Locally robust density estimation and near-parametric asymptotics

S. Penev[1*] and K. Naito[2]

[1] *The University of New South Wales, Australia ; s.penev@unsw.edu.au*
[2] *Shimane University, Matsue, Japan; naito@riko.shimane-u.ac.jp*
[*] *Presenting author*

Keywords. *Risk; Robustness; Bregman divergence; Power divergence; Kernel.*

The original application of local likelihood as a truly semiparametric method in density estimation was proposed in papers by Loader [1996] and Hjort and Jones [1996]: "The estimators run the gamut from a fully parametric fit to almost fully nonparametric with only a single smoothing parameter to be chosen". They also give an interpretation of the procedure as one that minimizes, at each value of the argument, the locally weighted Kullback-Leibler divergence between the "true" and the model density. The infusion of local adaptation to the global likelihood by considering maximization of an expression of the form

$$\sum_{i=1}^{n} K\big(\frac{x_i - t}{h}\big) \log\big\{g(x_i, \theta)\big\} \tag{1}$$

with $K(\frac{x-t}{h})$ being some suitable kernel function centered at t and with bandwidth h, and $g(x, \theta)$ is a nominal parametric density. In any version of these modifications,the resulting local maximum likelihood estimator $\hat{\theta}_{t,h}$ could be substituted to obtain the density estimator $g(x, \hat{\theta}_{t,h})$. As opposed to the global parametric model where the substitution of the global maximum likelihood estimator $\hat{\theta}$ automatically results in a density $g(x, \hat{\theta})$, this is not the case with $g(x, \hat{\theta}_{t,h})$ and one needs to normalize to get a shape-preserving density by $\hat{g}_h(x) = g(x, \hat{\theta}_{x,h}) / \int g(t, \hat{\theta}_{t,h}) dt$.

1 From likelihoods to Bregman divergences

When the ideal parametric model does not confidently hold, other divergences are used to replace the Kullback-Leibler divergence. These divergences have been demonstrated to possess good robustness properties relative to maximum likelihood methods. Specific applications for robust density estimation have been considered in Windham [1995] and in Basu et. al. [1998], with Bregman divergence type measures, parameterised by one "robustness control" parameter $\lambda \geq 0$, with $\lambda = 0$ corresponding to no efficiency loss. However none of the mentioned works deals with **local** versions of the robust divergence measures which are our main object.

Since the class of Bregman divergences is very large (and not all of them have found useful applications), we focus on a particular class of them. Starting with the Box-Cox transformation $G_\lambda(x) = \begin{cases} \frac{1}{\lambda}(x^\lambda - 1), \lambda > 0 \\ \log x \end{cases}$ we define $U_\lambda(x) = x(G_\lambda(x) - 1)$. Note that $U_\lambda(x) \to_{\lambda \to 0} x \log x - x$ which is the $U_\lambda(.)$ function that is used to define the von Neumann divergence. The only paper known to us where the case of **large** h has been analysed is Eguchi and Copas [1998] but it is completely devoted to the local likelihood method. In a nutshell, the results of Eguchi and Copas [1998] show that with respect to the relative entropy risk minimization, there is a benefit of using the local likelihood: little localization "always helps". We demonstrate both theoretically and numerically that such type of statement is also true with respect to robustness: with respect to the Bregman distance based risk minimization, little localization of the globally robust estimator"always helps".

We believe that the localisation proposed here offers a new view towards robustness. In the standard robustness approach (Huber and Ronchetti [2009]) the main focus is on modifying non-robust estimators of *parameters* of certain model density when it is believed that the data was not necessarily generated from the model density because there was contamination. The inference part is essentially finalized once the parameters have been estimated. In our approach we estimate the local features of the density that has generated the data. Our estimated $g(x, \hat{\theta}(x))$ does not in general belong to the class $g(x, \theta), \theta \in \Theta$ and is giving a better idea about the "true" density that has generated the data. On the other hand, there are similarities, too: the belief that the true density is "not too far away" from a model density $g(x, \theta_0)$ is common for both approaches.

References

Basu, A., Harris, I., Hjort, N. and Jones, M.C. (1998) Robust and Efficient Estimation by Minimizing a Density Power Divergence. *Biometrika*, 85, 549–559.

Eguchi, S. and Copas, J. (1998) A Class of Local Likelihood Methods and Near-Parametric Asymptotics. *J. R. Statist. Soc. B*, 60, 709–724.

Hjort, N. and Jones, M.C. (1996) Locally Parametric Nonparametric Density Estimation. *The Annals of Statistics*, 24, 1619–1647.

Huber, P. and Ronchetti, E. (2009) *Robust Statistics*, 2nd Edition. Wiley, New York.

Loader, C. (1996) Local Likelihood Density Estimation. *The Annals of Statistics*, 24, 1602–1618.

Windham, M. (1995) Robustifying Model Fitting. *J. R. Statist. Soc. B*, 57, 599–609.