Modeling high-dimensional data with a multivariate Bernoulli distribution

Mireille Boutin[1], **Evzenie Coupkova**[2] **and Marco Morosin**[1]

[1]*Eindhoven University of Technology, Netherlands*
[2]*Purdue University, West Lafayette, USA*

Abstract

Recent experiments have uncovered that several high-dimensional datasets have a high probability of forming binary clusters after projecting the data on a random one-dimensional subspace. We present a probability model for the high-dimensional data that could explain this phenomenon. The model consists of a discrete skeleton formed by several Bernoulli random variables arranged as the vertices of a parallelotope, on which noise is added. While clusters in high dimension are difficult to observe or may not exist at all, the groupings of points drawn from this distribution can be easily identified. These groupings can be used in place of clustering, especially when the dataset is small, and their statistical significance can be tested. This structure allows for semantic grouping of datapoints based on different criteria and provides a binary, compressed representation of the data in which each bit represents the binary class for one criterion.

Keywords

High-dimensional data, Clustering, Small data, Data binarization, Random projection, Data model.

References

[1] Boutin, M., Coupkova, E. (2026). A New Model for Natural Groupings in High-Dimensional Data. In: Nielsen, F., Barbaresco, F. (eds), *Geometric Science of Information*. GSI 2025. Lecture Notes in Computer Science, vol 16033. Springer.