1: Recreation of [Marti et al., 2016, Table 1] with statistical distances written as explicit functions of ρ_1 and ρ_2 and more precise figures.

does not lead to these counter-intuitive clusters." It is thus quite unclear which class of divergences the authors are actually advocating for. Were the four authors in complete agreement with each other on this point?[5]

At the end of Section 3.1 of the paper, the authors write that "computing the Wasserstein distance between two probability measures amounts to finding the most correlated copula associated with these measures", suggesting that they believe the Wasserstein distance to be an intuitively reasonable measure of association for multivariate time series. However, this is something of a red herring: any two multivariate distributions $\mathbb{P}$ and $\mathbb{Q}$ have their *own* copulas, both of which are unrelated to the "most correlated copula" between $\mathbb{P}$ and $\mathbb{Q}$ that $W_p(\mathbb{P} \,||\, \mathbb{Q})$ seeks to find. That said, their information-geometric arguments explaining the unintuitive divergence calculations between the three Gauss copulas R_A, R_B, R_C appear to be sound. Amari and Matsuda [2022], whose Section 1 briefly compares Wasserstein geometry and information geometry, puts it more directly: "[Wasserstein geometry] reflects the metric of the underlying manifold X on which the probability distributions are defined. [...] On the other hand, information geometry is constructed independently of the metric of X."

In contrast to [Tran, 2018], the present paper seems to be written for non-experts. The idea of clustering multivariate time series in general was clearly motivated by the financial applications mentioned in the introduction — which is not surprising, given that three of the four authors work in capital management. Numerous other concepts are motivated by literal questions (such as "What is a relevant distance to measure the resemblance of copulas?") and the technical details are usually kept to a minimum, despite the fact that the third author has also written a highly technical "elementary" introduction to information geometry [Nielsen, 2020]. As with Tran [2018], the paper needs editing: for example, in noting the advantages of Marti et al. [2016], the authors list "non-parametric" and "robust to noise" twice. The confusion between the 2-Wasserstein and Fisher-Rao distances discussed above is a much graver example.

[5] A quote extracted from a subsequent publication [Marti et al., 2017] makes the authors' position more explicit: "In (Marti et al., 2016a), authors illustrate in a parametric setting using Gaussian copulas that common divergences (such as Kullback-Leibler, Jeffreys, Hellinger, Bhattacharyya) are not relevant for clustering these distributions, especially when dependence is high."



(a) Fisher-Rao (b) KL-divergence (c) Jeffreys

(d) Hellinger (e) Bhattacharrya (f) W_2

Figure 1: Contour plots of $(\rho_1, \rho_2) \mapsto \mathcal{D}(C_{\rho_1} \mid\mid C_{\rho_2})$ for the six distances in Table 1

4 Discussion

The two papers we have discussed attempt combine concepts from both copulas and information geometry in somewhat different ways. Tran [2018] derives a variational Bayes algorithm that integrates a fixed copula into an approximating family, and applies information-geometric principles (namely the generalized Pythagorean theorem for Bregman divergences, specialized to KL-divergences) to derive the optimal choices of approximating marginal densities. Marti et al. [2016] computes Fisher-Rao and f-divergences distances between bivariate Gauss copulas and uses information-geometric principles to explain counterintuitive comparisons to the 2-Wasserstein distance.

If one believes the paper of Tran [2018] is worthy of more attention, there are a number of future directions. Aside from correcting some of the faults discussed in Subsection 2.3, we think that more examples of the algorithm in action would be helpful, as both of the author's case studies involve only bivariate Gaussian distributions. Despite the repeated use of the term "variational Bayes" throughout the paper, Tran [2018] does not actually include any Bayesian applications, and it would be interesting to see how the algorithm performs on real-world data in a Bayesian setting. Tran also remarks that one could in principle devise an alternative algorithm that finds $\widetilde{f}_{\boldsymbol{\theta}}$ such that $\mathcal{D}_{\mathrm{KL}}(\widetilde{c}(\boldsymbol{u}) \mid\mid c(F(\widetilde{F}^{\leftarrow}(\boldsymbol{u}))))$ approximates $\mathcal{D}_{\mathrm{KL}}(c(\boldsymbol{u}) \mid\mid \widetilde{c}(\boldsymbol{u}))$, which is equivalent to finding each of the exact marginals f_h; however, this would involve "copula's explicit analysis [sic]" and is left for future work.

The very short paper of Marti et al. [2016] serves its purpose as a simple comparison study between Fisher-Rao-based divergences and the 2-Wasserstein distance; however, there is much more to be explored. Ostensibly, this work was motivated by clustering of multivariate time series, and it would be helpful to cluster real-world time series together using the various divergences. A real-world application of particular interest would be to analyze financial time series from the "Great Recession" or other extreme economic downturns, when financial returns within asset classes become highly correlated with one another.[6] Further, most — possibly all — of the divergences compared by the authors could at least be approximated for other copula families using Monte Carlo sampling; this is especially true for those divergences that can be written as expectations with respect to one of the the

[6] Readers familiar with the history of the "Great Recession" will recall that the Gauss copula turned out to be a notoriously poor choice for modelling such extremes [McNeil et al., 2015].

underlying copulas. It would be interesting to see if the same counterintuitive clustering results hold for Archimedean copulas or other well-known non-Gauss copulas, particularly comprehensive ones — although if they did hold, they could no longer be explained by the negativity or flatness of the induced curvature of the space of covariance matrices.

On the theoretical side, the main point of Marti et al. [2016] is that the intuitive closeness of the copulas C_{R_B} and C_{R_C} fails because they are both "close" to the comonotonicity copula M, which is not absolutely continuous. In principle, the same should hold for bivariate Gauss copulas "close" to the *countermonotonicity copula* $W(u_1, u_2) = \max\{u_1 + u_2 - 1, 0\}$, and verifying this using such copulas with correlation coefficient ρ close to -1 would lend credence to the authors' argument. It would also be interesting to redo the calculations using regularized versions of the comonotonity and countermonotonicity copulas, such as those developed by Björnham et al. [2016], who regularized them in a way that makes them absolutely continuous.

Beyond the ideas covered in these two papers, there is much to explore. One future path involves non-parametrics. Non-parametric estimation is an important component of modern copula theory, where key concepts include rank-based estimators and their roles in empirical copula processes (see Chen and Huang [2007]). As noted previously, some work has also been done on non-parametric information geometry, and it would be interesting to see how compatible the two areas are, and whether the latter work might support the former theory.

We also believe that the idea of exploiting statistical divergences between copulas — be they Bregman divergences, f-divergences, or another class entirely — has further potential. For example, Lalancette and Zimmerman [2022] have introduced a new family of copulas $\mathcal{C}$ parameterized by the set of probability densities $\mathcal{F}_{[0,1]}$ supported on the unit interval. Those authors have shown that a number of popular but challenging desiderata for copulas (such as the ability to sample from them and compute certain concordance measures such as Spearman's rho) are very easy for this family; some evidence suggests that these copulas are a suitable model for angular data (i.e., data on the torus). We wonder whether this family can be viewed as an "approximating class" (in the sense of variational Bayes) for more general multivariate distributions; that is, given some statistical divergence $\mathcal{D}(\cdot \mid\mid \cdot)$ and a "true" copula C, one could form the either of the equivalent variational problems

$$\widetilde{C_f} = \underset{C_f \in \mathcal{C}}{\arg\min}\, \mathcal{D}(C_f \mid\mid C) \iff \widetilde{f} = \underset{f \in \mathcal{F}_{[0,1]}}{\arg\min}\, \mathcal{D}(C_f \mid\mid C).$$

If the variational problem admits a solution (be it exact or approximate), it could potentially be applied to real-world statistical problems by replacing C with its empirical counterpart.

Finally, why do [Tran, 2018] and Marti et al. [2016] appear to be the only papers to date that explicitly explore the intersection between information geometry and copulas? While it is true that neither copula theory nor information geometry is a "standard topic" in statistics, one might have expected more attention to their intersection. We surmise that the main issue is the general intractability of copulas. For most parametric copulas, the analytical expressions of the densities are quite complicated. Given how few closed-form expressions there are for divergences, even between members of "standard" parametric families (e.g., between two Gamma densities), it is not surprising that almost no results exist for copulas. For example, the Frank copula — among the most popular of the Archimedean copulas — has a density of the form [Hofert et al., 2012]

$$c_\theta(\boldsymbol{u}) = \left(\frac{\theta}{1 - e^{-\theta}}\right)^{d-1} \frac{\mathrm{Li}_{-(d-1)}\left(q_\theta(\boldsymbol{u})\right)}{q_\theta(\boldsymbol{u})} \exp\left(-\theta \sum_{h=1}^{d} u_j\right)$$

where $q_\theta(\boldsymbol{u}) = (1 - e^{-\theta})^{1-d} \prod_{h=1}^{d}(1 - e^{-\theta u_h})$ and $\mathrm{Li}_s(z) = \sum_{k \geqslant 1} z^k / k^s$. Merely *computing* this density at a single $\boldsymbol{u} \in [0, 1]^d$ is challenging. Determining an f-divergence $\mathcal{D}_f(C_{\theta_1} \mid\mid C_{\theta_2})$ between two Frank copulas via Equation 12 is clearly not viable. Most other commonly-used absolutely continuous copulas suffer from the same problems.

It is interesting to note that far more work has been done in the intersection of copulas and optimal transport — a field distinct from information geometry, but one with a number of deep theoretical connections to it [Amari, 2016, Khan and Zhang, 2022]. For example, the papers by Marti et al. [2017], Bartl et al. [2017], Chi et al. [2022], Mordant and Segers [2022] appear among those written only within the past five years. We speculate that one reason for this is that (as previously mentioned in Subsection 3.2) Wasserstein distances are defined on spaces of probability measures, rather than on spaces of probability densities (or on the statistical manifold implied by a parametric family of densities). As such, the ease of computing such distances for copulas relies less on the tractability of integrals involving the underlying copula densities which, as noted above, can be notoriously challenging to evaluate, especially in high dimensions.

Another possible reason is that copulas are simply not as appreciated as they could be. When approximating multivariate data, many statisticians are content to use multivariate normal distributions, or perhaps the Gauss copula if the marginals are significantly non-Gaussian. This does work reasonably well when the dependence within the data is not too extreme. Perhaps the intersection of the cohort that works with data requiring more exotic copulas (most notably in finance and extreme value theory) and the cohort that is proficient in information geometry is too small to produce much research.

References

[1] M Sklar. Fonctions de répartition à n dimensions et leurs marges. *Publ. inst. statist. univ. Paris*, 8:229–231, 1959.

[2] Roger B Nelsen. *An introduction to copulas*. Springer Science & Business Media, 2007.

[3] Alexander J McNeil, Rüdiger Frey, and Paul Embrechts. *Quantitative risk management: concepts, techniques and tools-revised edition*. Princeton university press, 2015.

[4] Marius Hofert and David Pham. Densities of nested archimedean copulas. *Journal of Multivariate Analysis*, 118:37–52, 2013.

[5] Jan Górecki, Marius Hofert, and Ostap Okhrin. Outer power transformations of hierarchical archimedean copulas: Construction, sampling and estimation. *Computational Statistics & Data Analysis*, 155:107109, 2021.

[6] Erich Peter Klement, Radko Mesiar, and Endre Pap. Archimax copulas and invariance under transformations. *Comptes Rendus Mathematique*, 340(10):755–758, 2005.

[7] Nihat Ay, Jürgen Jost, Hông Vân Lê, and Lorenz Schwachhöfer. *Information geometry*, volume 64. Springer, 2017.

[8] Giovanni Pistone and Carlo Sempi. An infinite-dimensional geometric structure on the space of all the probability measures equivalent to a given one. *The annals of statistics*, pages 1543–1561, 1995.

[9] Jun Zhang. Nonparametric information geometry: From divergence function to referential-representational biduality on statistical manifolds. *Entropy*, 15(12):5384–5418, 2013.

[10] Giovanni Pistone. Nonparametric information geometry. In *International Conference on Geometric Science of Information*, pages 5–36. Springer, 2013.

[11] Shun-ichi Amari. *Information geometry and its applications*, volume 194. Springer, 2016.

[12] Shun-Ichi Amari. α-divergence is unique, belonging to both f-divergence and bregman divergence classes. *IEEE Transactions on Information Theory*, 55(11):4925–4931, 2009.

[13] Viet Hung Tran. Copula variational bayes inference via information geometry. *arXiv preprint arXiv:1803.10998*, 2018.

[14] Carsten Peterson. A mean field theory learning algorithm for neural networks. *Complex systems*, 1: 995–1019, 1987.

[15] Geoffrey E Hinton and Drew Van Camp. Keeping the neural networks simple by minimizing the description length of the weights. In *Proceedings of the sixth annual conference on Computational learning theory*, pages 5–13, 1993.

[16] Michael I Jordan, Zoubin Ghahramani, Tommi S Jaakkola, and Lawrence K Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

[17] Dustin Tran, David Blei, and Edo M Airoldi. Copula variational inference. *Advances in neural information processing systems*, 28, 2015.

[18] Michael Stanley Smith, Rubén Loaiza-Maya, and David J Nott. High-dimensional copula variational approximation through transformation. *Journal of Computational and Graphical Statistics*, 29(4):729–743, 2020.

[19] David Gunawan, Robert Kohn, and David Nott. Flexible variational bayes based on a copula of a mixture of normals. *arXiv preprint arXiv:2106.14392*, 2021.

[20] Jinjin Chi, Jihong Ouyang, Ang Zhang, Xinhua Wang, and Ximing Li. Fast copula variational inference. *Journal of Experimental & Theoretical Artificial Intelligence*, 34(2):295–310, 2022.

[21] Frank Nielsen. What is an information projection. *Notices of the AMS*, 65(3):321–324, 2018.

[22] Jian Ma and Zengqi Sun. Mutual information is copula entropy. *Tsinghua Science & Technology*, 16(1): 51–54, 2011.

[23] Radford M Neal and Geoffrey E Hinton. A view of the em algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models*, pages 355–368. Springer, 1998.

[24] Brendan J Frey and Nebojsa Jojic. A comparison of algorithms for inference and learning in probabilistic graphical models. *IEEE Transactions on pattern analysis and machine intelligence*, 27(9):1392–1416, 2005.

[25] Gautier Marti, Sébastien Andler, Frank Nielsen, and Philippe Donnat. Optimal transport vs. fisher-rao distance between copulas for clustering multivariate time series. In *2016 IEEE Statistical Signal Processing Workshop (SSP)*, pages 1–5. IEEE, 2016.

[26] Gautier Marti, Frank Nielsen, and Philippe Donnat. Optimal copula transport for clustering multivariate time series. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2379–2383. IEEE, 2016.

[27] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[28] Yu Qiao and Nobuaki Minematsu. A study on invariance of f-divergence and its application to speech recognition. *IEEE Transactions on Signal Processing*, 58(7):3884–3890, 2010.

[29] Shun-ichi Amari and Andrzej Cichocki. Information geometry of divergence functions. *Bulletin of the polish academy of sciences. Technical sciences*, 58(1):183–195, 2010.

[30] Lenaic Chizat, Gabriel Peyré, Bernhard Schmitzer, and François-Xavier Vialard. An interpolating distance between optimal transport and fisher–rao metrics. *Foundations of Computational Mathematics*, 18(1):1–44, 2018.

[31] Salem Said, Lionel Bombrun, Yannick Berthoumieu, and Jonathan H Manton. Riemannian gaussian distributions on the space of symmetric positive definite matrices. *IEEE Transactions on Information Theory*, 63(4):2153–2170, 2017.

[32] Asuka Takatsu. Wasserstein geometry of gaussian measures. *Osaka Journal of Mathematics*, 48(4): 1005–1026, 2011.

[33] Frédéric Barbaresco. Geometric radar processing based on fréchet distance: information geometry versus optimal transport theory. In *2011 12th International Radar Symposium (IRS)*, pages 663–668. IEEE, 2011.

[34] Gautier Marti, Sébastien Andler, Frank Nielsen, and Philippe Donnat. Exploring and measuring non-linear correlations: Copulas, lightspeed transportation and clustering. In *NIPS 2016 Time Series Workshop*, pages 59–69. PMLR, 2017.

[35] Shun-ichi Amari and Takeru Matsuda. Wasserstein statistics in one-dimensional location scale models. *Annals of the Institute of Statistical Mathematics*, 74(1):33–47, 2022.

[36] Frank Nielsen. An elementary introduction to information geometry. *Entropy*, 22(10):1100, 2020.

[37] Oscar Björnham, Niklas Brännström, and Leif Persson. Absolutely continuous copulas obtained by regularization of the frechét–hoeffding bounds. *arXiv preprint arXiv:1603.03693*, 2016.

[38] Song Xi Chen and Tzee-Ming Huang. Nonparametric estimation of copula functions for dependence modelling. *Canadian Journal of Statistics*, 35(2):265–282, 2007.

[39] Michaël Lalancette and Robert Zimmerman. A new family of smooth copulas with arbitrarily irregular densities. *arXiv preprint arXiv:2204.04336*, 2022.

[40] Marius Hofert, Martin Mächler, and Alexander J McNeil. Likelihood inference for archimedean copulas in high dimensions under known margins. *Journal of Multivariate Analysis*, 110:133–150, 2012.

[41] Gabriel Khan and Jun Zhang. When optimal transport meets information geometry. *Information Geometry*, pages 1–32, 2022.

[42] Daniel Bartl, Michael Kupper, Thibaut Lux, Antonis Papapantoleon, and Stephan Eckstein. Marginal and dependence uncertainty: bounds, optimal transport, and sharpness. *arXiv preprint arXiv:1709.00641*, 2017.

[43] Jinjin Chi, Bilin Wang, Huiling Chen, Lejun Zhang, Ximing Li, and Jihong Ouyang. Approximate continuous optimal transport with copulas. *International Journal of Intelligent Systems*, 37(8):5354–5380, 2022.

[44] Gilles Mordant and Johan Segers. Measuring dependence between random vectors via optimal transport. *Journal of Multivariate Analysis*, 189:104912, 2022.