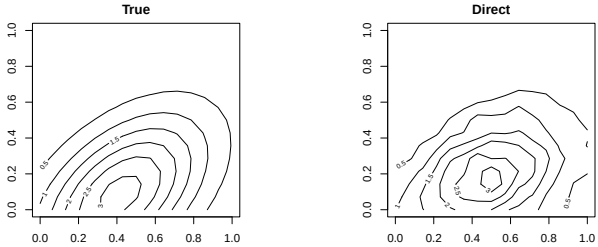
[4EAM-04] 直接型カーネル密度比推定量の多変量への拡張

奥野彰文[1,2]

[1]統計数理研究所, [2]理研AIP

何をしたのか

- 直接型カーネル密度比推定は1次元で定義されている (Ćwik and Mielniczuk, 1989).
- 2次元以上でも使えるように拡張した.



(a) 真値　　(b) 提案法

Figure: 等高線 ($n = m = 10000$).

密度比とは

$\mathbf{X} \sim F, \mathbf{Y} \sim G$として，その密度を$f, g$とするとき，

$$r(\mathbf{z}) = \frac{f(\mathbf{z})}{g(\mathbf{z})}$$

を密度比と呼ぶ．たとえばトレーニング/テストデータの共変量分布の密度比を使うと，

$$\hat{\theta} := \frac{1}{n} \sum_{i=1}^{n} \frac{p_{\text{test}}(\mathbf{x}_i)}{p_{\text{train}}(\mathbf{x}_i)} \{y_i - f_\theta(\mathbf{x}_i)\}^2$$

などによりテストデータにフィットした推定ができる．共変量シフト (Shimodaira, 2000) やそれを使った転移学習 (Weiss et al., 2016 など) が盛んに研究されていて，密度比をうまく推定することが特に重要.

間接的な密度比推定

密度関数 f, g をそれぞれ推定して，その比をとる

$$\widehat{r}_{\mathsf{Indirect}}(\mathbf{z}) := \frac{\widehat{f}(\mathbf{z})}{\widehat{g}(\mathbf{z})}$$

を間接的 (Indirect) な密度比推定量と呼ぶ．$\widehat{f}, \widehat{g}$ は例えばカーネル密度推定量 (KDE):

$$\hat{p}(\mathbf{z}) := \frac{1}{nh^d} \sum_{i=1}^{n} K\left(\frac{\mathbf{z} - \mathbf{X}_i}{h}\right)$$

などを使って推定できる．ただこれらは密度の計算を介するので，もっと直接的な方法が最近は好まれている．

直接的な密度比推定 (パラメトリック)

$r(\mathbf{z}) = f(\mathbf{z})/g(\mathbf{z})$を推定するために，密度比を表すパラメトリックモデル $r_\theta(\mathbf{z})$ を直接的に学習する方法が盛んに研究されている．典型的な方法としては，F, G に従う$\{\mathbf{X}_i\}$と$\{\mathbf{Y}_j\}$の判別を介したロジスティック損失

$$\ell(\theta) := -\frac{1}{n}\sum_{i=1}^{n} \log \frac{r_\theta(\mathbf{X}_i)}{1 + r_\theta(\mathbf{X}_i)} - \frac{1}{m}\sum_{j=1}^{m} \log \frac{1}{1 + r_\theta(\mathbf{Y}_j)}$$

の最小化などを介してr_θを学習する．国内だと杉山先生のグループ(+2010年前後の金森先生, 山田先生, 鈴木先生, ...)がたくさんの研究論文を出している．

- 実用を考えるとこの方向性がベターと思いますが，ところで...

直接的な密度比推定 (カーネル，一次元)

$r(\mathbf{z})$を推定する，カーネル密度比推定

$$\widehat{r}_{\mathrm{Direct}}(z) := \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{\widehat{G}(z) - \widehat{G}(X_i)}{h}\right), \quad \widehat{G}(z) := \frac{1}{n}\sum_{j=1}^{m} \mathbb{1}(Y_j \le z).$$

がĆwik and Mielniczuk (1989)によって提案された.

- Gijbels and Mielniczuk (1995), Chen et al. (2009): 理論
- Motoyama and Maesono (2018): 改善した推定量
- Moriyama and Maesono (2020): ハザード比の推定に応用
- Igarashi (2020): 端でのバイアス補正

のようにいろいろ修正が出てきたが，全て $d = 1$ 次元に限定されていた．一般の d 次元に拡張できるのか，前から気になっていた.

直接的な密度比推定 (カーネル，多次元)

今回の研究では多次元拡張を提案しています.

$$\widehat{r}_{\mathsf{Direct}}(\mathbf{z}) := \frac{1}{nh^d}\sum_{i=1}^{n} K\left(\frac{\widehat{\mathbf{H}}(\mathbf{z}) - \widehat{\mathbf{H}}(\mathbf{X}_i)}{h}\right), \quad \widehat{\mathbf{H}}(\mathbf{z}) = (\widehat{H}_1(\mathbf{z}), \widehat{H}_2(\mathbf{z}), \ldots, \widehat{H}_d(\mathbf{z})),$$

ただし $\widehat{H}_\ell(\mathbf{z}) : \mathbb{R}^d \to \mathbb{R}$ はユーザの指定する，条件付き累積分布関数

$$H_\ell(\mathbf{z}) = \mathbb{P}(\mathsf{Y}_\ell \le z_\ell \mid \mathsf{Y}_1 = z_1, \mathsf{Y}_2 = z_2, \ldots, \mathsf{Y}_{\ell-1} = z_{\ell-1}) : \mathbb{R}^d \to [0,1]$$

の推定量．例えば Hall et al. (1999) などのノンパラ推定量がある.

なぜこんな謎推定量で密度比がわかるのか？

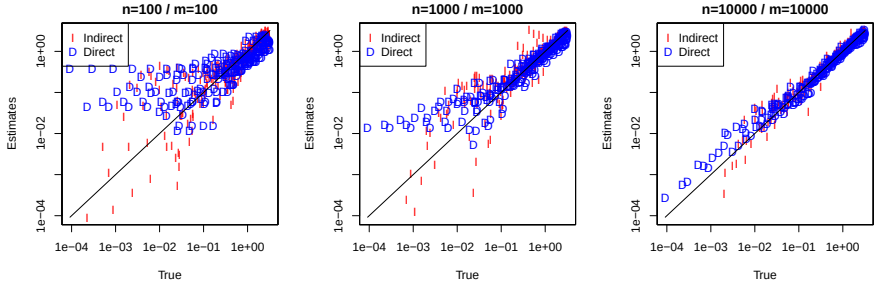
$\hat{r}_{\mathrm{Direct}}(\mathbf{z})$ は $\hat{\mathbf{V}} = \hat{\mathbf{H}}(\mathbf{X})$ の，点 $\hat{\mathbf{v}} = \hat{\mathbf{H}}(\mathbf{z})$ でのカーネル密度推定量である．
密度 (Radon-Nikodym微分) をいい感じに展開すると，

$$
\begin{aligned}
\frac{\mathrm{d}\mathbb{P}_{\mathbf{V}}}{\mathrm{d}\mu}(\mathbf{H}(\mathbf{z})) &= \lim_{\eta \searrow 0} \frac{\mathbb{P}(\mathbf{V} \in B_\eta(\mathbf{v}))}{\mu(B_\eta(\mathbf{v}))}\Bigg|_{\mathbf{v}=\mathbf{H}(\mathbf{z})} \\
&= \lim_{\eta \searrow 0} \frac{1}{\mu(B_\eta(\mathbf{v}))} \int_{B_\eta(\mathbf{v})} \frac{f(\mathbf{H}^{-1}(\mathbf{u}))}{g(\mathbf{H}^{-1}(\mathbf{u}))} \mathrm{d}\mu(\mathbf{u}) \Bigg|_{\mathbf{v}=\mathbf{H}(\mathbf{z})} \\
&= \frac{f(\mathbf{z})}{g(\mathbf{z})}.
\end{aligned}
$$

数値実験

$F = N(\mu_X, \Sigma_X), G = N(\mu_Y, \Sigma_Y)$とする．ただし

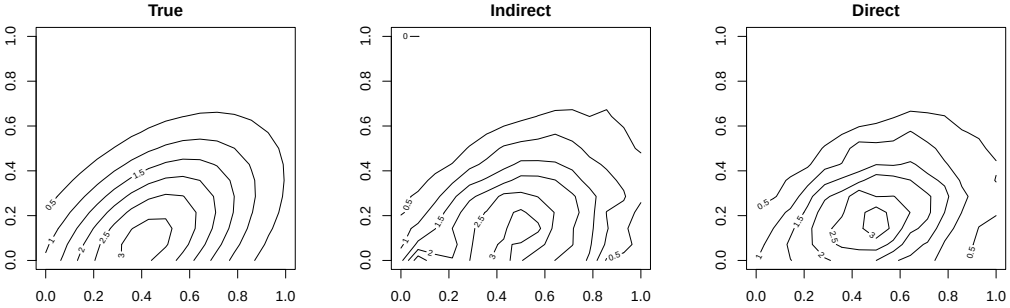
$$\mu_X = \begin{pmatrix} 0 \\ -0.5 \end{pmatrix}, \quad \Sigma_X = \begin{pmatrix} 0.3 & 0.1 \\ 0.1 & 0.3 \end{pmatrix}, \quad \mu_Y = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \quad \Sigma_Y = \begin{pmatrix} 0.5 & 0.1 \\ 0.1 & 0.5 \end{pmatrix}.$$



(a) $n = m = 100$ (b) $n = m = 1000$ (c) $n = m = 10000$

Figure: 真の密度比 $r(\mathbf{z})$ と間接/直接密度比推定量.

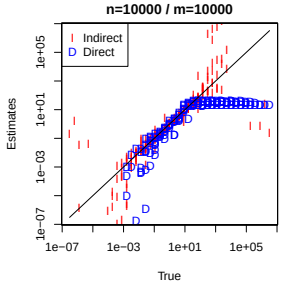
数値実験



(a) True (b) Indirect (c) Direct

Figure: 等高線 ($n = m = 10000$).

- 真の密度比が大きい場合の直接カーネル密度比推定は厳しい.



なぜならば：推定量が上からバウンドされている.

$$\hat{r}^{\dagger}(\mathbf{z}) \leq h^{-d} \sup_{\mathbf{z} \in \mathbb{R}^d} |K(\mathbf{z})|, \quad \mathbf{z} \in \mathbb{R}^d.$$

- 直接型+ノンパラ法を支持する積極的な理由は(現状)ない.

まとめ

- Ćwik and Mielniczuk (1989)で提案された単変量の直接型カーネル密度比推定量を多変量に拡張した.
- Akifumi Okuno. A multivariate adaptation of direct kernel estimation of density ratio. arXiv:2311.12380

質問やコメントは okuno@ism.ac.jp まで.