Michele Allegra[1], Maria d'Errico[2], Elena Facco[2], Alessandro Laio[2,3] and Alex Rodriguez[2]

1 Institut de Neurosciences de la Timone (INT), CNRS, Marseille, France
2 Scuola Internazionale Superiore di Studi Avanzati, Trieste, Italy
3 International Centre for Theoretical Physics, Trieste, Italy

Reconstructing the topography of multidimensional probability landscapes

Exploratory data analysis (EDA) has the goal of revealing patterns in a data set. Persi Diaconis warned that this practice can come close to "magical thinking" if there is no attempt at checking whether the structures found could have arisen by chance. Any such attempt at statistical validation unavoidably requires to introduce some sort of statistical assumption on the data.
We first developed a method for EDA, Density Peak Clustering (DPC) [1]. At the beginning, DPC was fully exploratory and hence lacked a procedure for statistical validation of the results. The effort to put DPC on a rigorous statistical footing has yielded a complex procedure to reconstruct a probability density [2], its intrinsic dimension [3,4], and its peaks [5] in a high-dimensional space. The whole procedure builds on a set of minimal, but effective statistical assumptions on the data, and requires several steps of model selection and estimation.

References

[1] A. Rodriguez, and A. Laio, *Clustering by fast search and find of density peaks*, Science 344(691), 1492 (2014).
[2] A. Rodriguez, M. d'Errico, E. Facco and A. Laio, *Computing the free energy without collective variables*, JCTC 14(3), 1206 (2018).
[3] E. Facco, M. d'Errico, A. Rodriguez and A. Laio, *Estimating the intrinsic dimension of datasets by a minimal neighborhood information*, Scientific Reports 7(1), 12140 (2017).
[4] M. Allegra, E. Facco, A. Laio and A. Mira, *Data classification based on the local intrinsic dimension,* in prep., (2018).
[5] M. d'Errico, E. Facco, A. Laio and A. Rodriguez, *Automatic topography of high-dimensional data sets by non-parametric Density Peak clustering*, arXiv:1802.10549 (2018).