# Locally robust density estimation and near-paramtric asymptotics

Kanta Naito (Shimane University, Japan)
Spiridon Penev (University of New South Wales, Sydney, Australia)

Let $X_1, \ldots, X_n$ be a sample drawn from a $d$ variate density $f$. Consider the density estimation problem, where we aim at obtaining a density estimator with local adaptation. Also we expect the estimator to have a robustness property in some sense. To do this we utilize a parametric $d$ variate density $g(x, \theta)$, where $\theta \in \Theta \subset \mathbb{R}^p$. Estimation is implemented at a target point $t$ and localization is induced by using a kernel $K(z)$ which is a smooth unimodal integrable function symmetric around $z = 0$.

For estimation of $f(t)$ at the target point $t$, an estimator of $\theta$ is defined as

$$\hat{\theta}_\lambda(t) = arg \max_{\theta \in \Theta} \int_{\mathbb{R}^d} \rho_\lambda(t, x, \theta) dF_n(x),$$

where $F_n$ is the empirical distribution function and

$$\rho_\lambda(t, x, \theta) = \left(\frac{\lambda + 1}{\lambda}\right) K\left(\frac{x - t}{h}\right) \{g(x, \theta)^\lambda - 1\} - \int_{\mathbb{R}^d} K\left(\frac{s - t}{h}\right) g(s, \theta)^{\lambda+1} ds \tag{1}$$

with $0 < \lambda \leq 1$ and the bandwidth $h > 0$. This $\rho_\lambda(t, x, \theta)$ is a localized version of a loss function based on the Bregman divergence associated with the Box-Cox transformation. The additional parameter $\lambda$ represents the power in the Box-Cox transformation and the divergence with $\lambda = 0$ reduces to the Kullback-Leibler divergence. The robustness property is expected to hold for small positive $\lambda$.

Let $g(t, \hat{\theta}_\lambda(t))$ be a pilot estimator of $f(t)$ obtained by using local estimator $\hat{\theta}_\lambda(t)$. Our proposed estimator of $f(t)$ is its normalized version defined as

$$\hat{f}(t) = \frac{g(t, \hat{\theta}_\lambda(t))}{\displaystyle\int_{\mathbb{R}^d} g(s, \hat{\theta}_\lambda(s)) ds}.$$

The behavior of $\hat{f}$ is evaluated by the power divergence risk, with respect to which $\hat{f}$ asymptotically, as $n$ and $h$ both increase, performs better than the usual plug-in parametric estimator $g(\cdot, \hat{\theta}_\lambda)$, where $\hat{\theta}_\lambda$ is a robust estimator of $\theta$ obtained through the global loss optimization:

$$\hat{\theta}_\lambda = arg \max_{\theta \in \Theta} \int_{\mathbb{R}^d} \rho_\lambda(x, \theta) dF_n(x).$$

Here $\rho_\lambda(x, \theta)$ is a global version of $\rho_\lambda(t, x, \theta)$ in (1) with eliminating the kernel localization (or $h \to \infty$.) This means that there is a benefit of using the localized divergence: such a localization always helps even in a robust setting.