Sample Moments Integrating Normal Kernel (SMINK) density and regression estimators

1 SMINK density estimation

Let $X_1, ..., X_n$ be a random sample from a k-variate absolutely continuous distribution with density $f(x)$, expectation μ, and non-singular variance matrix Σ. Let $x^{(i)}$ be the i-th component of x, and $X_{i,j}$ the i-th component of X_j. The SMINK density estimator of $f(x)$ takes the form:

$$\widehat{f}(x|\gamma) = \frac{1}{n}\sum_{j=1}^{n}\left(\prod_{i=1}^{k} I(x^{(i)} \neq X_{i,j})\right)\widehat{f}_{n,j}(x|\gamma),$$

where

$$\begin{aligned}\widehat{f}_{n,j}(x|\gamma) &= \left(\left(\gamma\sqrt{2\pi}\right)^k \det\left(\widehat{\Sigma}\right)\right)^{-1} \qquad (1)\\ &\times \exp\left[-\frac{1}{2}\left(x - \sqrt{1-\gamma^2}X_j - \left(1-\sqrt{1-\gamma^2}\right)\overline{X}\right)'\widehat{\Sigma}^{-1}\right.\\ &\times \left.\left(x - \sqrt{1-\gamma^2}X_j - \left(1-\sqrt{1-\gamma^2}\right)\overline{X}\right)/\gamma^2\right],\end{aligned}$$

with $\overline{X} = (1/n)\sum_{j=1}^{n} X_j$ and $\widehat{\Sigma} = (1/n)\sum_{j=1}^{n}\left(X_j - \overline{X}\right)\left(X_j - \overline{X}\right)'$.

For $\gamma \in (0,1]$ we have

$$\int x\widehat{f}(x|\gamma)dx = \overline{X}, \quad \int xx'\widehat{f}(x|\gamma)dx = \frac{1}{n}\sum_{j=1}^{n} X_j X_j'.$$

Let $\xi_n \in (0,1]$ be a sequence of non-random numbers such that $\lim_{n\to\infty}\xi_n = 0$, $\lim_{n\to\infty}\sqrt{n}\xi_n^k = \infty$. In particular, let

$$\xi_n = \left(\sqrt{n}\right)^{-\alpha/k}, \; 0 < \alpha < 1. \qquad (2)$$

EasyReg International will ask you to specify α (the default value is 0.5).

Moreover, let γ_n be a sequence of (random) numbers such that $\gamma_n \in [\xi_n, 1]$ (a.s), $(p)\lim_{n\to\infty} \gamma_n = 0$. Then $\widehat{f}(x|\gamma_n)$ is uniformly consistent:

$$p \lim_{n\to\infty} \left| \widehat{f}(x|\gamma_n) - f(x) \right| = 0. \tag{3}$$

Furthermore, if we choose

$$\gamma_n = \arg \min_{\gamma \in [\xi_n, 1]} \widehat{Q}(\gamma),$$

where

$$\widehat{Q}(\gamma) = \int \widehat{f}(x|\gamma)^2 dx - 2\frac{1}{n}\sum_{j=1}^{n} \widehat{f}(X_j|\gamma),$$

then $\int \left(\widehat{f}(x|\gamma_n) - f(x)\right)^2 dx$ is (approximately) minimal, and (3) carries over.

2 SMINK regression estimation

Now let $(Y_1, X_1), ..., (Y_n, X_n)$ be a random sample from a $k+1$-variate absolutely continuous distribution with density $f_{y,x}(y, x)$ and marginal density $f_x(x)$, where $Y_j \in \mathbb{R}$. As is well known, the conditional expectation function $g(x) = E[Y_j|X_j = x]$ takes the form

$$g(x) = \frac{\int_{-\infty}^{\infty} y f_{y,x}(y, x) dy}{f_x(x)}.$$

This suggests to estimate $g(x)$ nonparametrically by

$$\widehat{g}(x|\gamma_1, \gamma_2) = \frac{\int_{-\infty}^{\infty} y \widehat{f}_{y,x}(y, x|\gamma_1) dy}{\widehat{f}_x(x|\gamma_2)},$$

where $\widehat{f}_{y,x}(y, x|\gamma_1)$ and $\widehat{f}_x(x|\gamma_2)$ are SMINK density estimators. It can be shown that $\widehat{g}(x|\gamma_1, \gamma_2)$ takes the form

$$\widehat{g}(x|\gamma_1, \gamma_2) = \frac{\left(\widehat{\beta}_0 + \widehat{\beta}_1' x\right) \widehat{f}_x(x|\gamma_1) + \left(\sqrt{1-\gamma_1^2}\right) \frac{1}{n}\sum_{j=1}^{n} \widehat{U}_j . I_j(x) . \widehat{f}_{n,j}(x|\gamma_1)}{\widehat{f}_x(x|\gamma_2)},$$

where $\widehat{\beta}_0$ and $\widehat{\beta}_1$ are the OLS estimators of the constant and the slope parameters, respectively, of the linear regression of Y_j on X_j, $\widehat{U}_j = Y_j - \widehat{\beta}_0 - \widehat{\beta}_1' X_j$ is the OLS residual, $I_j(x) = \prod_{i=1}^{k} I(x^{(i)} \neq X_{i,j})$, and $\widehat{f}_{n,j}(x|\gamma)$ is defined in (1). Note that $\widehat{g}(x|1,1) = \widehat{\beta}_0 + \widehat{\beta}_1' x$, hence the linear regression estimator is a special case of a SMINK regression estimator.

Next, choose ξ_n the same as in (2), and let

$$\gamma_{2,n} = \arg \min_{\gamma \in [\xi_n, 1]} \widehat{Q}_2(\gamma),$$

where

$$\widehat{Q}_2(\gamma) = \int \widehat{f}_x(x|\gamma)^2 dx - 2\frac{1}{n}\sum_{j=1}^{n} \widehat{f}_x(X_j|\gamma).$$

If we choose

$$\gamma_{1,n} = \arg \min_{\gamma \in [\xi_n, 1]} \widehat{Q}_1(\gamma),$$

where

$$\widehat{Q}_1(\gamma) = \frac{1}{n}\sum_{j=1}^{n} (Y_j - \widehat{g}(X_1|\gamma, \gamma_{2,n}))^2 \widehat{f}_x(X_j|\gamma_{2,n})^2,$$

then $\widehat{g}(x|\gamma_{1,n}, \gamma_{2,n})$ is uniformly consistent:

$$p \lim_{n\to\infty} \sup_{f_x(x)>\varepsilon} |g(x) - \widehat{g}(x|\gamma_{1,n}, \gamma_{2,n})| = 0$$

for all $\varepsilon \in (0, \sup_x f_x(x))$.

Reference: Bierens, H.J. (1983), "Sample Moments Integrating Normal Kernel Estimators of Multivariate Density and Regression Functions", *Sankhya*, **45**, *Series B*, 160-192.