DENSITY KERNELS ON UNORDERED SETS FOR KERNEL-BASED SIGNAL PROCESSING

Frederic Desobry[1], *Manuel Davy*[2] *and William J. Fitzgerald*[1]

[1] Signal Processing Group, Univ. of Cambridge Engineering Department, Cambridge, UK
[2] LAGIS/CNRS/INRIA-FUTURS Sequel, Cité Scientifique, BP 48, 59651 Villeneuve d'Ascq, France

ABSTRACT

Algorithms involved in applications such as speaker recognition or image classification need to be able to process data which are sets of vectors with variable size. As opposed to the standard setting for kernel methods, where the data are individual vectors, it is difficult to build a reliable reproducing kernel between such sets of unordered vectors. Most effective techniques rely on the design of kernel calculated on densities estimated independently on each set of vectors; however, this calculation can be numerically tricky: therefore these techniques either use poor estimates such as histograms, or assume unjustified restrictive conditions. In this paper, we improve on the existing framework and design kernels between densities, where these are estimated using an effective nonparametric technique, namely the Akaike-Parzen-Rosenblatt (APR) estimate. Closed-form expressions are obtained for positive definite kernels, and simulation results illustrate the soundness of the approach.

***Index Terms*—** Kernel-based algorithm, density estimation, nonstationary signal classification

1. INTRODUCTION

Various practical situations involve the comparison of two sets of vectors. Important examples are, among others, speaker recognition where one speaker can be represented by a set of vectors in the 20-30 dimensional space of cepstral coefficients, or image processing where the number of texture features and interest points may vary from one object to another. More generally, the design of learning algorithms for classification, recognition, anomaly detection, is made possible by the ability to compare two sets (of vectors) with variable size, as occurs naturally with such representations as the *bag-of-pixels* in image processing or the *bag-of-words* in text processing.

A first possible approach consists of comparing each vector in one set to each vector in the other set using a pairwise similarity measure, such as the kernels used in reproducing kernel Hilbert space (RKHS) methods. A second possible approach considers each of the two sets of vectors as a single data object, and a higher level kernel is designed to compare these two objects. In the first situation, the similarity between the two sets is summarized in the *kernel matrix*; in the second, the similarity is summarized by a single value, which is required in many applications. In this paper, we propose a family of such kernels.

We now introduce some assumptions and notations. All vectors are assumed to lie in $\mathcal{X} \subset \mathbb{R}^p$ as in previous works in the same vein. The cardinality of the two sets may not be the same. In the following, we denote by $\boldsymbol{x} = \{x_1, \dots x_n\}$ with size n and $\boldsymbol{x}' = \{x'_1, \dots x'_{n'}\}$ with size n' the two sets to be compared. We propose an original methodology to design a (reproducing) kernel between $\boldsymbol{x}$ and $\boldsymbol{x}'$, denoted $k(\boldsymbol{x}, \boldsymbol{x}')$. Such a kernel can be used in any RKHS method.

Table 1: Examples of (reproducing) kernels $k(\cdot, \cdot)$ between two densities.

Kernels	Definitions
Entropy [3]	$\exp\left(-h\left(\frac{f+f'}{2}\right) + \frac{h(f)+h(f')}{2}\right)$ with $h(f) = -\int_{\mathcal{X}} f \ln f$
Inverse generalized variance [3]	$\dfrac{1}{\det\Sigma\left(\frac{F+F'}{2}\right)}$
Symmetric-χ^2 [1]	$\displaystyle\int \frac{ff'}{f+f'}$
Hellinger [1]	$\displaystyle\int \sqrt{ff'}$
Jensen-Shannon [1]	$-\int \left(f \log \frac{f}{f+f'} + f' \log \frac{f'}{f+f'}\right)$
Expected likelihood [2]	$\displaystyle\int ff'$
Total variation [1]	$\displaystyle\int \min(f, f')$
L_2 Exponentiated	$\displaystyle\int \exp(-\|f - f'\|^2/h_k)$

1.1. Related work

Considering a set of vectors as a single object is a common idea in learning problems. Standard approaches consist of choosing heuristically a representative for the set, e.g., its barycenter, the vector closest to the barycenter, or by fitting some parametric probability density function (pdf) model. Then, the two sets are compared with a norm or a distance in the space of parameters.

It is often the case, however, that no model nor simple pdf can represent the data well enough. Other existing solutions then consist of building the kernel from standard similarity measures between pdfs estimated using histograms: this is the approach followed in [1], where the positive definiteness of several kernels based on classical metrics (symmetric-χ^2, Jensen, total variation, etc.) is established. The Bhattacharrya kernel was already proposed in [2], though in a less general framework. Instead of comparing the densities, one can compare the probability measures: when these overlap, their sum is expected to be more concentrated. [3] defines semi-group kernels and prove the positive definiteness of kernels based on entropy and on inverse generalized variance. Table 1 summarizes all these kernels.

Histograms are known to provide poor density estimates, but computing a similarity measure between more complex estimates is often both costly and numerically unstable. Hence authors resort to approximations, some of which assume Gaussianity in feature space [2], or work within a tractable family of pdfs such as the exponential family [3]. The approach then steps back to the simpler estimation of a small number of parameters, generally the mean and variance. Aside these approaches, one may quote [4], that uses the structure of the statistical manifold to build a diffusion kernel. However, though developed in a theoretically sound and elegant framework, its positive definiteness is not established. Similarly, the Kullback-Leibler kernel derived in [5] is not positive-definite. The kernel in [6] is defined as a product of cosines, and is known to scale

1-4244-0728-1/07/$20.00 ©2007 IEEE

poorly with the number of learning vectors.

1.2. Contributions

We follow the approach of [1] as a starting point. We propose to use a more reliable nonparametric density estimate than a histogram, namely the Akaike-Parzen-Rosenblatt (APR) estimate [7]. The main advantage of this estimate is that its convergence rate is much faster than that of histograms. However, our approach is feasible only if it leads to a computationally tractable solution. We show how to build kernels between APR estimates that are both positive definite and indeed computationally tractable. Our approach is theoretically sound as no unnecessary restrictive assumption is made over the data, it is computationally tractable as no optimization is required to evaluate the kernel, and it admits simple yet efficient model selection rules. Most of all, it is not limited to small size sets, as is the case with other approaches.

1.3. Paper organization

The remainder of this paper is organized as follows. We describe and justify our approach in Section 2. We also introduce our kernels between sets of vectors $K(\boldsymbol{x}, \boldsymbol{x}')$, and their closed-form expressions are given as functions of the parameters of the APR estimate. In Section 3, we demonstrate these kernels in the problem of classifying nonstationary signals; simulation results are also presented for manifold-inspired learning applied to handwritten digits recognition. Conclusions and perspectives for future work are provided in Section 5.

2. COMPUTING KERNELS BETWEEN APR ESTIMATES

Assume that the learning vectors x_i's (for $i = 1, \ldots, n$) are distributed according to some unknown continuous probability density function (pdf) f. We restrict the presentation to so-called *translation invariant/spherical* kernels and use the notation $l(x, x') = \widetilde{l}\,(r)\,/h^p$ with $r = \|x - x'\|/h$ when appropriate (h is the *kernel bandwidth*). Given a kernel $\widetilde{l}(\cdot)$, the APR density estimator of f is given by: $\hat{f}_n(x) = 1/(nh_n^p) \sum_{i=1}^n \widetilde{l}\,((x - x_i)/h_n)$.

The fundamental result in APR density estimation is that provided $\widetilde{l}$ is normalized, that is, $\int \widetilde{l} = 1$, and L_1-integrable, that is, $\int |\widetilde{l}| < \infty$, necessary and sufficient conditions for consistency of the estimate are $h_n \longrightarrow 0$ and $nh_n^p \longrightarrow \infty$, where we recall that p is the dimension of $\mathcal{X}$. Remarkably, no additional condition needs to be imposed on the kernel: the APR estimate based on any normalized L_1 integrable kernel is consistent. In practice, we use the $o(n \log n)$ dual-tree implementation of [8] to achieve fast yet robust model selection and fast evaluations.

2.1. Example kernels

We now introduce two example APR-based kernels on sets: the expected likelihood kernel (first proposed in [2] along with the Gaussian assumption in $\mathcal{H}$), and the exponentiated L_2 kernel. We then show how to compute these kernels in a simple and efficient way.

2.1.1. Expected likelihood kernel

Given two estimates $\widehat{f}_n(x)$ and $\widehat{f'}_{n'}(x)$, obtained by independent APR estimations onto the sets $\boldsymbol{x}$ and $\boldsymbol{x}'$, the *expected likelihood kernel* is defined as $k(\boldsymbol{x}, \boldsymbol{x}') = \mathbb{E}_{\widehat{f'}_{n'}}\left[\widehat{f}_n(\cdot)\right] = \mathbb{E}_{\widehat{f}_n}\left[\widehat{f'}_{n'}(\cdot)\right]$, where $\mathbb{E}_f\left[\,\cdot\,\right]$ denotes expectation w.r.t. the density f. Using the expression of the APR estimate with kernel $l(\cdot, \cdot)$ yields

$$k(\boldsymbol{x}, \boldsymbol{x}') = \frac{1}{nn'} \langle \sum_{i=1}^{n} l(\cdot, x_i), \sum_{j=1}^{n'} l'(\cdot, x'_j) \rangle_{L_2} = \frac{1}{nn'} \sum_{i,j=1}^{n,n'} l_*(x_i, x'_j) \quad (1)$$

with $l'(x, x') = \frac{1}{h'^p} \widetilde{l}(\frac{x - x'}{h'})$, with $\langle \cdot, \cdot \rangle_{L_2(\mathcal{X})}$ the inner product on the space of square-integrable functions on $\mathcal{X}$, and where l_* results from the convolution of l with l': $l_* = l * l'$.

2.1.2. Exponentiated L₂ kernel

The exponentiated L_2 kernel is: $\exp\left[-\|\widehat{f}_n - \widehat{f'}_{n'}\|^2_{L_2}/h_k\right]$ where h_k is the kernel bandwidth parameter and $\|\cdot\|_{L_2}$ is the norm induced by the inner product $\langle \cdot, \cdot \rangle_{L_2}$. The evaluation of $k_2(\cdot, \cdot)$ requires the calculation of $\|\widehat{f}_n - \widehat{f'}'_{n'}\|^2_{L_2}$ which equals:

$$\frac{1}{n^2} \sum_{i,j}^{n} l_*(x_i, x_j) - \frac{2}{nn'} \sum_{i=1}^{n} \sum_{j=1}^{n'} l_*(x_i, x'_j + \frac{1}{n'^2} \sum_{i,j}^{n'} l_*(x'_i, x'_j) \quad (2)$$

2.2. Positive definiteness

It is straightforward to prove that the kernels defined above are positive definite, by embedding the set of densities in $L_2(\mathcal{X})$, in which case they are respectively the linear and Gaussian kernels over $L_2(\mathcal{X})$. It is both more interesting and more formal to see that these two kernels coincide with kernels defined directly between the sets $\boldsymbol{x}$ and $\boldsymbol{x}'$ as we now prove. The connection is inspired by the material in [9, Chapter 4] from which the below construction is taken; to simplify the exposition, we assume that $l = l'$ and that they are positive definite though both these assumptions are not required for the connection to hold.

Let $\mathcal{H}$ be a RKHS with kernel l^*. First step is to embed the space of signed measures $\mathcal{M}(\mathcal{X})$ in $\mathcal{H}$:

$$\begin{array}{ccccc} \mathbb{R}^p \times \ldots \mathbb{R}^p & \xrightarrow{\mu_n} & \mathcal{M}(\mathbb{R}^p) & \xrightarrow{\Gamma} & \mathcal{H} \\ \boldsymbol{x} = \{x_1, \ldots, x_n\} & \mapsto & \frac{1}{n} \sum\limits_{i=1}^{n} \delta_{x_i} & \mapsto & \text{representer of } \mu_n \text{ in } \mathcal{H} \end{array}$$

where the operator $\Gamma_{(\cdot)}$ acts on $\mu \in \mathcal{M}(\mathcal{X})$ as follows: $\Gamma_\mu = \int l^*(x, \cdot) d\mu(x)$, which leads us to define an inner product on $\mathcal{M}(\mathcal{X})$ as: $\langle \mu, \nu \rangle_{\mathcal{M}} \equiv \langle \Gamma_\mu, \Gamma_\nu \rangle_{\mathcal{H}}$. Then, plugging Eq. (**??**) in the expression of the inner product yields:

$$\begin{array}{rcl} \langle \mu, \nu \rangle_{\mathcal{M}} & = & \langle \Gamma_\mu, \Gamma_\nu \rangle_{\mathcal{H}} = \langle \int l^*(x, \cdot) d\mu(x), \int l^*(y, \cdot) d\nu(y) \rangle_{\mathcal{H}} \\ & = & \iint \langle l^*(x, \cdot), l^*(y, \cdot) \rangle_{\mathcal{H}} d\mu(x) d\nu(y) = \iint l^* d(\mu * \nu) \end{array}$$

Now, the measure we want to embed in $\mathcal{H}$ are the empirical measures on $\boldsymbol{x}$ and $\boldsymbol{x}'$. Applying Eq. (2.2) to $\mu = \delta_x$, $\nu = \delta_y$ yields:

$$\langle \delta_x, \delta_y \rangle_{\mathcal{M}} = \iint l^* d(\delta_x * \delta_y) = l^*(x, y)$$

therefore the inner product between the empirical measures on $\boldsymbol{x}$ and $\boldsymbol{x}'$ is $\frac{1}{nn'} \sum_{i=1}^{n} \sum_{j=1}^{n'} l^*(x_i, x'_j)$. Choosing $l^* = l * kl' (= l_*)$ establishes the connection for the kernel k, which proves that it is a positive definite kernel between sets.

Similar reasoning obviously holds for the exponentiated L_2 kernel, and more generally for positive definite spherical kernels expressing as $g(-\|\widehat{f}_n - \widehat{f'}_{n'}\|_{L_2})$ with g a function respecting e.g. Schoenberg condition (g is infinitely differentiable and such that $(-1)^i \frac{d^i g(r)}{dr^i} \geq 0$, for any integer $i \geq 0$) or Bochner condition (g is the inverse Fourier transform of a positive finite Borel measure).

We now focus on the calculation of l_*, which is the keystone of both kernels computation.

2.3. Kernel convolutions

We propose different approaches to calculate l_*. To start with, we consider the simplest case, where the APR estimate relies on the Gaussian kernel.

Gaussian kernel: As the convolution of a Gaussian kernel with itself yields a Gaussian kernel, in this case l_* is easy to compute, whatever the dimension p.

General spherical kernels: For more general cases, the kernel convolution can be computed efficiently for a number of translation invariant kernels (e.g., the trapezoidal kernel: $\frac{7+\cos(\|r\|)-8\cos^2(\|r\|)}{4\pi\|r\|^2}$). First, following [10], we note that

$$\widetilde{l}*\widetilde{l} = \mathcal{F}^{-1}\left(\mathcal{F}(\widetilde{l}*\widetilde{l})\right) = \mathcal{F}^{-1}\left(\mathcal{F}(\widetilde{l})^2\right) = \mathcal{F}^{-1}\left(\varphi^2\right) \qquad (3)$$

where $\mathcal{F}$ (resp. $\mathcal{F}^{-1}$) is the Fourier transform (resp. inverse Fourier transform) and with φ the Fourier transform of $\widetilde{l}$. This approach is computationally affordable as l is spherical: the complexity of the FFT is $O(n \log n)$. As an example, we provide the expression of $l_*(x_i, x_j)$:

$$l_*(x_i, x_j) = \int \widetilde{l}_*\,(x/h)\,\delta_{(x_i - x_j)}(x)dx = \widetilde{l}_*\,(\cdot/h) * \delta_{(x_i-x_j)}(\cdot) \quad (4)$$

Applying the Fourier transform yields:

$$\mathcal{F}\left(l_*(x_i, x_j)\right)(u) = hh'\varphi(uh)\varphi(uh')\exp\left(\mathrm{j}u(x_i - x_j)\right) \qquad (5)$$

with φ the Fourier transform of $\widetilde{l}$ and $\mathrm{j}^2 = -1$. Finally the inverse Fourier transform of Eq. (5) yields $l(x_i, x_j)$. More efficiently, the sum $\frac{1}{n'}\sum_{j=1}^{n} l_*(x_i, x_j)$ itself is computed by replacing $\delta_{(x_i-x_j)}(\cdot)$ by $\frac{1}{n'}\sum_{j=1}^{n}\delta_{(x_i-x_j)}(\cdot)$ in the last line of Eq. (4). We refer to [10] for technical details concerning the implementation of this approach, including dealing with boundary effects.

2.4. Kernel between APR estimates

In this Section we deal with choices for g. Due to the introduction of the Fourier transform φ of the kernel, it is natural to define the kernel directly by its Fourier transform. This proves to be specifically interesting for model selection purposes; it is common in the RKHS learning literature to require that the kernel is parameterized by a small number of parameters, typically 2 at most, so that splitting techniques (cross-validation) are computationally affordable when solving a learning problem. In our case, we can define hierarchies of kernels $\{k_s(\cdot,\cdot); (s, h)\}$ indexed over the kernel bandwidth h and its order s. We recall that the s^{th}-moment of a kernel, $s \geq 2$, is its first nonzero moment; s is related to the smoothness of the kernel, hence of the regressor based on k; for further details, we refer to [11]. Due to Bochner theorem, constraining kernels in the hierarchy to be positive definite is simple, as one just needs to check that the φ_s's are nonnegative. Two examples of hierarchies of kernels are the Hall-Marron class: $\varphi_s(\cdot) = \exp(-(\cdot)^s)$, or its modification as the Devroye-Lugosi class: $\varphi_s(\cdot) = 2\left(1 - \frac{1}{2}(\cdot)\right)\exp\left(-\frac{1}{2}(\cdot)^s\right) - (1 - (\cdot))\exp\left(-(\cdot)^s\right)$.

Another possible approach is to define higher-order kernels recursively, from an initial kernel k_{s_0}, generally for $s_0 = 2$. A useful recursive formula may be found in the density estimation literature: if the order of the kernels k_s and $k_{s'}$ respectively is s and s', then the order of $k_{s+s'}$ defined as $\widetilde{k}_{s+s'} = \widetilde{k}_s + \widetilde{k}_{s'} - \widetilde{k}_s * \widetilde{k}_{s'}$ is $(s + s')$. A simple study yields the following sufficient conditions for positive definiteness:

Proposition 2.1 *Assume that k_s is positive definite and s.t. $\widetilde{k}_s$ integrates to 1. Then k_{3s} s.t. $\widetilde{k}_{3s} = 3\widetilde{k}_s - 3\widetilde{k}_s * \widetilde{k}_s + \widetilde{k}_s * \widetilde{k}_s * \widetilde{k}_s$ is p.d.; if $\varphi_s < 2$ then k_{2s} s.t. $\widetilde{k}_{2s} = 2\widetilde{k}_s - \widetilde{k}_s * \widetilde{k}_s$ is p.d.*

A hierarchy of kernels may be induced from e.g. the Gaussian kernel (in the space domain this time) and an optimal couple (s, h) may be chosen by any splitting or resampling technique.

3. APPLICATION TO CLASSIFICATION

3.1. Unordered sets classification by Support Vector Machines

Let $\mathcal{S}$ denote the set of sets of vectors in $\mathcal{X}$. Let $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N\}$ be a set in $\mathcal{S}$, that is a collection of sets of vectors in $\mathcal{X}$; to each set $\boldsymbol{x}_j$, there corresponds a label y_j, chosen in a finite set of labels $\mathcal{Y}$. The objective of classification is to estimate a function $\mathsf{f} : \mathcal{S} \rightarrow \mathcal{Y}$ such that $\mathsf{f}(\boldsymbol{x}) = y$ for any set $\boldsymbol{x}$, with y its true label. We concentrate here on the 2-class problem: $y = \pm 1$, and the decision function is expressed with no loss of generality as $\text{sign}(\mathsf{f}(\cdot))$. Support Vector Machines (SVMs) for classification [12] solve this problem by minimizing a regularized risk, which trades fit to the data (in terms of the L_1 norm – the hinge loss) for complexity of f, chosen in a RKHS.

3.2. Nonstationary signal classification

Nonstationary signal classification consists of classifying time series according to their time-varying frequency contents using either time-frequency or time-scale representations. Here, we follow a time-frequency approach. The Time-Frequency Representations (TFRs) we considered are derived from the Wigner-Ville representation

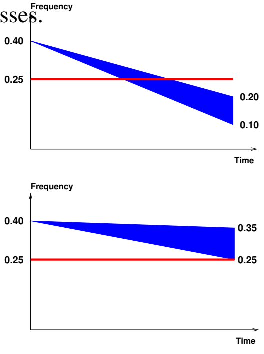
$$\mathcal{W}_s(t, \xi) \;=\; \int_{-\infty}^{\infty} s(t + \frac{\tau}{2})\, s(t - \frac{\tau}{2})\; e^{-\mathrm{j}2\pi\xi\tau} \qquad (6)$$

where $s(t)$ is a given time series (t denotes continuous time). A given TFR is computed from $\mathcal{W}_s(t, \xi)$ by a 2-D convolution with a *time-frequency kernel* $\phi(t, \xi)$. In particular, the smoothed pseudo Wigner-Ville representation is obtained by using $\phi(t, \xi) = h(t)g(\xi)$, where h (resp. g) is a time (resp. frequency) smoothing window. In practice, discretized versions of these continuous TFRs are used.

In order to enable comparison with alternative approaches, we consider the problem described in [13] and references therein. The learning set is composed of signals $x(t)$ which are the sum of two components, each being a sine wave with linear frequency modulation (*linear chirp*):

$$s(t) = A\sin\left[2\pi(a_0 + b_1 t)\right] + B\sin\left[2\pi(b_0 + b_1 t + b_2 t^2)\right] + \varepsilon(t) \quad (7)$$

for $t = 0, \ldots, T-1$, where the noise samples $\varepsilon(t)$ are distributed i.i.d. according to $\mathcal{N}(0, \sigma_\varepsilon^2)$. The parameters for the signals are $\{a_0, a_1, b_0, b_1, b_2, A, B, \sigma_\varepsilon\}$; both classes share the same parameters: $A = B = 1$, $a_0, b_0 \sim \mathcal{U}(0, 1)$, $a_1 = 0.25$, $b_1 = 0.40$ and $\sigma_\varepsilon^2 = 2$, except for b_2: $b_2 \sim \mathcal{U}\left(-0.30/2(T-1), -0.20/2(T-1)\right)$ for class 1 and $b_2 \sim \mathcal{U}\left(-0.15/2(T-1), -0.05/2(T-1)\right)$ for class 2. Fig. 1 plots the idealized TFRs of signals from the two classes.



Approach	Error (%)
Wigner distribution	22.30
MCMC classification (*)	5.24
Ambiguity plane (*)	4.56
Time-Frequency (*)	2.25
SVM linear (*)	1.98
SVM Gaussian (*)	1.40
SVM APR kernel	1.40

Fig. 1: Idealized time-frequency representations of the two classes of signal to be classified. The decreasing chirp may be anywhere in the blue triangle, and the two classes differ by the end frequency of this chirp (left). Classification results (error rates) for the nonstationary signal classification problem. Several previous approaches are compared to that developed in this paper, see [13] for details. (*): Approaches which optimize the TFR of the signals (much more costly from the computational perspective) (right).

We adopt the following approach: the signals to be classified are first mapped to the Time-Frequency domain using the smoothed pseudo Wigner-Ville representation. As opposed to state-of-the-art methods in the field, here we do not optimize the time-frequency kernel: this saves a lot of computations. Binarization is then applied to the TFRs. Coefficients with low amplitude are set to zero (they correspond to noise only) while high amplitude coefficients are set to one (they correspond to signal). In our experiments, threshold was chosen *a priori* and set to the value 25% of the peak amplitude in the TFR. This thresholding operations yields a sparse (t, ξ)-image. Each image is thus converted to a set of vectors, each vector being composed of the t and ξ coordinates of a non-zero pixel.

The kernel for the APR estimate is a Gaussian kernel with bandwidth $0.46\hat{\sigma}_n n^{-1/3}$ (this value was obtained by cross-validating a set of values picked uniformly around Deheuvels rule-of-thumb bandwidth value). A 2-class C-SVM is then trained with the exponentiated L_2 kernel and with cross-validation based model selection rules leading to the parameter $h_k = 10^{-3}$. Classification of a test sets of signals generated according to the model defined in Eq. (7) led to an error rate of 1.4%, which compares favorably with state-of-the-art approaches (See table in Fig 1). It should be noted that most state-of-the art methods require optimization of the time-frequency kernel $\phi(t, \xi)$, which requires much heavier computations than the implementation proposed here.

4. APPLICATION 2: MANIFOLD-INSPIRED LEARNING

4.1. Manifold-inspired learning

Manifold learning methods assume that the Euclidean distance may not be the most relevant way of comparing the data. Instead, learning is performed on a(n estimated) manifold of $\mathcal{X}$ where the data lies. Such methods include spectral clustering, and other nonlinear generalization of the Principal Component Analysis (PCA) such as the Locally Linear Embedding (LLE), ISOmetric feature MAPping [14] and Kernel Principal Component Analysis [12]. In this Section we focus on a smoothed version of KPCA performed on objects which are sets, and we compare it to the traditional ISOMAP algorithm on a standard dataset.

It is possible to define a smoothed extension of KPCA by solving the following optimization problem:

$$\min_{\alpha_1,\ldots,\alpha_n \in \mathbb{R}} \sum_{i=1}^{n} \|k(\boldsymbol{x}_i, \cdot) - \sum_{j=1}^{n} \alpha_j k(\boldsymbol{x}_j, \cdot)\|_{\mathcal{H}}^2 + \beta \sum_{i=1}^{n} |\alpha_i| \qquad (8)$$

with $\mathcal{H}$ a RKHS with kernel k and $\beta \geq 0$ a regularization parameter which can be optimally tuned with a sampling technique such as cross-validation (see, *e.g.*, [15]). The problem defined in Eq. (8) admits many solutions (up to n when k is nonsingular). It can be proved that this problem resort to the generalized eigenvalue problem consisting in diagonalizing $(\mathsf{k}_n + \beta I_n)$, where the kernel matrix $\mathsf{k}_n = [k(\boldsymbol{x}_i, \boldsymbol{x}_j)]_{i,j=1,\ldots,n}$ is calculated on the $\boldsymbol{x}_i$'s and with I_n the n-dimension identity matrix. Each vector $[\alpha_1, \ldots, \alpha_n]$ represents an empirical estimate of a principal curve for the data $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$, and all directions are ordered starting with the largest eigenvalue.

4.2. Manifold learning for handwritten digits recognition

We illustrate the smoothed KPCA algorithm on unordered sets of vectors by performing dimensionality reduction on 500 2's of the MNIST handwritten digits database, as was done in [14]. Each data is an image with size 28×28 pixels. Each pixel is quantified to be either white or black. Here, the unordered sets of vectors are composed of 2-D vectors, where each vector contains the x and y coordinate of black pixels in the 28×28 image. Hence, each image is represented by a set of vectors, which cardinality varies from one image to another. In Fig. 2, we compare the variance residuals obtained for our approach to those obtained with an implementation of the ISOMAP algorithm using 7 neighbors and the L_2-distance (no noticeable change was obtained with other distances). As it takes into account the structure of the image, our approach is able to recover a structure with lower dimension that the ISOMAP based on the vectorization of the image. The elbow corresponding to the dimensionality is more identifiable, and the residual variance converges faster to 0.

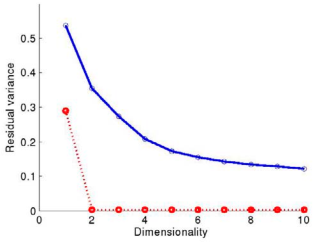


Fig. 2: Residual variance for ISOMAP (full line) and a smoothed KPCA on sets (dotted line). The latter exhibits increased performance both in terms of convergence to 0 of the residual variance, and of the estimation of dimensionality.

5. CONCLUSION

In this paper, we derive kernels between sets of vectors as similarities between APR density estimates. The approach we propose is computationally attractive, scales well in front of large dimensional data and large learning sets and relies on no restrictive assumption. Simulation results confirm the soundness of the approach.

6. REFERENCES

[1] M. Hein and O. Bousquet, "Hilbertian metrics and positive definite kernels on probability measures," in *AISTATS 2005*, 2005.

[2] T. Jebara, R.I. Kondor, and A. Howard, "Probability product kernels," *Journal of Machine Learning Research*, vol. 5, pp. 819–844, 2004.

[3] M. Cuturi, K. Fukumizu, and J.P. Vert, "Semigroup kernels on measures," *Journal of Machine Learning Research*, 2005.

[4] J. Lafferty and G. Lebanon, "Diffusion kernels on statistical manifolds," *Journal of Machine Learning Research*, vol. 6, pp. 129–163, 2005.

[5] P.J. Moreno, P.P. Ho, and N. Vasconcelos, "A kullback-leibler divergence based kernel for svm classification in multimedia applications," in *NIPS*, 2004.

[6] L. Wolf and A. Shashua, "Learning over sets using kernel principal angles," *Journal of Machine Learning Research*, vol. 4, pp. 913–931, Oct. 2003.

[7] B.W. Silverman, *Density estimation for statistics and data analysis*, Chapman and Hall, 1986.

[8] A. Ihler, "Kernel density estimation class (matlab)," .

[9] A. Berlinet and C. Thomas-Agnan, *Reproducing kernel Hilbert spaces in Probability and Statistics*, Kluwer Academic Press, 2004.

[10] B.W. Silverman, "Kernel density estimation using the fast fourier transform," *Applied Statistics*, vol. 31, pp. 93–99, 1982.

[11] P. Hall and J.S. Marron, "Choice of kernel order in density estimation," *The Annals of Statistics*, vol. 16(1), pp. 161–173, 1988.

[12] B. Schlkopf and A.J. Smola, *Learning with Kernels*, The MIT Press, 2002.

[13] M. Davy, A. Gretton, A. Doucet, and P. Rayner, "Optimized support vector machines for nonstationary signal classification," *IEEE Signal Processing Letters*, 2002.

[14] J.B. Tenenbaum, V. de Silva, and J.C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, pp. 2319–2323, 2000.

[15] J.O. Ramsay and B.W. Silverman, *Functional Data Analysis*, Springer, 1997.