Journal of Machine Learning Research 15 (2014) 283-287 Submitted 10/12; Revised 9/13; Published 1/14

Information Theoretical Estimators Toolbox

Zoltán Szabó[*] ZOLTAN.SZABO@GATSBY.UCL.AC.UK
Gatsby Computational Neuroscience Unit
Centre for Computational Statistics and Machine Learning
University College London
Alexandra House, 17 Queen Square, London - WC1N 3AR

Editor: Balázs Kégl

Abstract

We present ITE (information theoretical estimators) a free and open source, multi-platform, Matlab/Octave toolbox that is capable of estimating many different variants of entropy, mutual information, divergence, association measures, cross quantities, and kernels on distributions. Thanks to its highly modular design, ITE supports additionally (i) the combinations of the estimation techniques, (ii) the easy construction and embedding of novel information theoretical estimators, and (iii) their immediate application in information theoretical optimization problems. ITE also includes a prototype application in a central problem class of signal processing, independent subspace analysis and its extensions.

Keywords: entropy-, mutual information-, association-, divergence-, distribution kernel estimation, independent subspace analysis and its extensions, modularity, Matlab/Octave, multi-platform, GNU GPLv3 ($\geq$)

1. Introduction

Since the pioneering work of Shannon (1948), *entropy*, *mutual information*,[1] *association*, *divergence* measures and *kernels on distributions* have found a broad range of applications in many areas of machine learning (Beirlant et al., 1997; Wang et al., 2009; Villmann and Haase, 2010; Basseville, 2013; Póczos et al., 2012; Sriperumbudur et al., 2012). 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.

A central problem based on information theoretical objectives in signal processing is independent subspace analysis (ISA; Cardoso 1998), a cocktail party problem with *independent groups*. One of the most relevant and fundamental hypotheses of the ISA research

[*] The authors would like to thank the anonymous reviewers for their valuable suggestions. 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).

[1] The Shannon mutual information is also known in the literature as the special case of total correlation or multi-information when two variables are considered.

is the *ISA separation principle* (Cardoso, 1998): the ISA task can be solved by ICA (ISA with one-dimensional sources, Hyvärinen et al. 2001; Cichocki and Amari 2002; Choi et al. 2005) followed by clustering of the ICA elements. This principle (i) forms the basis of the state-of-the-art ISA algorithms, (ii) can be used to design algorithms that scale well and efficiently estimate the dimensions of the hidden sources, (iii) has been recently proved (Szabó et al., 2007) and (iv) can be extended to different linear-, controlled-, post nonlinear-, complex valued-, partially observed systems, as well as to systems with nonparametric source dynamics. For a recent review on the topic and ISA applications, see Szabó et al. (2012).

Although there exist many exciting applications of information theoretical measures, to the best of our knowledge, available packages in this domain focus on (i) discrete variables, or (ii) quite specialized applications/information theoretical estimation methods. Our **goal** is to fill this serious gap by coming up with a (i) highly modular, (ii) free and open source, (iii) multi-platform toolbox, the ITE (information theoretical estimators) package, which focuses on *continuous* variables and

1. is capable of estimating *many* different kind of entropy, mutual information, association, divergence measures, distribution kernels based on nonparametric methods.[2]
2. offers a *simple and unified framework* to (i) easily construct new estimators from existing ones or from scratch, and (ii) transparently use the obtained estimators in information theoretical optimization problems,
3. with a *prototype application* in ISA and its extensions.

2. Library Overview

Below we provide a brief overview of the ITE package:

Information Theoretical Measures: The ITE toolbox is capable of estimating numerous important information theoretical quantities including

Entropy: Shannon-, Rényi-, Tsallis-, complex-, Φ-, Sharma-Mittal entropy,

Mutual information: generalized variance, kernel canonical correlation analysis, kernel generalized variance, Hilbert-Schmidt independence criterion, Shannon-, L_2-, Rényi-, Tsallis-, Cauchy-Schwartz quadratic-, Euclidean distance based quadratic-, complex-, χ^2 mutual information; copula-based kernel dependency, multivariate version of Hoeffding's Φ, Schweizer-Wolff's σ and κ, distance covariance and correlation, approximate correntropy independence measure,

Divergence: Kullback-Leibler-, L_2-, Rényi-, Tsallis-, Cauchy-Schwartz-, Euclidean distance based-, Jensen-Shannon-, Jensen-Rényi-, Jensen-Tsallis-, K-, L-, Pearson χ^2-, f-divergences; Hellinger-, Bhattacharyya-, energy-, (non-)symmetric Bregman-, J-distance; maximum mean discrepancy,

Association measure: multivariate (conditional) extensions of Spearman's ρ, (centered) correntropy, correntropy induced metric, correntropy coefficient, centered correntropy induced metric, multivariate extension of Blomqvist's β, lower and upper tail dependence via conditional Spearman's ρ,

Cross quantity: cross-entropy,

[2] It is highly advantageous to apply nonparametric approaches: the 'opposite' plug-in type methods—estimating the underlying densities—scale poorly as the dimension is increasing.

Distribution kernel: expected-, Bhattacharyya-, probability product-, (exponentiated) Jensen-Shannon-, (exponentiated) Jensen-Tsallis-, exponentiated Jensen-Rényi kernel.

Independent Process Analysis (IPA): ITE offers solution methods for independent subspace analysis (ISA) and its extensions to different linear-, controlled-, post nonlinear-, complex valued-, partially observed systems, as well as to systems with nonparametric source dynamics; combinations are also possible. The solutions are based on the ISA separation principle and its generalizations (Szabó et al., 2012).

Quick Tests: Beyond IPA, ITE provides quick tests to study the efficiency of the estimators. These tests cover (i) analytical value vs. estimation, (ii) positive semi-definiteness of Gram matrices defined by distribution kernels and (iii) image registration problems.

Modularity: The core idea behind the design of ITE is modularity. The modularity is based on the following four pillars:

1. The estimation of many information theoretical quantities can be reduced to k-nearest neighbor-, minimum spanning tree computation, random projection, ensemble technique, copula estimation, kernel methods.
2. The ISA separation principle and its extensions make it possible to decompose the solutions of the IPA problem family to ICA, clustering, ISA, AR (autoregressive)-, ARX- (AR with exogenous input) and mAR (AR with missing values) identification, gaussianization and nonparametric regression subtasks.
3. Information theoretical identities can relate numerous entropy, mutual information, association, cross- and divergence measures, distribution kernels (Cover and Thomas, 1991).
4. ISA can be formulated via information theoretical objectives (Szabó et al., 2007):

$$J_{\text{I}}(\mathbf{P}) = I\left(\mathbf{y}^1, \ldots, \mathbf{y}^M\right), \qquad J_{\text{Irecursive}}(\mathbf{P}) = \sum_{m=1}^{M-1} I\left(\mathbf{y}^m, \left[\mathbf{y}^{m+1}, ..., \mathbf{y}^M\right]\right),$$

$$J_{\text{sumH}}(\mathbf{P}) = \sum_{m=1}^{M} H\left(\mathbf{y}^m\right), \qquad J_{\text{sum-I}}(\mathbf{P}) = -\sum_{m=1}^{M} I\left(y_1^m, ..., y_{d_m}^M\right),$$

$$J_{\text{Ipairwise}}(\mathbf{P}) = \sum_{m_1 \neq m_2} I\left(\mathbf{y}^{m_1}, \mathbf{y}^{m_2}\right), \ J_{\text{Ipairwise1d}}(\mathbf{P}) = \sum_{m_1 \neq m_2}^{M} \sum_{i_1=1}^{d_{m_1}} \sum_{i_2=1}^{d_{m_2}} I\left(y_{i_1}^{m_1}, y_{i_2}^{m_2}\right),$$

where the minimizations are w.r.t. the optimal clustering ($\mathbf{P}$) of the ICA elements.

Dedicated Subtask Solvers, Extension: The ITE package offers dedicated solvers for the obtained subproblems detailed in '*Modularity*:1-2'. Thanks to this flexibility, extension of ITE can be done effortlessly: it is sufficient to add a new switch entry in the subtask solver.

Base and Meta Estimators: One can *derive* new, *meta* (entropy, mutual information, association, divergence, cross quantity, distribution kernel) estimators in ITE from existing base or meta ones by '*Modularity*:3'. The calling syntax of base and meta methods are completely identical thanks to the underlying unified template structure followed by the estimators. The ITE package also supports an indicator for the importance of multiplicative constants.[3]

We illustrate how easily one can estimate information theoretical quantities in ITE:

```
>Y1 = rand(3,1000); Y2 = rand(3,2000);  %data of interest
>mult = 1;                   %multiplicative constant is important
>co = D_initialization('Jdistance',mult);%initialize the estimator
>D = D_estimation(Y1,Y2,co);            %estimation
```

Next, we demonstrate how one can construct meta estimators in ITE. We consider the definitions of the initialization and the estimation of the J-distance. The KL-divergence, which is symmetrised in J-distance, is estimated based on the existing k-nearest neighbor technique.

```
function [co] = DJdistance_initialization(mult)
co.name = 'Jdistance';          %name of the estimator
co.mult = mult;                 %importance of multiplicative const.
co.member_name = 'KL_kNN_k';    %method used for KL estimation
co.member_co = D_initialization(co.member_name,mult); %initialization
```

```
function [D_J] = DJdistance_estimation(X,Y,co)
D_J = D_estimation(X,Y,co.member_co) + D_estimation(Y,X,co.member_co);
```

ISA Objectives and Optimization: Due to the unified syntax of the estimators, one can formulate and solve information theoretical optimization problems in ITE in a high-level view. Our example included in ITE is ISA (and its extensions) whose objective can be expressed by entropy and mutual information terms, see '*Modularity*:4'. The unified template structure in ITE makes it possible to use *any* of the estimators (base/meta) in these cost functions.

A further attractive aspect of ITE is that even in case of unknown subspace dimensions, it offers well-scaling approximation schemes based on spectral clustering methods. Such methods are (i) robust and (ii) scale excellently, a single general desktop computer can handle about a million observations—in our case estimated ICA elements—within several minutes (Yan et al., 2009).

3. Availability and Requirements

The ITE package is *self-contained*, it only needs a Matlab or an Octave environment[4] with standard toolboxes. ITE is *multi-platform*, it has been extensively tested on Windows and Linux; since it is made of standard Matlab/Octave and C++ files, it is expected to work on alternative platforms as well.[5] ITE is released under the free and open source GNU GPLv3 ($\geq$) license. The accompanying source code and the documentation of the toolbox has been enriched with numerous comments, examples, detailed instructions for extensions, and pointers where the interested user can find further mathematical details about the embodied techniques. The ITE package is available at `https://bitbucket.org/szzoli/ite/`.

[3] In many applications, it is completely irrelevant whether we estimate, for example, $H(\mathbf{y})$ or $cH(\mathbf{y})$, where $c = c(d)$ is a constant depending only on the *dimension* of $\mathbf{y} \in \mathbb{R}^d$ (d), but *not on the distribution* of $\mathbf{y}$. Such 'up to proportional factor' estimations can often be carried out more efficiently.

[4] See `http://www.mathworks.com/products/matlab/` and `http://www.gnu.org/software/octave/`.

[5] On Windows (Linux) we suggest using the Visual C++ (GCC) compiler.

References

Michéle Basseville. Divergence measures for statistical data processing - an annotated bibliography. *Signal Processing*, 93:621–633, 2013.

Jan Beirlant, Edward J. Dudewicz, László Győrfi, and Edward C. van der Meulen. Nonparametric entropy estimation: An overview. *International Journal of Mathematical and Statistical Sciences*, 6:17–39, 1997.

Jean-François Cardoso. Multidimensional independent component analysis. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 1941–1944, 1998.

Seungjin Choi, Andrzej Cichocki, Hyung-Min Park, and Soo-Yound Lee. Blind source separation and independent component analysis. *Neural Information Processing - Letters and Reviews*, 6:1–57, 2005.

Andrzej Cichocki and Shun-ichi Amari. *Adaptive Blind Signal and Image Processing*. John Wiley & Sons, 2002.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley and Sons, New York, USA, 1991.

Aapo Hyvärinen, Juha Karhunen, and Erkki Oja. *Independent Component Analysis*. John Wiley & Sons, 2001.

Barnabás Póczos, Zoubin Ghahramani, and Jeff Schneider. Copula-based kernel dependency measures. In *International Conference on Machine Learning*, pages 775–782, 2012.

Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948.

Bharath K. Sriperumbudur, Kenji Fukumizu, Arthur Gretton, Bernhard Schölkopf, and Gert R. G. Lanckriet. On the empirical estimation of integral probability metrics. *Electronic Journal of Statistics*, 6:1550–1599, 2012.

Zoltán Szabó, Barnabás Póczos, and András Lőrincz. Undercomplete blind subspace deconvolution. *Journal of Machine Learning Research*, 8:1063–1095, 2007.

Zoltán Szabó, Barnabás Póczos, and András Lőrincz. Separation theorem for independent subspace analysis and its consequences. *Pattern Recognition*, 45:1782–1791, 2012.

Thomas Villmann and Sven Haase. Mathematical aspects of divergence based vector quantization using Fréchet-derivatives. Technical report, University of Applied Sciences Mittweida, 2010.

Quing Wang, Sanjeev R. Kulkarni, and Sergio Verdú. Divergence estimation for multidimensional densities via k-nearest-neighbor distances. *IEEE Transactions on Information Theory*, 55:2392–2405, 2009.

Donghui Yan, Ling Huang, and Michael I. Jordan. Fast approximate spectral clustering. In *International Conference on Knowledge Discovery and Data Mining*, pages 907–916, 2009.