Construction of principal curves by incorporating estimates of the data distribution

[1]Junping Zhang, [2]Uwe Kruger

The construction of principal curves relies on the arc-length as a global index for representing the middle of the data distribution. This work shows that under the assumption that the stochastic uncertainty of each observation is i.i.d. and follows a Gaussian distribution, existing work does provide a unique estimate of the principal curve that converges to the actual generating curve as the number of samples tends to go to infinity. However, the presented work outlines that a departure from this assumption does not guarantee a consistent estimation of the generating curve.

Given that the distribution function of the uncertainty is not being considered in the iterative calculation of the principal curve, the presented work discusses the incorporation of a kernel density estimation into the construction of a principal curve. The benefit of blending an estimator of the data distribution into this calculation is that the arc-length parameter does not only reflect the global structure of data distribution but also embody density property of the data and the stochastic uncertainty. We show the effectiveness of the density-based principal curve algorithm through a number of experimental results and analysis the statistical properties of the estimated curves.

[1] Associate Professor, Shanghai Key Laboratory of Intelligent Information Processing, Department of Computer Science and Engineering, Fudan University, 200433 Shanghai, P.R. China. Email: jpzhang@fudan.edu.cn, Tel.: +86-21-55664712, Fax.: +86-21-65654253.

[2] Associate Professor, Department of Electrical Engineering, The Petroleum Institute, Abu Dhabi, United Arab Emirates, Tel.: +971-2-607-5150, Fax.: +971-2-607-5200.