Information Theoretical Estimators Toolbox*

Zoltán Szabó
Gatsby Computational Neuroscience Unit, CSML, University College London
zoltan.szabo@gatsby.ucl.ac.uk

Since the pioneering work of Shannon, *entropy*, *mutual information*, *association*, *divergence* measures and *kernels on distributions* have found a broad range of applications in many areas of machine learning. Entropies provide a natural notion to quantify the *uncertainty* of random variables, mutual information and association indices measure the *dependence* among its arguments, divergences and kernels offer efficient tools to define the 'distance' and the inner product of probability measures, respectively.

Recent advances in the field have shown that many of these quantities can be estimated *nonparametrically* in a *statistically consistent* way. Nonparametric nature is highly important since our goal is to estimate *functionals* of distributions and *not* the underlying distributions, densities. From practical point of view this means that the performance of nonparametric methods is becoming superior to plug-in type approaches as the dimension is increasing. For certain kernel based estimators it can also be proved that their efficiency is in fact *dimension independent*.

An other challenge one has to face with is invariance to monotone increasing transformations of the random variables. Indeed, for example in feature selection it is highly desirable to construct and apply estimators that do not depend on the *unit* chosen, i.e., to achieve *scale-invariance*. As it has been shown recently, one can formulate consistent estimators (i) possessing the dimensional independent and the scale-invariance properties simultaneously, (ii) that are also *robust to outliers*.

There exist numerous relations among information theoretical quantities. One of the most well-known such relation is that mutual information can be expressed in terms of Shannon entropy, or Kullback-Leibler divergence. From estimation point of view, such relations (once identified) can be readily used to build *meta* estimators from *base* ones.

Despite the large number of successful applications and the recent theoretical contributions, existing software packages in the domain either focus on discrete variables, or quite specialized applications/information theoretical estimation methods. Such limitations of the available solutions highly encumber their reusage, restrict their extension with the latest advances, and put restraints to collaborative development.

In order to cope with these challenges, we have recently released the ITE (information theoretical estimators) toolbox. ITE

- is designed to be *modular* and highly *reusable*:
  - Beside the large number of covered state-of-the-art nonparametric techniques, it enables the user to easily construct novel estimators from existing ones (meta estimators) or from scratch (base ones).
  - The unified design of ITE makes it possible to immediately and transparently apply the obtained estimators (base/meta) in information theoretical optimization problems.
  - ITE also provides a prototype application in independent subspace analysis and its extensions. The application demonstrates how different information theoretical objective functions and their optimization can be handled in a high-level way.
- is multi-platform, written in Matlab/Octave, open source and released under the GPLv3 ($\geq$) license.
- is available on Bitbucket (https://bitbucket.org/szzoli/ite/) endowed with the Mercurial distributed source control management tool enabling collaborative development. ITE is linked on MLOSS (http://mloss.org/software/view/428/).

*This work was supported by the Gatsby Charitable Foundation. The research was carried out as part of the EITKIC_12-1-2012-0001 project, which is supported by the Hungarian Government, managed by the National Development Agency, financed by the Research and Technology Innovation Fund and was performed in cooperation with the EIT ICT Labs Budapest Associate Partner Group. (www.ictlabs.elte.hu). The Project was supported by the European Union and co-financed by the European Social Fund (grant agreement no. TÁMOP 4.2.1/B-09/1/KMR-2010-0003).