Sensitivity of centrality measures to estimation of adjacency structure: a study of mutual information estimators

M. Vejmelka J. Hlinka D. Hartman M. Paluš

Institute of Computer Science, Academy of Sciences of the Czech Republic,
Pod vodárenskou věží 2, Prague 8, Czech Republic

This study was supported by the Czech Science Foundation project No. P103/11/J068.

Complex network analysis: centrality measures

In network analysis, multiple measures of centrality have been proposed [3] to characterize the importance of a node with respect to the rest of the network. One of these is betweenness centrality (BC), which is a quantity reflecting the number of shortest paths between all pairs of nodes in the network, which pass through a given node. More precisely, the betweenness centrality of a node is computed using the formula

$$C_B(n) = \sum_{i,j \in V, i \neq j, i \neq n, j \neq n} \frac{\sigma_{ij}(n)}{\sigma_{ij}}, \qquad (1)$$

where V is the set of nodes, σ_{ij} is the number of shortest paths between the nodes i and j and $\sigma_{ij}(n)$ is the number of shortest paths from i to j which also pass through n.

Constructing networks from multivariate time series

When investigating a complex process, we can usually observe time series of relevant quantities at different locations but cannot directly map all interactions within the process. The first problem in network analysis in such cases is to build the network. The standard way of constructing unweighted undirected networks proceeds by computing the dependencies (connectivities) between the time series corresponding to each node and converting the strongest dependencies (in this work the top 0.5%) into edges.

Estimating dependency using mutual information

Mutual information (MI) quantifies statistical dependencies between variables. Many estimators of MI based on different features of time series have been proposed. Among these, we study those based on equiquantal binning (EQQ) [6], equidistant binning (EQD) [5] and permutation entropy (PERM) [1]. Mutual information between processes X and Y using the EQQ, EQD and PERM estimators is computed according to the formula:

$$I(X, Y) = \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p_{xy} \log \frac{p_{xy}}{p_x p_y}, \qquad (2)$$

where $\mathcal{X}, \mathcal{Y}$ are the sets of bins and p_{xy} is the probability that process X is in bin x and simultaneously Y is in bin y, p_x and p_y are defined similarly. The estimators differ among themselves in the procedure of determining bins corresponding to each sample of the time series. Details can be found in the respective referenced publications. We also compare selected results to those of the k-nearest neighbors (kNN) estimator [4].

Sensitivity analysis of BC on NCEP/NCAR data

We have analyzed the monthly mean surface air temperatures (756 months, from Jan 1948 to Dec 2010) from the NCEP/NCAR reanalysis project [2]. The North and South pole have been removed from the data and the series were converted into temperature anomalies by subtracting the respective monthly means from each month. The construction of the network from the NCEP/NCAR data proceeded using selected estimators of MI according to the procedure described above. Two types of tests were conducted to examine the reliability and agreement of BC computed from networks estimated using different MI estimators:

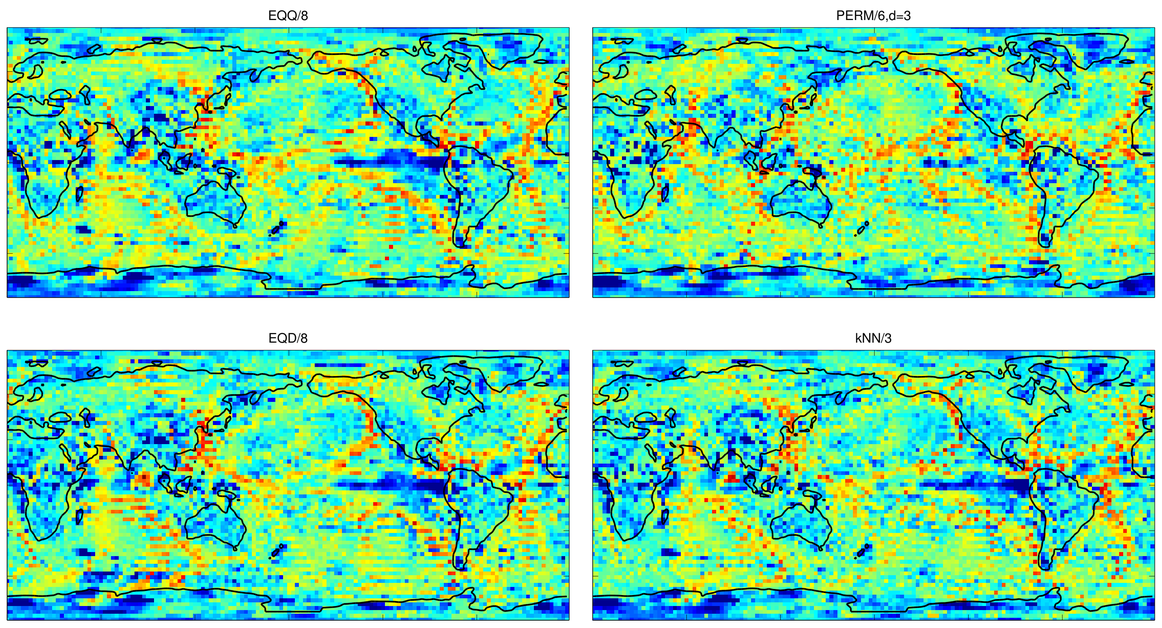
- using bootstrap re-sampling of the original data,
- using different levels of white gaussian noise added to the time series.

Bootstrap testing results

	EQQ/4	EQQ/8	EQQ/16	EQD/4	EQD/8	EQD/16
EQQ/4	0.63	0.72	0.64	0.43	0.59	0.61
EQQ/8	0.61	0.65	0.74	0.43	0.6	0.64
EQQ/16	0.58	0.63	0.63	0.42	0.57	0.61
EQD/4	0.41	0.41	0.4	0.5	0.58	0.53
EQD/8	0.53	0.54	0.53	0.48	0.6	0.72
EQD/16	0.55	0.57	0.55	0.46	0.58	0.6

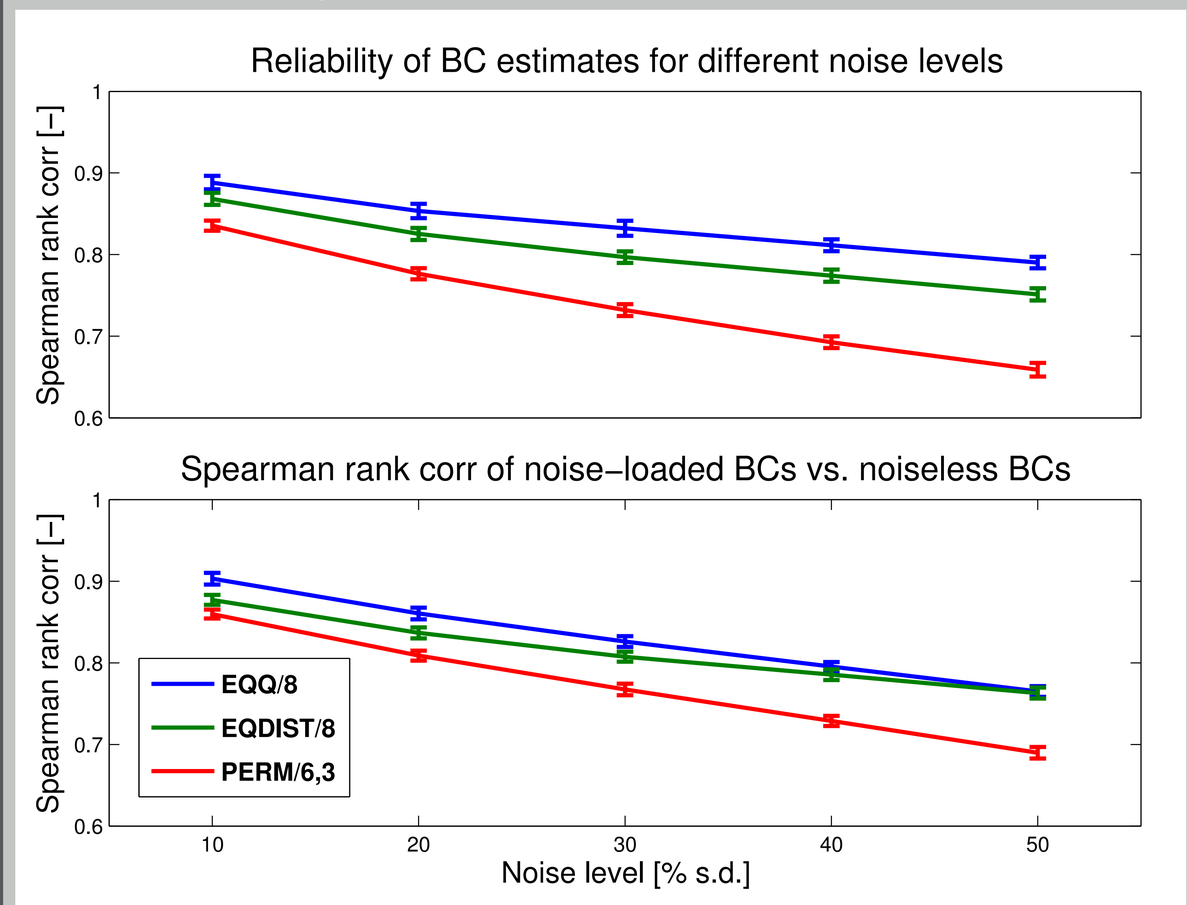
The diagonal contains average pairwise Spearman rank correlations of BCs computed from 100 bootstrap samples. The upper triangle contains correlations between pairs of BC vectors computed from the same bootstrap realization using different MI estimators and the lower triangle contains correlations between different bootstrap realizations and different MI estimators (which can be compared to the diagonal). All s.d. ≈ 0.01.

Betweenness centrality from different estimators of MI



Additive white Gaussian noise results

After standardizing each time series in the NCEP/NCAR data, 10%-50% (s.d.) of white Gaussian noise were added to all timeseries to investigate the reliability of the resulting BC estimates. For each noise level 100 realizations were generated and processed.



Conclusions

- Reliability of the BCs computed from different bootstrap samples is surprisingly low. Studies interpreting differences between different MI estimators (or even other dependency measures) should consider this effect.
- BC values in the NCEP/NCAR network constructed from estimates of dependence are quite reliable with respect to perturbations in the data modeled by additive white gaussian noise.
- For finite samples, there may be large differences between estimates of dependence using different MI estimators.

References

[1] **C. Bandt and B. Pompe**. Permutation entropy: A natural complexity measure for time series. *Physical Review Letters*, 88:174102, 2002.

[2] **R. Kistler et al**. The NCEP-NCAR 50-year reanalysis: Monthly means CD-ROM and documentation. *Bulletin of the American Meteorological Society*, 82:247–268, 2001.

[3] **L. C. Freeman**. Centrality in social networks: Conceptual clarification. *Social Networks*, 1(3):215–239, 1979.

[4] **A. Kraskov, H. Stögbauer, and P. Grassberger**. Estimating mutual information. *Physical Review E*, 69:066138, 2004.

[5] **R. Moddemeijer**. On estimation of entropy and mutual information of continuous distributions. *Signal Processing*, 16(3):233–246, 1989.

[6] **M. Paluš**. Testing for nonlinearity using redundancies: Quantitative and qualitative aspects. *Physica D*, 80:186–205, 1995.

http://ndw.cs.cas.cz/ {vejmelka,hlinka,hartman,palus}@cs.cas.cz