Principal Curves における MIC の活用について

北海道大学情報基盤センター　水田正弘

１．はじめに

近年、非線形な構造に対する「相関係数」として多くの指標が提案されている。多次元データ解析における多くの手法では、相関係数が基本になっていることを勘案すると、これらの指標は、非線形なデータ解析法の構築のための有用な道具となる可能性がある。そこで、非線形な「相関係数」として MIC (Maximal Information Coefficient)、データ解析の手法として Principal Curves を例にとり検討する。

２．MIC (Maximal Information Coefficient)について

2011 年に Reshef 他により提案された MIC は、21 世紀の相関係数として注目されている。2 つの変数に「関係」がある場合、データが存在している領域を適切なメッシュで区切ったとき、領域に含まれるデータの個数にバラツキが存在する。このバラツキを相互情報量で評価する。ただし、メッシュの個数にはデータ数に依存した制限を加えている。MIC は 0 から 1 の値を取り、1 に近いほど、変数間の関連性が高いと判断される。MIC の問題点の指摘や他の「相関係数」も研究・提案されている。

３．Principal Curves について

Principal Curves は、多変量データの分布の中央を通る、媒介変数表現された非線形な曲線である。サンプルに対する Principal Curves は、初期曲線を設定した後、データ点から最も近い曲線上の点を求める Projection Step と、曲線上に射影されるデータ点の平均(正確には条件付期待値)を求める Expectation Step を繰り返す。指定すべき複数のパラメータがあり、適切に収束しないことも多い。データ点と曲線との距離の二乗により当てはまりの良さを評価することがある。その値が悪い場合、手法の収束が不適切なのか、データに対してあてはまる曲線が本質的に存在しないのかを判断することが困難である。そこで、MIC を利用してその判断をする方法を考察する。

参考文献

[1] David N. Reshef, Yakir A. Reshef, Hilary K. Finucane, Sharon R. Grossman, Gilean McVean (2011) Detecting Novel Associations in Large Data Sets, *Science Vol. 334, Issue 6062, pp. 1518-1524 DOI: 10.1126/science.1205438*

[2] Trevor Hastie, Werner Stuetzle (1989) Principal Curves, *Journal of the American Statistical Association*, 84:406, 502-516

[3] 水田正弘,馬場康維 Mizuta, M. (1993) Principal Curves と数量化III類を用いた質的データの１次元構造の抽出、統計数理41, 1-11.