Bernoulli **30**(4), 2024, 2623–2643
https://doi.org/10.3150/23-BEJ1687

Minimum information dependence modeling

TOMONARI SEI[1,a] and KEISUKE YANO[2,b]

[1]*The University of Tokyo, 7-3-1 Hongo, Bunkyo-ku, Tokyo, 113-8656, Japan,* [a]*sei@mist.i.u-tokyo.ac.jp*
[2]*The Institute of Statistical Mathematics, 10-3 Midori cho, Tachikawa City, Tokyo, 190-8562, Japan,* [b]*yano@ism.ac.jp*

We propose a method to construct a joint statistical model for mixed-domain data to analyze their dependence. Multivariate Gaussian and log-linear models are particular examples of the proposed model. It is shown that the functional equation defining the model has a unique solution under fairly weak conditions. The model is characterized by two orthogonal parameters: the dependence parameter and the marginal parameter. To estimate the dependence parameter, a conditional inference together with a sampling procedure is proposed and is shown to provide a consistent estimator. Illustrative examples of data analyses involving penguins and earthquakes are presented.

Keywords: Conditional inference; copula; earthquake data; graphical model; mixed-domain; Monte Carlo method

1. Introduction

In multivariate analysis, there are a lot of statistical models describing dependence such as copula models, regression models, log-linear models and Gaussian graphical models. Such models are quite powerful and frequently used in applications. However, these approaches depend to varying degrees on the characteristics of the data domain. For example, in copula modeling, the variables are assumed to be real-valued and transformed into $[0,1]$ by monotone transformation. In generalized linear models, the domains of explanatory variables are arbitrary via quantification, but the conditional density of the response variable has to be specified as Gaussian or Poisson, for example. A log-linear model assumes the domain to be discrete.

In this paper, we propose a method to construct dependence models without using domain characteristics. The joint density function that we suppose takes the form

$$p(x_1,x_2;\theta)=e^{\theta^\top h(x_1,x_2)}A_1(x_1;\theta)A_2(x_2;\theta) \tag{1}$$

in bivariate cases, and is defined similarly in multivariate cases. Here, the domains of x_1 and x_2 are arbitrary as long as they have base measures. The function $h(x_1,x_2)$ represents the dependence of the variables. The functions A_1 and A_2 are determined by the marginal distributions of x_1 and x_2. It is proved that a density function satisfying the marginal constraints exists and is unique for every value of the parameter θ under fairly weak conditions (Theorem 1). This makes the model quite flexible since the parametric or nonparametric forms of the marginal distributions can be separately designed. We will call A_1 and A_2 adjusting functions. See Section 2 for a precise definition.

We show the usefulness of our model by enumerating various examples. In particular, models for continuous, discrete, and any other type of variables can be jointly designed up to the same cost as homogeneous data. There has been a similar attempt to construct a mixed-variable model based on univariate conditional exponential families (Yang et al. (2015)), which, however, does not avoid the restriction on the parameter space of joint exponential families. The conditional Gaussian families for mixed data discussed in Lauritzen (1996) and Whittaker (1990) are tractable but restrictive.

The adjusting functions in the proposed model (1) cannot be written in a closed form except for limited cases. This means a naive likelihood analysis is intractable. However, we can perform conditional

inference (Cox and Hinkley, 1974, Reid, 1995) given the marginal empirical distribution, as described in Section 3. It is shown that the conditional likelihood has almost the same information as the full likelihood and is not affected by nuisance parameters that characterize the marginal distribution (Theorem 2). These aspects are similar to Fisher's exact test for contingency tables (e.g., Choi, Blume and Dupont (2015), Little (1989)), which fixes marginal frequencies for testing of independence. Given the marginal empirical distribution, the problem of estimation reduces to that of an exponential family on a set of permutations. See Mukherjee (2016) for properties of exponential families on permutations.

A closed form for the conditional likelihood remains unknown. To deal with this, we propose a sampling method for the conditional distribution in a Markov chain Monte Carlo (MCMC) manner (Subsection 3.3) and a pseudo likelihood method (Subsection 3.4) that make the conditional inference tractable.

The idea behind our method is the minimum information copula model (e.g., Bedford, Daneshkhah and Wilson (2016), Bedford and Wilson (2014), Meeuwissen and Bedford (1997), Piantadosi, Howlett and Borwein (2012)), in which the joint density function of real-valued data is determined by uniform marginals and fixed values of expectations of some statistics. In a study related to the minimum information copula model, Geenens (2020) constructed a bivariate discrete model written as $p(x_1,x_2) = c(x_1,x_2)A_1(x_1)A_2(x_2)$, where the marginal distributions of $c(x_1,x_2)$ are discrete uniform distributions.

The paper is organized as follows. In Section 2, we define our model together with practical examples and establish an existence and uniqueness theorem (Theorem 1). In Section 3, we develop a conditional inference approach for the dependence parameter and prove its validity under mild conditions (Theorems 2–5). In Section 4, we present simulation studies for the inference. We conclude the paper in Section 5 with some future work. In Appendix A of Sei and Yano (2024), we provide more examples of our model. In Appendix B of Sei and Yano (2024), we summarize useful properties of the model, including information geometry (e.g., Amari and Nagaoka (2000), Csiszár (1975)), and relationship to the optimal transport and Schrödinger problems (e.g., Haasler et al. (2021), Léonard (2012), Peyré and Cuturi (2019)). Appendix C of Sei and Yano (2024) gives all proofs of the results. Appendix D of Sei and Yano (2024) provides illustrative examples of data analyses involving penguins and earthquakes, respectively.

2. Minimum information dependence model

In this section, we introduce the minimum information dependence model with its existence guarantee and present several examples.

2.1. Definition

Let $(\mathcal{X}_i,\mathcal{F}(\mathcal{X}_i),\mathrm{d}x_i)$ for $i=1,\dots,d$ be a measure space and denote their product space by $\mathcal{X}=\prod_{i=1}^d \mathcal{X}_i$ and $\mathrm{d}x=\prod_{i=1}^d \mathrm{d}x_i$. For index i, use the notation $-i$ to indicate the removal of the i-th coordinate, e.g., $x_{-i}=(x_j)_{j\neq i}$, $\mathcal{X}_{-i}=\prod_{j\neq i}\mathcal{X}_j$, and $\mathrm{d}x_{-i}=\prod_{j\neq i}\mathrm{d}x_j$.

Let $r_1(x_1;\nu),\dots,r_d(x_d;\nu)$ be statistical models of marginal densities on $\mathcal{X}_1,\dots,\mathcal{X}_d$, respectively, where ν denotes parameters characterizing the marginal densities. We can assign, if necessary, independent parameters to each r_i as $r_i(x_i;\nu_i)$ by setting $\nu=(\nu_1,\dots,\nu_d)$. It is also possible to deal with infinite-dimensional parameters.

We consider a class of probability density functions

$$p(x;\theta,\nu) = \exp\left(\theta^\top h(x) - \sum_{i=1}^{d} a_i(x_i;\theta,\nu) - \psi(\theta,\nu)\right) \prod_{i=1}^{d} r_i(x_i;\nu), \tag{2}$$

where $\theta \in \mathbb{R}^K$ is a K-dimensional parameter representing the dependence, and $h : \mathcal{X} \to \mathbb{R}^K$ is a given function. The functions $a_i(x_i;\theta,\nu)$ and $\psi(\theta,\nu)$ are simultaneously determined by constraints

$$\int p(x;\theta,\nu)\mathrm{d}x_{-i} = r_i(x_i;\nu), \quad i = 1,\ldots,d, \text{ and} \tag{3}$$

$$\int \sum_{i=1}^{d} a_i(x_i;\theta,\nu)p(x;\theta,\nu)\mathrm{d}x = 0. \tag{4}$$

The equation (3), which specifies the marginal distributions, is essential in our modeling. The equation (4) is assumed just for identifiability of $\psi(\theta,\nu)$, because for fixed $a_i(x_i;\theta,\nu)$s and fixed $\psi(\theta,\nu)$ and for any $c \in \mathbb{R}$, $(a_i(x_i;\theta,\nu) + c)$s and $(\psi(\theta,\nu) - dc)$ yield the same probability density function as that with $a_i(x_i;\theta,\nu)$s and $\psi(\theta,\nu)$. With these constraints, we shall see that the functions $\sum_{i=1}^{d} a_i(x_i;\theta,\nu)$ and $\psi(\theta,\nu)$ are unique if they exist. If each term of (4) is integrable, the equation (4) is equivalent to the equation $\sum_{i=1}^{d} \int a_i(x_i;\theta,\nu)r_i(x_i;\nu)\mathrm{d}x_i = 0$ that does not involve $p(x;\theta,\nu)$, under the marginal condition (3). Note that the density (2) is reduced to the independent model $\prod_{i=1}^{d} r_i(x_i;\nu)$ if $\theta = 0$.

Definition 1. A statistical model (2) together with the constraints (3) and (4) is called a *minimum information dependence model*. The parameter θ is called the *canonical parameter*, ν is the *marginal parameter*, $h(x)$ comprises the *canonical statistics*, $a_i(x_i;\theta,\nu)$s are the *adjusting functions* and $\psi(\theta,\nu)$ is the *potential function*.

Throughout the paper, we assume that the canonical statistics $h_k(x)$, $k = 1,\ldots,K$, are linearly independent modulo additive functions. That is, if θ satisfies

$$\theta^\top h(x) + \sum_{i=1}^{d} A_i(x_i) = 0,$$

with $A_i(x_i)$ not depending on x_{-i} $(i = 1,\ldots,d)$, then $\theta = 0$.

The canonical statistics are not sufficient statistics for θ in the full likelihood because the adjusting functions contain θ and x, but are sufficient statistics in the conditional likelihood; this point will be clarified in Section 3.

We state several useful properties of the minimum information dependence model. First, the derivative of the potential function with respect to θ is $\mathbb{E}[h(X)]$. Second, the potential function $\psi(\theta,\nu)$ is shown to be strictly convex with respect to θ. Third, the value of θ is directly linked to the total correlation (Watanabe, 1960) $\mathbb{E}[\log\{p(X;\theta,\nu)/\prod_{j=1}^{d} r_j(X_j;\nu)\}]$, a measure of strength of the association, as

$$\mathbb{E}\left[\log \frac{p(X;\theta,\nu)}{\prod_{j=1}^{d} r_j(X_j;\nu)}\right] = \theta^\top \nabla_\theta \psi(\theta,\nu) - \psi(\theta,\nu),$$

where this follows from the equation (4) and Lemma S.2 of Sei and Yano (2024). Finally, the parameters θ and ν are mutually orthogonal with respect to the Fisher information metric; see Appendices B.1 and B.2 of Sei and Yano (2024) for details of these properties.

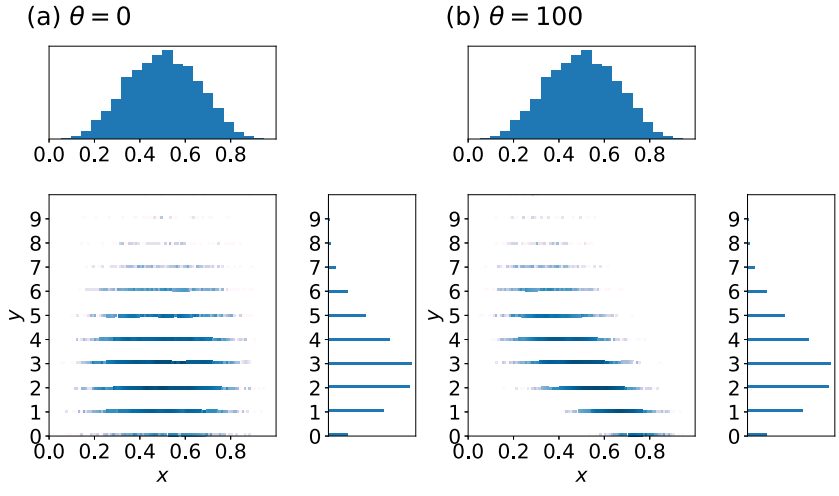

Figure 1. Two-dimensional histograms of 10000 samples from the minimum information dependence model with Beta(10, 10) and Po(3) marginals. The canonical statistic $h(x, y)$ is given by $h(x, y) = x/(y + 1)$. The joint histogram and marginal histograms are plotted. (a) Joint histogram with $\theta = 0$. (b) Joint histogram with $\theta = 100$.

As a final remark in this subsection, our model is a generalization of the minimum information copula model proposed by Bedford and Wilson (2014), particular cases of which appear in Meeuwissen and Bedford (1997) and Piantadosi, Howlett and Borwein (2012). In this copula model, the space $\mathcal{X}_i$ is the interval $[0, 1] \subset \mathbb{R}$ and the marginal density functions $r_i(x_i)$ are assumed to be the uniform density on $[0, 1]$. This copula model is derived from the maximum entropy principle (Jaynes (1957)), or equivalently, from the minimum information principle, which is the origin of the name. We discuss a similar property for our model in Appendix B of Sei and Yano (2024). Yet, copula models are, in general, intended to be used together with the probability integral transform that makes the variables have uniform marginal distributions. In contrast, our model specifies the marginal model without transforming the variables, which allows arbitrary $\mathcal{X}_i$ rather than $\mathbb{R}$. Figure 1 displays an example of two-dimensional histograms of samples from the minimum information dependence model for mixed variables (discrete and $[0, 1]$) with negative correlation. This example shows that the minimum information dependence model easily expresses arbitrary dependence between variables in arbitrary product spaces.

Also, the minimum information dependence model naturally contains the existing dependence models for specific sample spaces (Amari (2001), Holland and Wang (1987), Jansen (1997), Jones, Pewsey and Kato (2015)). Several examples highlighting the connection between the minimum information dependence model and the existing models are presented in Subsections 2.2 and 2.4, and Appendix A of Sei and Yano (2024). Remark 1 discusses the difference between the minimum information dependence models and copula models.

2.2. An illustrative example

We give an elementary example for illustrative purposes. More practical examples are provided in Subsection 2.4. Further examples related to the total positivity (e.g., Holland and Wang (1987), Kurowicka and van Horssen (2015)) and circulas (Jones, Pewsey and Kato (2015)) are presented in Appendix A of Sei and Yano (2024).

Example 1 (Gaussian model; Jansen (1997)). The bivariate Gaussian distribution is a minimum information dependence model. This fact was pointed out by Jansen (1997) in an argument of the copula theory. Suppose that the mean vector is zero for simplicity. If the variance and correlation parameters are denoted as σ_1^2, σ_2^2 and ρ, respectively, then the density is written as

$$p(x_1, x_2; \sigma_1, \sigma_2, \rho) = e^{\theta x_1 x_2 - a_1(x_1;\theta,\nu) - a_2(x_2;\theta,\nu) - \psi(\theta,\nu)} \phi(x_1; \sigma_1^2) \phi(x_2; \sigma_2^2),$$

where $\theta = (\sigma_1\sigma_2)^{-1}\rho(1-\rho^2)^{-1}$ is the canonical parameter, $x_1 x_2$ is the canonical statistic, $\nu = (\sigma_1^2, \sigma_2^2)$ is the marginal parameter and $\phi(x_i; \sigma_i^2)$ denotes the univariate normal density. The adjusting functions and the potential function are

$$a_i(x_i;\theta,\nu) = \left(\frac{1}{2(1-\rho^2)} - \frac{1}{2}\right)\left(\left(\frac{x_i}{\sigma_i}\right)^2 - 1\right), \quad i = 1,2,$$

$$\psi(\theta,\nu) = \frac{1}{2}\log(1-\rho^2) + \frac{1}{1-\rho^2} - 1,$$

respectively, where $\rho = \rho(\theta,\nu)$ is the unique solution of $\theta = (\sigma_1\sigma_2)^{-1}\rho(1-\rho^2)^{-1}$. More explicitly, $\rho = \rho(\theta,\nu) = 2\theta\sigma_1\sigma_2/\{1 + (1 + 4\theta^2\sigma_1^2\sigma_2^2)^{1/2}\}$. The canonical parameter θ ranges over $\mathbb{R}$ whereas ρ ranges over $(-1,1)$. The one-to-one relationship between ρ and θ for a given ν is a consequence of a more general result (see Appendix B.1 of Sei and Yano (2024)). We will show that higher-dimensional Gaussian models also have a similar structure in Subsection 2.4.

Remark 1. Here we clarify the difference between minimum information dependence models and copula models. Assume that $\mathcal{X}_i = \mathbb{R}$ for $i = 1, \ldots, d$. A copula model is built by using the Sklar theorem and the change of variables as

$$p_{\text{copula}}(x; c, \nu) = c(R_1(x_1;\nu), \ldots, R_d(x_d;\nu)) \prod_{i=1}^{d} r_i(x_i;\nu),$$

where $c(\cdot)$ is a copula density, $R_i(\cdot;\nu)$ $(i = 1, \ldots, d)$ is the distribution function of the i-th variable for $i = 1, \ldots, d$, and $r_i(\cdot;\nu)$ $(i = 1, \ldots, d)$ is the density function of the i-th variable. In contrast, the minimum information dependence model is built as

$$p(x;\theta,\nu) = \exp\left(\theta^\top h(x) - \sum_{i=1}^{d} a_i(x_i;\theta,\nu) - \psi(\theta,\nu)\right) \prod_{i=1}^{d} r_i(x_i;\nu).$$

Two models have the form of the dependence part and the product of marginals. But, the dependence part in the copula model has the composite form $c \circ R$, whereas that of our model has the exponential of the additive form: the term $\theta^\top h(x)$ independent of marginals plus the terms $-\sum_{i=1}^{d} a_i(x_i;\theta,\nu) - \psi(\theta,\nu)$ depending on the marginals. So, the coincidence between copula and minimum information dependence models depends on the marginal distributions.

One example of the coincidence is a Gaussian model (Gaussian copula with Gaussian marginals). Another example is the bi-variate minimum information dependence copula with uniform marginals. In contrast, one example of the discordance is the Farlie–Gumbel–Morgenstern copula with Gaussian marginals that is included in copula models but is excluded in our models.

In Example 1, the adjusting functions and potential functions are explicitly represented. However, such cases are exceptional. To generate a rich class of distributions, we establish an existence and uniqueness theorem for the functions in the following subsection.

2.3. Existence and uniqueness

We find tractable conditions for the existence and uniqueness of the adjusting functions and potential function. Fix θ and ν and drop the dependence on these parameters in this subsection. Then, the inner product $\theta^\top h(x)$ is replaced with a scalar function $H(x) = \theta^\top h(x)$. The marginal parameter ν is also abbreviated as $r_i(x_i)$.

Denote the product density of $r_i(x_i)$ by $p_0(x) = \prod_{i=1}^d r_i(x_i)$ that corresponds to the density function for $H = 0$. Denote the set of integrable functions with respect to a measure μ by $L_1(\mu)$.

Definition 2. We say that a function $H \in L_1(p_0(x)\mathrm{d}x)$ is *feasible* if there exist measurable functions $\{a_i(x_i) : i = 1, \dots, d\}$ and a real number $\psi \in \mathbb{R}$ such that the function $p(x) = e^{H(x) - \sum_{i=1}^d a_i(x_i) - \psi} p_0(x)$ satisfies $\int p(x)\mathrm{d}x_{-i} = r_i(x_i)$ for each $i = 1, \dots, d$ and $\int \sum_{i=1}^d a_i(x_i) p(x)\mathrm{d}x = 0$.

Definition 3. We say that H is *strongly feasible* if H is feasible and each a_i belongs to $L_1(r_i(x_i)\mathrm{d}x_i)$.

Definition 4. We say that H is *moderately feasible* if there exist $\{b_i \in L_1(r_i(x_i)\mathrm{d}x_i) : i = 1, \dots, d\}$ such that

$$\int e^{H(x) - \sum_{i=1}^d b_i(x_i)} p_0(x)\mathrm{d}x < \infty. \tag{5}$$

The only difference between feasibility and strong feasibility is the integrability of $a_i(x_i)$. Strong feasibility is convenient for theoretical discussions whereas feasibility is sufficient for the modeling and the inference. Strong feasibility implies moderate feasibility by Definition 4. If H is feasible and the adjusting functions are integrable, then H is strongly feasible.

Our first main result is that moderate feasibility is a sufficient condition of feasibility. The proof, which is based on the results of Borwein, Lewis and Nussbaum (1994), is given in Appendix C.1 of Sei and Yano (2024).

Theorem 1. *If a function $H \in L_1(p_0(x)\mathrm{d}x)$ is moderately feasible, then H is feasible. Furthermore, if H is feasible, then $\sum_{i=1}^d a_i(x_i)$ and ψ are unique.*

It is usually easy to find integrable functions b_is satisfying the inequality (5); we provide concrete examples in Subsection 2.4.

For moderate feasibility, we obtain a couple of corollaries. First, by applying Hölder's inequality to the integral in (5), we immediately obtain the following corollary.

Corollary 1. *The set of moderately feasible functions is convex.*

Further, we provide the following two useful criteria for moderately feasibility. The proofs are given in Appendix C.2.

Corollary 2. *Suppose that for each $i = 1, \dots, d$, the space $\mathcal{X}_i$ is equal to $\mathbb{R}$, and the marginal density function r_i has finite moments of any order. Then, any polynomial function $H(x)$ is moderately feasible.*

Corollary 3. *Suppose that for each $i = 1,\dots,d$, $(\mathcal{X}_i, d_i)$ is a metric space and there exists $x_i^0 \in \mathcal{X}_i$ for which $d_i(x_i, x_i^0) \in L_1(r_i(x_i)\mathrm{d}x_i)$. Then, any Lipschitz function $H(x)$ with respect to $d_{(p)}(x,y) := \|(d_i(x_i,y_i))_{i=1,\dots,d}\|_p$ for some $p \in [1,\infty]$ is moderately feasible, where $\|\cdot\|_p$ is the p-norm.*

2.4. Practical examples

In this subsection, we provide various examples of the minimum information dependence model by applying Theorem 1. Denote the canonical part of the model by $H_\theta(x) = \theta^\top h(x)$.

Example 2 (Gaussian). For each $i = 1,\dots,d$, let $r_i(x_i;\nu)$ be the Gaussian density $\phi(x_i;\mu_i,\sigma_i^2)$, where $\nu = (\mu_1,\dots,\mu_d,\sigma_1,\dots,\sigma_d)$. Then, a quadratic function $H_\theta(x) = \sum_{i<j}\theta_{ij}x_ix_j$ is feasible for any real vector $\theta = (\theta_{ij}) \in \mathbb{R}^{d(d-1)/2}$ because the condition of Corollary 2 is satisfied. The obtained model $p_\theta(x)$ is simply the multivariate normal density. Indeed, there exists a unique positive definite matrix $\Sigma = (\sigma_{ij})$ such that $\sigma_{ii} = \sigma_i^2$ and $(\Sigma^{-1})_{ij} = \theta_{ij}$ (Dempster (1972)). We also point out that the covariance selection model (Gaussian graphical model) is specified by the set of edges (i,j) such that $\theta_{ij} = 0$. See Lauritzen (1996) and Whittaker (1990) for details of the covariance selection model.

The following example deals with three-dimensional interaction. We emphasize that it is not easy to construct such a model if we use exponential families.

Example 3 (Three-dimensional interaction). Let $d = 3$ and r_i be the standard normal density for $i = 1,2,3$ and define $H_\theta(x_1,x_2,x_3) = \theta x_1x_2x_3$, where θ is a real parameter. The function is feasible by Corollary 2. The obtained distribution can describe Simpson's paradox for continuous variables, that is, the conditional correlation coefficient between the first and second variables given the third variable depends on the value of the third variable; Gaussian distributions do not have this property. As $\theta \to \infty$, the distribution tends to a distribution supported on $\{x \in \mathbb{R}^3 \mid |x_1| = |x_2| = |x_3|,\ x_1x_2x_3 \geq 0\}$ that is the optimal coupling between the three marginal distributions:

$$\text{Minimize } -\int x_1x_2x_3\, p(x_1,x_2,x_3)\mathrm{d}x$$

$$\text{subject to } p \text{ with marginals equal to } r_i\text{s}.$$

We can see this convergence in more detail through the relationship with the entropic optimal transport problem discussed in Appendix B.3 of Sei and Yano (2024).

Example 4 (count data). Suppose that $(\mathcal{X}_i, \mathrm{d}x_i)$ for $i = 1,2$ is the set of non-negative integers with the counting measure. For each i, let $r_i(x_i;\nu_i) = (\nu_i^{x_i}/x_i!)e^{-\nu_i}$ be the Poisson distribution with mean $\nu_i > 0$ and define $H_\theta(x_1,x_2) = \theta x_1x_2$ for $\theta \in \mathbb{R}$. Then we can see that H_θ is feasible for any $\theta \in \mathbb{R}$ from Corollary 2. The parameter θ controls the correlation between the two variables. The range of the Pearson correlation coefficient is a proper subset of $(-1,1)$ that depends on $\nu = (\nu_1,\nu_2)$. Table 1 shows numerical evaluation of the lower and upper bounds of the correlation coefficient for several values of ν_2 with ν_1 set to 1, where the Sinkhorn–Knopp algorithm (Sinkhorn and Knopp (1967)) is applied for this evaluation; see Appendix B.3 of Sei and Yano (2024) for details of the computation. Geenens (2020) discusses another construction of bivariate distributions with Poisson marginals and a negative association.

Table 1. Ranges of the Pearson correlation coefficient for the Poisson marginal model. Parameter ν_1 is set to 1.

	ν_2				
	0.25	0.5	1	2	4
upper bound	0.82	0.87	0.99	0.94	0.93
lower bound	−0.50	−0.67	−0.74	−0.81	−0.87

Example 5 (mixed variables). In our model, the domains $\mathcal{X}_1,\dots,\mathcal{X}_d$ are not necessarily identical. Suppose that $d = 2$, $\mathcal{X}_1 = \mathbb{R}$ and $\mathcal{X}_2 = \{0,1\}$ for simplicity. (The argument presented herein is easily generalized to other domains and higher-dimensional cases.) Then, the minimum information dependence model is $p(x_1,x_2;\theta,\nu) = \exp(\theta^\top h(x_1,x_2) - \widetilde{a}_1(x_1;\theta,\nu) - \widetilde{a}_2(x_2;\theta,\nu) - \psi(\theta,\nu))$, where $\widetilde{a}_i(x_i;\theta,\nu) = a_i(x_i;\theta,\nu) - \log r_i(x_i;\nu)$. The interaction between the continuous and discrete variables is described by the canonical statistic $h(x_1,x_2)$.

The model is interpreted as regression models in two ways as follows. First, let x_1 be the explanatory variable and x_2 be the response variable. The conditional density of $x_2 = 1$ given x_1 is written as

$$p(x_2 = 1|x_1;\theta,\nu) = \frac{1}{1+\exp(-\theta^\top u(x_1) - \alpha(\theta,\nu))},$$

where $u(x_1) = h(x_1,1) - h(x_1,0)$ and $\alpha(\theta,\nu) = -\widetilde{a}_2(1;\theta,\nu) + \widetilde{a}_2(0;\theta,\nu)$. This is the logistic regression model, except that the intercept term $\alpha(\theta,\nu)$ depends on the other regression coefficient θ. Since ν is a nuisance parameter, we can treat $\alpha(\theta,\nu)$ as a nuisance parameter. More precisely, it is shown that there is a one-to-one correspondence between $\alpha(\theta,\nu)$ and $r_2(1;\nu)$, given θ, by Theorem 1. Thus, the proposed model is equivalent to the logistic model in this sense, and the difference is which of α and r_2 is specified first.

Next, let x_1 be the response variable and x_2 be the explanatory variable. The conditional density of x_1 given x_2 is

$$p(x_1|x_2;\theta,\nu) = \exp(\theta^\top h(x_1,x_2) - \psi_2(\theta,\nu|x_2))m(x_1;\theta,\nu),$$

where $m(x_1;\theta,\nu) := e^{-a_1(x_1;\theta,\nu)}r_1(x_1;\nu)$ is the base measure independent from x_2, and $\psi_2(\theta,\nu|x_2)$ is the normalizing constant that makes $\int p(x_1|x_2)\mathrm{d}x_1 = 1$. This is just a generalized linear model except that the base measure depends on the regression coefficient θ. From Theorem 1, the base measure has a one-to-one correspondence with the marginal density for a given θ.

An important consequence here is that the two regression models of opposite direction are derived from a common joint density function. This is in contrast to the traditional regression approach for mixed variables (e.g., Chapter 6 of Lauritzen (1996) and Chapter 11 of Whittaker (1990)). Yang et al. (2015) proposed a mixed-variable model based on univariate conditional exponential families, but the model cannot avoid the restriction on the parameter space due to the integrability. The difference between our model and the model proposed by Yang et al. (2015) is that Yang et al. (2015)'s model specifies the base measure a priori instead of the marginal densities, while our model specifies marginal densities instead of the base measure.

Remark 2 (Choice of h). The choice of canonical statistics h is one of important issues for practitioners. This section and Appendix A give several examples of h. The standard choice of h has the form $h(x) = \prod_{i=1}^d h_i(x_i)$ with $h_i : \mathcal{X}_i \to \mathbb{R}$. There are several ways to choose h_i: The first is to use polynomial functions in a suitable sense. The second is to use monotone functions in a suitable sense. The third is to employ known embedding functions; for example, see Appendix D.2. In any case, the conditional inference in Section 3 is applicable, and the selection of h using the conditional likelihood is possible.

3. Inference

In this section, we consider the inference for θ using the conditional likelihood as well as the sampling algorithm for the minimum information dependence model.

3.1. Decomposition of the likelihood

Suppose that $x(1),\dots,x(n)$ are independent and identically distributed (i.i.d.) according to a density in a minimum information dependence model (2). Denote the components of $x(t)$ as $x(t)=(x_i(t))_{i=1}^d$.

We decompose the likelihood function into a marginal part and a dependent part using an order and a rank. By the well-ordering principle, we can define a total order $\le_i$ on $\mathcal{X}_i$ for each $i=1,\dots,d$. Using the ordering is convenient for the following description and, as we shall see in Lemma 1, the choice does not affect the inference. Also, the following remark provides a standard choice of the ordering.

Remark 3 (Observational order). Given the n observations $(x(t))_{t=1}^n$, we can use the "observational order", $x_i(t)\le_i x_i(s)$ if $t\le s$, as if they are predetermined. The observational order makes implementation easier since it does not require sorting.

For each $i=1,\dots,d$, define the set of i-th marginal values by

$$M_i(1)\le_i\cdots\le_i M_i(n), \tag{6}$$

where for each i, $M_i=(M_i(1),\dots,M_i(n))$ are the n observations $(x_i(t))_{t=1}^n$ sorted by the predetermined order $\le_i$. We call it the *marginal order statistic*. Define the *rank statistic* by a permutation $\pi_i=(\pi_i(t))_{t=1}^n\in\mathbb{S}_n$ such that $x_i(t)=M_i(\pi_i(t))$, where we denote the symmetric group of degree n by $\mathbb{S}_n$. If there are ties of observations, we choose π with equal probability over the set of permutations giving the same observations. Denote the vector of all statistics as $M=(M_1,\dots,M_d)\in\prod_{i=1}^d\mathcal{X}_i^n$ and $\pi=(\pi_1,\dots,\pi_d)\in\mathbb{S}_n^d$. For each $t=1,\dots,n$, the t-th observation $x(t)$ is recovered from M and π, and is written as

$$x(t)=(M\circ\pi)(t)=(M_i(\pi_i(t)))_{i=1}^d.$$

Example 6. For illustrative purposes, consider discrete spaces $\mathcal{X}_1=\{\mathsf{a},\mathsf{b},\mathsf{c}\}$ and $\mathcal{X}_2=\{1,2\}$ equipped with the alphabetical and numeric orders, respectively. Suppose that we have the 4 observations $x(1)=(\mathsf{c},2)$, $x(2)=(\mathsf{c},1)$, $x(3)=(\mathsf{b},2)$, $x(4)=(\mathsf{a},1)$. Then, the marginal order statistics are $M_1=(\mathsf{a},\mathsf{b},\mathsf{c},\mathsf{c})$ and $M_2=(1,1,2,2)$. The rank statistic π_1 is an element of $\{(3,4,2,1),(4,3,2,1)\}$ with equal probability $1/2$. Suppose that we choose $\pi_1=(3,4,2,1)$. Similarly, choose $\pi_2=(3,1,4,2)$ from four possible permutations. Then, the observations are recovered from M and π as

$$\begin{aligned}
(M\circ\pi)(1)&=(M_1(\pi_1(1)),M_2(\pi_2(1)))=(M_1(3),M_2(3))=(\mathsf{c},2),\\
(M\circ\pi)(2)&=(M_1(\pi_1(2)),M_2(\pi_2(2)))=(M_1(4),M_2(1))=(\mathsf{c},1),\\
(M\circ\pi)(3)&=(M_1(\pi_1(3)),M_2(\pi_2(3)))=(M_1(2),M_2(4))=(\mathsf{b},2),\\
(M\circ\pi)(4)&=(M_1(\pi_1(4)),M_2(\pi_2(4)))=(M_1(1),M_2(2))=(\mathsf{a},1).
\end{aligned}$$

The following lemma implies that the full likelihood is decomposed as the product of the conditional likelihood independent of the marginal parameter ν and the marginal likelihood.

Lemma 1. *The likelihood function is decomposed as*

$$L(M,\pi;\theta,\nu) := \prod_{t=1}^{n} p((M\circ\pi)(t);\theta,\nu) = f(\pi|M;\theta)g(M;\theta,\nu),$$

where

$$f(\pi|M;\theta) = \frac{e^{\sum_{t=1}^n \theta^\top h((M\circ\pi)(t))}}{\sum_{\widetilde{\pi}\in\mathbb{S}_n^d} e^{\sum_{t=1}^n \theta^\top h((M\circ\widetilde{\pi})(t))}} \quad \text{and} \quad g(M;\theta,\nu) = \sum_{\widetilde{\pi}\in\mathbb{S}_n^d} L(M,\widetilde{\pi};\theta,\nu).$$

Further, the conditional likelihood does not depend on the choice of the ordering:

$$f(\pi \mid M;\theta) = \frac{e^{\sum_{t=1}^n \theta^\top h(x(t))}}{\sum_{\widetilde{\pi}\in\mathbb{S}_n^d} e^{\sum_{t=1}^n \theta^\top h((x_i(\widetilde{\pi}_i(t)))_{i=1}^d)}}.$$

Proof. For the former assertion, it is sufficient to show that the conditional distribution of π is

$$\frac{L(M,\pi;\theta,\nu)}{\sum_{\widetilde{\pi}\in\mathbb{S}_n^d} L(M,\widetilde{\pi};\theta,\nu)} = \frac{e^{\sum_{t=1}^n \theta^\top h((M\circ\pi)(t))}}{\sum_{\widetilde{\pi}\in\mathbb{S}_n^d} e^{\sum_{t=1}^n \theta^\top h((M\circ\widetilde{\pi})(t))}}.$$

Indeed, the adjusting-function part of $\log L(M,\pi;\theta,\nu)$ is

$$\sum_{j=1}^{d}\sum_{t=1}^{n} a_j(M_j(\pi_j(t));\theta,\nu) = \sum_{j=1}^{d}\sum_{t=1}^{n} a_j(M_j(t);\theta,\nu)$$

and does not depend on π, which proves the former assertion. The latter assertion follows immediately from the identity $x(t) = (M\circ\pi)(t)$. □

Remark 4. Let us mention that the choice or the randomization in defining rank statistics does not impact on the inference based on the conditional likelihood. In fact, the latter part of Lemma 1 shows the conditional likelihood is independent from the choice of total orders $(\leq_i)_{i=1}^d$ and is not affected by the presence of ties, which is different from the inference of discrete copula models (Genest and Nešlehová, 2007).

3.2. Conditional maximum likelihood estimation

In this subsection, on the basis of the conditional likelihood, we propose a method of estimating θ and show that this yields a consistent estimator.

We first point out that the log conditional likelihood ratio is almost the same as the log likelihood ratio and thus utilizing the conditional likelihood for the inference of θ is reasonable. To show this, we make the following assumptions.

Assumption 1. The following hold:

(1) Let $(\mathcal{X}_i, d_i, \mathrm{d}x_i)$ for $i = 1,\dots,d$ be metric measure spaces. There exist $\kappa > 0$ and $\alpha \geq 0$ such that for each $i = 1,\dots,d$, the ε-covering number $\mathcal{N}(\mathcal{X}_i, d_i, \varepsilon)$ is bounded as $\mathcal{N}(\mathcal{X}_i, d_i, \varepsilon) \leq \kappa\varepsilon^{-\alpha}$ for sufficiently small $\varepsilon > 0$.

(2) The parameter space Θ of canonical parameter θ is bounded.
(3) The canonical statistic $h(x)$ is Lipschitz continuous with respect to x with Lipschitz constant L_h, where $\mathcal{X}$ is endowed with the 2 product metric $d_{\mathcal{X}}(x,y) := (\sum_{i=1}^{d} d_i^2(x_i,y_i))^{1/2}$.

Assumption 1 (1) states that each $\mathcal{X}_i$ is a totally bounded space with a finite upper Minkowski–Bouligand dimension (e.g., Falconer (2014)), and is quite general. Assumption 1 (2) is a usual assumption. Assumption 1 (3) ensures the existence of $p(x;\theta,\nu)$ by Corollary 3 and that of an approximator of $p(x;\theta,\nu)$.

Under these assumptions, we obtain the following theorem implying that the log conditional likelihood ratio (per sample) approaches to the log likelihood ratio (per sample) uniformly in θ as the sample size gets larger.

Theorem 2. *Let θ_0 be the true canonical parameter in the parameter space Θ of θ, and let ν_0 be the true marginal parameter in the parameter space* N *of ν. Suppose that $x(1),\dots,x(n)$ are i.i.d. according to $p(x;\theta_0,\nu_0)$.*

(1) *Under Assumption 1, there exists a positive constant C not depending on n such that we have*

$$\mathbb{E}\left[\sup_{\theta\in\Theta}\frac{1}{n}\left|\log\frac{\prod_{t=1}^{n}p(x(t);\theta_0,\nu_0)}{\prod_{t=1}^{n}p(x(t);\theta,\nu_0)}-\log\frac{f(\pi\mid M;\theta_0)}{f(\pi\mid M;\theta)}\right|\right]\le C\varepsilon_n, \tag{7}$$

where

$$\varepsilon_n := \max\left\{n^{-\frac{(d-1)\alpha+1}{2(d\alpha+1)}}(\log n)^{-\frac{\alpha}{2(d\alpha+1)}}, n^{-\frac{1}{d\alpha+1}}(\log n)^{\frac{1}{d\alpha+1}}\right\}.$$

(2) *In addition, under Assumption 1 and under the additional assumption that for $i=1,\dots,d$, the marginal density $r_i(x_i;\nu)$ is log-Lipschitz continuous with Lipschitz constant $L_{r_i,\nu}$. where $\sup_{\nu\in\mathrm{N}} L_{r_i,\nu}<\infty$, there exists a positive constant C not depending on n such that for any subset* $\mathrm{S}\subset\mathrm{N}$ *containing ν_0, we have*

$$\mathbb{E}\left[\sup_{\theta\in\Theta,\nu\in\mathrm{S}}\frac{1}{n}\left|\log\frac{\prod_{t=1}^{n}p(x(t);\theta_0,\nu_0)}{\prod_{t=1}^{n}p(x(t);\theta,\nu)}-\log\frac{f(\pi\mid M;\theta_0)}{f(\pi\mid M;\theta)}\right|\right]\le C\left\{\sup_{\nu\in\mathrm{S}}\widetilde{D}(\nu_0,\nu)+\varepsilon_n\right\}, \tag{8}$$

where $\widetilde{D}(\nu_0,\nu) := \sum_{i=1}^{d}\{\|r_i(\cdot;\nu_0)-r_i(\cdot;\nu)\|_1 + D(r_i(\cdot;\nu_0),r_i(\cdot;\nu))\}$ with $D(\cdot,\cdot)$ the Kullback–Leibler divergence.

The proof of the theorem is given in Appendix C.3 of Sei and Yano (2024). The main ingredients of the proof are the approximation using ε-net, the Stirling approximation (e.g., Robbins (1955)), the ℓ_1 and the Kullback–Leibler deviation inequalities for the multinomial distribution (Weissman et al. (2003) and Agrawal (2020)), the recent quantitative stability result for the entropic optimal transport with respect to the marginals (e.g., Eckstein and Nutz (2022)), and the Pythagorean theorem for the minimum information dependence model (Theorem S.1 in Appendix B.2 of Sei and Yano (2024)).

Theorem 2 implies that the marginal order statistics M are almost ancillary for the dependence parameter θ, that is, they contain little information about θ:

$$\mathbb{E}\left[\sup_{\theta\in\Theta}\left|\frac{1}{n}\log\frac{g(M;\theta_0,\nu_0)}{g(M;\theta,\nu_0)}\right|\right]\to 0.$$

In the literature, there are several results related to Theorem 2. In the bivariate Gaussian model, the sample variances are shown to be almost ancillary for the correlation parameter; see Example 2.30 of

Cox and Hinkley (1974). In contingency tables, Choi, Blume and Dupont (2015) and Little (1989) have also shown that the marginal frequency is an almost ancillary statistic for the dependence parameter.

Henceforth, fix the marginal order statistics M and denote $h_*(\pi) = \sum_{t=1}^n h((M \circ \pi)(t)) \in \mathbb{R}^K$ for simplicity. The conditional likelihood is then expressed as

$$f(\pi|M;\theta) = \frac{e^{\theta^\top h_*(\pi)}}{\sum_{\widetilde{\pi}\in\mathbb{S}_n^d} e^{\theta^\top h_*(\widetilde{\pi})}} \tag{9}$$

and forms an exponential family with a canonical parameter θ and a sufficient statistic $h_*(\pi)$. Exponential families on permutations without conditioning and their limiting behavior were investigated by Mukherjee (2016), where the limit of the model was shown to be a minimum information copula model under suitable conditions.

Definition 5. The conditional maximum likelihood estimate (CLE) $\widehat{\theta}$ is a maximizer of the conditional likelihood (9).

Using Theorem 2, we obtain the consistency of CLE. To show this, we make an additional assumption.

Assumption 2. There exists a $p(\cdot;\theta_0,\nu_0)$-square-integrable function $A(\cdot)$ such that

$$\left|\sum_{i=1}^d a_i(x_i;\theta,\nu_0) - \sum_{i=1}^d a_i(x_i;\theta',\nu_0)\right| < A(x)\|\theta-\theta'\| \quad \text{for} \quad \theta \neq \theta' \in \Theta.$$

This assumption ensures that the log-likelihood of the minimum information dependence model forms Glivenko–Cantelli class (e.g., van der Vaart and Wellner (1996)), and is expected by the quantitative stability result for the entropic optimal transport with respect to the cost (e.g., Eckstein and Nutz (2022)).

Then, we obtain the following consistency result of CLE $\widehat{\theta}$.

Corollary 4. *Under Assumptions 1 and 2, we have $\widehat{\theta} \to \theta_0$ in probability.*

We then consider the asymptotic variance of $\widehat{\theta}$. Let $\Psi(\theta)$ be the potential function of the conditional likelihood (9) as an exponential family, that is, $\Psi(\theta) = \Psi(\theta|M) = \log\left(\sum_{\widetilde{\pi}\in\mathbb{S}_n^d} e^{\theta^\top h_*(\widetilde{\pi})}\right)$. Then the conditional likelihood (9) can be written as $f(\pi|M) = \exp(\theta^\top h_*(\pi) - \Psi(\theta))$. Denote the derivative with respect to the parameter as $\partial_j = \partial/\partial\theta_j$. The expectation parameter and Fisher information matrix are

$$\mu_j(\theta) = \partial_j\Psi(\theta) = E[h_{*j}(\pi)|M] \text{ and} \tag{10}$$

$$G_{jk}(\theta) = \partial_j\partial_k\Psi(\theta) = E[\{h_{*j}(\pi) - \mu_j(\theta)\}\{h_{*k}(\pi) - \mu_k(\theta)\}|M], \tag{11}$$

respectively, where $h_{*j}(\pi)$ is the j-th element of $h_*(\pi)$ and the expectation is taken with respect to $f(\pi|M;\theta)$.

For 2 by 2 contingency tables, CLE was shown to be asymptotically normal and efficient by Harkness (1965), based on the asymptotic form of the non-central hypergeometric distribution (Hannan and Harkness (1963)); see also Kou and Ying (1996). For general contingency tables, the following theorem holds.

Theorem 3 (Haberman (1977), Theorem 4.1). *Suppose that $\mathcal{X}_1,\ldots,\mathcal{X}_d$ are finite. Then, the asymptotic distribution of $\sqrt{n}(\widehat{\theta}-\theta)$ is $\mathrm{N}(0,g^{jk}(\theta))$, where $g^{jk}(\theta)$ is the inverse of the Fisher information matrix $g_{jk}(\theta)$ with respect to θ of the model (2). The conditional Fisher information matrix in (11) satisfies $G_{jk}(\theta)/n \to g_{jk}(\theta)$ in probability.*

In Appendix B.1, we show that the Fisher information matrix g_{jk} is the Hessian matrix of the potential function ψ and actually coincides with that given by Haberman (1977). We also provide a useful expression for g_{jk} using the back-fitting algorithm (Buja, Hastie and Tibshirani (1989)). Furthermore, in Appendix B.2, the canonical parameter θ and the marginal parameter ν are shown to be orthogonal to each other.

We expect that the same property as Theorem 3 holds for infinite sample spaces. A numerical study supporting this claim is provided in Section 4. Note that even the $\sqrt{n}$-consistency of the maximum likelihood estimator (MLE) for the exponential families on permutations has not been shown in the literature. Mukherjee (2016) showed the MLE to be consistent, but obtained a $\sqrt{n}$-consistency result only for the pseudo-likelihood estimator of Besag. Although the asymptotic normality in general cases is not proved, it is reasonable to use the inverse of the conditional information matrix $G_{jk}(\widehat{\theta})/n$ as an estimate of the asymptotic covariance of $\widehat{\theta}$. We also suggest using the likelihood ratio (or score/Wald) test and Akaike's information criterion assuming asymptotic normality.

From the theory of discrete exponential families (e.g., Rinaldo, Fienberg and Zhou (2009)), we obtain the following lemma.

Lemma 2. *Let $P=\mathrm{conv}(\{h_*(\widetilde{\pi}) \mid \widetilde{\pi}\in\mathbb{S}_n^d\})\subset\mathbb{R}^K$ be the convex hull of the range of the sufficient statistic. Suppose that the interior of P is nonempty. Then, CLE $\widehat{\theta}$ exists if and only if $h_*(\pi)\in\mathrm{int}(P)$. The estimator is unique and satisfies $\mu_j(\widehat{\theta})=h_{*j}(\pi)$, whenever it exists.*

The condition $\mathrm{int}(P)\neq\varnothing$ in the lemma is difficult to check because it requires the computation of all possible values of h_*. We provide a tractable sufficient condition in Subsection 3.4.

Since any exponential family is log-concave with respect to the canonical parameter, we can use, in principle, any convex programming solver to obtain CLE. However, a critical issue here is to calculate the normalizing constant in the denominator of (9). To overcome this difficulty, we propose a sampling approach and a pseudo likelihood approach in the following subsections.

3.3. Estimation via Monte Carlo

To perform the conditional inference, we need to compute the expectations of several statistics under the conditional distribution (9). We propose a sampling method in the Metropolis–Hastings manner described in Table 2, which is quite easy to implement. We call the method the exchange algorithm. Note that the method is essentially the same as those for contingency tables (e.g., Diaconis and Sturmfels (1998)).

Note that the marginal order statistics M are preserved during the procedure. The state space of the Markov chain is $\mathbb{S}_n^d$. The following lemma is obtained immediately from the construction, and the proof is omitted.

Lemma 3. *The chain $(\pi^{(l)})_{l=1}^{\infty}$ from the exchange algorithm is ergodic with its stationary distribution $f(\pi|M;\theta)$ in (9).*

Table 2. Exchange algorithm.

Input: An initial permutation $\pi^{(0)} \in \mathbb{S}_n^d$ and the number of samples L.
Output: L samples $(\pi^{(l)} \in \mathbb{S}_n^d)_{l=1}^L$ of permutations.
Step 1: Initialize $\pi \leftarrow \pi^{(0)}$ and $l = 1$.
Step 2: Select $1 \le i \le d$ and $1 \le s < t \le n$ uniformly at random.
Let $\tau_{st}^i \in \mathbb{S}_n^d$ be the transposition between s and t with respect to the i-th variable.
Compute the conditional likelihood ratio $\rho = \frac{f(\pi\circ\tau_{st}^i \mid M;\theta)}{f(\pi \mid M;\theta)} = \frac{e^{H_s(\pi\circ\tau_{st}^i)+H_t(\pi\circ\tau_{st}^i)}}{e^{H_s(\pi)+H_t(\pi)}}$, where $H_t(\pi) = \theta^\top h((M\circ\pi)(t))$.
Step 3: Generate a random number u uniformly distributed on $[0,1]$.
If $u \le \min(1,\rho)$, then update π to $\pi \circ \tau_{st}^i$.
Step 4: Let $\pi^{(l)} \leftarrow \pi$ and $l \leftarrow l+1$.
Go to Step 2 if $l \le L$, and output $(\pi^{(l)})_{l=1}^L$ otherwise.

Remark 5. It is natural to consider a sampling method for $p(x;\theta,\nu)$. A naive method is just to generate n random elements $(x_i(t))_{t=1}^n$ according to the marginal distribution $r_i(x_i;\nu)$ for each $i = 1,\dots,d$, independently, and then to employ the exchange algorithm.

However, the method is not exact. Indeed, if $n = 1$, then the procedure generates a sample from the independent model $\prod_{i=1}^d r_i(x_i;\nu)$, not from the correct distribution. It is expected that the distribution of the sample generated in this way converges to the correct distribution as $n \to \infty$. This observation is supported by Theorem 2 because the target marginal density $g(M;\theta,\nu)$ and the independent counterpart $g(M;0,\nu)$ are asymptotically equivalent as $n \to \infty$.

We can compute CLE $\widehat{\theta}$ via MCMC in line with Geyer and Thompson (1992). More specifically, let θ be a current estimate of $\widehat{\theta}$ and let $\{\pi^{(l)}\}_{l=1}^L$ denote a sample from the conditional likelihood in (9) obtained by the exchange algorithm. The quantities $\mu_j(\theta)$ and $G_{jk}(\theta)$ are approximated by $\check{\mu}_j = \sum_{l=1}^L h_{*j}(\pi^{(l)})/L$ and $\check{G}_{jk} = \sum_{l=1}^L (h_{*j}(\pi^{(l)}) - \check{\mu}_j)(h_{*k}(\pi^{(l)}) - \check{\mu}_k)/L$. Then, the estimate is updated by Fisher's scoring method: $\theta_j \leftarrow \theta_j + \sum_{k=1}^K \check{G}^{jk}\{h_{*k}(\pi) - \check{\mu}_k\}$, where $\check{G}^{jk}$ is the inverse matrix of $\check{G}_{jk}$. The procedure is repeated until convergence. As noted by Geyer and Thompson (1992), the MCMC samples can be reused at every step by importance sampling.

The standard error of $\widehat{\theta}_j$ is estimated by $(\check{G}^{jj}/n)^{1/2}$. Hypothesis testing and model selection based on the likelihood ratio statistic (or score/Wald statistic) are also available, where the statistic is computed from the MCMC samples.

3.4. Besag's pseudo likelihood

An alternative to CLE is Besag's pseudo likelihood estimator for permutations (see Mukherjee (2016)), which does not require the normalizing constant in (9). The *pseudo likelihood estimator* (PLE) is defined as a maximizer of

$$\prod_{i=1}^d \prod_{1\le s<t\le n} f(\pi_i(s),\pi_i(t) \mid (\pi_j(u))_{(u,j)\neq(s,i),(t,i)}, M;\theta),$$

where $\pi = (\pi_i(t)) \in \mathbb{S}_n^d$ is the observed rank statistic and

$$f(\pi_i(s),\pi_i(t) \mid (\pi_j(u))_{(u,j)\neq(s,i),(t,i)}, M;\theta) = \frac{1}{1+e^{\theta^\top(h_*(\pi\circ\tau_{st}^i)-h_*(\pi))}}.$$

Recall that τ_{st}^i denotes the transposition between s and t with respect to the i-th variable (see Subsection 3.3). A similar form of PLE for copula models was proposed by Chen and Sei (2022) in the framework of scoring rules.

The expression of the pseudo likelihood coincides with the likelihood of a logistic regression model, where the explanatory variable is $u_{st}^i = h_*(\pi) - h_*(\pi \circ \tau_{st}^i) \in \mathbb{R}^K$ and the response variable is $y_{st}^i = 1$ for all i and (s,t). Hence, any software package for logistic regression can be utilized.

As a consequence of regarding PLE as MLE of a logistic regression model, we obtain a condition for the existence of PLE.

Lemma 4 (Albert and Anderson (1984)). *Let* $Q = \mathrm{conv}(\{h_*(\pi \circ \tau_{st}^i) \mid 1 \le i \le d, 1 \le s < t \le n\})$. *Suppose that* $\mathrm{int}(Q) \neq \varnothing$. *Then, PLE exists if and only if* $h_*(\pi) \in \mathrm{int}(Q)$.

The condition $h_*(\pi) \in \mathrm{int}(Q)$ is a sufficient condition for $h_*(\pi) \in \mathrm{int}(P)$ in Lemma 2 because $Q \subset P$. In other words, we obtain the following theorem that is practically useful since we first try to find PLE, and if that succeeds, we can then proceed to computing CLE.

Theorem 4. *If PLE exists, then CLE exists.*

Further, we obtain the consistency of PLE. For $i = 1, \dots, d$, and for $(s,t) \in \{1, \dots, n\}^2$ with $s \neq t$, let $u_{st}^i = h_*(\pi) - h_*(\pi \circ \tau_{st}^i) \in \mathbb{R}^K$.

Theorem 5. *In addition to Assumptions 1 (2)–(3), we assume that*

- *for any* $v \in \mathbb{R}^K$ *and* $\theta \in \Theta$, $\mathbb{E}\left[\sum_{i=1}^d \{v^\top u_{12}^i\}^2 \{\cosh(\theta^\top u_{12}^i/2)\}^{-2}\right] > 0$, *and*
- *for each* $i = 1, \dots, d$, *there exists* x_i^0 *such that* $\mathbb{E} d_i^2(X_i(1), x_i^0) < \infty$.

Then, PLE converges to θ_0 *in probability.*

The proof is given in Appendix C.5 and employs the theory of U-statistics (in particular, the theory of $M_{m=2}$-estimators); see Bose and Chatterjee, 2018, de la Peña and Giné, 1999). The assumptions in the theorem are quite mild. The first additional assumption ensures the global uniqueness of the expected pseudo likelihood; Hyvärinen (2006) discusses a similar assumption for PLE in Boltzman machines. The second additional assumption ensures the moments of marginals.

Note that in Boltzman machines, PLE can be regarded as the contrastive divergence learning (Hinton, 2002) that is a surrogate of MLE via MCMC; see Hyvärinen (2006). The same argument is applicable to our case, which, together with Theorems 4–5, suggests that PLE is adopted as the initial value of Fisher's scoring method.

Remark 6 (Computation time). Generally, as the computation of CLE includes MCMC iteration in each optimization step, the computation of PLE is faster than that of CLE. This aspect is confirmed in the subsequent simulation studies.

Finally, by applying the theory of $M_{m=2}$-estimators (Theorem 2.3 of Bose and Chatterjee (2018); see also Appendix C.6), we can estimate the asymptotic variance of PLE $\widehat{\theta}_{\mathrm{PLE}}$ by the sandwich estimator

$(4/n)\widehat{J}_{\mathrm{PLE}}^{-1}\widehat{I}_{\mathrm{PLE}}\widehat{J}_{\mathrm{PLE}}^{-1}$, where

$$\widehat{I}_{\mathrm{PLE}} = \frac{1}{n}\sum_{s=1}^{n}\left(\frac{1}{n}\sum_{t\neq s,t=1}^{n}\sum_{i=1}^{d}\frac{u_{st}^i}{1+e^{\widehat{\theta}_{\mathrm{PLE}}^\top u_{st}^i}}\right)\left(\frac{1}{n}\sum_{t\neq s,t=1}^{n}\sum_{i=1}^{d}\frac{(u_{st}^i)^\top}{1+e^{\widehat{\theta}_{\mathrm{PLE}}^\top u_{st}^i}}\right) \text{ and}$$

$$\widehat{J}_{\mathrm{PLE}} = \frac{2}{n(n-1)}\sum_{1\le s<t\le n}\sum_{i=1}^{d}\frac{u_{st}^i(u_{st}^i)^\top}{\{1+e^{\widehat{\theta}_{\mathrm{PLE}}^\top u_{st}^i}\}^2}.$$

4. Simulation studies

In this section, we provide several numerical studies for the inference based on CLE and PLE.

4.1. Gaussian cases

We first examine the performance of CLE developed in Section 3 by applying it to the Gaussian model. We generated a random sample $\{x(t)\}_{t=1}^n$ of size $n = 50$ from the 4-dimensional centered Gaussian distribution with the covariance matrix $\sigma_{ij} = (1/2)^{|i-j|}$. For estimation, we assumed the full Gaussian model, which is a minimum information dependence model (Example 2). The parameter of interest is $\theta = (\theta_k)_{k=1}^6 = (-\sigma^{12}, -\sigma^{13}, -\sigma^{14}, -\sigma^{23}, -\sigma^{24}, -\sigma^{34})$, where σ^{ij} denotes the inverse of σ_{ij}. The true value of θ is set to $\theta = (2/3, 0, 0, 2/3, 0, 2/3)$. The tolerance for solving the conditional likelihood equation was set to 10^{-2} and the MCMC length L was adaptively increased from an initial value $L = 150n = 7500$. We used PLE as an initial value for the Fisher scoring method on CLE. Furthermore, the scoring method is restarted if a component of θ becomes a huge value at some step due to variability of MCMC. These rules are practically effective and adopted in subsequent examples as well.

We repeated the same experiment 1,000 times. Table 3 shows the root mean square errors of CLE and MLE together with those of PLE defined in Section 3.4. CLE has almost the same performance as MLE. PLE is also competitive but slightly worse. Note that the values on θ_5 and θ_6 are close to those on θ_2 and θ_1, respectively, by symmetry of the covariance structure.

We also examined other covariance matrices of the form $\sigma_{ij} = \rho^{|i-j|}$ for $\rho \in \{0, 1/4, 1/2, 3/4\}$ and $\sigma_{ij} = (1-\rho)\delta_{ij} + \rho$ for $\rho \in \{1/4, 1/2, 3/4\}$. The root mean squares of the norm $\|\widehat{\theta} - \theta\|$ are summarized in Table 4, where the number of experiments is 200 in each case.

The mean computational time (resp. standard deviation) for CLE and PLE per each experiment was 6.2 (3.3) and 2.0 (0.2) in seconds, respectively. We find that PLE is faster and (numerically) more stable.

Table 3. Root mean square errors of three estimators. The model is Gaussian. The true parameter value is $\theta = (2/3, 0, 0, 2/3, 0, 2/3)$.

	θ_1	θ_2	θ_3	θ_4	θ_5	θ_6
CLE	0.283	0.253	0.225	0.314	0.251	0.285
MLE	0.285	0.254	0.226	0.313	0.251	0.286
PLE	0.300	0.262	0.235	0.337	0.260	0.306

Table 4. The root mean squares of the norm $\|\widehat{\theta} - \theta\|$ for the three estimators. The model is Gaussian. The covariance structures are $\sigma_{ij} = \rho^{|i-j|}$ for the auto-regressive model and $\sigma_{ij} = (1-\rho)\delta_{ij} + \rho$ for the exchangeable model.

	Auto-regressive model				Exchangeable model		
	$\rho = 0$	$\rho = 1/4$	$\rho = 1/2$	$\rho = 3/4$	$\rho = 1/4$	$\rho = 1/2$	$\rho = 3/4$
CLE	0.415	0.507	0.647	1.410	0.477	0.688	1.392
MLE	0.416	0.509	0.653	1.422	0.479	0.693	1.398
PLE	0.433	0.529	0.694	1.494	0.507	0.734	1.516

4.2. Three-dimensional interaction

We next consider a three-dimensional interaction model

$$h(x_1, x_2, x_3) = (x_1x_2, x_1x_3, x_2x_3, x_1x_2x_3)^\top$$

on $\mathcal{X} = \prod_{i=1}^3 \mathcal{X}_i = \mathbb{R}^3$. The true parameter value of θ is set to $\theta = (1, 0, 0, -1)$ and the true marginal densities are set to the standard normal distribution. For the simulation, we first generate a "population" of size $N = 10^3$ by the sampling algorithm in Remark 5 with the number of iterations $L = 150N$. Then a sample of size $n = 100$ is randomly selected from the population without replacement. The tolerance for CLE and the step length of MCMC are set to the same values as the Gaussian case. We repeated the same experiment 1,000 times.

Table 5 shows the root mean square errors of CLE and PLE together with the coverage probability of the 95% confidence intervals constructed from CLE. As expected, CLE has smaller error than PLE. The confidence intervals are almost exact or slightly conservative.

We also examined other parameter values $\theta = (a, 0, 0, -a)$ for $a \in \{0, 1, 2\}$. The root mean squares of the norm $\|\widehat{\theta} - \theta\|$ are summarized in Table 6.

4.3. Mixed variables

Finally, we study a case with continuous and discrete variables. As in Example 5, the marginal distributions are set to Beta(10, 10) and Poisson(3), respectively. The canonical statistic is $h(x_1, x_2) = x_1/(1 + x_2)$ and the true parameter values are set to $\theta \in \{0, 10, 100\}$. The sampling method is the same as the preceding subsection except that the sample size is $n = 50$. The tolerance for solving CLE was set to 10^{-5}.

Table 7 shows the root mean square errors, biases and standard deviations of CLE and PLE. CLE has better performance than PLE.

Table 5. Root mean square error of two estimators and coverage probabilities of the 95% confidence intervals based on CLE. The model is the three-dimensional interaction model and the true parameter value is $\theta = (1, 0, 0, -1)$.

	θ_1	θ_2	θ_3	θ_4
error of CLE	0.214	0.159	0.153	0.168
error of PLE	0.260	0.190	0.189	0.250
coverage	0.953	0.952	0.965	0.968

Table 6. The root mean squares of the norm $\|\widehat{\theta} - \theta\|$ for CLE and PLE. The model is the three-dimensional interaction model. The true parameter values are $\theta = (a, 0, 0, -a)$, where $a \in \{0, 1, 2\}$.

	$a = 0$	$a = 1$	$a = 2$
CLE	0.211	0.353	0.556
PLE	0.230	0.459	0.738

Table 7. Root mean square errors, biases and standard deviations of the two estimators. The model is $h(x_1, x_2) = x_1/(1 + x_2)$ with the beta and Poisson marginals. The true parameter value is set to $\theta \in \{0, 10, 100\}$.

	$\theta = 0$			$\theta = 10$			$\theta = 100$		
	RMSE	bias	SD	RMSE	bias	SD	RMSE	bias	SD
CLE	8.19	0.08	8.18	9.40	1.14	9.33	29.26	4.96	28.84
PLE	9.03	0.12	9.03	10.34	1.89	10.17	32.34	7.31	31.51

5. Future directions

We here address potential future directions of our work.

The interpretability of the canonical parameter θ have to be more clarified. In Section 2, we have clarified the connection between the total correlation and θ. In the application to Earthquake data in Appendix D.2, we have demonstrated that the estimation of θ not only provided the existence of the dependence but also identified the pattern of the dependence. For clearer interpretation of θ, the connection to the partial correlation has to be more investigated.

The asymptotic properties of CLE have to be clarified. The asymptotic normality and efficiency for contingency tables are described in Theorem 3. We expect that the same properties are valid even for infinite sample spaces. The limiting behavior of the approximate sampling algorithm in Remark 5 as $n \to \infty$ is also under investigation.

In Theorem 4, we found that CLE of θ exists under suitable conditions. However, if the dimension of θ is high, the estimator may not exist and some regularization will be necessary. A simple method of regularization is to assume a (conditional) conjugate prior density

$$p(\theta|M) = \exp(\lambda_0 \mu_0^\top \theta - \lambda_0 \Psi(\theta))$$

for the exponential family (9), where $\Psi(\theta)$ is the potential function of the conditional likelihood in (9), and $\mu_0 \in \mathbb{R}^K$ and $\lambda_0 > 0$ are hyper-parameters. The maximum a posteriori estimator exists if μ_0 is an interior point of P defined in Lemma 2. In practice, we can select μ_0 as the sample mean of randomly generated vectors in $h_*(\mathbb{S}_n^d)$. The hyper-parameter λ_0 may be set to $1/n$ as a rule of thumb. Investigating the PLE with regularization would be another important issue.

In our estimation procedure, we need to perform MCMC sampling as stated in Subsection 3.3. The proposed algorithm was to exchange two elements of permutations. This is just one particular move. A mover producing a shorter mixing time should be found. Parallel algorithms will also be valuable.

We focused on estimation of the canonical parameter θ representing the dependence. However, estimation of the marginal parameters ν is also important in some cases. For example, if our goal is to predict future observations, then an estimator of the marginal parameters is necessary. A simple procedure to estimate ν is to perform the maximum likelihood estimation assuming the independent model $\prod_{j=1}^d r_j(x_j; \nu)$. The obtained estimator $\widehat{\nu}$ is consistent by Theorem 2. However, $\widehat{\nu}$ is not asymptotically efficient in general, as observed in the bivariate Gaussian model with a common variance.

In our analysis, we assumed that the data were completely observed. In practice, handling data missing is a common challenge. To deal with missing data, we can extend the domain $\mathcal{X}_j$ to $\mathcal{X}_j \cup \{\mathsf{NA}\}$ for each $j = 1, \ldots, d$, where NA indicates missing. Then, the proposed method is applicable whenever we specify the canonical statistic $h(x)$ including the missing indicator. For example, $h(x_1, x_2) = x_1 I_{\{x_2 = \mathsf{NA}\}}$ represents a missing effect of the second variable x_2 on the first variable x_1. A missing data analysis is important future work.

Finally, there are two mathematical open problems. First, the properties of the potential function $\psi(\theta)$ in the model (2) are unknown except for convexity. We conjecture that $\psi(\theta)$ is analytic at every interior point, by analogy with the standard theory of exponential families. However, $\psi(\theta)$ is defined by functional equations, which makes the problem complicated. The second open problem is the equivalence of the three feasibility conditions discussed in Subsection 2.3. The problem is related to the closedness of sum spaces of L_1-functions; see Rüschendorf and Thomsen (1993).

Acknowledgements

We thank the AE and three referees for their constructive comments that have improved the quality of this paper. We thank Yici Chen, Hironori Fujisawa, Masayuki Kano, Hisahiko Kubo, Michiko Okudo, and Akifumi Okuno for helpful discussions. We used the earthquake catalog provided by the Japan Meteorological Agency (Japan Meteorological Agency (2022)) and used GMT software package (Wessel and Smith (1998)) to create the maps. This work is supported by JST CREST (JPMJCR1763), JSPS KAKENHI (19K20222, 21H05205, 21K11781, 21K12067), and MEXT (JPJ010217). The R and Python codes are available at https://doi.org/10.5281/zenodo.8012980.

Supplementary Material

Supplement to "Minimum information dependence modeling" (DOI: 10.3150/23-BEJ1687SUPP; .pdf). This provides more examples of our model, useful properties of the model, all proofs of the results, and application to real data.

References

Japan Meteorological Agency (2022). The seismological bulletin of Japan. Available at https://www.data.jma.go.jp/svd/eqev/data/bulletin/index_e.html.

Agrawal, R. (2020). Finite-sample concentration of the multinomial in relative entropy. *IEEE Trans. Inf. Theory* **66** 6297–6302. MR4173539 https://doi.org/10.1109/TIT.2020.2996134

Albert, A. and Anderson, J.A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* **71** 1–10. MR0738319 https://doi.org/10.1093/biomet/71.1.1

Amari, S. (2001). Information geometry on hierarchy of probability distributions. *IEEE Trans. Inf. Theory* **47** 1701–1711. MR1842511 https://doi.org/10.1109/18.930911

Amari, S. and Nagaoka, H. (2000). *Methods of Information Geometry*. *Translations of Mathematical Monographs* **191**. Providence, RI: Amer. Math. Soc.. MR1800071 https://doi.org/10.1090/mmono/191

Bedford, T., Daneshkhah, A. and Wilson, K.J. (2016). Approximate uncertainty modeling in risk analysis with vine copulas. *Risk Anal.* **36** 792–815.

Bedford, T. and Wilson, K.J. (2014). On the construction of minimum information bivariate copula families. *Ann. Inst. Statist. Math.* **66** 703–723. MR3224606 https://doi.org/10.1007/s10463-013-0422-0

Borwein, J.M., Lewis, A.S. and Nussbaum, R.D. (1994). Entropy minimization, *DAD* problems, and doubly stochastic kernels. *J. Funct. Anal.* **123** 264–307. MR1283029 https://doi.org/10.1006/jfan.1994.1089

Bose, A. and Chatterjee, S. (2018). *U-Statistics, M_m-Estimators and Resampling*. *Texts and Readings in Mathematics* **75**. Singapore: Springer. MR3837543 https://doi.org/10.1007/978-981-13-2248-8

Buja, A., Hastie, T. and Tibshirani, R. (1989). Linear smoothers and additive models. *Ann. Statist.* **17** 453–555. MR0994249 https://doi.org/10.1214/aos/1176347115

Chen, Y. and Sei, T. (2022). A proper scoring rule for minimum information copulas. Available at arXiv:2204.03118.

Choi, L., Blume, J.D. and Dupont, W.D. (2015). Elucidating the foundations of statistical inference with 2 x 2 tables. *PLoS ONE* **10** e0121263. https://doi.org/10.1371/journal.pone.0121263

Cox, D.R. and Hinkley, D.V. (1974). *Theoretical Statistics*. London: CRC Press. MR0370837

Csiszár, I. (1975). *I*-divergence geometry of probability distributions and minimization problems. *Ann. Probab.* **3** 146–158. MR0365798 https://doi.org/10.1214/aop/1176996454

Dempster, A.P. (1972). Covariance selection. *Biometrics* **28** 157–175. MR3931974

de la Peña, V.H. and Giné, E. (1999). *Decoupling: From Dependence to Independence*. *Probability and Its Applications (New York)*. New York: Springer. MR1666908 https://doi.org/10.1007/978-1-4612-0537-1

Diaconis, P. and Sturmfels, B. (1998). Algebraic algorithms for sampling from conditional distributions. *Ann. Statist.* **26** 363–397. MR1608156 https://doi.org/10.1214/aos/1030563990

Eckstein, S. and Nutz, M. (2022). Quantitative stability of regularized optimal transport and convergence of Sinkhorn's algorithm. *SIAM J. Math. Anal.* **54** 5922–5948. MR4506579 https://doi.org/10.1137/21M145505X

Falconer, K. (2014). *Fractal Geometry: Mathematical Foundations and Applications*, 3rd ed. Chichester: Wiley. MR3236784

Geenens, G. (2020). Copula modeling for discrete random vectors. *Depend. Model.* **8** 417–440. MR4193472 https://doi.org/10.1515/demo-2020-0022

Genest, C. and Nešlehová, J. (2007). A primer on copulas for count data. *Astin Bull.* **37** 475–515. MR2422797 https://doi.org/10.2143/AST.37.2.2024077

Geyer, C.J. and Thompson, E.A. (1992). Constrained Monte Carlo maximum likelihood for dependent data. *J. Roy. Statist. Soc. Ser. B* **54** 657–699. MR1185217

Haasler, I., Singh, R., Zhang, Q., Karlsson, J. and Chen, Y. (2021). Multi-marginal optimal transport and probabilistic graphical models. *IEEE Trans. Inf. Theory* **67** 4647–4668. MR4306289 https://doi.org/10.1109/tit.2021.3077465

Haberman, S.J. (1977). *The Analysis of Frequency Data: Statistical Research Monographs*. Chicago: Univ. Chicago Press.

Hannan, J. and Harkness, W. (1963). Normal approximation to the distribution of two independent binomials, conditional on fixed sum. *Ann. Math. Stat.* **34** 1593–1595. MR0160281 https://doi.org/10.1214/aoms/1177703893

Harkness, W.L. (1965). Properties of the extended hypergeometric distribution. *Ann. Math. Stat.* **36** 938–945. MR0182073 https://doi.org/10.1214/aoms/1177700066

Hinton, G.E. (2002). Training products of experts by minimizing contrastive divergence. *Neural Comput.* **14** 1771–1800. https://doi.org/10.1162/089976602760128018

Holland, P.W. and Wang, Y.J. (1987). Dependence function for continuous bivariate densities. *Comm. Statist. Theory Methods* **16** 863–876. MR0886560 https://doi.org/10.1080/03610928708829408

Hyvärinen, A. (2006). Consistency of pseudolikelihood estimation of fully visible Boltzmann machines. *Neural Comput.* **18** 2283–2292. MR2256106 https://doi.org/10.1162/neco.2006.18.10.2283

Jansen, M.J.W. (1997). Maximum entropy distributions with prescribed marginals and normal score correlations. In *Distributions with Given Marginals and Moment Problems* (V. Beneš and J. Štěpán, eds.). 87–92. Springer.

Jaynes, E.T. (1957). Information theory and statistical mechanics. *Phys. Rev. (2)* **106** 620–630. MR0087305

Jones, M.C., Pewsey, A. and Kato, S. (2015). On a class of circulas: Copulas for circular distributions. *Ann. Inst. Statist. Math.* **67** 843–862. MR3390169 https://doi.org/10.1007/s10463-014-0493-6

Kou, S.G. and Ying, Z. (1996). Asymptotics for a 2×2 table with fixed margins. *Statist. Sinica* **6** 809–829. MR1422405

Kurowicka, D. and van Horssen, W.T. (2015). On an interaction function for copulas. *J. Multivariate Anal.* **138** 127–142. MR3348837 https://doi.org/10.1016/j.jmva.2014.12.012

Lauritzen, S.L. (1996). *Graphical Models*. *Oxford Statistical Science Series* **17**. New York: Oxford University Press. MR1419991

Léonard, C. (2012). From the Schrödinger problem to the Monge-Kantorovich problem. *J. Funct. Anal.* **262** 1879–1920. MR2873864 https://doi.org/10.1016/j.jfa.2011.11.026

Little, R.J.A. (1989). Testing the equality of two independent binomial proportions. *Amer. Statist.* **43** 283–288.

Meeuwissen, A.M.H. and Bedford, T. (1997). Minimally informative distributions with given rank correlation for use in uncertainty analysis. *J. Stat. Comput. Simul.* **57** 143–174. MR1463104 https://doi.org/10.1080/00949659708811806

Mukherjee, S. (2016). Estimation in exponential families on permutations. *Ann. Statist.* **44** 853–875. MR3476619 https://doi.org/10.1214/15-AOS1389

Peyré, G. and Cuturi, M. (2019). Computational optimal transport: With applications to data science. *Found. Trends Mach. Learn.* **11** 355–607.

Piantadosi, J., Howlett, P. and Borwein, J. (2012). Copulas with maximum entropy. *Optim. Lett.* **6** 99–125. MR2886586 https://doi.org/10.1007/s11590-010-0254-2

Reid, N. (1995). The roles of conditioning in inference. *Statist. Sci.* **10** 138–157. MR1368097

Rinaldo, A., Fienberg, S.E. and Zhou, Y. (2009). On the geometry of discrete exponential families with application to exponential random graph models. *Electron. J. Stat.* **3** 446–484. MR2507456 https://doi.org/10.1214/08-EJS350

Robbins, H. (1955). A remark on Stirling's formula. *Amer. Math. Monthly* **62** 26–29. MR0069328 https://doi.org/10.2307/2308012

Rüschendorf, L. and Thomsen, W. (1993). Note on the Schrödinger equation and I-projections. *Statist. Probab. Lett.* **17** 369–375. MR1237783 https://doi.org/10.1016/0167-7152(93)90257-J

Sei, T. and Yano, K. (2024). Supplement to "Minimum information dependence modeling." https://doi.org/10.3150/23-BEJ1687SUPP

Sinkhorn, R. and Knopp, P. (1967). Concerning nonnegative matrices and doubly stochastic matrices. *Pacific J. Math.* **21** 343–348. MR0210731

van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes*. *Springer Series in Statistics*. New York: Springer. MR1385671 https://doi.org/10.1007/978-1-4757-2545-2

Watanabe, S. (1960). Information theoretical analysis of multivariate correlation. *IBM J. Res. Develop.* **4** 66–82. MR0109755 https://doi.org/10.1147/rd.41.0066

Weissman, T., Ordentlich, E., Seroussi, G., Verdu, S. and Weinberger, M. (2003). Inequalities for the L_1 deviation of the empirical distribution. Technical Report HPL-2003-97R1, Hewlett-Packard Labs.

Wessel, P. and Smith, W. (1998). New, improved version of generic mapping tools released. *Eos Trans. AGU* **79** 409–410.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. *Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics*. Chichester: Wiley. MR1112133

Yang, E., Ravikumar, P., Allen, G.I. and Liu, Z. (2015). Graphical models via univariate exponential family distributions. *J. Mach. Learn. Res.* **16** 3813–3847. MR3450553

Received June 2023 and revised October 2023