Applied Soft Computing xxx (2017) xxx–xxx

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

Sensitivity maps of the Hilbert–Schmidt independence criterion

Adrián Pérez-Suay*, Gustau Camps-Valls

Image Processing Laboratory (IPL), Universitat de València, Catedrático A. Escardino, 46980 Paterna, València, Spain

ARTICLE INFO

Article history:
Received 30 September 2016
Received in revised form 6 March 2017
Accepted 11 April 2017
Available online xxx

Keywords:
Kernel methods
Dependence estimation
Sensitivity analysis
Random features
Visualization
Feature selection
Causal inference

ABSTRACT

Kernel dependence measures yield accurate estimates of nonlinear relations between random variables, and they are also endorsed with solid theoretical properties and convergence rates. Besides, the empirical estimates are easy to compute in closed form just involving linear algebra operations. However, they are hampered by two important problems: the high computational cost involved, as two kernel matrices of the sample size have to be computed and stored, and the interpretability of the measure, which remains hidden behind the implicit feature map. We here address these two issues. We introduce the sensitivity maps (SMs) for the Hilbert–Schmidt independence criterion (HSIC). Sensitivity maps allow us to explicitly analyze and visualize the relative relevance of both examples and features on the dependence measure. We also present the randomized HSIC (RHSIC) and its corresponding sensitivity maps to cope with large scale problems. We build upon the framework of random features and the Bochner's theorem to approximate the involved kernels in the canonical HSIC. The power of the RHSIC measure scales favourably with the number of samples, and it approximates HSIC and the sensitivity maps efficiently. Convergence bounds of both the measure and the sensitivity map are also provided. Our proposals are illustrated in many synthetic illustrative examples, and challenging real problems of dependence estimation, feature selection, and causal inference from empirical data. The methods allow estimating and visualizing HSIC in large scale data regimes, while still yielding closed-form analytical solutions.

1. Introduction

The problem of estimating statistical dependencies between random variables is ubiquitous in Science and Engineering, and the basis to discover causal relations from empirical data. Many methods exist to this purpose. Very often one traditionally resorts to Pearson's correlation, but the measure can only identify linear associations between random variables. Other measures of dependence, such as the Spearman's rank or the Kendall's tau criteria, assume monotonically increasing variable relations, and can be better suited in problems exhibiting such relations. All of them, however, can be computed for pairs of variables only, and thus the multidimensional problem of dependence estimation is tackled by repeating the test for all pairwise combinations, and then summarizing the 'dependence matrix' into an *ad hoc* overall statistic.

In recent years, we have witnessed the introduction of an increasing amount of nonlinear dependence measures. Among the vast amount of criteria, kernel dependence methods exhibit some good properties [1]. They typically reveal (1) good robustness properties in high dimensional and low number of samples settings; (2) criteria are not restricted to estimate pairwise dependencies, but capture higher-order relations between (multidimensional) random variables; (3) the empirical estimates are very simple to implement in closed form and only involve kernel matrix computation and linear algebra operations; (4) there is a well-founded theoretical background to study and characterize them, and fast converge rates to the true measure can be derived; and (5) one can actually derive *p*-values associated to the empirical measure. In this paper, we focus on improving the family of kernel dependence estimates in terms of *computational efficiency* and *interpretability*.

The principle underlying kernel-based dependence estimation is to define covariance and cross-covariance operators in reproducing kernel Hilbert spaces (RKHS) [2], and derive statistics from these operators capable of measuring dependence between functions therein. In [3] the largest singular value of the kernel canonical correlation analysis (KCCA) – which uses both covariance and cross-covariances – was used as a statistic to test independence. Later, in [4], the constrained covariance (COCO) statistic was proposed, which uses the largest singular value of the cross-covariance operator: high efficiency was obtained with virtually no regularization needed. A variety of empirical kernel quantities derived from bounds on the mutual information that hold near independence

* Corresponding author.

http://dx.doi.org/10.1016/j.asoc.2017.04.024
1568-4946/

were also proposed: namely the kernel Generalised Variance (kGV) and the Kernel Mutual Information (kMI) [5,1].

Among the most interesting kernel dependence methods, we find the Hilbert–Schmidt Independence Criterion (HSIC) [6]. The method consists of measuring cross-covariances in a proper RKHS, and generalizes several measures, such as COCO, by using the entire spectrum of the cross-covariance operator, not just the largest singular value. The HSIC empirical estimator is very easy to compute, has good theoretical properties [6,1], and yields very good results in practice, e.g. HSIC has been successfully used for ranking [7], clustering [8], dimensionality reduction [9], screening [10], image representation for classification [11], sensitivity analysis [12,10], as well as feature selection from satellite images [13] and gene expression [14].

Kernel dependence estimates such as HSIC however face two main challenges: (1) the measure is hardly interpretable in geometric terms, as it is based on implicit mappings reproduced via reproducing kernel functions; and (2) the method scales poorly with the number of examples, as it involves computing and storing kernel matrices of the sample size. We will tackle these two important limitations in this paper, illustrating the methodology for the particular case of HSIC. Specifically, the contributions are summarized as follows:

- In order to analyze and visualize the kernel dependence measure, we propose to derive *sensitivity maps* (SMs) of the estimate. Sensitivity maps allow us to explicitly analyze and visualize the relative relevance of *both* examples and features on the dependence measure [15]. Our inspiration is a probabilistic approach to derive sensitivity maps for Support Vector Machines (SVM) in neuroimage applications [16], which has been recently extended to the field of Gaussian Processes (GPs) visualization in geoscience problems [17,18]. In both cases, the goal was to study the sensitivity (relevance, impact) of features on the learned *supervised* model. In our case, however, we deal with the more challenging *unsupervised* scenario of scrutinizing kernel-based dependence measures. For this, we develop the SMs to visualize and study HSIC dependence measure quantitatively. We will show that the SMs provide a *vector field* that allows us to identify both examples and features most affecting the measure of dependence.
- In order to alleviate the high computational burden involved in both HSIC and its SM, we here introduce the randomized HSIC (RHSIC), and derive an efficient SM that still preserves the appealing closed-form computation property. Essentially, we replace the involved kernels in HSIC by explicit mappings generated via linear projections on random features. This approximation builds upon the framework of random features originally introduced in [19,20] and the Bochner's theorem [21,22]. We want to highlight that introducing the RHSIC is not incidental, but capitalizes on the fact that still permits to derive sensitivity maps in a very efficient, closed-form manner.

The remainder of the paper is organized as follows. Section 2 fixes notation, briefly introduces the HSIC estimate, and presents the randomized HSIC for computational efficiency. We also discuss on the computational gain and on the convergence rates for the estimate, the decision threshold and the associated p-values. Section 3 introduces the sensitivity maps for both HSIC and its randomized version. Section 4 shows experiments of the performance of RHSIC and the properties of the sensitivity maps. In particular, we give empirical evidence of performance of the sensitivity maps in both synthetic examples that allow us to understand dependence measures, and challenging real problems of dependence estimation, feature ranking, and causal inference from empirical data. We emphasize the usefulness of the sensitivity maps for data visualization, and point out the relation to the field of leveraging points. Section 5 concludes this paper with some remarks and future research lines. Source code and demos are given for the interested reader, and some theoretical properties of convergence of the randomized measure and its sensitivity are given in Appendices A and B.

2. Efficient HSIC dependence estimation

To fix notation, let us consider two spaces $\mathcal{X} \subseteq \mathbb{R}^{d_x}$ and $\mathcal{Y} \subseteq \mathbb{R}^{d_y}$, on which we jointly sample observation pairs $(\mathbf{x}, \mathbf{y})$ from distribution $\mathbb{P}_{\mathbf{xy}}$. The covariance matrix can be defined as

$$\mathcal{C}_{\mathbf{xy}} = \mathbb{E}_{\mathbf{xy}}(\mathbf{x}\mathbf{y}^\top) - \mathbb{E}_{\mathbf{x}}(\mathbf{x})\mathbb{E}_{\mathbf{y}}(\mathbf{y}^\top), \tag{1}$$

where $\mathbb{E}_{\mathbf{xy}}$ is the expectation with respect to $\mathbb{P}_{\mathbf{xy}}$, $\mathbb{E}_{\mathbf{x}}$ is the expectation with respect to the marginal distribution $\mathbb{P}_{\mathbf{x}}$ (hereafter, we assume that all these quantities exist), and $\mathbf{y}^\top$ is the transpose of $\mathbf{y}$. The covariance matrix encodes all first order dependencies between the random variables. A statistic that efficiently summarizes the content of this matrix is its Hilbert–Schmidt norm. The square of this norm is equivalent to the squared sum of its eigenvalues γ_i:

$$\|\mathcal{C}_{\mathbf{xy}}\|^2_{\mathrm{HS}} = \sum_i \gamma_i^2. \tag{2}$$

This quantity is zero if and only if there exists no first order dependence between $\mathbf{x}$ and $\mathbf{y}$. Note that the Hilbert Schmidt norm is limited to the detection of second order relations, and thus more complex (higher-order effects) cannot be captured.

2.1. Kernel dependence estimation

Let us define a (possibly non-linear) mapping $\boldsymbol{\phi} : \mathcal{X} \to \mathcal{F}$ such that the inner product between features is given by a positive definite (p.d.) kernel function $K_x(\mathbf{x}, \mathbf{x}') = \langle \boldsymbol{\phi}(\mathbf{x}), \boldsymbol{\phi}(\mathbf{x}') \rangle$. The feature space $\mathcal{F}$ has the structure of a reproducing kernel Hilbert space (RKHS). Let us now denote another feature map $\boldsymbol{\psi} : \mathcal{Y} \to \mathcal{G}$ with associated p.d. kernel function $K_y(\mathbf{y}, \mathbf{y}') = \langle \boldsymbol{\psi}(\mathbf{y}), \boldsymbol{\psi}(\mathbf{y}') \rangle$. Then, the cross-covariance operator between these feature maps is a linear operator $\mathcal{C}_{\mathbf{xy}} : \mathcal{G} \to \mathcal{F}$ such that $\mathcal{C}_{\mathbf{xy}} = \mathbb{E}_{\mathbf{xy}}[(\phi(\mathbf{x}) - \mu_x) \otimes (\psi(\mathbf{y}) - \mu_y)]$, where $\otimes$ is the tensor product, $\mu_x = \mathbb{E}_{\mathbf{x}}[\boldsymbol{\phi}(\mathbf{x})]$, and $\mu_y = \mathbb{E}_{\mathbf{y}}[\boldsymbol{\psi}(\mathbf{y})]$. See more details in [23,24]. The squared norm of the cross-covariance operator, $\|\mathcal{C}_{\mathbf{xy}}\|^2_{\mathrm{HS}}$, is called the Hilbert–Schmidt Independence Criterion (HSIC) and can be expressed in terms of kernels [6]. Given the sample datasets $\mathbf{X} \in \mathbb{R}^{n \times d_x}$, $\mathbf{Y} \in \mathbb{R}^{n \times d_y}$, with n pairs $\{(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_n, \mathbf{y}_n)\}$ drawn from the joint $\mathbb{P}_{\mathbf{xy}}$, an empirical estimator of HSIC is [6]:

$$\mathrm{HSIC}(\mathcal{F}, \mathcal{G}, \mathbb{P}_{\mathbf{xy}}) = \frac{1}{n^2}\mathrm{Tr}(\mathbf{K}_x\mathbf{H}\mathbf{K}_y\mathbf{H}) = \frac{1}{n^2}\mathrm{Tr}(\mathbf{H}\mathbf{K}_x\mathbf{H}\mathbf{K}_y), \tag{3}$$

where $\mathrm{Tr}(\cdot)$ is the trace operation (the sum of the diagonal entries), $\mathbf{K}_x$, $\mathbf{K}_y$ are the kernel matrices for the input random variables $\mathbf{x}$ and $\mathbf{y}$, respectively, and $\mathbf{H} = \mathbf{I} - \frac{1}{n}\mathbb{1}\mathbb{1}^\top$ centers the data in the feature spaces $\mathcal{F}$ and $\mathcal{G}$, respectively.

2.2. The randomized HSIC

An outstanding result in the recent kernel methods literature makes use of a classical definition in harmonic analysis to improve approximation and scalability [19,20]. The Bochner's theorem [21,22] states that a continuous kernel $K(\mathbf{x}, \mathbf{x}') = K(\mathbf{x} - \mathbf{x}')$ on $\mathbb{R}^d$ is positive definite (p.d.) if and only if K is the Fourier transform of a non-negative measure. If a shift-invariant kernel K is properly scaled, its Fourier transform $p(\mathbf{w})$ is a proper probability distribution. This property is used to approximate kernel functions and

Please cite this article in press as: A. Pérez-Suay, G. Camps-Valls, Sensitivity maps of the Hilbert–Schmidt independence criterion, Appl. Soft Comput. J. (2017), http://dx.doi.org/10.1016/j.asoc.2017.04.024