Detecting Novel Associations in Large Data Sets

https://science.sciencemag.org/content/sci/334/6062/1518.full.pdf

▼ What did the authors try to accomplish?

- Present a measure of dependence for two-variable relationships: the maximal
information coefficient (MIC)
- Give ability to examine all potentially interesting relationships in a data set, independent of their form

▼ What were the key elements of the approach?

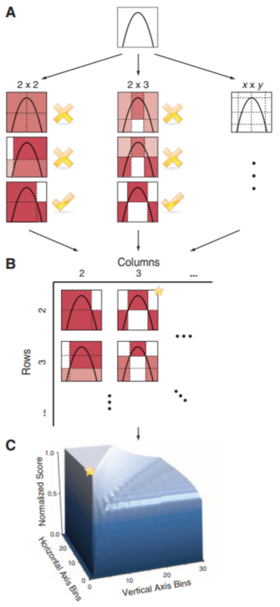
MIC

- based on the idea that if a relationship exists between two variables, then a
grid can be drawn on the scatterplot of the two variables that partitions the data to encapsulate that relationship.

Procedure

1. For each pair (x,y), the MIC algorithm finds the x-by-y grid with the highest
induced mutual information (A)
2. The algorithm normalizes the mutual information scores and compiles a matrix that stores, for each resolution, the best grid at that resolution and its normalized
score (B)
3. The normalized scores form the characteristic matrix, which can be visualized as a surface; MIC corresponds to the highest point on this surface (C)

In this example, there are many grids that achieve the highest score. The star in (B) marks a sample grid achieving this score, and the star in (C) marks that grid’s corresponding location on the surface.



- captures a wide range of associations both functional and not, and for functional relationships provides a score that roughly equals the coefficient of
determination (R^2) of the data relative to the regression function.
- gives rise to a larger family of statistics, which we refer to as MINE, or maximal
information-based nonparametric exploration. MINE statistics can be used not only to identify interesting associations, but also to characterize them according to properties such as nonlinearity and monotonicity
- in [0, 1}
- symmetric
- invariant under order-preserving transformations
- A characteristic matrix with a high maximum indicates a strong relationship, a symmetric characteristic matrix indicates a monotonic relationship. We can thus detect deviation from monotonicity with the maximum asymmetry score (MAS), defined as the maximum over M of |m_x,y − m_y,x|. MAS is useful, for example, for detecting periodic relationships with unknown frequencies that vary over time

More formally:

For a grid G, let I_G denote the mutual information of the probability distribution induced on the boxes of G, where the probability of a box is

proportional to the number of data points falling inside the box. The (x,y)-th entry m_x,y of the characteristic matrix
equals max{I_G}/log min{x,y}, where the maximum is taken over all x-by-y grids G. MIC is the maximum of m_x,y over ordered pairs (x,y) such that xy < B, where B is a function of sample size; we usually set B=n^0.6

▼ What can you use yourself?

- MIC
- Implementation: minepy.MINE

▼ What other references do you want to follow?

- **Distance Correlation:** G. Székely, M. Rizzo, Ann. Appl. Stat. 3, 1236 (2009)
- **Maximal Correlation:** L. Breiman, J. H. Friedman, J. Am. Stat. Assoc. 80, 580
  (1985).
- **Principal curve-based methods:** “Principal curve-based methods” refers to mean-squared error relative to the principal curve, and CorGC, the principal curve-based measure of dependence of Delicado et al.