

---

# Measuring statistical dependencies via maximum norm and characteristic functions

---

Anonymous Author(s)

Affiliation

Address

email

## Abstract

1        In this paper we focus on the problem of statistical dependence estimation. We pro-  
2        pose statistical dependence measure based on the maximum-norm of the absolute  
3        value of difference between joint and product-marginal characteristic functions,  
4        and its iterative estimation algorithm. The proposed measure is differentiable,  
5        can be efficiently applied to high-dimensional data, and integrated into modern  
6        machine learning pipelines. We also conduct experiments both with simulated and  
7        real data, which reveal that the proposed measure can exploit statistical dependence  
8        in non-linear data sets more efficiently, comparing to the previous work in this line  
9        of research, and that it can improve real-data classification accuracy, when applied  
10       for feature extraction and regularisation.

## 11    1    Introduction

12    The measurement of statistical dependence plays important role in various empirical learning methods  
13    (e.g. hypothesis testing [1], feature selection and extraction [2, 3], information bottleneck methods  
14    [4], causal inference [5], self-supervised learning [6], representation learning [7], among others).  
15    Historically, earliest statistical dependence estimation ideas (e.g. conditional probability) share  
16    nearly-common origin with the beginning of formal statistical reasoning itself. During last two  
17    centuries ideas of correlation and (relative) entropy (including various generalizations) were proposed  
18    and became very popular in numerous applications and theoretical developments. However, with  
19    the increasing growth of machine and deep learning, new statistical dependence estimation methods,  
20    that are robust, applicable to noisy, high-dimensional, structured data, and which can be efficiently  
21    integrated with modern machine learning and deep learning methods are helpful for the development  
22    both of the theory and application.

23    In this article we focus on quantitative estimation of statistical dependencies, using characteristic  
24    functions. We begin with the short review of some important previous dependence estimation ap-  
25    proaches (Section 2), devoting special attention to ones based on characteristic functions (Section 2.1).  
26    Afterwards, in (Section 3), we formulate the proposed measure, its empirical estimator, and conduct  
27    preliminary theoretical analysis, which are the main theoretical contribution of our paper. Section 4  
28    is devoted to experiments with simulated and real data sets, where we apply the proposed statistical  
29    measure for feature extraction and deep neural network (DNN) regularisation. Finalizing Section 5  
30    discusses and concludes this article.

## 31    2    Previous Work

32    During recent years, various approaches have been used in order to construct statistical dependence  
33    estimation methods. For example, information theory (mutual information [8] and generalisations),

reproducing kernel Hilbert spaces (Hilbert-Schmidt independence criterion [1]), characteristic functions (distance correlation [9, 10]), and other (e.g. [11] copula-based kernel dependence measures, integral-probability-metric-reliant Sobolev independence criterion [12]). Further we will focus on characteristic-function-based methods.

## 2.1 Characteristic-function-based methods

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

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

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

$$\widehat{\phi}_X(\alpha) := \frac{1}{n} \sum_{j=1}^n e^{i\langle \alpha, x_j \rangle}. \quad (2)$$

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

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

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

$$\widehat{\phi}_{X,Y}(\alpha, \beta) := \frac{1}{n} \sum_{j=1}^n e^{i(\langle \alpha, x_j \rangle + \langle \beta, y_j \rangle)}. \quad (4)$$

Uniqueness theorem states that two random variables  $X$  and  $Y$  have the same distribution if and only if their CF's are identical [?]. Therefore, CF's can be regarded as an alternative description of distribution. Roughly speaking, CF can be regarded as Fourier transform of probability density function (PDF).

If cumulative distribution function (CDF) of  $(X, Y)$ ,  $F_{X,Y}(x, y)$ ,  $x \in \mathbb{R}^{d_X}$  and  $y \in \mathbb{R}^{d_Y}$  factorises as  $F_X(x)F_Y(y)$  for all  $x$  and  $y$ ,  $X$  and  $Y$  are called independent (the same holds for probability density function, PDF). However, this criterion is impractical due to need of evaluation of potentially high-dimensional CDF or PDF, and often alternative independence criteria are more useful. Let us define

$$\Delta_{X,Y}(\alpha, \beta) := \phi_{X,Y}(\alpha, \beta) - \phi_X(\alpha)\phi_Y(\beta), \quad (5)$$

an its empirical counterpart:

$$\widehat{\Delta}_{X,Y}(\alpha, \beta) := \widehat{\phi}_{X,Y}(\alpha, \beta) - \widehat{\phi}_X(\alpha)\widehat{\phi}_Y(\beta). \quad (6)$$

In terms of CF's, statistical independence of  $X$  and  $Y$  is equivalent to  $\forall \alpha \in \mathbb{R}^{d_X}, \forall \beta \in \mathbb{R}^{d_Y}, \Delta_{X,Y}(\alpha, \beta) = 0$  [13].

Historically,  $\Delta_{X,Y}(\alpha, \beta)$  was used as the basis (first in [9] for one-dimensional case, and afterwards extended and developed by [10] for bivariate multidimensional random vectors) for construction of statistical independence tests and measures. Distance covariance and distance correlation, proposed by [10] relies on weighted  $L^2$ -norm analysis of (5). They select weighting function in such a way, that dependence measure can be expressed in terms of correlation of data-dependent distances. Recent result of [14] generalises [10] to multivariable case. [15] proposed computationally efficient algorithm for estimation of distance correlation measure, reducing computational complexity from  $O(n^2)$  to  $O(n \cdot \log n)$ , where  $n$  is sample size.

**Motivation and Connection To Previous Work** Taking  $\Delta_{X,Y}(\alpha, \beta) = 0$  (5) as the criterion of statistical independence we view the work [10] from the perspective of weighted  $L^p$  spaces, measuring statistical dependence by the corresponding  $L^p$ -norms of (5).

Taking into account that [10] in high dimensions is affected with the curse of dimensionality [16], we focus on the limit case  $p \rightarrow \infty$  ( $L^\infty$  space), which is associated to the supremum norm. This norm has several potential advantages.

We hypothesise, that its locality could be exploited to detect statistical independence more efficiently, comparing to case  $p = 2$ . In addition, numerically calculation of  $L^\infty$  norm would not require to directly calculate norm integral, since norm of  $L^p$  converges to supremum norm when  $p \rightarrow \infty$ . Also, from practical point of view maximization is convenient, because it is efficiently implemented in modern deep learning frameworks (e.g. Pytorch [17]). In addition, in our opinion it is worth to note, that applications of characteristic functions in machine learning are quite scarce, despite that they provide quite convenient theoretical proxy to access distributions.

### 3 Proposed Measure

The above considerations serves as the basis for constructing of a novel dependence measure, which we further refer to as Kac independence measure (KacIM). Having two random vectors  $X$  and  $Y$ , KacIM is defined as

$$\kappa(X, Y) := \max_{\alpha \in \mathbb{R}^{d_X}, \beta \in \mathbb{R}^{d_Y}} |\Delta_{X,Y}(\alpha, \beta)|. \quad (7)$$

#### 3.1 Basic Properties

**Theorem 1.** *KacIM (7) has the following properties:*

1.  $\kappa(X, Y) = \kappa(Y, X)$ ,
2.  $0 \leq \kappa(X, Y) \leq 1$ ,
3.  $\kappa(X, Y) = 0$  iff  $X \perp Y$ .

*Proof.* Property 1. is obvious from definition (7) (commutativity of addition and multiplication), and property 2. directly follows from Cauchy inequality and that absolute value of CF is bounded by 1:

$$\begin{aligned} |\phi_{X,Y}(\alpha, \beta) - \phi_X(\alpha)\phi_Y(\beta)|^2 &= \mathbb{E}_{X,Y} |(e^{i\alpha^T X} - \phi_X(\alpha))(e^{i\beta^T Y} - \phi_Y(\beta))|^2 \leq \\ &\mathbb{E}_{X,Y} |(e^{i\alpha^T X} - \phi_X(\alpha))|^2 |(e^{i\beta^T Y} - \phi_Y(\beta))|^2 = (1 - |\phi_X(\alpha)|^2)(1 - |\phi_Y(\beta)|^2). \end{aligned}$$

Proof of property 3. directly follows from properties of CF's (see e.g. [18], Corollary 14.1)<sup>1</sup>.  $\square$

Although (7) is not scale invariant in general, scale invariance can be achieved by assuming standardization of  $X$  and  $Y$ .

#### 3.2 Estimation

Having i.i.d. observations  $(x_j, y_j)$ ,  $j = 1, 2, \dots, n$ , an empirical estimator of (7) is defined via corresponding ECF's (4) and (2):

$$\hat{\kappa}(X, Y) := \max_{\alpha, \beta} |\hat{\Delta}_{X,Y}(\alpha, \beta)| = \max_{\alpha, \beta} |\widehat{\phi_{X,Y}}(\alpha, \beta) - \widehat{\phi_X}(\alpha)\widehat{\phi_Y}(\beta)|. \quad (8)$$

By *Levy continuity theorem* [?] ECF converges to CF. Therefore empirical estimator (8) converges (?) into KacIM (7) (in what sense, check?).

It can be calculated iteratively by Algorithm 1 (Pytorch [17] implementation can be accessed from [https://github.com/povidanius/kac\\_independence\\_measure](https://github.com/povidanius/kac_independence_measure)).

Algorithm 1 requires to initialise  $\alpha$  and  $\beta$ , and optimiser. In our implementation we use uniform initialisation of parameters, and decoupled weight decay regularization optimizer [19]. We also empirically observed that normalisation of parameters  $\alpha$  and  $\beta$  on to unit sphere increases estimation stability. After the estimation of KacIM via Algorithm 1, the evaluation the estimator has computation complexity  $O(n)$ , where  $n$  is sample size.

<sup>1</sup>This property also is known as Kac's theorem [13]. Although it is quite simple mathematical fact, this provides the basis of the proposed measure's name.

---

**Algorithm 1** KacIM estimation

---

**Require:** Number of iterations  $N$ , gradient-based optimiser  $GradOpt([parameters], \cdot)$ , initial  $\alpha \in \mathbb{R}^{d_x}, \beta \in \mathbb{R}^{d_y}$ .  
**for** iteration=1 to  $N$  **do**  
    Sample data batch  $(X, Y) := (x_i, y_i)_{i=1}^{n_b}$ .  
    Normalize  $(X, Y)$  to zero mean and unit variance (scale invariance).  
    Calculate  $\hat{\Delta}_{\alpha, \beta}(X, Y)$ .  
    Perform one maximization iteration of  $\hat{\Delta}(X, Y)$  via  $\alpha, \beta \rightarrow GradOpt([\alpha, \beta], \hat{\Delta}_{\alpha, \beta}(X, Y))$ .  
**end for**

---

### 3.3 Interpretation and connection to the related approaches

**Interpretation as maximum covariance.** Since (7) can be reformulated as

$$\kappa(X, Y) = \max_{\alpha, \beta} |\text{cov}(e^{i\alpha^T X}, e^{i\beta^T Y})|, \quad (9)$$

by Euler's formula, it corresponds to the maximum pseudocovariance between complex exponents  $e^{i\alpha^T X} = \cos(\alpha^T X) + i \cdot \sin(\alpha^T X)$  and  $e^{i\beta^T Y} = \cos(\beta^T Y) + i \cdot \sin(\beta^T Y)$ . Since  $\text{var}(e^{i\alpha^T X}) = \phi(2\alpha) - \phi(\alpha)^2$ , one can also define the normalised version of KacIM (refine or remove this):

$$\kappa_{norm}(X, Y) = \max_{\alpha, \beta} |\text{corr}(e^{i\alpha^T X}, e^{i\beta^T Y})| = \max_{\alpha, \beta} \frac{|\text{cov}(e^{i\alpha^T X}, e^{i\beta^T Y})|}{\sqrt{|\phi_X(2\alpha) - \phi_X^2(\alpha)| |\phi_Y(2\beta) - \phi_Y^2(\beta)|}}. \quad (10)$$

**Interpretation in Gaussian case.** In special case when both  $X$  and  $Y$  are zero mean Gaussian random vectors we have:

$$\kappa(X, Y) = \max_{\alpha, \beta} |e^{-\frac{1}{2}(\alpha^T \Sigma_x \alpha + \beta^T \Sigma_y \beta)} (e^{-\alpha^T \Sigma_{x,y} \beta} - 1)|. \quad (11)$$

Assuming constant  $\alpha^T \Sigma_x \alpha$  and  $\beta^T \Sigma_y \beta$ , the maximization corresponds to the maximization of  $\alpha^T \Sigma_{x,y} \beta$ , which coincides with canonical correlation analysis [21]. Here  $\Sigma_x$ ,  $\Sigma_y$ , and  $\Sigma_{x,y}$  are covariance matrices of  $X$ ,  $Y$ , and cross-covariance matrix between  $X$  and  $Y$ , respectively.

**Mutual information** For the neural estimation of mutual information its variational (Donsker-Varadhan) representation  $I(X, Y) = \max_{\theta} \mathbb{E}_{X,Y} f(x, y|\theta) - \log \mathbb{E}_X \mathbb{E}_Y e^{f(x,y|\theta)}$  [20] is often used, since it allows to avoid density estimation (here  $f(x, y|\theta)$  is neural network with parameters  $\theta$ ). The estimation is also iterative process, similar to Algorithm 1. In this case, optimisation is conducted over the space of neural network parameters, which often is substantially larger than the number of parameters needed to estimate KacIM (i.e.  $d_x + d_y$  parameters).

## 4 Experiments

Further we will conduct empirical investigation of KacIM in order to demonstrate that it can measure non-linear statistical dependencies, and that it can be practically useful as a component of cost functions (we investigate feature extraction, and regularisation problems).

### 4.1 Generated data

**Non-linear statistical dependence detection.** We begin with simulated multivariate data with additive and multiplicative noise.

Figure 1 reflects KacIM values during iterative adaptation (200 iterations). In the case of independent data, both  $x_i$  and  $y_i$  ( $d_x = 512, d_y = 4$ ) are sampled from gaussian distribution, independently. In the case of dependent data, an additive noise and multiplicative noise, the dependent variable is generated according to  $y_i = \sin(Px_i) + \cos(Px_i) + \lambda \epsilon_i$  ( $\lambda = 1.00$ ) 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, both in additive and multiplicative cases, due to independence, estimator (8) is resistant to maximisation, and oscillates near zero. On the other hand, when the data is not independent, the condition (5) is violated and maximization of estimator (8) is possible.

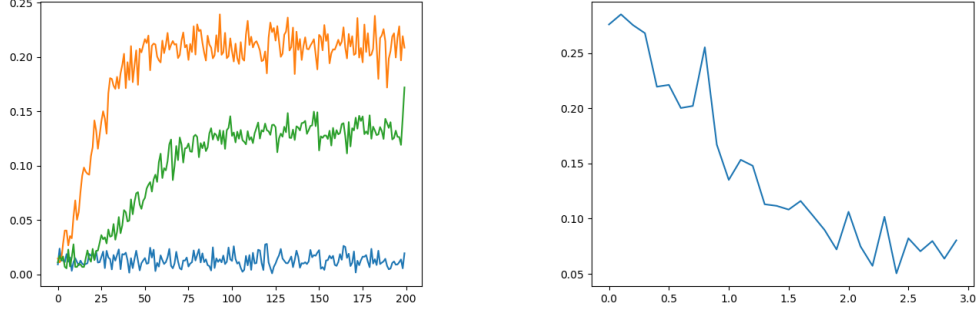


Figure 1: Left figure: KacIM evaluation for independent data (blue), additive (orange) and multiplicative (green) noise scenarios ( $x$  axis - iteration, and  $y$  - corresponding value of KacIM). Right figure: noise level ( $x$  axis) vs final iteration KacIM value ( $y$  axis). KacIM values for larger noise levels saturates as in tail of graph

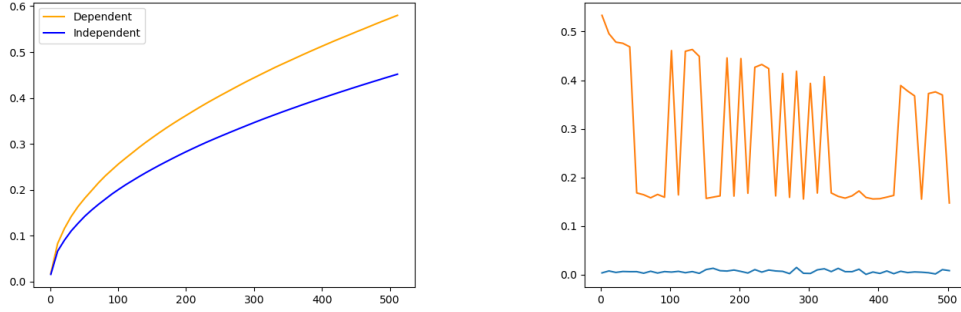


Figure 2: The dimension of data is on the  $x$  axis, and on  $y$  axis is evaluation of distance correlation (left) and KacIM (right). Blue graph corresponds of independent data of dimension, indicated by  $x$  axis, and orange one corresponds to dependent data.

137 **Noise variance effect** In this simulation we use the same additive noise setting as in previous  
 138 paragraph, but evaluate all noise levels  $\lambda \in [0.1, 3.0]$ , with step 0.1. Figure 1 empirically shows that  
 139 value of KacIM negatively correlates with noise level, and therefore the proposed measure is able not  
 140 only to detect whether independence is present, but also to quantitatively evaluate it.

141 **Comparison with distance correlation** We also evaluated distance correlation [10] on the same  
 142 generated samples of data, comparing it with KacIM. From Figure. 2 we see that as data dimensionality  
 143 grows, for independent data, the values of measure not only is significantly larger than zero, but it  
 144 also grows like values of measure of dependent data. This empirically demonstrates that distance  
 145 correlation is affected by the curse of dimensionality. On the other side, KacIM even for larger  
 146 dimensions oscillates near zero for independent data, and significantly deviates from zero for dependent  
 147 data case, as indicated in right component of Figure. 2.

## 148 4.2 Feature extraction

149 Previous work in the field of supervised feature extraction, which rely on dependency-based cost  
 150 functions, include [2, 3, 22] (HSIC),....().

151 Let us denote by  $T := (x_i, y_i)_{i=1}^N$  a supervised-learning dataset of  $N$  pairs of  $d_x$ -dimensional inputs  
 152  $x_i$ , and  $d_y$ -dimensional one-hot-encoded outputs  $y_i$ .

Dataset	$N/d_x/n_c$	Raw	KacIMFE	NCA
isolet	(7797,617,26)	0.9261	0.9437	<b>0.9477</b>
madelon	(2600,500,2)	<b>0.6015</b>	0.5484	0.5685
prnn-viruses	(61,18,4)	0.6452	0.9265	0.9355
ionosphere	(351,34,2)	0.8807	0.9278	<b>0.9375</b>
micro-mass	(360,1300,10)	0.8778	<b>0.9282</b>	0.8944
clean1	(476,168,2)	0.7689	<b>0.9888</b>	0.9790
robot-failures-lp2	(47,90,5)	0.4583	0.6067	0.5833
waveform-5000	(5000,40,3)	<b>0.8692</b>	0.8017	0.8516
spambase	(4601,57,2)	0.6906	0.8285	<b>0.8705</b>
gina-agnostic	(3468,970,2)	<b>0.8512</b>	0.7894	0.8080
scene	(2407,299,2)	0.8895	<b>0.9707</b>	0.9336
tokyo1	(959,44,2)	0.7250	0.8995	<b>0.9062</b>
one-hundred-plants-shape	(1600,64,100)	0.1013	<b>0.4913</b>	0.4688

Table 1: Classification accuracies.  $N$  denotes full data set size,  $d_x$  - input dimensionality, and  $n_c$  - number of classes. In this table feature dimension is equal to a half of original input dimension. Best accuracies that are also statistically significant (Wilcoxon’s signed rank test [25], 25 runs,  $p$ -value threshold 0.01) are indicated in bold text.

In feature extraction experiments we will use a set of classification data sets from OpenML [23], which cover different domains. We use KacIM in order to conduct supervised linear feature extraction by seeking

$$W^* = \arg \max_W \kappa(Wx, y) - \alpha \text{Tr}\{(W^T W - I)^T (W^T W - I)\}, \quad (12)$$

where the regularisation term, controlled by multiplier  $\alpha \geq 0$ , enforces semi-orthogonality of projection matrix  $W^*$ , and  $\text{Tr}\{\cdot\}$  denotes matrix trace operator.

In all the experiments (12) the cost function is optimised iteratively (250 iterations), simultaneously optimising parameters of KacIM ( $\alpha$  and  $\beta$ ) and projection matrix  $W$ . After the optimisation, the feature extraction is conducted by  $f(x) = W^*x$ , where  $x$  is original input vector, and  $f$  are corresponding feature vector.

We randomly split all the datasets in training and testing sets of equal size. In our experiments we set  $\alpha$  to 1.0 to quickly ensure orthogonal projection matrices, and further proceed to dependence maximization stage. In order to quantitatively evaluate features, we use logistic regression classifier accuracy, measured on the testing set.

We use two baselines: raw features (RAW column in Table 1) and neighborhood component analysis [24] (NCA column in Table 1). The purpose of these experiments is to provide the preliminary evaluation of the applicability of KacIM for feature extraction, hence we use rather basic cost function and comparative baselines.

The classification accuracies, reported in Table 1 demonstrate that KacIM-based feature extraction procedure (KacIMFE column) indeed allows to increase classification accuracy when applied to real data sets from different domains, including high-dimensional and ill-defined ones (e.g. *micro-mass* dataset). In contrast to our feature extraction approach, NCA explicitly optimises for classification accuracy, rather than more abstract dependency of features  $f(x)$  with the dependent variable  $y$ .

### 4.3 Regularisation

In regularisation experiments we investigate skin lesion classification task. It is a binary classification data set, consisting of 10605 images, which should be classified as benign or malignant (e.g. Figure 3).

We use ResNet18 backbone model (pretrained on ImageNet) with added classification head. Further we train this model with batches of 128 elements. We denote our classification network as  $f(\phi(x|\theta_0)|\theta_1)$ , where  $\theta_0$  are parameters of ResNet18,  $\theta_1$  are classification head parameters, and  $x$  is  $224 \times 224$  input image. For optimisation we use decoupled weight decay regularization optimizer [19]

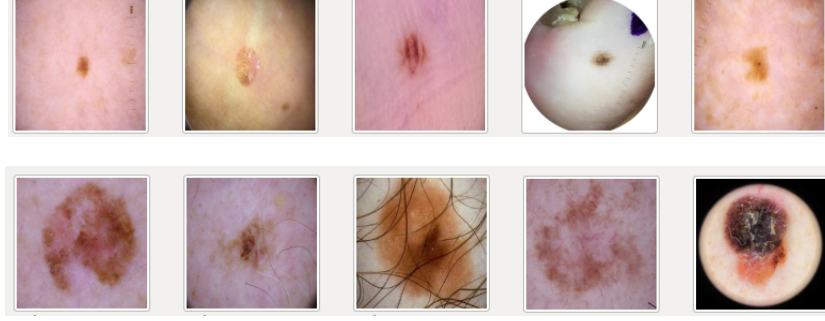


Figure 3: Top figure - benign moles, bottom figure - malignant tumors.

Mode	Average accuracy (%)
Without regularisation	93.01
With regularisation	<b>93.34</b>

Table 2: Melanoma classification accuracy comparison of regularised and not regularised model. Bold text indicates that model with regulariser was more accurate (Wilcoxon’s signed rank test [25], 30 runs,  $p$ -value threshold 0.04)

182 with learning rate set to 0.0002, and weight decay parameter set to 0.00001 (3 epochs). The internal  
 183 learning rate of estimator (seeefalga) was set to 0.07 and weight decay parameters to 0.01.

184 We will investigate additive regularizer, which maximises dependency of bottleneck the feature  $\phi(x|\theta_0)$   
 185 and target variable  $y$  (one-hot encoding):

$$Cost(\theta_0, \theta_1, W) := (1 - \rho)CE(f(\phi(x|\theta_0)|\theta_1), y) - \rho\kappa(\phi(x|\theta_0), y), \quad (13)$$

186 where  $CE(., .)$  is cross-entropy loss,  $\rho \geq 0$  is regularisation parameter (in our experiments  $\rho = 0.2$ ).  
 187 During backward pass, this regularizer is designed to directly transfer information from  $y$  to the  
 188 output ResNet18  $\phi(., |\theta_0)$ , and we hypothesise that this could provide possibility to learn more  
 189 discriminative features.

190 In each experiment we train classifier 30 times with randomly splitted training and testing data  
 191 (9000 images for training, and 1605 for testing). The average accuracies reported in Table 2, that  
 192 application (13) slightly (but with statistical significance) increased classification accuracy.

## 193 5 Conclusion

194 In this article we propose statistical dependence measure, KacIM, which corresponds to the  $L^\infty$  norm  
 195 of the absolute value of difference between joint characteristic function and the product of marginal  
 196 ones. The proposed measure, in theory can detect non-linear statistical dependence between a pairs  
 197 of random variables of possibly different dimension, extended to various directions (e.g. kernels,  
 198 multiple variables), applied to several machine learning tasks (e.g. feature extraction, regularisation,  
 199 among others). On the other side, it raises a corresponding set of unanswered questions, both  
 200 theoretical and empirical.

201 For example, although it converges, the variance of the estimator sometimes is high and it is still  
 202 remains unclear how to control it, also the interpretability when it approaches its maximal value  
 203 remains insufficiently clear. However empirical experiments with simulated data reveals, that  
 204 increasing independence between two random variables is reflected in a decreasing trend on the  
 205 estimated values of the proposed dependence measure(e.g. Figure 1).

206 Therefore, parameter initialization, meta-parameter (e.g. stopping criteria, batch size) selection are  
 207 needed in order to evaluate it efficiently.

208 Beside demonstrated applications in Section 4, the proposed measure is differentiable and thereby can  
 209 be integrated with various modern deep-learning methods, applied to high-dimensional and structured

210 data. We see exploration and comparative analysis of KacIM in causality, information bottleneck  
 211 theory, self-supervised learning, and other modern problems, where dependence measures define a  
 212 criterion of optimisation, as future work.

## 213 6 Acknowledgements

214 We sincerely thank Dr. Pranas Vaitkus, Dr. Linas Petkevičius, Dr. Aleksandras Voicikas, and  
 215 colleagues from Neurotechnology for discussions. We also thank Neurotechnology for supporting  
 216 this research.

## 217 References

## 218 References

- 219 [1] Arthur Gretton, Olivier Bousquet, Alex Smola, and Bernhard Schölkopf. Measuring Statistical Dependence  
 220 with Hilbert-Schmidt Norms. In *ALT*, 2005.
- 221 [2] Daniušis, P. and Vaitkus, P. Supervised Feature Extraction Using Hilbert-Schmidt Norms. In Corchado,  
 222 Emilio and Yin, Hujun, editor, *Intelligent Data Engineering and Automated Learning - IDEAL 2009*, pages  
 223 25–33, Berlin, Heidelberg, 2009. Springer Berlin Heidelberg.
- 224 [3] P. Daniušis, Pr. Vaitkus, and L. Petkevičius. Hilbert–Schmidt component analysis. *Lithuanian mathematical*  
 225 *journal*, 57(A):7–11, Dec. 2016.
- 226 [4] Kurt Wan-Duo Ma, J. P. Lewis, and W. Kleijn. The hsic bottleneck: Deep learning without back-propagation.  
 227 *ArXiv*, abs/1908.01580, 2020.
- 228 [5] Patrik Hoyer, Dominik Janzing, Joris M Mooij, Jonas Peters, and Bernhard Schölkopf. Nonlinear causal  
 229 discovery with additive noise models. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors,  
 230 *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc., 2009.
- 231 [6] Yazhe Li, Roman Pogodin, Danica J. Sutherland, and Arthur Gretton. Self-supervised learning with kernel  
 232 dependence maximization. In *Thirty-Fifth Conference on Neural Information Processing Systems*, 2021.
- 233 [7] Ruggero Ragonesi, Riccardo Volpi, Jacopo Cavazza, and Vittorio Murino. Learning Unbiased Representations via Mutual Information Backpropagation. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 2723–2732, 2021.
- 236 [8] Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory 2nd Edition (Wiley Series in Telecommunications and Signal Processing)*. Wiley-Interscience, July 2006.
- 238 [9] Andrey Feuerverger. A Consistent Test for Bivariate Dependence. *International Statistical Review / Revue Internationale de Statistique*, 61(3):419–433, 1993.
- 240 [10] 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.
- 242 [11] Barnabás Póczos, Zoubin Ghahramani, and Jeff G. Schneider. Copula-based Kernel Dependency Measures. *ArXiv*, abs/1206.4682, 2012.
- 244 [12] Mroueh, Youssef and Sercu, Tom and Rigotti, Mattia and Padhi, Inkit and Nogueira dos Santos, Cicero. Sobolev Independence Criterion. In *Advances in Neural Information Processing Systems 32*, editor = H. Wallach and H. Larochelle and A. Beygelzimer and F. d'Alché-Buc and E. Fox and R. Garnett, pages 9505–9515. Curran Associates, Inc., 2019.
- 248 [13] Johan Kustermans J. Martin Lindsay Michael Schuermann Uwe Franz David Applebaum, B.V. Rajarama Bhat. Quantum Independent Increment Processes I: From Classical Probability to Quantum Stochastic Calculus. 2005.
- 251 [14] Björn Böttcher, Martin Keller-Ressel, and René Schilling. Distance multivariance: New dependence measures for random vectors, 10 2018.
- 253 [15] Arin Chaudhuri and Wenhao Hu. A fast algorithm for computing distance correlation. *Computational Statistics and Data Analysis*, 135:15–24, 2019.



- [16] Dominic Edelmann, Konstantinos Fokianos, and Maria Pitsillou. An Updated Literature Review of Distance Correlation and Its Applications to Time Series. *International Statistical Review*, 87(2):237–262, August 2019.
- [17] 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. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc., 2019.
- [18] Jacod Jean. *Probability essentials / Jean Jacod, Philip Protter*. Universitext. Springer, Berlin Heidelberg New York, 2nd edition edition, 2003.
- [19] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *ICLR*, 2019.
- [20] Mohamed Ishmael Belghazi, Aristide Baratin, Sai Rajeshwar, Sherjil Ozair, Yoshua Bengio, Aaron Courville, and Devon Hjelm. Mutual Information Neural Estimation. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 531–540. PMLR, 10–15 Jul 2018.
- [21] Bernhard Schölkopf, Alexander J. Smola, and Francis Bach. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press, 2018.
- [22] Yin Zhang and Zhi-Hua Zhou. Multilabel dimensionality reduction via dependence maximization. *ACM Trans. Knowl. Discov. Data*, 4(3), oct 2010.
- [23] Joaquin Vanschoren, Jan N. van Rijn, Bernd Bischl, and Luis Torgo. Openml: Networked science in machine learning. *SIGKDD Explorations*, 15(2):49–60, 2013.
- [24] Jacob Goldberger, Geoffrey E Hinton, Sam Roweis, and Russ R Salakhutdinov. Neighbourhood Components Analysis. In L. Saul, Y. Weiss, and L. Bottou, editors, *Advances in Neural Information Processing Systems*, volume 17. MIT Press, 2004.
- [25] Wilcoxon, Frank. *Individual Comparisons by Ranking Methods*, pages 196–202. Springer New York, New York, NY, 1992.

## A Appendix

Optionally include extra information (complete proofs, additional experiments and plots) in the appendix. This section will often be part of the supplemental material.