Adaptive kernel density estimation in L^2-norm using artificial data

Alejandro Pereira
Instituto de Estadística, Universidad de Valparaíso

Karine Bertin
Instituto de Ingeniería Matemática, Universidad de Valparaíso

Abstract

Estimating the common density of a sample is one of the most useful steps in any data analysis. Among the non-parametric techniques used for this goal, the Kernel Density Estimator (KDE) is perhaps the most widely used. These estimators depend on a kernel function K and a bandwidth h. In this work we study an adaptive data driven method to select the bandwidth h, as introduced in Goldenshluger and Lepski (2013).

More precisely, we are interested in estimating the density f of a random variable Y that satisfies $Y = m(X)$ where X is another random variable and $m : \mathbb{R} \mapsto \mathbb{R}$ is an unknown function. We observe two i.i.d. samples generated from theses variables. The first one is quite difficult to obtain and rather small: $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, that satisfies $Y_i = m(X_i)$. The second one (independent of the first one) that is simpler to obtain $\{X_{n+1}, \ldots, X_N\}$ and that can be as large as the statistician needs.

To estimate f, we can use two approaches. In the classical approach, we use the sample $Y_1, \ldots, Y_n$. In the artificial data approach (see Felber, Kohler and Krzyżak, 2015), we estimate the function m by $\hat{m}$ using $(X_1, Y_1), \ldots, (X_n, Y_n)$, then construct the artificial data $\hat{Y}_{n+1} = \hat{m}(X_{n+1})$,…,$\hat{Y}_{n+N} = \hat{m}(X_{n+N})$ and finally estimate f using these artificial data.

In this work, we prove that kernel estimators using artificial data achieves a faster convergence rates when compared to the same estimator in the classical approach. Moreover, we propose a Goldenshluger-Lepski method to select the bandwidth in the artificial data approach and prove that it converges at optimal rate of convergence. Finally, we perform a simulation study and compare the results via the MISE criterion.

References

1. Goldenshluger, A., Lepski, O. (2013). On adaptive minimax density estimation on $\mathbb{R}^d$. *Probability Theory and Related Fields* **159**, 479-543.
2. Felber, T., Kohler, M., Krzyżak, A. (2015). Adaptive density estimation based on real and artificial data. *Journal of Nonparametric Statistics* **27**, 1-18.