Reliable estimation of causally conditioned entropy from multivariate time-series data

MATTEK9 or 9+10 project proposal.

It is generally hard to assess the causal dependency and cross-dependency within and between observed variables, since correlation does not imply causation [1]. For the same reason, standard statistical tests are usually insufficient when establishing causal relationships and it could be necessary to construct probabilistic models [2,3].

The objective of this project is to better understand the causal relationships within observed data. In particular, the goal is to be able to reliably estimate quantities such as causally conditioned entropy and directed information from data measurements. These measures can, for example, be used to quantify the directional information flow in general feedback systems [4] such as networked control systems [5] or the human brain [6]. In the former case, the minimum amount of information required to e.g., stabilize an unable system over a network can be assessed [5]. In the latter case, the neuronal activity due to information processing in the brain as a response to external stimuli can be indirectly assessed from EEG measurements [6].

The mutual information, the directed information, and their causally conditioned counterparts, can all be expressed as a difference of entropies. In particular, let X and Y be random variables that both have well-defined probability density functions, then their mutual information can be written as a function of non-conditional entropies, i.e.:

(1) $I(X;Y) = h(X) - h(X|Y) = h(X) + h(Y) - h(X,Y),$

where h(.) denotes the differential entropy. Similar trick applies to other informational measures. Thus, it is sufficient to have a good estimator of entropy in order to be able to get a good estimate of related informational measures.

Proposed project scope:

1) The k-nearest neighbor (KNN) entropy estimator is widely used in theory and practice, and known to be asymptotically efficient for a range of input signals and under somewhat mild technical conditions [7,8]. Unfortunately, it is also known that the KNN has a bias, which at least for finite time series, to a certain degree depends upon the distribution and the dimensionality of the input. From (1), we notice that the entropy estimator needs to be applied on data of different dimensionality, since h(X), and h(X,Y) are both needed. Due to differences in the bias, the resulting estimate of the difference (1), can become negative too inaccurate.

2) The KNN estimator also appears to be unreliable when used on high-dimensional data series with a small number of samples [7,8]. In our case, we could for example focus on EEG data and exploit the distribution of the data to improve the estimator. One way to approach the problem could be to change approximation of the local distribution around sample points in the KNN estimator as was done in [7]. Specifically, in [7], a Gaussian assumption was utilized. Depending upon the particular type of EEG data, other distributions might be more suitable.

3) The KNN estimator uses either a uniform or non-uniform embedding. It was recently shown that a non-uniform embedding is much better when having short time series [9]. There is no exact theory for chosing the best embedding, and one could possible devise one that better exploits the structure and dependencies within the data. The work of [9] is documented by a Matlab toolbox. This toolbox provides state-of-the-art knn-based entropy estimation on EEG signals.

Supervisors:

- Jan Østergaard (jo@es.aau.dk), Department of Electronic Systems
- Rasmus Waagepetersen (rw@math.aau.dk), Institut for matematiske fag.

References:

1. C.Granger, “Investigating causal relations by econometric models and cross-spectral methods,”*Econometrica*, vol. 37, no. 3, pp. 424 – 438, 1969.
2. P. Suppes, *A Probabilistic Theory of Causality*. North-Holland, Amsterdam, 1970.
3. J. E. H. R. Motwani and J. D. Ullman, *Introduction to Automata Theory, Languages, and Computation*. Addison-Wesley, 2nd ed., 2001.
4. M. S. Derpich, E. I. Silva, and J. Østergaard, "Fundamental Inequalities and Identities Involving Mutual and Directed Informations in Closed-Loop Systems". Electronically available on arxiv.org
5. E. I. Silva, M. S. Derpich, J. Østergaard, and M.A. Encina, "A Characterization of the Minimal Average Data Rate that Guarantees a Given Closed-Loop Performance"**,** IEEE Transactions on Automatic Control, Vol. 61, Issue 8, pp. 2171 - 2186, 2016.
6. M. Wibral, R. Vicente, J.T. Lizier, “Directed information measures in neuroscience”, Springer, 2014.
7. Damiano Lombardi, Sanjay Pant. A non-parametric k-nearest neighbor entropy estimator. Physical Reviev E, 2016, 10.1103/PhysRevE.93.013310 . hal-01272527
8. S. Singh, B. Poczos, “Analysis of the k-nearest neighbor distances with application to entropy estimation”. Electronically available on arxiv.org.
9. http://www.lucafaes.net/code/ITS_Toolbox_v2.0.pdf