Minimum information dependence modeling for mixed domain data

Keisuke Yano (the Institute of Statistical Mathematics),
Tomonari Sei (the University of Tokyo)

Abstract

In this talk, we introduce a joint statistical model for mixed-domain data that is proposed by [1]. The proposed model contains multivariate Gaussian and log-linear models. We show the existence and uniqueness of the proposed model under fairly weak conditions. To estimate the dependence parameter in our model, we present a conditional inference together with a sampling procedure and show it provides a consistent estimator of the dependence parameter.

1 Minimum information dependence model

Let $(\mathcal{X}_i, \mathcal{F}(\mathcal{X}_i), \mathrm{d}x_i)$ for $i = 1, \ldots, d$ be measure spaces 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 $r_1(x_1; \nu), \ldots, r_d(x_d; \nu)$ be statistical models of marginal densities on $\mathcal{X}_1, \ldots, \mathcal{X}_d$, 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 $\nu = (\nu_1, \ldots, \nu_d)$.

We consider a class of probability density functions

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

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

$$\int p(x; \theta, \nu) \mathrm{d}x_{-i} = r_i(x_i; \nu), \quad i = 1, \ldots, d, \text{ and} \tag{2}$$

$$\int \sum_{i=1}^d a_i(x_i; \theta, \nu) p(x; \theta, \nu) \mathrm{d}x = 0. \tag{3}$$

Note that the density (1) is reduced to the independent model $\prod_{i=1}^d r_i(x_i; \nu)$ if $\theta = 0$.

Definition 1. A statistical model (1) together with the constraints (2) and (3) 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 *normalizing functions* and $\psi(\theta, \nu)$ is the *potential function*.

Figure 1 displays a two-dimensional histogram of samples from the minimum information dependence model for mixed variables (discrete and $[0,1]$) with negative correlation, which shows that the minimum information dependence model easily expresses the dependence between mixed-domain variables.

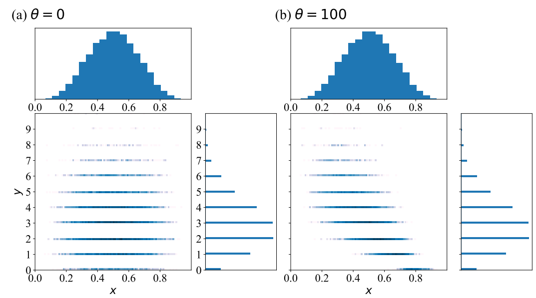

Figure 1: Two-dimensional histograms of 10000 samples from the minimum information dependence model with the Beta Beta(10, 10) and Poisson Po(3) marginals. The canonical statistic $h(x, y)$ is given by $h(x, y) = x/(y + 1)$. The joint histogram and marginal histograms are plotted. (a) Joint histogram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

Let $p_0(x) := \prod_{i=1}^{d} r_i(x_i; \nu)$. We say that a function $H \in L_1(p_0(x)\mathrm{d}x)$ is *feasible* if there exist measurable functions $\{a_i(x_i) : i = 1, \ldots, d\}$ 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

$$\int p(x)\mathrm{d}x_{-i} = r_i(x_i; \nu) \qquad \text{for each } i = 1, \ldots, d \quad \text{and}$$
$$\int \sum_{i=1}^{d} a_i(x_i) p(x)\mathrm{d}x = 0.$$

Theorem 1 (Theorem 1 of [1])**.** A function $H \in L_1(p_0(x)\mathrm{d}x)$ is feasible if there exist $\{b_i \in L_1(r_i(x_i)\mathrm{d}x_i) : i = 1, \ldots, d\}$ such that

$$\int e^{H(x)-\sum_{i=1}^{d} b_i(x_i)} p_0(x)\mathrm{d}x < \infty. \tag{4}$$

Furthermore, if H is feasible, then $\sum_{i=1}^{d} a_i(x_i)$ and ψ are unique.

2 Conditional inference

To analyse the dependence, we need to estimate the dependence parameter θ. However, the maximum likelihood estimate requires the values of a_is and ψ, and a_is cannot be written in a closed form except for limited cases. So, we propose a conditional maximum likelihood estimator of θ that does not require the values of a_is and ψ.

Suppose that $x(1), \ldots, x(n)$ are independent and identically distributed (i.i.d.) from the minimum information dependence model (1). 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 $\leq_i$ on $\mathcal{X}_i$

for each $i = 1, \ldots, d$. Using the ordering is convenient for the following description and the choice does not affect the inference. We denote the symmetric group of degree n by $\mathbb{S}_n$.

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

$$M_i(1) \leq_i \cdots \leq_i M_i(n), \tag{5}$$

where for each i, $M_i = (M_i(1), \ldots, M_i(n))$ are the n observations $(x_i(t))_{t=1}^n$ sorted by the predetermined order $\leq_i$. 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 $x_i(t) = M_i(\pi_i(t))$. 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 $M = (M_1, \ldots, M_d) \in \prod_{i=1}^d \mathcal{X}_i^n$ and $\pi = (\pi_1, \ldots, \pi_d) \in \mathbb{S}_n^d$. For each $t = 1, \ldots, n$, the t-th observation $x(t)$ is recovered from M and π, and is written as

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

Using the marginal order statistic and the rank statistic, we have the following likelihood decomposition.

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

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

By denoting $h_*(\pi) = \sum_{t=1}^n h((M \circ \pi)(t)) \in \mathbb{R}^K$, the conditional likelihood is then expressed as

$$f(\pi | M; \theta) = \frac{e^{\theta^\top h_*(\pi)}}{\sum_{\widetilde{\pi} \in \mathbb{S}_n^d} e^{\theta^\top h_*(\widetilde{\pi})}}, \tag{6}$$

where the conditional likelihood does not involve $a_i(x_i; \theta, \nu)$s and $\psi(\theta, \nu)$.

Definition 2. The conditional maximum likelihood estimate $\widehat{\theta}$ is a maximizer of the conditional likelihood (6).

We obtain the following consistency result of the conditional maximum likelihood estimator $\widehat{\theta}$.

Theorem 2 (Corollary 4 of [1])**.** Under Assumptions 1 and 2 in [1], we have $\widehat{\theta} \to \theta_0$ in probability.

References

[1] T. Sei and K. Yano. Minimum information dependence modeling, 2022. arXiv:2206.06792.