Statistical independence measure based on maximum norm of joint and product-marginal characteristic functions

povilas.daniusis, povilasd@neurotechnology.com

November 2021

Abstract

In this paper we propose statistical independence measure based on the maximum norm of difference between joint and product of marginal characteristic functions. We discuss simulated examples, and applications for feature selection/extraction, causal inference and conduct corresponding empirical experiments with diverse collection of data sets from different domains.

1 Introduction

Estimation of statistical independence, both qualitatively and quantitatively, plays important role in various statistical and machine learning methods (e.g. hypothesis testing, feature selection and extraction [**?**], information bottleneck methods [**?**], cost function / reinforcement learning reward design, causal inference [**?**], among others). In this article we will focus on quantitative estimations of statistical independence. Therefore, earliest statistical dependence estimation ideas (e.g. conditional probability) likely share nearly-common origin with the beginning of formal statistical reasoning itself. During last two centuries ideas of correlation and (relative) entropy (including various generalizations) were proposed and became very popular in numerous applications and theoretical developments. However, with the increasing popularity of statistical machine learning, new statistical dependence estimation methods, that are robust, applicable to noisy, high-dimensional, structured data, and which can be efficiently integrated with modern machine learning methods are helpful for the development both of the theory and application.

In this study we will begin with the short review of some important previous dependence estimation approaches (Section 2), devoting special attention to ones based on characteristic functions (Section 2.1). Afterwards we formulate new characteristic function-based statistical dependence measure and its empirical estimator (Section 3), including its extension into reproducing kernel Hilbert spaces (RKHS'es), which are the main theoretical contribution of our paper. Section 4 is devoted to experiments with simulated and real data sets, and finalizing Section 5 concludes this article.

2 Previous work

Shannon mutual information [1] and generalizations [2], Hilbert-Schmidt independence criterion [3] and generalizatios [?], [4] copula-based kernel dependence measures.

2.1 Characteristic-function-based methods

Characteristic function of d_X-dimensional random vector X defined in some probability space $(\Omega_X, \Sigma_X, \mathbb{P}_X)$ is defined as

$$\phi_X(\alpha) = \mathbb{E}_X e^{i\alpha^T X}, \tag{1}$$

where $i = \sqrt{-1}$, $\alpha \in R^{d_X}$. Having n i.i.d. realisations of X, corresponding empirical characteristic function is defined as

$$\widehat{\phi_X}(\alpha) = \frac{1}{n}\sum_{j=1}^{n} e^{i<\alpha, x_j>}. \tag{2}$$

Having pair of two random vectors (X, Y) defined in another probability space $(\Omega_{X,Y}, \Sigma_{X,Y}, \mathbb{P}_{X,Y})$ joint characteristic function is defined as:

$$\phi_{X,Y}(\alpha, \beta) = \mathbb{E}_{X,Y} e^{i(\alpha^T X + \beta^T Y)}, \tag{3}$$

where $\alpha \in R^{d_X}$ and $\beta \in R^{d_Y}$. Similarly, having n i.i.d. realisations of (X, Y), joint empirical characteristic function is defined as

$$\widehat{\phi_{X,Y}}(\alpha, \beta) = \frac{1}{n}\sum_{j=1}^{n} e^{i(<\alpha, x_j> + <\beta, y_j>)} \tag{4}$$

In terms of characteristic functions, statistical independence of X and Y is equivalent to $\forall\alpha \in R^{d_x}, \forall\beta \in R^{d_y}$,

$$\phi_{X,Y}(\alpha, \beta) = \phi_X(\alpha)\phi_Y(\beta), \tag{5}$$

where d_y are dimensions of X and Y, respectively.

Among other applications, based on characteristic functions independence tests (e.g. [5]) and measures were proposed. For example, [6] proposes bivariate dependence test and measures (*distance covariance* and *distance correlation*) for random vectors based on weighted L^2-distance between joint characteristic function and product of marginal-ones. [7] generalises their work to multivariate case and proposes *distance multivariance* and derivative dependence measure, called *total distance multivariance*. Our motivation stems from the fact that evaluation of [6] measures in high dimensional cases may be prone to curse of dimensionality. Partial - [8]

3 Proposed Independence Measure

This motivates the construction of a novel dependence measure, which we further refer to as Kac independence measure (KacIM):

$$\kappa(X,Y) = \max_{\alpha\in\mathbb{R}^{d_X},\beta\in\mathbb{R}^{d_Y}} |\phi_{X,Y}(\alpha,\beta) - \phi_X(\alpha)\phi_Y(\beta)|. \tag{6}$$

In contrary to [6] the proposed (6) measure relies on maximum norm of difference between joint and product-marginal characteristic functions.

3.1 Properties

Theorem 1. *Independence measure* (6) *has the following properties:*

- $\kappa(X,Y) = \kappa(Y,X)$
- $0 \le \kappa(X,Y) \le 1.$

Proof. Proof Symmetry is obvious from definition (6) (commutativity of addition and multiplication), and second property directly follows from Cauchy inequality and that absolute value of characteristic function is bounded by 1:

$$|\phi_{X,Y}(\alpha,\beta) - \phi_X(\alpha)\phi_Y(\beta)|^2 = \mathbb{E}_{X,Y}|(e^{i\alpha^TX} - \phi_X(\alpha))(e^{i\beta^TY} - \phi_Y(\beta))|^2 =$$

$$\mathbb{E}_{X,Y}|(e^{i\alpha^TX} - \phi_X(\alpha))|^2|(e^{i\beta^TY} - \phi_Y(\beta))|^2 = (1-|\phi_X(\alpha)|^2)(1-|\phi_Y(\beta)|^2).$$

□

3.2 Estimation

Having i.i.d. standartized data (x_j, y_j), $j = 1, 2, ..., n$ an empirical scale-invariant estimator of KacIM (6) is defined via corresponding empirical characteristic functions (4) and (2):

$$\hat{\kappa}(X,Y) = \max_{\|\alpha\|=\|\beta\|=1} |\widehat{\phi_{X,Y}}(\alpha,\beta) - \widehat{\phi_X}(\alpha)\widehat{\phi_Y}(\beta)|. \tag{7}$$

Empirical estimator (7) also is symmetric and and bounded (Theorem 1) . Normalisation of parameters α and β on to unit sphere is included due to stability issues (really this is the reason?). The estimator (7) can be calculated by using Algorithm 1 [1].

In our implementation we use decoupled weight decay regularization optimizer [10].

[1] Pytorch [9] implementationcan be accessed from `https://github.com/povidanius/kac_independence_measure`

Algorithm 1 KacIM estimator computation algorithm

```
Require: data batch $(x, y)$, gradient-based optimiser $GradOpt(loss)$
  Normalize $(x, y)$ to zero mean and unit variance (scale invariance).
  Calculate KacIM estimator $\hat{\kappa}(x, y)$, without maximization step (i.e. using current $\alpha, \beta$).
  Perform one maximization iteration of computed $\hat{\kappa}(x, y)$ via $\alpha, \beta \to GradOpt(\hat{\kappa}(x, y))$.
```

3.3 Kernel version

Having two RKHS'es, defined by feature mapping $(x, y) \to (l(x, .), l(y, .)$, where $k : \mathbb{R}^{d_x} \times \mathbb{R}^{d_x} \to \mathbb{R}$ and $l : \mathbb{R}^{d_y} \times \mathbb{R}^{d_y} \to \mathbb{R}$ (see [11]). Then, estimation of kernel-KIM $(\hat{\kappa}_{k,l}(X, Y))$ can be reformulated as maximization of :

$$|\frac{1}{n}\sum_{j=1}^{n} e^{i(<\alpha,k(x_j,.)>+<\beta,l(y_j>,.))} - \frac{1}{n^2}\sum_{j=1}^{n} e^{i<\alpha,k(x_j,.)>}\sum_{k=1}^{n} e^{i<\beta,l(y_k,.)>}|, \tag{8}$$

and representer theorem[**?**] implies

$$\hat{\kappa}_{k,l}(X, Y) = \max_{\|\alpha\|=\|\beta\|=1} |\frac{1}{n} 1^T e^{i(\alpha^T K+\beta^T L)} - \frac{1}{n^2}(1^T e^{i(\alpha^T K)})(1^T e^{i(\beta^T L)})|, \tag{9}$$

where K and L are Gram matrices, corresponding to x_i and y_i. Note that the number of parameters of $\hat{\kappa}_{k,l}(X, Y)$ is dimension-idnependent and is equal to $2n_b$, where n_b is batch size. Also, kernel-KIM can be applied for structured data, via corresponding positive defined kernels.

4 Experiments

Dependence measures have board area of applications. For example, regularization [**?**, **?**], feature selection and extraction [12], information bottleneck methods [13], causal inference [14], among others. Further we will conduct empirical investigation of KacIM. Starting with simple illustrative simulations, we will reformulate some key ideas in aforementioned topics for KacIM, and experimentally investigate corresponding empirical scenarios.

4.1 Generated data

We begin with simple example, which demonstrates the efficiency of KacIM for simulated multivariate data with additive and multiplicative noise.

In Figure 4.1 reflects KacIM values during iterative adaptation (500 iterations). In the case of independent data, both x_i and y_i ($d_x = 1024$, $d_y = 32$) are sampled from gaussian distribution, independently. In the case of dependent data, an additive noise (left graph) and multiplicative noise (right graph), the

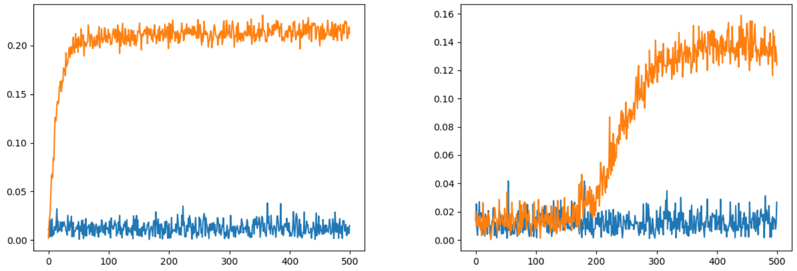

Figure 1: Dependence detection in additive (left) and multiplicative (right) noise scenarios.

dependent variable is generated according to $y_i = sin(Px_i) + cos(Px_i) + \lambda\epsilon_i$ ($\lambda = 0.15$) and $y_i = (sin(Px_i) + cos(Px_i))\epsilon_i$, respectively, where P is $d_x \times d_y$ random projection matrix, $\epsilon_i \sim N(0, 1)$ and $\epsilon_i \perp x_i$.

When data is independent (blue graph), both in additive and multiplicative cases, due to independence, estimator (7) is resistant to maximization, and oscillates near zero. On the other hand, when data is not independent (orange graph), the condition of Kac theorem is violated and maximization of estimator (7) is possible.

4.2 Influence of noise variance and data scale

Noise variance In this simulation we use the same additive noise setting as in previous paragraph, but evaluate all noise levels $\lambda \in [0.1, 3.0]$, with step 0.1. Figure 4.2 empirically shows that value of KacIM negatively correlates with noise level, and therefore the proposed measure is able not only to detect whether independence is present, but also to quantitatively evaluate it, which enables to use KacIM to derive cost functions for vairous other learning-based algorithms. In addition, we empirically

Scale invariance In data scale experiments we investigate behaviour of $\kappa(rx, ry)$, where r is scale parameter.

4.3 Feature Extraction

We conduct linear feature extraction by seeking

$$W^* = arg \max_W \kappa(Wx, y). \tag{10}$$

Afterwards, feature extraction is conducted by $f = W^*x$ and k-nearest neighbor classification with Euclidean distance is performed, comparing unmodified inputs x and features of all possible dimensions up to d_x.

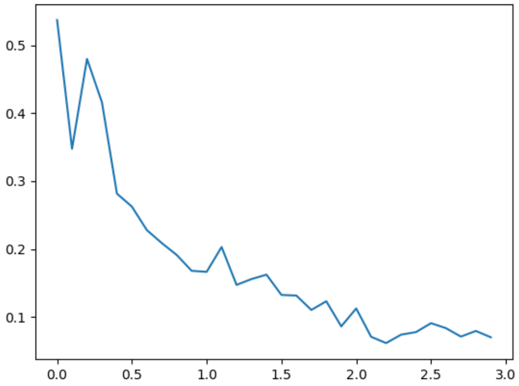

Figure 2: Noise level (x axis) vs final iteration KacIM value (y axis). KacIM values for larger noise levels saturates as in tail of graph.

5 Discussion

In this article we propose statistical dependence measure, KacIM, which relies on simple fact that statistical independence is equivalent to the decomposability of joint characteristic function into the product of marginal ones. Although we formulated and analysed KacIM for bivariate vectorial case, similarly it can be generalised for multivariate case. In addition, since characteristic functions are defined for matrices, graphs, and other objects [**?**], likely KacIM can be extended to those objects as well, which is potential direction of future research of KacIM.

Empirical analysis show, that KacIM can detect and measure statistical independence for non-linearly related, high-dimensional data. (...)

6 Acknowledgements

References

[1] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, July 2006.

[2] José M. Amigó, Sámuel G. Balogh, and Sergio Hernández. A brief review of generalized entropies. *Entropy*, 20(11), 2018.

[3] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring statistical dependence with hilbert-schmidt norms. In *ALT*, 2005.

[4] Barnabás Póczos, Zoubin Ghahramani, and Jeff G. Schneider. Copula-based kernel dependency measures. *ArXiv*, abs/1206.4682, 2012.

[5] Andrey Feuerverger. A consistent test for bivariate dependence. *International Statistical Review / Revue Internationale de Statistique*, 61(3):419–433, 1993.

[6] Gábor J. Székely, Maria L. Rizzo, and Nail K. Bakirov. Measuring and testing dependence by correlation of distances. *The Annals of Statistics*, 35(6):2769 – 2794, 2007.

[7] Björn Böttcher, Martin Keller-Ressel, and René Schilling. Distance multivariance: New dependence measures for random vectors, 10 2018.

[8] Gábor J. Székely and Maria L. Rizzo. Partial distance correlation with methods for dissimilarities. *arXiv: Methodology*, 2013.

[9] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.

[10] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019.

[11] Bernhard Schlkopf, Alexander J. Smola, and Francis Bach. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond.* The MIT Press, 2018.

[12] P. Daniušis and P. Vaitkus. Supervised feature extraction using hilbert-schmidt norms. In Emilio Corchado and Hujun Yin, editors, *Intelligent Data Engineering and Automated Learning - IDEAL 2009*, pages 25–33, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.

[13] Kurt Wan-Duo Ma, J. P. Lewis, and W. Kleijn. The hsic bottleneck: Deep learning without back-propagation. *ArXiv*, abs/1908.01580, 2020.

[14] Patrik Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal discovery with additive noise models. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2009.