M.S. Tikhov (Lobachevsky State University of Nizhny Novgorod, Russia)

Multivariate k-Nearest Neighbor Distribution Function Estimates in Dose-Effect Relationship.

Let $\mathcal{X}^{(n)} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n)$ and $\mathcal{U}^{(n)} = (\boldsymbol{U}_1, \ldots, \boldsymbol{U}_n)$ be independent, identically distributed random p-vectors with bounded continuous densities $f(\boldsymbol{x})$ and $g(\boldsymbol{u})$, with distribution functions f $F(\boldsymbol{x}), G(\boldsymbol{u})$, and $W_j = I(\boldsymbol{X}_j < \boldsymbol{U}_j)$. We are observing the sample $\{(W_j, \boldsymbol{U}_j), j = 1, .., n\}$. It is necessary to estimate the unknown distribution function $F(\boldsymbol{x})$. Consider an estimate of $F(\boldsymbol{x})$ given by $F_n(\boldsymbol{x}) = Fg_n(\boldsymbol{x})/g_n(\boldsymbol{x})$, where

$$g_n(\boldsymbol{x}) = \frac{1}{nR_n^p} \sum_{j=1}^{n} K\left(\frac{\boldsymbol{x} - \boldsymbol{U}_j}{R_n}\right), \quad Fg_n(x) = \frac{1}{nR_n^p} \sum_{j=1}^{n} W_j K\left(\frac{\boldsymbol{x} - \boldsymbol{U}_j}{R_n}\right),$$

$F_n(\boldsymbol{x}) = Fg_n(\boldsymbol{x})/g_n(\boldsymbol{x})$, $R_n = R_n(\boldsymbol{x})$ is the Euclidean distance between $\boldsymbol{x}$ and kth nearest neighbor of $\boldsymbol{x}$ among the $\boldsymbol{U}_j$'s, $K(\boldsymbol{x})$ is *non-negative kernel*, and $k = k(n)$ is a sequence of positive integers such that $kn^{-4/5} = 1 + o(1)$ as $n \to \infty$, $S_r = \{\boldsymbol{z} : ||\boldsymbol{z} - \boldsymbol{x}|| < r\}$. Given a differentiable function F, let $D_\alpha F$ denote the partial derivative of F with respect x_α. If F is twice continuously differentiable and $K(\boldsymbol{x})$ has finite second order moments let $Q\left(K\right)(\boldsymbol{x}) = \sum_{\alpha,\beta} D_\alpha D_\beta F(\boldsymbol{x}) \int u_\alpha u_\beta K\left(\boldsymbol{u}\right) d\boldsymbol{u}$.

Theorem. *Let the function $K(\boldsymbol{x})$ be bounded and assume that the functions F and g are three times continuously differentiable with $\int ||\,\boldsymbol{u}||^2 K(\boldsymbol{u})\, d\boldsymbol{u} < \infty, \int \boldsymbol{u} K(\boldsymbol{u})\, d\boldsymbol{u} =$* $\mathbf{0}$. *Further, let $1 - \mathbf{P}(S_r) = O(r^{-\delta})$ for some $\delta > 0$ as $r \to \infty$. Consider a point $\boldsymbol{x}$ with $f(\boldsymbol{x}) > 0$ and $f(\boldsymbol{x})$ continuously differentiable in a neighborhood of $\boldsymbol{x}$. Then*

$$\mathbf{E}(Fg_n(\boldsymbol{x})) = Fg(\boldsymbol{x}) + \frac{c^{\frac{2}{p}} Q(Fg)(\boldsymbol{x})}{2\pi (Fg(Q(Fg)(\boldsymbol{x})))^{\frac{2}{p}}} \left(\frac{k}{n}\right)^{\frac{2}{p}} + \frac{\pi^{\frac{2}{p}}}{c} \frac{Fg(\boldsymbol{x})}{k} \int_{||\boldsymbol{u}||=1} K(\boldsymbol{x})\, d\Sigma(\boldsymbol{u}) + o(1),$$

$$\sqrt{k}\,(F_n(\boldsymbol{x}) - \mathbf{E}(F_n(\boldsymbol{x}))) \xrightarrow{d} N(0, F(\boldsymbol{x})(1 - F(\boldsymbol{x})) \int K^2(\boldsymbol{u})\, d\boldsymbol{u}), n \to \infty, c = \Gamma\left(\tfrac{p+2}{2}\right).$$

where Σ is the uniform distribution on the surface of the p-sphere of unit radius.

REFERENCES

[1] Y. P. Mack, M. Rosenblatt, Multivariate k-Nearest Neighbor Density Estimates, *J. Multivar. Analysis*, **9** (1979), 1-15.