修士論文 要旨

最小情報コピュラのスコアマッチング

数理情報学専攻 48206222　陳　懿慈
指導教員　清　智也 教授

1　はじめに

周辺分布が一様分布である多次元分布はコピュラと呼ばれる。その中で、期待値の制約を満たしかつ独立分布と Kullback-Leibler(KL)-ダイバージェンスの意味で一番近いものを最小情報コピュラという。最小情報コピュラは金融モデル [2] や洪水への応用 [3] で使われる。最小情報コピュラは存在すれば、以下のような指数型分布族と似た形で書けることが知らている。

定理 1. (Bedford and Wilson (2014)[1])
最小情報コピュラは存在するのであれば、$\theta_i \in \mathbb{R}$ と可測関数 $a(x)$ と $b(y)$ によって、以下の形で表される：

$$c(x,y) = \exp\left(\sum_i \theta_i h_i(x,y) + a(x) + b(y)\right)$$

また、最小情報コピュラは、可測関数 $a(x), b(y) \in \mathbb{R}$ の定数差の任意性を除いて、一意である。

$a(x), b(y)$ は、関数 $h_i(x,y)$ と θ_i が定まれば、定数差を除いて一意に定まることから、規格化関数という。また、規格化関数を求めるのは一般的には難しい。

推定を行う際、スコア関数というモデルの良し悪しを測る関数がよく使われる。一番良く知られいているスコアの一つに KL-ダイバージェンスに対応する KL-スコアがある。スコアには様々な性質を持つ物が含まれており、使用目的に沿った性質を持つスコアを使うことで、推定の精度を上げたり、計算量を抑えたりすることができる。例えば、データにノイズが乗っている場合に、ノイズに対してロバストな推定を行えるスコアが提案されている [5]。また、モデルの規格化定数を求めるのが難しい場合、規格化定数を計算せずともスコアが求まる Hyvärinen スコア [4] が提案されている。

本論文は、最小情報コピュラの規格化関数を計算せずに推定を行えるスコアを提案する。このスコアは、漸近一致性を持つので、データを十分取れば誤差は真の値にいくらでも近づく。さらに、提案スコアはパラメータに対して凸であり、勾配法で最適化が簡単に行える。

2　既存研究

一般の密度関数の推定に関して、規格化定数の計算が難しい問題を解決するために、定数倍しても値が変わらない (このような性質を 0-斉次性という)Hyvärinen スコアが提案されている。0-斉次性を持つスコアは Hyvärinen スコア以外にも多く存在し、簡単に作ることができる [6]。しかし、多変量分布 $q(x_1, \cdots, x_n)$ に対して、定数だけでなく任意の関数 $f(x_i)$ を掛けても値が変わらない (このような性質を拡張 0-斉次性という) スコアは提案されていない。拡張 0-斉次性を持つスコアが作れると、規格化関数が求まらない最小情報コピュラの推定も簡単に行えることになる。

3　本研究で提案するアルゴリズム

以上の背景を踏まえ，以下のスコアを提案した。

定理 2. 密度関数 $q(x,y)$ に対して、

$$q^{ij} := q(x^i, y^j) \qquad x^i \in \mathcal{X},\ y^j \in \mathcal{Y},\ i,j \in \mathbb{N}$$

とする。任意の密度関数 $q(x,y)$ に対して、

$$S(x^1, y^1, x^2, y^2, q) = -\log\left(\frac{q^{11}q^{22}}{q^{11}q^{22} + q^{12}q^{21}}\right)$$

を独立な二点 $(x^1, y^1), (x^2, y^2)$ に対するスコアとすると、このスコアは拡張 0-斉次である。

さらに、このスコアは漸近一致性を持つ。

定理 3. 真のパラメータを θ_0 とし、最小情報コピュラの独立同分布のデータ点を

$$x_1^1, y_1^1, x_1^2, y_1^2, \cdots, x_n^1, y_n^1, x_n^2, y_n^2$$

とする。提案スコア S に対して、

$$\arg\min_\theta -\frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{q_i^{11}q_i^{22}}{q_i^{11}q_i^{22} + q_i^{12}q_i^{21}}\right) \xrightarrow[\text{a.s.}]{n\to\infty} \theta_0$$
$$q_k^{ij} := q(x_k^i, y_k^j)$$

が成り立つ。

よって、経験分布を代入したスコア $\hat{S}$ に対して、$\hat{\theta} := \arg\min_\theta \hat{S}$ が推定値になる。この際、$\hat{S}$ は θ に関して凸なので、最適化は勾配法で簡単に行える。

4　数値実験

まず、簡単な例として、正規コピュラの実験結果を述べる。正規コピュラは最小情報コピュラの中で規格化関数が計算できる数少ない例の一つである。よって、正規コピュラに関しては、最尤推定が可能である。最尤推定と提案手法を比較した結果は図 1 のとおりである。

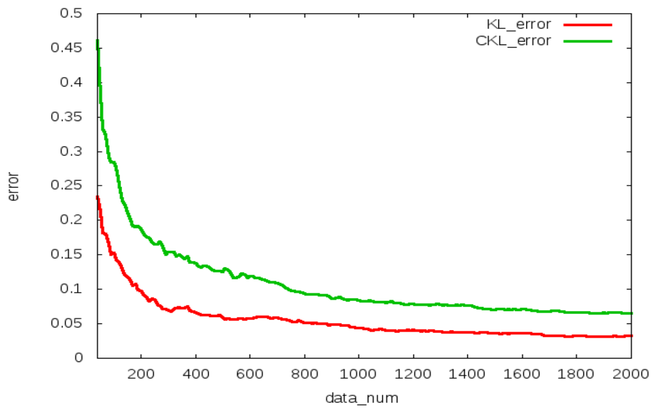

図 1. 正規コピュラのパラメータ推定

両軸対数グラフにプロットし直し、線形フィッティングしたのが図 2 である。フィッティング直線 (青線、紫線) の傾きは概ね -0.5 である。

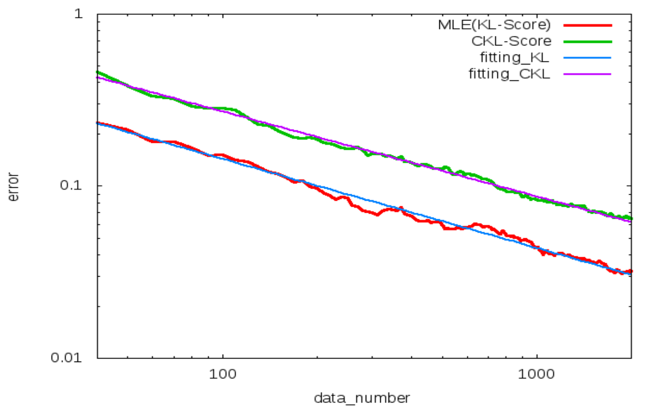

図 2. 正規コピュラのパラメータ推定 (対数グラフ)

一般の最小情報コピュラに対してもパラメータ推定を行った。一般の最小情報コピュラは規格化関数が求まらない為、サンプリングが難しい。本論文では [7] で提案された近似サンプリングでデータを生成した。$\theta = 5.0, 10.0$ と $h(x,y) = xy, x^2y$ の 4 パターンで実験した。推定結果を両軸対数グラフにプロットしたのが図 3 である。直線でフィッチングした場合の傾きは全部概ね -0.5 である。

以上の実験より、提案スコアが最小情報コピュラに対して、拡張 0-斉次性を持ち、推定が簡単に行えることが

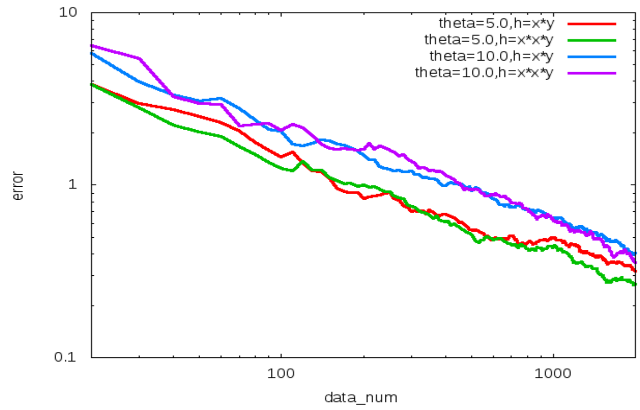

図 3. 一般の最小情報コピュラのパラメータ推定 (対数グラフ)

確認できた。このスコアは、最小情報コピュラにおいて計算し難い規格化関数を求めずとも推定ができるという点が最大の強みである。また、誤差の収束が中心極限定理通り $\frac{1}{\sqrt{N}}$ で 0 に収束していることも、対数グラフの線形フィッティングの傾きが概ね -0.5 であることから確認できる。

参考文献

[1] Tim Bedford and Kevin Wilson. On the construction of minimum information bivariate copula families. *Annals of the Institute of Statistical Mathematics*, 66, 08 2014.

[2] Omid Chatrabgoun, Amin Hosseinian-Far, Victor Chang, Nigel G. Stocks, and Alireza Daneshkhah. Approximating non-Gaussian Bayesian networks using minimum information vine model with applications in financial modelling. *Journal of Computational Science*, 24:266–276, 2018.

[3] Alireza Daneshkhah, Renji Remesan, Omid Chatrabgoun, and Ian P. Holman. Probabilistic modeling of flood characterizations with parametric and minimum information pair-copula model. *Journal of Hydrology*, 540:469–487, 2016.

[4] Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6(Apr):695–709, 2005.

[5] Takafumi Kanamori and Hironori Fujisawa. Robust estimation under heavy contamination using unnormalized models. *Biometrika*, 102(3):559–572, 05 2015.

[6] Matthew Parry, A. Philip Dawid, and Steffen Lauritzen. Proper local scoring rules. *Ann. Statist.*, 40(1):561–592, 02 2012.

[7] Tomonari Sei and Keisuke Yano. Minimum information dependence modeling for arbitrary product spaces. unpublished manuscript, 2022.