Parametric estimation using nearest neighbours

Parametric family $\mathcal{F} = \{f_\theta \mid \theta \in \Theta\}$. We are interested in finding θ given a sample $\mathcal{X} = \{X_1, \ldots, X_n\}$ distributed according to f_θ.

Given a proposed parameter θ', the Hellinger distance between $f_{\theta'}$ and f_θ (θ fixed) is

$$\mathcal{H}^2(\theta') = \frac{1}{2} \int \left(\sqrt{f_{\theta'}}(x) - \sqrt{f_\theta}(x) \right)^2 \, \mathrm{d}x. \tag{1}$$

Expanding the square, we obtain

$$\mathcal{H}^2(\theta') = 1 - \int \sqrt{f_{\theta'}}(x) \sqrt{f_\theta}(x) \, \mathrm{d}x \doteq 1 - \mathcal{B}(\theta'), \tag{2}$$

owing to $\int f_{\theta'} = \int f_\theta = 1$. The integral $\mathcal{B}(\theta')$ is the 'Bhattacharyya affinity coefficient' [2] between the distribution $f_{\theta'}$ and ground truth f_θ. Minimising $\mathcal{H}^2(\theta')$ is obviously equivalent to maximising $\mathcal{B}(\theta')$ – so we are looking for making $f_{\theta'}$ as similar as possible to f_θ in the Bhattacharyya sense.

As $\mathcal{B}(\theta')$ involves the unknown f_θ, it needs to be estimated.

Here we propose to use [1, Lemma 1], allowing consistent estimation of quantities of type $\int \phi \sqrt{f}$. Specifically, we estimate $\mathcal{B}(\theta')$ by

$$\widehat{\mathcal{B}}(\theta') \doteq \frac{2}{\sqrt{\pi}} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sqrt{f_{\theta'}}(X_i) \sqrt{V_i}, \tag{3}$$

where, V_i is the volume of the ball centred at X_i whose radius is the distance from X_i to its closest neighbour in the sample.

We choose for parameter θ' the maximiser:

$$\widehat{\theta} \doteq \arg\max_{\theta' \in \Theta} \widehat{\mathcal{B}}(\theta'). \tag{4}$$

For a regression setting with normally distributed iid errors ϵ_i, we have

$$y_i = \alpha \, x_i + \epsilon_i \tag{5}$$

$$\widehat{\mathcal{B}}(\alpha) \doteq \frac{2}{\sqrt{\pi}} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \sqrt{N(y_i - \alpha \, x_i, \sigma)} \sqrt{E_i}, \tag{6}$$

References

[1] Carlos Aya-Moreno, Gery Geenens, and Spiridon Penev. "Shape-preserving wavelet-based multivariate density estimation". In: *Journal of Multivariate Analysis* 168 (2018), pp. 30–47.

[2] Anil Bhattacharyya. "On a measure of divergence between two multinomial populations". In: *Sankhyā: the indian journal of statistics* (1946), pp. 401–406.