Robust Nonparametric Copula Based Dependence Estimators

Barnabás Póczos
School of Comp. Sci.
Carnegie-Mellon University
Pittsburgh, PA 15213
bapoczos@cs.cmu.edu

Sergey Krishner
Dept. of Statistics
Purdue University
West Lafayette, IN, 47907
skirshne@purdue.edu

Dávid Pál
Google, Inc.
New York, NY 10011
dpal@google.com

Csaba Szepesvári
Dept of Comp. Sci.
University of Alberta
Edmonton, AB, Canada
szepesva@ualberta.ca

Jeff Schneider
School of Comp. Sci.
Carnegie-Mellon University
Pittsburgh, PA 15213
schneide@cs.cmu.edu

A fundamental problem in statistics is the estimation of dependence between random variables. While information theory provides standard measures of dependence (e.g. Shannon-, Rényi-, Tsallis-mutual information (MI)), it is still unknown how to estimate these quantities from i.i.d. samples in the most efficient way. Dependence estimators have numerous applications in real-world problems. Among others, they have been used in feature selection [1], clustering [2], causality detection [3], optimal experimental design [4, 5], fMRI data processing [6], prediction of protein structures [7], boosting, facial expression recognition [8], independent component and subspace analysis [9, 10, 11, 12], and image registration [13, 14, 15].

Density estimation over a high-dimensional domain is known to suffer from the curse of dimensionality. Therefore, it is of great importance to know which functionals of densities can be estimated efficiently in a *direct* way, *without* estimating the density. It has been shown that copula methods provide a natural framework to estimate MI in a consistent way. They completely *avoid density estimation* and only use *rank statistics*. This is an important property, which leads to remarkable robustness to outliers [16]. Upper bounds on the convergence rates have also been derived for these MI estimators [17]. It is somewhat surprising that MI can be consistently estimated using rank statistics only, since the same does not hold for the *less informative* Pearson correlation coefficient. Furthermore, with copula methods we can also define and estimate other dependence measures such as the Schweizer-Wolff σ measure (SW) [18]. Below we review these estimators [16, 17, 19, 20].

MI Estimators The Rényi MI of d real-valued random variables[1] $\mathbf{X} = (X^1, X^2, \dots, X^d)$ with joint density $f : \mathbb{R}^d \to \mathbb{R}$ and marginal densities $f_i : \mathbb{R} \to \mathbb{R}$, $1 \le i \le d$, is defined for any real parameter α using

$$I_\alpha(\mathbf{X}) \doteq I_\alpha(f) = \frac{1}{\alpha - 1} \log \int_{\mathbb{R}^d} f^\alpha(x^1, x^2, \dots, x^d) \left(\prod_{i=1}^d f_i(x^i) \right)^{1-\alpha} \mathrm{d}(x^1, x^2, \dots, x^d),$$

assuming the underlying integrals exist. By definition, $I_1 = \lim_{\alpha \to 1} I_\alpha$, which is the well-known Shannon MI. Given an i.i.d. sample $\mathbf{X}_{1:n} = (\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n)$ from a distribution with density f, our goal is to estimate $I_\alpha(\mathbf{X})$. The main idea we are going to use is that by means of a copula transformation we can reduce the MI estimation problem to estimating entropies, a problem that has been studied previously. The main observation is that

$$I_\alpha(\mathbf{X}) = I_\alpha(F_1(X^1), F_2(X^2), \dots, F_d(X^d)) = -H_\alpha(F_1(X^1), F_2(X^2), \dots, F_d(X^d)),$$

[1] We use superscript for indexing dimension coordinates.

where H_α stands for the Rényi entropy, and F_i is the cumulative distribution function (c.d.f.) of X^i. The problem is, of course, that F_i is not known and need to be estimated from the sample. To this end, we will use the empirical c.d.f.'s: $\widehat{F}_j(x) \doteq \frac{1}{n}|\{i : 1 \le i \le n,\ x \le X_i^j\}|$, for $x \in \mathbb{R},\ 1 \le j \le d$. Let $\mathbf{F}(x^1, x^2, \ldots, x^d) \doteq (F_1(x^1), F_2(x^2), \ldots, F_d(x^d))$ and $\widehat{\mathbf{F}}(x^1, x^2, \ldots, x^d) \doteq (\widehat{F}_1(x^1), \widehat{F}_2(x^2), \ldots, \widehat{F}_d(x^d))$. The joint distribution of $\mathbf{F}(\mathbf{X}) = (F_1(X^1), F_2(X^2), \ldots, F_d(X^d))$ and the sample $(\widehat{\mathbf{Z}}_1, \widehat{\mathbf{Z}}_2, \ldots, \widehat{\mathbf{Z}}_n) = (\widehat{\mathbf{F}}(\mathbf{X}_1), \widehat{\mathbf{F}}(\mathbf{X}_2), \ldots, \widehat{\mathbf{F}}(\mathbf{X}_n))$ are called the copula and empirical copula, respectively [21]. We estimate the Rényi mutual information I_α by $\widehat{I}_\alpha(\mathbf{X}_{1:n}) \doteq -\widehat{H}_\alpha(\widehat{\mathbf{Z}}_1, \widehat{\mathbf{Z}}_2, \ldots, \widehat{\mathbf{Z}}_n)$, where $\widehat{H}_\alpha$ is a Rényi entropy estimator, for which there are efficient methods available, for example k nearest-neighbor-graph based estimators [22], and Euclidean graph optimization algorithms [14, 23, 24]. The following theorem states that $\widehat{I}_\alpha$ is strongly consistent. Upper bounds on the rate of convergence can also be derived [19].

Theorem 1 (Consistency of $\widehat{I}_\alpha$)**.** *Let $d \ge 3$ and $\alpha = 1 - p/d \in (1/2, 1)$. Let μ be an absolutely continuous distribution over $\mathbb{R}^d$ with density f. If $\mathbf{X}_{1:n} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)$ is an i.i.d. sample from μ then*

$$\lim_{n\to\infty} \widehat{I}_\alpha(\mathbf{X}_{1:n}) = I_\alpha(f) \qquad \textit{a.s.}$$

Robustness Inspired by Tukey's finite-sample influence curve [25], define $\Delta_n(\boldsymbol{x}) \doteq |\widehat{I}_\alpha(\boldsymbol{X}_{1:n}, \boldsymbol{x}) - \widehat{I}_\alpha(\boldsymbol{X}_{1:n})|$, the amount of change caused in the estimate by adding a single observation $\boldsymbol{x}$ to the sample $\boldsymbol{X}_{1:n}$. We would like $\Delta_n(\boldsymbol{x}) = o(1)$ to hold a.s. independently of $\boldsymbol{x}$ as this indicates that the effect of a single sample becomes negligible as $n \to \infty$. We have the following result on the robustness of $\widehat{I}_\alpha$.

Theorem 2 (Robustness)**.** *When we use Euclidean graphs with the so-called smoothness property [23] (e.g. minimum spanning trees) for the entropy estimation after the empirical copula transformation, then $\Delta_n(\boldsymbol{x}) = O(n^{-\alpha})$ holds a.s., uniformly in $\boldsymbol{x}$.*

SW Estimators Here we show how the so-called "Schweizer-Wolff σ" can be estimated using empirical copulas. For simplicity, we present the estimator only for two variables; the extension to several random variables is straightforward. Let a pair of random variables $\left(X^1, X^2\right) \in \mathbb{R}^2$ be distributed according to a probability distribution with copula distribution $C(u,v) = \mathbb{P}(F_1(X^1) < u \wedge F_2(X^2) < v)$. The Schweizer-Wolff σ is defined as the L_1 distance between the copula C and the product copula $\Pi(u,v) \doteq uv$:

$$\sigma \doteq 12 \int_{\mathbf{I}^2} |C\left(u,v\right) - uv| \,\mathrm{d}u \,\mathrm{d}v.$$

The measure σ has a range of $[0,1]$, with an important property that $\sigma = 0$ if and only if the corresponding variables are mutually independent. Assume now that we are given N i.i.d. samples, $\mathbf{X}_{1:n} = (\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)$, where $X_i = (X_i^1, X_i^2)$. Our goal is to estimate σ using the sample $\mathbf{X}_{1:n}$.

Since the true copula C is not known, we estimate it again from the empirical copula C_N, i.e., the empirical c.d.f of $(\widehat{\mathbf{F}}(\mathbf{X}_1), \widehat{\mathbf{F}}(\mathbf{X}_2), \ldots, \widehat{\mathbf{F}}(\mathbf{X}_n))$. This is given by $C_N\left(\frac{i}{N}, \frac{j}{N}\right) = \frac{1}{N}\{\#\text{ of }\left(X_k^1, X_k^2\right) \text{ s.t. } X_k^1 \le X_i^1 \text{ and } X_k^2 \le X_j^2\}$. Using the empirical copula, a natural way to estimate σ is as follows:

$$s = \frac{12}{N^2 - 1} \sum_{i=1}^{N} \sum_{j=1}^{N} \left| C_N\left(\frac{i}{N}, \frac{j}{N}\right) - \frac{i}{N} \times \frac{j}{N} \right|. \tag{1}$$

In [20], this estimator was used for independent component analysis (ICA). To the best of our knowledge, this is currently the most robust ICA algorithm [20].

Numerical results In our presentation we will show applications on image registration, and independent subspace analysis. We will empirically demonstrate the robustness properties of the copula based estimators, and will compare them to other standard methods.

Finally, we note that there are other interesting dependence measures, such as the kernel mutual information [26] and the Székely's distance based correlation [27]. It would be important to know whether these dependence measures could be related to copula methods, as well.

References

[1] H. Peng and C. Ding. Feature selection based on mutual information: Criteria of max-dependency, max-relevance, and min-redundancy. *IEEE Trans On Pattern Analysis and Machine Intelligence*, 27, 2005.

[2] M. Aghagolzadeh, H. Soltanian-Zadeh, B. Araabi, and A. Aghagolzadeh. A hierarchical clustering based on mutual information maximization. In *in IEEE ICIP*, pages 277–280, 2007.

[3] K. Hlaváckova-Schindler, M. Paluŝb, M. Vejmelkab, and J. Bhattacharya. Causality detection based on information-theoretic approaches in time series analysis. *Physics Reports*, 441:1–46, 2007.

[4] J. Lewi, R. Butera, and L. Paninski. Real-time adaptive information-theoretic optimization of neurophysiology experiments. In *Advances in Neural Information Processing Systems*, volume 19, 2007.

[5] B. Póczos and A. Lőrincz. Identification of recurrent neural networks by Bayesian interrogation techniques. *Journal of Machine Learning Research*, 10:515–554, 2009.

[6] B. Chai, D. B. Walther, D. M. Beck, and L. Fei-Fei. Exploring functional connectivity of the human brain using multivariate information analysis. In *NIPS*, 2009.

[7] C. Adami. Information theory in molecular biology. *Physics of Life Reviews*, 1:3–22, 2004.

[8] C. Shan, S. Gong, and P. W. Mcowan. Conditional mutual information based boosting for facial expression recognition. In *British Machine Vision Conference (BMVC)*, 2005.

[9] E. Learned-Miller and J. W. Fisher. ICA using spacings estimates of entropy. *Journal of Machine Learning Research*, 4:1271–1295, 2003.

[10] B. Póczos and A. Lőrincz. Independent subspace analysis using geodesic spanning trees. In *ICML*, pages 673–680, 2005.

[11] M. M. Van Hulle. Constrained subspace ICA based on mutual information optimization directly. *Neural Computation*, 20:964–973, 2008.

[12] Z. Szabó, B. Póczos, and A. Lőrincz. Undercomplete blind subspace deconvolution. *Journal of Machine Learning Research*, 8:1063–1095, 2007.

[13] J. Kybic. Incremental updating of nearest neighbor-based high-dimensional entropy estimation. In *Proc. Acoustics, Speech and Signal Processing*, 2006.

[14] A. O. Hero, B. Ma, O. Michel, and J. Gorman. Applications of entropic spanning graphs. *IEEE Signal Processing Magazine*, 19(5):85–95, 2002.

[15] A. O. Hero, B. Ma, O. Michel, and J. Gorman. Alpha-divergence for classification, indexing and retrieval, 2002. Communications and Signal Processing Laboratory Technical Report CSPL-328.

[16] B. Póczos, S. Kirshner, and Cs. Szepesvári. REGO: Rank-based estimation of Rényi information using Euclidean graph optimization. In *AISTATS 2010*, 2010.

[17] D. Pál, B. Póczos, and Cs. Szepesvári. Estimation of Rényi entropy and mutual information based on generalized nearest-neighbor graphs. In *Proceedings of the Neural Information Processing Systems (NIPS)*, 2010.

[18] B. Schweizer and E. F. Wolff. On nonparametric measures of dependence for random variables. *The Annals of Statistics*, 9, 1981.

[19] D. Pál, Cs. Szepesvári, and B. Póczos. Estimation of Rényi entropy and mutual information based on generalized nearest-neighbor graphs, 2010. http://arxiv.org/abs/1003.1954, Technical Report of the NIPS 2010 paper.

[20] S. Kirshner and B. Póczos. ICA and ISA using Schweizer-Wolff measure of dependence. In *International Conference on Machine Learning (ICML)*, pages 464–471. ACM Press, 2008.

[21] J. Dedecker, P. Doukhan, G. Lang, J.R. Leon, S. Louhichi, and C Prieur. *Weak Dependence: With Examples and Applications*, volume 190 of *Lecture notes in Statistics*. Springer, 2007.

[22] N. Leonenko, L. Pronzato, and V. Savani. Estimation of entropies and divergences via nearest neighbours. *Tatra Mt. Mathematical Publications*, 39, 2008.

[23] J. E. Yukich. *Probability Theory of Classical Euclidean Optimization Problems*. Springer, 1998.

[24] J. M. Steele. *Probability Theory and Combinatorial Optimization*. Society for Industrial and Applied Mathematics, 1997.

[25] J. W. Tukey. *Exploratory Data Analysis. Mimeograph*. Addison-Wesley, 1970.

[26] A. Gretton, R. Herbrich, and A. Smola. The kernel mutual information. In *Proc. ICASSP*, 2003.

[27] G. Székely, M. Rizzo, and N. Bakirov. Measuring and testing dependence by correlation of distances. *Annals of Statistics*, 35(6):2769–2794, 2007.