2024 年 5 月 24 日 統計数理研究所 オープンハウス

コピュラに基づくダイバージェンスを用いた適合度検定

加藤 昇吾　　統計基盤数理研究系 准教授

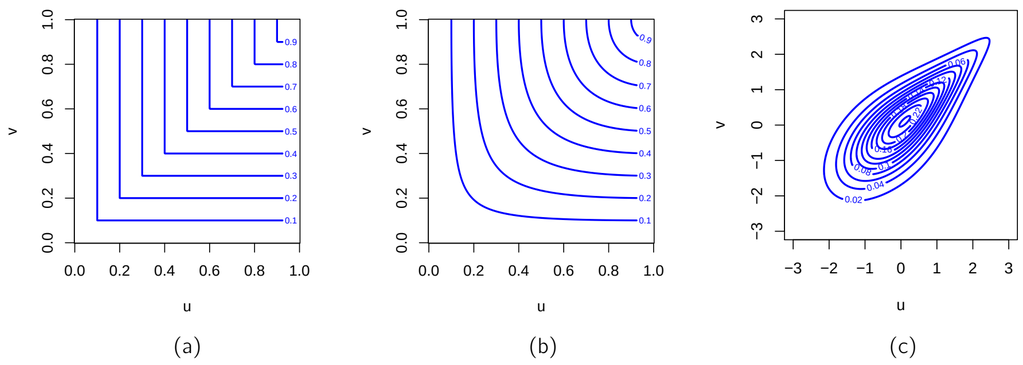
はじめに

コピュラ

d次元コピュラとは，それぞれの1次元周辺分布が$[0,1]$上の一様分布となるd次元分布関数のことをいう．

d次元コピュラを変換することにより，任意の1次元周辺分布関数を持つd次元分布関数を得ることが可能（スクラーの定理）．

$\Longrightarrow$　コピュラにより，柔軟なモデリングが可能となる．



(a)　(b)　(c)

図1．(a),(b): 2次元コピュラの等高線プロットの例，
(c): 標準正規分布を周辺分布に持つコピュラの確率密度関数の等高線プロット．

研究の目的

コピュラのための適合度検定としては，Cramér–von Mises距離に基づく検定（Cramér–von Mises検定）がしばしば用いられる．

$\Longrightarrow$　この検定は裾部分での感度が低い問題が知られている．

本研究では，裾部分での重みを調節可能な新たなダイバージェンスを定義し，それに基づく適合度検定を提案する．

なお本研究は，江口真透名誉教授（統計数理研究所）・吉羽要直教授（東京都立大学）との共同研究である．

コピュラに基づくダイバージェンスと適合度検定への応用

コピュラに基づくダイバージェンス

確率密度関数に基づくβダイバージェンス(Basu *et al.*, 1998; Minami & Eguchi, 2002)から着想を得て，以下のダイバージェンスを定義する．

定義　C_0, C_1: d次元コピュラ とする．
このとき，C_0とC_1に対して，新しいダイバージェンスを以下で定義する：

$$D_\beta(C_0, C_1) = \int_{\mathcal{H}} \left\{ \frac{C_0(\boldsymbol{u})^{\beta+1}}{\beta(\beta+1)} + \frac{C_1(\boldsymbol{u})^{\beta+1}}{\beta+1} - \frac{C_0(\boldsymbol{u})C_1(\boldsymbol{u})^{\beta}}{\beta} \right\} \mathrm{d}W(\boldsymbol{u}), \tag{1}$$

ただし，W: $\mathcal{H}$を台（support）とするd次元コピュラ測度，
$\beta \in \mathbb{R} \setminus \{-1, 0\}$: チューニングパラメータ である．

また，$\beta = 0, -1$のときは，以下のように定義する．

(i) $\beta = 0$ (累積Kullback–Leiblerダイバージェンス)：

$$D_0(C_0, C_1) = \lim_{\beta \to 0} D_\beta(C_0, C_1) = -\int_{\mathcal{H}} \left\{ C_0(\boldsymbol{u}) \log\left(\frac{C_1(\boldsymbol{u})}{C_0(\boldsymbol{u})} \right) + C_0(\boldsymbol{u}) - C_1(\boldsymbol{u}) \right\} \mathrm{d}W(\boldsymbol{u}),$$

(ii) $\beta = -1$ (累積Itakura–Saitoダイバージェンス)：

$$D_{-1}(C_0, C_1) = \lim_{\beta \to -1} D_\beta(C_0, C_1) = \int_{\mathcal{H}} \left\{ -\log\left(\frac{C_1(\boldsymbol{u})}{C_0(\boldsymbol{u})} \right) + \frac{C_1(\boldsymbol{u})}{C_0(\boldsymbol{u})} - 1 \right\} \mathrm{d}W(\boldsymbol{u}).$$

特別な場合　$\beta = 1$のとき，提案したダイバージェンス(1)は，Crámer–von Mises距離と等しくなる：

$$D_1(C_0, C_1) = \frac{1}{2} \int_{\mathcal{H}} \{C_0(\boldsymbol{u}) - C_1(\boldsymbol{u})\}^2 \mathrm{d}W(\boldsymbol{u}).$$

ダイバージェンスの性質　ダイバージェンス(1)は，以下を満たす：

(i) 任意の$C_0, C_1 \in \mathcal{C}$に対して，$D_\beta(C_0, C_1) \geq 0$,

(ii) $D_\beta(C_0, C_1) = 0 \iff C_0(\boldsymbol{u}) = C_1(\boldsymbol{u})$ for all $\boldsymbol{u} \in \mathcal{H}$.

ただし，$\mathcal{C}$は全てのd次元コピュラの集合を表す．

適合度検定への応用

定義　$\boldsymbol{U}_1, \ldots, \boldsymbol{U}_n \sim i.i.d.$ コピュラG,　C_0: コピュラ, とする．

以下の仮説検定を考える：

$$H_0 : G = C_0 \quad \text{vs.} \quad H_1 : G \neq C_0.$$

この仮説検定のための検定統計量を以下で定義する：

$$T = 2nD_\beta(\hat{C}, C_0).$$

ここで，$\hat{C}$は$(\boldsymbol{U}_1, \ldots, \boldsymbol{U}_n)$の経験分布関数を表す．$T$が大きいときに$H_0$を棄却し，小さいときに$H_0$を採択する．

検定統計量Tの性質

- $\beta = 1$　$\Longrightarrow$　Crámer–von Mises検定量．
- 累積Kullback–Leiblerダイバージェンス($\beta = 0$)や累積Itakura–Saitoダイバージェンス($\beta = -1$)に基づく検定量も特別な場合として含む．

検定統計量Tの漸近分布　以下を仮定する：

$W(\boldsymbol{u})$: $\{\boldsymbol{v}_1, \ldots, \boldsymbol{v}_k\} \subset \mathcal{H}$を台とする分布関数，つまり，

$$W(\boldsymbol{u}) = \frac{1}{k} \sum_{j=1}^{k} \mathcal{I}(\boldsymbol{u} \leq \boldsymbol{v}_j).$$

ただし，$\mathcal{I}$は定義関数を表す．このとき，帰無仮説H_0の下で，以下が検定統計量Tに対して成り立つ：

$$T \xrightarrow{d} \frac{1}{k} \sum_{j=1}^{k} \lambda_j \chi_j^2 \quad (n \to \infty).$$

ここに，$\chi_1^2, \ldots, \chi_k^2 \sim i.i.d.$ 自由度1のχ^2分布,
$\lambda_1, \ldots, \lambda_k$: $\Sigma_T = (\sigma_T(\boldsymbol{v}_i, \boldsymbol{v}_j))_{1 \leq i,j \leq k}$の固有値,
$\sigma_T(\boldsymbol{v}_i, \boldsymbol{v}_j) = \{C_0(\boldsymbol{v}_i)C_0(\boldsymbol{v}_j)\}^{(\beta-1)/2}\{C_0(\boldsymbol{v}_i \wedge \boldsymbol{v}_j) - C_0(\boldsymbol{v}_i)C_0(\boldsymbol{v}_j)\}$.

シミュレーション

$\boldsymbol{U}_1, \ldots, \boldsymbol{U}_n \sim i.i.d.$ パラメータθのFrankコピュラ.

$$H_0 : \theta = 10 \quad \text{vs.} \quad H_1 : \theta \neq 10.$$

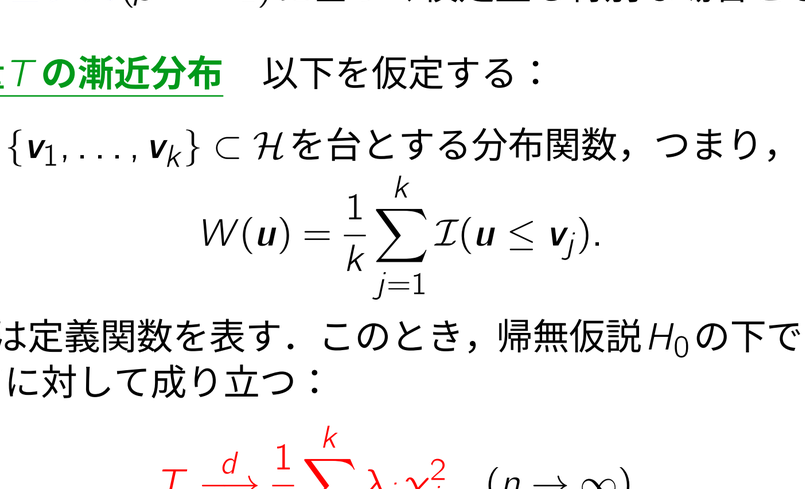


図2．検定統計量Tに基づく検定の検出力のプロット．有意水準は0.05, $k = 100$とし，検出力は10000個のサンプルから計算されている．(x軸はβ, y軸は検出力を表す．)