Density estimation based on U-divergence

Osamu Komori　Prediction and Knowledge Discovery Research Center,　Project Researcher

1 U-divergence

Let $U : \mathbf{R}^+ \to \mathbf{R}$ be a convex and strictly increasing function with the derivative u and the inverse function $\xi = u^{-1}$. Then for real-valued functions f and $g : \mathbf{R}^p \to \mathbf{R}^+$, the U-divergence is given as a special case of the Bregman divergence (**?**):

$$D_U(g, f) = \int d(\xi(g(\boldsymbol{x})), \xi(f(\boldsymbol{x})))d\boldsymbol{x}, \tag{1}$$

where

$$d(g', f') = U(f') - \{u(g')(f' - g') + U(g')\}. \tag{2}$$

Note that $D_U(g, f)$ is non-negative because of the convexity of U. The equality holds if and only if $f = g$ (a.e. $\boldsymbol{x}$). It is also simply expressed as

$$D_U(g, f) = C_U(g, f) - H_U(g), \tag{3}$$

where

$$C_U(g, f) = -\int g(\boldsymbol{x})\xi(f(\boldsymbol{x}))d\boldsymbol{x} + \int U(\xi(f(\boldsymbol{x})))d\boldsymbol{x} \tag{4}$$

$$H_U(g) = -\int g(\boldsymbol{x})\xi(g(\boldsymbol{x}))d\boldsymbol{x} + \int U(\xi(g(\boldsymbol{x})))d\boldsymbol{x} \ (= C_U(g, g)), \tag{5}$$

and $C_U(g, f)$ and $H_U(g)$ are called the U-cross entropy and U-entropy, respectively.

U-loss function with volume-mass-one

The U-loss function for observations $D = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, which derived from the cross-entropy in (4), is defined as

$$L_U(f) = -\frac{1}{n}\sum_{i=1}^{n} \xi(f(\boldsymbol{x}_i)) + \int U(\xi(f(\boldsymbol{x})))d\boldsymbol{x}. \tag{6}$$

Then, we consider the following variant:

$$\mathcal{L}_U(f) \equiv L_U(u(U^{-1}(f))) \tag{7}$$

$$= -\frac{1}{n}\sum_{i=1}^{n} U^{-1}(f(\boldsymbol{x}_i)) + 1 \tag{8}$$

The point is that the second integral term in (6) is restricted to be 1, which we call volume-mass-one. Here we consider $U(t) = (1 + \beta t)^{(1+\beta)/\beta}/(1 + \beta)$ with $\beta > 0$.

2 Algorithm

1. Set $f_0(\boldsymbol{x}) = 0$.
2. For $k = 1, \ldots, K$,
   (a) Initialize $\pi = \pi_0\ (\ll 1)$, $\boldsymbol{\Sigma} = \boldsymbol{I}$ and $\boldsymbol{\mu} = \underset{\boldsymbol{\mu} \in D}{\mathrm{argmin}}\Big\{\mathcal{L}_\beta\Big((1-\pi)f_{k-1}^{1+\beta} + \pi\phi(\boldsymbol{\mu}, \boldsymbol{I})\Big)\Big\}$, where $\boldsymbol{I}$ is the $p \times p$ identity matrix; ϕ is the basis function in $\mathcal{D}_\beta$. Define

$$\mathcal{R}_{\boldsymbol{\mu},\boldsymbol{\Sigma}} = \left\{ i \;\middle|\; \frac{\beta}{2(1+\beta)}(\boldsymbol{x}_i - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i - \boldsymbol{\mu}) < 1,\ \boldsymbol{x}_i \in D \right\}. \tag{9}$$

   (b) For $\boldsymbol{x}_i$ such that $i \in \mathcal{R}_{\boldsymbol{\mu},\boldsymbol{\Sigma}}$, calculate

$$q(\boldsymbol{x}_i) = \frac{\pi\phi(\boldsymbol{x}_i)}{(1-\pi)f_{k-1}(\boldsymbol{x}_i)^{1+\beta} + \pi\phi(\boldsymbol{x}_i)} \tag{10}$$

$$\boldsymbol{\mu}_q = \frac{\sum_{\mathcal{R}_{\boldsymbol{\mu},\boldsymbol{\Sigma}}} q(\boldsymbol{x}_i)^{\frac{1}{1+\beta}}\boldsymbol{x}_i}{\sum_{\mathcal{R}_{\boldsymbol{\mu},\boldsymbol{\Sigma}}} q(\boldsymbol{x}_i)^{\frac{1}{1+\beta}}}. \tag{11}$$

   where $\sum_{\mathcal{R}_{\boldsymbol{\mu},\boldsymbol{\Sigma}}}$ is the summation of i over $\mathcal{R}_{\boldsymbol{\mu},\boldsymbol{\Sigma}}$.

   (c) Update $\boldsymbol{\mu} = \boldsymbol{\mu}_q$ and go to step (d) if $\mathcal{R}_{\boldsymbol{\mu},\boldsymbol{\Sigma}} \subset \mathcal{R}_{\boldsymbol{\mu}_q,\boldsymbol{\Sigma}}$; otherwise go back to step (b).
   (d) For $\boldsymbol{x}_i$ such that $i \in \mathcal{R}_{\boldsymbol{\mu},\boldsymbol{\Sigma}}$, update $q(\boldsymbol{x}_i)$ as in (10) and calculate

$$\boldsymbol{\Sigma}_q = \frac{2 + (2+p)\beta}{2(1+\beta)} \frac{\sum_{\mathcal{R}_{\boldsymbol{\mu},\boldsymbol{\Sigma}}} q(\boldsymbol{x}_i)^{\frac{1}{1+\beta}}(\boldsymbol{x}_i - \boldsymbol{\mu})(\boldsymbol{x}_i - \boldsymbol{\mu})'}{\sum_{\mathcal{R}_{\boldsymbol{\mu},\boldsymbol{\Sigma}}} q(\boldsymbol{x}_i)^{\frac{1}{1+\beta}}}. \tag{12}$$

   (e) Update $\boldsymbol{\Sigma} = \boldsymbol{\Sigma}_q$ and go to step (f) if $\mathcal{R}_{\boldsymbol{\mu},\boldsymbol{\Sigma}} \subset \mathcal{R}_{\boldsymbol{\mu},\boldsymbol{\Sigma}_q}$; otherwise go back to step (d).
   (f) For $\boldsymbol{x}_i$ such that $i \in \mathcal{R}_{\boldsymbol{\mu},\boldsymbol{\Sigma}}$, update $q(\boldsymbol{x}_i)$ as in (10) and calculate

$$\pi_q = \frac{A_2^{1+\beta}}{A_1^{1+\beta} + A_2^{1+\beta}}, \tag{13}$$

   where

$$A_1 = \sum_{\mathcal{R}_{\boldsymbol{\mu},\boldsymbol{\Sigma}}} (1 - q(\boldsymbol{x}_i))^{\frac{1}{1+\beta}} f_{k-1}(\boldsymbol{x}_i)^{\beta} \tag{14}$$

$$A_2 = \sum_{\mathcal{R}_{\boldsymbol{\mu},\boldsymbol{\Sigma}}} q(\boldsymbol{x}_i)^{\frac{1}{1+\beta}}\phi(\boldsymbol{x}_i)^{\frac{\beta}{1+\beta}}, \tag{15}$$

   and update $\pi = \pi_q$, and $q(\boldsymbol{x}_i)$ as in (10).
   (g) Repeat the steps from (b) to (f) until the values of $\boldsymbol{\mu}$, $\boldsymbol{\Sigma}$ and π converges, and set them to be $\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k, \pi_k$, respectively.
   (h) Update f_{k-1} with $\phi_k(\boldsymbol{x}) = \phi_\beta(\boldsymbol{x}, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$ and π_k as

$$f_k = \Big\{(1-\pi_k)f_{k-1}^{1+\beta} + \pi_k\phi_k\Big\}^{\frac{1}{1+\beta}}. \tag{16}$$

3. Output $\hat{f} = f_K$.

Theorem 2.1 *The empirical loss $\mathcal{L}_\beta(f_k)$ in the boosting algorithm is monotonically decreasing with respect to k. That is, for $k = 1, \ldots, K$,*

$$\mathcal{L}_\beta(f_k) \le \mathcal{L}_\beta(f_{k-1}). \tag{17}$$

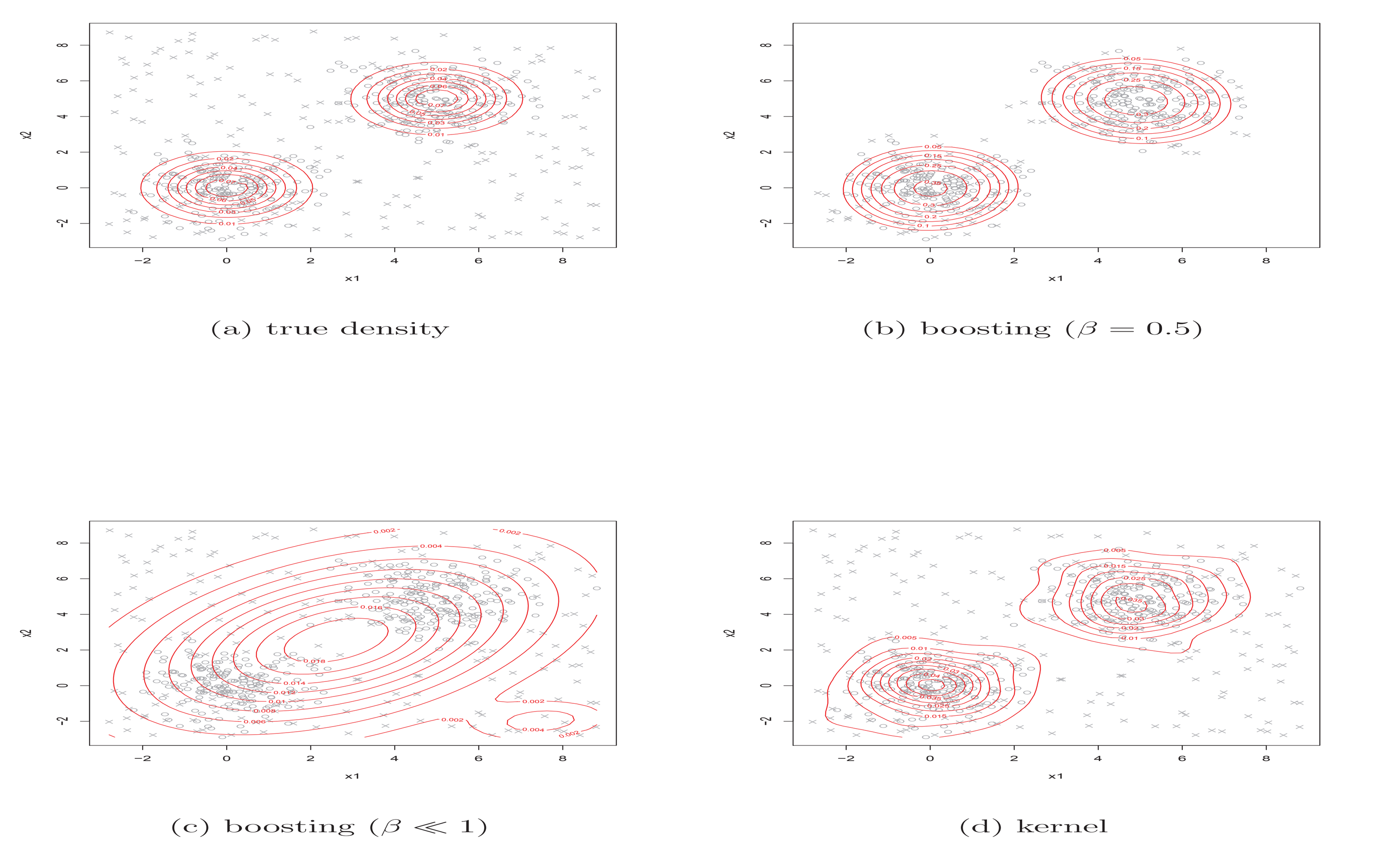

Fig1. Contour plots for the true density (a) and density estimators by three methods (b), (c) and (d). Observations from the normal distributions are denoted by circles; noisy observations are denoted by cross marks. Observations that are not used in the estimation are deleted in the panel (b).