Kernel-based Learning from Infinite Dimensional 2-way Tensors

Marco Signoretto[1], Lieven De Lathauwer[2], and Johan A. K. Suykens[1]

[1] Katholieke Universiteit Leuven, ESAT-SCD/SISTA
Kasteelpark Arenberg 10, B-3001 Leuven (BELGIUM)
[2] Group Science, Engineering and Technology
Katholieke Universiteit Leuven, Campus Kortrijk
E. Sabbelaan 53, 8500 Kortrijk (BELGIUM)

Abstract. In this paper we elaborate on a kernel extension to tensor-based data analysis. The proposed ideas find applications in supervised learning problems where input data have a natural 2−way representation, such as images or multivariate time series. Our approach aims at relaxing linearity of standard tensor-based analysis while still exploiting the structural information embodied in the input data.

1 Introduction

Tensors [8] are multidimensional $N-$way arrays that generalize the ordinary notions of *vectors* (first-order tensors or $1-$way arrays) and *matrices* (second-order tensors or $2-$way arrays). They find natural applications in many domains since many types of data have intrinsically many dimensions. Gray-scale images, for example, are commonly represented as second order tensors. Additional dimensions may account for different illuminations conditions, views and so on [13]. An alternative representation prescribes to *flatten* the different dimensions namely to represent observations as high dimensional vectors. This way, however, important structure might be lost. Exploiting a natural $2-$way representation, for example, retains the relationship between the row-space and the column-space and allows to find structure preserving projections more efficiently [7]. Still, a main drawback of tensor-based learning is that it allows the user to construct models which are linear in the data and hence fail in the presence of nonlinearity. On a different track kernel methods [11],[12] lead to flexible models that have been proven successful in many different context. The core idea in this case consists of mapping input points represented as $1-$way arrays $\{x^l\}_{l=1}^n \subset \mathbb{R}^p$ into a high dimensional inner-product space $(\mathcal{F}, \langle\cdot,\cdot\rangle)$ by means of a *feature map* $\phi : \mathbb{R}^p \to \mathcal{F}$. In this space, standard linear methods are then applied [1]. Since the feature map is normally chosen to be nonlinear, a linear model in the feature space corresponds to a nonlinear rule in $\mathbb{R}^p$. On the other hand, the so called *kernel trick* allows to develop computationally feasible approaches regardless of the dimensionality of $\mathcal{F}$ as soon as we know $k : \mathbb{R}^p \times \mathbb{R}^p \to \mathbb{R}$ satisfying $k(x, y) = \langle\phi(x), \phi(y)\rangle$.

When input data are $N-$way arrays $\{X^l\}_{l=1}^n \subset \mathbb{R}^{p_1 \times p_2 \times \cdots \times p_N}$, nonetheless, the use of kernel methods requires to perform flattening first. In light of this, our main contribution consists of an attempt to provide a kernel extension to tensor-based data analysis. In particular, we focus on $2-$way tensors and propose an approach that aims at relaxing linearity of standard tensor-based models while still exploiting the structural information embodied in the data. In a nutshell, whereas vectors are mapped into high dimensional vectors in standard kernel methods, our proposal corresponds to mapping matrices into high dimensional matrices that retain the original $2-$way structure. The proposed ideas find applications in supervised learning problems where input data have a natural $2-$way representation, such as images or multivariate time series.

In the next Section we introduce the notation and some basic facts about $2-$way tensors. In Section 3 we illustrate our approach towards an operatorial representation of data. Successively on Section 4 we turn into a general class of supervised learning problems where such representations are exploited and provide an explicit algorithm for the special case of regression and classification tasks. Before drawing our conclusions in Section 6 we present some encouraging experimental results (Section 5).

2 Data Representation through 2-way Tensors

In this Section we first present the notation and some basic facts about 2-way tensors in Euclidean spaces. In order to come up a with kernel-based extension we then discuss their natural extensions towards infinite dimensional spaces.

2.1 Tensor Product of Euclidean Spaces and Matrices

For any $p \in \mathbb{N}$ we use the convention of denoting the set $\{1, \ldots, p\}$ by $\mathbb{N}_p$. Given two Euclidean spaces $\mathbb{R}^{p_1}$ and $\mathbb{R}^{p_2}$ their *tensor product* $\mathbb{R}^{p_1} \bigotimes \mathbb{R}^{p_2}$ is simply the space of linear mappings from $\mathbb{R}^{p_2}$ into $\mathbb{R}^{p_1}$. To each pair $(a, b) \in \mathbb{R}^{p_1} \times \mathbb{R}^{p_2}$ we can associate $a \otimes b \in \mathbb{R}^{p_1} \bigotimes \mathbb{R}^{p_2}$ defined for $c \in \mathbb{R}^{p_2}$ by

$$(a \otimes b)c = \langle b, c \rangle a \tag{1}$$

where $\langle b, c \rangle = \sum_{i \in \mathbb{N}_{p_2}} b_i c_i$. It is not difficult to show that any $X \in \mathbb{R}^{p_1} \bigotimes \mathbb{R}^{p_2}$ can be written as a linear combination of *rank 1* operators (1). Furthermore, as is well known, any such element X can be identified by a matrix in $\mathbb{R}^{p_1 \times p_2}$. Correspondingly $\mathbb{R}^{p_1 \times p_2}$ or $\mathbb{R}^{p_1} \otimes \mathbb{R}^{p_2}$ denote essentially the same space and we may equally well write X to mean the operator or the corresponding matrix. Finally, the Kronecker (or tensor) product between $A \in \mathbb{R}^{w_1} \bigotimes \mathbb{R}^{p_1}$ and $B \in \mathbb{R}^{w_2} \bigotimes \mathbb{R}^{p_2}$, denoted by $A \otimes B$ is the linear mapping $A \otimes B : \mathbb{R}^{p_1} \bigotimes \mathbb{R}^{p_2} \rightarrow \mathbb{R}^{w_1} \bigotimes \in \mathbb{R}^{w_2}$ defined by

$$(A \otimes B)X = AXB^\top \tag{2}$$

where $B^\top$ denotes the adjoint (transpose) of B. This further notion of tensor product also features a number of properties. If X is a rank 1 operator $a \otimes b$, for example, then it can be verified that $(A \otimes B)(a \otimes b) = Aa \otimes Bb$.

2.2 Extension to Hilbert Spaces and Operators

Instead of Euclidean spaces, we now consider more general Hilbert spaces (HSs) $(\mathcal{H}_1, \langle\cdot,\cdot\rangle_{\mathcal{H}_1})$, $(\mathcal{H}_2, \langle\cdot,\cdot\rangle_{\mathcal{H}_2})$. The definitions and properties recalled above have a natural extension in this setting. In the general case, however, additional technical conditions are required to cope with infinite dimensionality. We follow [14, Supplement to Chapter 1] and restrict ourselves to Hilbert-Schmidt operators. Recall that a bounded operator $A : \mathcal{H}_2 \to \mathcal{H}_1$ has adjoint A^* defined by the property $\langle Ax, y\rangle_{\mathcal{H}_1} = \langle x, A^*y\rangle_{\mathcal{H}_2}$ for all $x \in \mathcal{H}_2,\ y \in \mathcal{H}_1$. It is of *Hilbert-Schmidt type* if

$$\sum_{i\in\mathbb{N}} \|Ae_i\|^2_{\mathcal{H}_1} < \infty \tag{3}$$

where $\|x\|^2_{\mathcal{H}_1} = \langle x, x\rangle_{\mathcal{H}_1}$ and $\{e_i\}_{i\in\mathbb{N}}$ is an orthonormal basis[3] of $\mathcal{H}_2$. The tensor product between $\mathcal{H}_1$ and $\mathcal{H}_2$, denoted by $\mathcal{H}_1 \bigotimes \mathcal{H}_2$, is defined as the space of linear operators of Hilbert-Schmidt type from $\mathcal{H}_2$ into $\mathcal{H}_1$. Condition (3) ensures that $\mathcal{H}_1 \bigotimes \mathcal{H}_2$ endowed with the inner-product

$$\langle A, B\rangle_{\mathcal{H}_1 \bigotimes \mathcal{H}_2} = \sum_{i\in\mathbb{N}} \langle Ae_i, Be_i\rangle_{\mathcal{H}_1} = \mathrm{trace}(B^*A) \tag{4}$$

is itself a HS. As for the finite dimensional case to each pair $(h_1, h_2) \in \mathcal{H}_1 \times \mathcal{H}_2$ we can associate $h_1 \otimes h_2$ defined by

$$(h_1 \otimes h_2)f = \langle h_2, f\rangle_{\mathcal{H}_2} h_1\ . \tag{5}$$

One can check that (5) is of Hilbert-Schmidt type and hence $h_1 \otimes h_2 \in \mathcal{H}_1 \bigotimes \mathcal{H}_2$. As for the finite dimensional case, elements of $\mathcal{H}_1 \bigotimes \mathcal{H}_2$ can be represented as sum of rank-1 operator (5). Finally let $A : \mathcal{H}_1 \to \mathcal{G}_1$ and $B : \mathcal{H}_2 \to \mathcal{G}_2$ be bounded Hilbert-Schmidt operators between HSs and suppose $X \in \mathcal{H}_1 \bigotimes \mathcal{H}_2$. The linear operator $X \to AXB^*$ is a mapping from $\mathcal{H}_1 \otimes \mathcal{H}_2$ into $\mathcal{G}_1 \otimes \mathcal{G}_2$. It is called Kronecker product between the *factors* A and B and denoted as $A \otimes B$. The sum of elements $A_1 \otimes B_1 + A_2 \otimes B_2$ corresponds to the mapping $X \to A_1XB_1^* + A_2XB_2^*$ and scalar multiplication reads $\alpha A \otimes B : X \to \alpha AXB^*$. With these operations the collection of tensor product operators we just defined can be naturally endowed with a vector space structure and further normed according to:

$$\|A \otimes B\| = \|A\|\|B\| \tag{6}$$

where $\|A\|$ and $\|B\|$ denote norms for the corresponding spaces of operators. One such norm is the Hilbert-Schmidt norm

$$\|A\| = \sqrt{\langle A, A\rangle_{\mathcal{H}_1 \otimes \mathcal{H}_1}} \tag{7}$$

where $\langle\cdot,\cdot\rangle_{\mathcal{H}_1\otimes\mathcal{H}_1}$ is defined as in (4). Another norm that recently attracted attention in learning is the trace norm (a.k.a. Schatten $1-$norm, nuclear norm

[3] If A is of Hilbert-Schmidt type then (3) actually holds for *any* basis.

or Ky Fan norm). For[4] $|A| = (A^*A)^{\frac{1}{2}}$ the trace norm of A is defined as:

$$\|A\|_\star = \text{trace}\left(|A|\right) \ . \tag{8}$$

3 Reproducing Kernels and Operatorial Representation

Our interest arises from learning problems where one wants to infer a mapping given a number of evaluations at data sites and corresponding output values. Hence we focus on the case where $(\mathcal{H}_1, \langle\cdot,\cdot\rangle_{\mathcal{H}_1})$ and $(\mathcal{H}_2, \langle\cdot,\cdot\rangle_{\mathcal{H}_2})$ are *Reproducing Kernel* HSs (RKHSs) [3] where such function evaluations are well defined. We briefly recall properties of such spaces and then turn into the problem of representing 2-way tensor input observations as high dimensional operators.

3.1 Reproducing Kernel Hilbert spaces

We recall that given an arbitrary set $\mathcal{X}$, a HS $(\mathcal{H}, \langle\cdot,\cdot\rangle)$ of functions $f : \mathcal{X} \to \mathbb{R}$ is a RKHS if for any $x \in \mathcal{X}$ the *evaluation functional* $L_x : f \mapsto f(x)$ is bounded. A function $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ is called a *reproducing kernel* of $\mathcal{H}$ if $k(\cdot, x) \in \mathcal{H}$ for any $x \in \mathcal{X}$ and $f(x) = \langle f, k(\cdot, x)\rangle$ holds for any $x \in \mathcal{X},\ f \in \mathcal{H}$. From the two requirements it is clear that $k(x, y) = \langle k(\cdot, y), k(\cdot, x)\rangle$ for any $(x, y) \in \mathcal{X} \times \mathcal{X}$. Hence, $\mathcal{H}$ is an instance[5] of the feature space $\mathcal{F}$ discussed in the introduction as soon as we let $\phi(x) = k(x, \cdot)$. The Moore-Aronszajn theorem [3] guarantees that any *positive definite kernel*[6] is uniquely associated to a RKHS for which it acts as reproducing kernel. Consequently, picking up a positive definite kernel such as the popular Gaussian RBF-kernel [11], implicitly amounts at choosing a function space with certain properties. Euclidean spaces $\mathbb{R}^p$ can be seen as specific instances of RKHSs. In fact the dual space[7] of a finite dimensional space is the space itself. Therefore, we may regard $\mathbb{R}^p$ as both the *input* space $\mathcal{X}$ and the space of linear functions $w(x) = \sum_{i\in\mathbb{N}_p} w_i x_i$. It is not difficult to check that the *linear kernel* $k(x, y) = \sum_{i\in\mathbb{N}_p} x_i y_i$ acts as reproducing kernel for this space.

3.2 2-way Operatorial Representation

So far we have defined tensor products and characterized the spaces of interest. We now turn into the problem of establishing a correspondence between an input matrix (a training or a test observation) $X \in \mathbb{R}^{p_1} \bigotimes \mathbb{R}^{p_2}$ with an element $\Phi_X \in \mathcal{H}_1 \bigotimes \mathcal{H}_2$. Notice that the standard approach in kernel methods corresponds to (implicitly) mapping $\text{vec}(X)$, where $\text{vec}(X) \in \mathbb{R}^{p_1 p_2}$ is a vector obtained for

[4] Given a positive operator T, by $T^{\frac{1}{2}}$ we mean the unique positive self-adjoint operator such that $T^{\frac{1}{2}}T^{\frac{1}{2}} = T$.

[5] Alternative feature space representations can be stated, see e.g. [5, Theorem 4].

[6] See e.g. [11] for a formal definition.

[7] The (continuous) dual of a space $\mathcal{X}$ is the space of all continuos linear mappings from $\mathcal{X}$ to $\mathbb{R}$.

example by concatenating the columns of X. On the contrary, our goal here is to construct Φ_X so that the structural information embodied in the original representation is retained. Recall that for $p = \min\{p_1, p_2\}$ the thin SVD [6] of a point X is defined as the factorization $X = U\Sigma V^\top$ where $U \in \mathbb{R}^{p_1 \times p}$ and $V \in \mathbb{R}^{p_2 \times p}$ satisfies $U^\top U = I_p$ and $V^\top V = I_p$ respectively and $\Sigma \in \mathbb{R}^{p \times p}$ has its only nonzero elements on the first $r = \mathrm{rank}(X)$ entries along the main diagonal. These elements are the ordered singular values $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$ whereas columns of U and V are called respectively left and right singular vectors. Equivalently

$$X = \sum_{i \in \mathbb{N}_r} \sigma_i u_i \otimes v_i \tag{9}$$

where $u_i \otimes v_i$ are rank-1 operators of the type (1) and the set $\{u_i\}_{i \in \mathbb{N}_r}$ and $\{v_i\}_{i \in \mathbb{N}_r}$ span respectively the column space $\mathcal{R}(X)$ and the row space $\mathcal{R}(X^\top)$. Let $\phi_1 : \mathbb{R}^{p_1} \to \mathcal{H}_1$ and $\phi_2 : \mathbb{R}^{p_2} \to \mathcal{H}_2$ be some feature maps. Based upon $\{u_i\}_{i \in \mathbb{N}_r}$ and $\{v_i\}_{i \in \mathbb{N}_r}$ we now introduce the *mode*-0 operator $\Gamma_U : \mathcal{H}_1 \to \mathbb{R}^{p_1}$ and the *mode*-1 operator $\Gamma_V : \mathcal{H}_2 \to \mathbb{R}^{p_2}$ defined, respectively, by

$$\Gamma_U h = \sum_{i \in \mathbb{N}_r} \langle \phi_1(u_i), h \rangle_{\mathcal{H}_1} u_i \quad \text{and} \quad \Gamma_V h = \sum_{i \in \mathbb{N}_r} \langle \phi_2(v_i), h \rangle_{\mathcal{H}_1} v_i \ . \tag{10}$$

Recall from Section 2.2 that by $\Gamma_U \otimes \Gamma_V$ we mean the Kronecker product between Γ_U and Γ_V, $\Gamma_U \otimes \Gamma_V : \mathcal{H}_1 \otimes \mathcal{H}_2 \to \mathbb{R}^{p_1} \otimes \mathbb{R}^{p_2}$. Under the assumption that $X \in \mathcal{R}(\Gamma_U \otimes \Gamma_V)$ we finally define $\Phi_X \in \mathcal{H}_1 \otimes \mathcal{H}_2$ by

$$\Phi_X := \arg\min \left\{ \|\Psi_X\|^2_{\mathcal{H}_1 \otimes \mathcal{H}_2} \ : \ (\Gamma_U \otimes \Gamma_V)\Psi_X = X, \ \Psi_X \in \mathcal{H}_1 \otimes \mathcal{H}_2 \right\} \ . \tag{11}$$

In this way X is associated to a minimum norm solution of an operatorial equation. Notice that the range $\mathcal{R}(\Gamma_U \otimes \Gamma_V)$ is closed in the finite dimensional space $\mathbb{R}^{p_1} \otimes \mathbb{R}^{p_2}$ and hence a solution Φ_X is guaranteed to exist. The following result

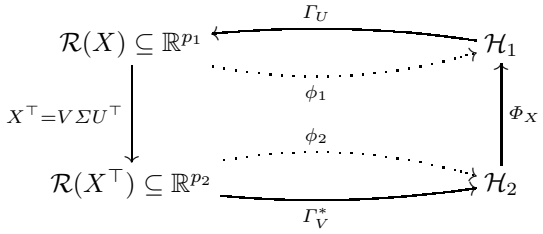


Fig. 1: A diagram illustrating the different spaces and mappings that we have introduced. The operator $\Phi_X \in \mathcal{H}_1 \otimes \mathcal{H}_2$ is the feature representation of interest.

that we state without proof due to space limitations, further characterizes Φ_X.

Theorem 1. *Let $A_U : \mathcal{H}_1 \to \mathbb{R}^r$ and $B_V : \mathcal{H}_2 \to \mathbb{R}^r$ be defined entry-wise as $(A_U h)_i = \langle \phi(u_i), h \rangle$ and $(B_V h)_i = \langle \phi(v_i), h \rangle$ respectively. The unique solution Φ_X of (11) is then given by*

$$\Phi_X = A_U^* Z B_V \tag{12}$$

where $Z \in \mathbb{R}^r \otimes \mathbb{R}^r$ *is any solution of*

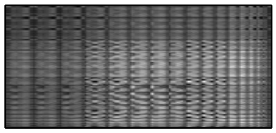
$$K_U Z K_V = \Sigma \tag{13}$$

where $(K_U)_{ij} = \langle \phi_1(u_i), \phi_1(u_j) \rangle$ *and* $(K_V)_{ij} = \langle \phi_2(v_i), \phi_2(v_j) \rangle$.

Fig. 2: An image (a) and its feature representation (b) for the case of 2−degree polynomial feature maps Φ_X was found based upon (12) and (13).



(a) A 19×18 image X



(b) Its 190×171 feature representation Φ_X (not in scale)

The approach can be easily understood for the case of polynomial kernel $k(x, y) = (\langle x, y \rangle)^d$ where $d > 1$ is an arbitrary degree [11]. Suppose this type of kernel is employed and $\phi_1,\ \phi_2$ in (10) denote the corresponding feature maps. Then K_U and K_V are identity matrices, $Z = \Sigma$ and

$$\Phi_X = \sum_{i \in \mathbb{N}_r} \sigma_i \phi_1(u_i) \otimes \phi_2(v_i)\ .$$

In particular when $d = 1$ (linear kernel), ϕ_1 and ϕ_2 denote the identity mapping and the latter formula corresponds to the factorization in (9).

4 Tensor-based Penalized Empirical Risk Minimization

We now turn into problem formulations where the generalized tensor-based framework presented might find application. Instead of working with matrix-shaped observations X (training or test points) the key idea consists in using their operatorial representation Φ_X.

4.1 A general class of supervised problems

We consider supervised learning and assume we are given a dataset consisting of input-output pairs $\mathcal{D} = \{(X_l, Y_l)\ :\ l \in \mathbb{N}_n\} \subset \mathcal{X} \times \mathcal{Y}$ where $\mathcal{X} \subset \mathbb{R}^{p_1} \otimes \mathbb{R}^{p_2}$ and $\mathcal{Y} \subset \mathbb{R}^{w_1} \otimes \mathbb{R}^{w_2}$. The situation where $\mathcal{Y} \subset \mathbb{R}^w$ or simply $\mathcal{Y} \subset \mathbb{R}$ is clearly a special case of this framework. Our goal is then to find a predictive operator

$$F : \Phi_X \mapsto \hat{Y} \tag{14}$$

mapping the operatorial representation Φ_X into a latent variable. This objective defines a rather broad class of problems that gives rise to different special cases. When the feature maps ϕ_1, ϕ_2 are simply identities, then Φ_X corresponds to X and we recover linear tensor models. In this case we have $F = A \otimes B : \mathbb{R}^{p_1} \otimes \mathbb{R}^{p_2} \to \mathbb{R}^{w_1} \otimes \mathbb{R}^{w_2}$, $F : X \mapsto AXB^\top$. This is the type of models considered, for example,

in [7]. Their problem is unsupervised and amounts to finding a pair of matrices $A \in \mathbb{R}^{w_1 \times p_1}$ and $B \in \mathbb{R}^{w_2 \times p_2}$ such that the mapping $X \mapsto AXB^\top$ constitutes a structure preserving projection onto a lower dimensional space $\mathbb{R}^{w_1 \times w_2}$. On the other hand for general feature maps ϕ_1, ϕ_2, we have $A \otimes B : \mathcal{H}_1 \otimes \mathcal{H}_2 \to \mathbb{R}^{w_1} \otimes \mathbb{R}^{w_2}$ and the predictive model becomes $A\Phi_X B^*$. For nonlinear feature maps, $A\Phi_X B^*$ defines a nonlinear model in X and thus we can account for possible nonlinearities. Here below for both the linear and nonlinear case we write Φ_l to mean Φ_{X_l}. Extending a classical approach, the problem of finding $A \otimes B$ can be tackled by penalized empirical risk minimization as:

$$\min \left\{ \sum_{l \in \mathbb{N}_n} c(Y_l, (A \otimes B)\Phi_l) + \frac{\lambda}{2} \|A \otimes B\|^2 \mid A : \mathcal{H}_1 \to \mathbb{R}^{w_1},\ B : \mathcal{H}_2 \to \mathbb{R}^{w_2} \right\} \tag{15}$$

where $c : (\mathbb{R}^{w_1} \otimes \mathbb{R}^{w_2}) \times (\mathbb{R}^{w_1} \otimes \mathbb{R}^{w_2}) \to \mathbb{R}^+$ is a *loss function* and the regularization term is based on the norm defined in (6) as $\|A \otimes B\| = \|A\|\|B\|$. Different norms for the factors are of interest. The use of Hilbert-Schmidt norm (7) corresponds to a natural generalization of the standard $2-$norm regularization used for learning functions [15]. However, recently there has been an increasing interest in vector-valued learning problems [9] and multiple supervised learning tasks [2]. In both these closely related class of problems the output space is $\mathbb{R}^w$. In this setting the nuclear norm (8) has been shown to play a key role. In fact, regularization via nuclear norm has the desirable property of favoring low-rank solutions [10].

Our next goal in this paper is to compare linear versus non-linear approaches in a tensor-based framework. Hence in the next Section we turn into the simpler case where outputs take values in $\mathbb{R}$. Before, we state a general representer theorem for the case where

$$c : \left(Y, \hat{Y}\right) \mapsto \frac{1}{2} \left\| Y - \hat{Y} \right\|_F^2 \tag{16}$$

and $\|\cdot\|_F$ denotes the Frobenius norm. The proof is not reported for space constraints.

Theorem 2 (Representer theorem). *Consider problem* (15) *where the loss is defined as in* (16) *,* $\|A \otimes B\| = \|A\|\|B\|$ *is such that* $\|A\|$ *is either the Hilbert-Schmidt norm* (7) *or the nuclear norm* (8) *and* B *is fixed. Then for any optimal solution* $\hat{A}$ *there exist a set of functions* $\{a_i\}_{i \in \mathbb{N}_{w_1}} \subset \mathcal{H}_1$ *such that for any* $i \in \mathbb{N}_{w_1}$

$$(\hat{A}h)_i = \langle a_i, h \rangle_{\mathcal{H}_1} \tag{17}$$

and for[8] $p = \min\{p_1, p_2\}$ *there is* $\alpha^i \in \mathbb{R}^{np}$ *so that*

$$a_i = \sum_{\substack{l \in \mathbb{N}_n \\ m \in \mathbb{N}_p}} \alpha_{lm}^i \phi_1(u_m^l) \ . \tag{18}$$

[8] Without loss of generality it is assumed that all the training matrices have rank p.

where u^l_m denotes the $m-$th left singular vector corresponding to the factorization of the $l-$th point $X_l = U_l \Sigma_l V_l^\top$.

A symmetric result holds if we fix A instead of B. This fact naturally gives rise to an alternating algorithm that we fully present for scalar outputs in the next Section.

4.2 The Case of Scalar Outputs

In this Section we focus on simple regression ($\mathcal{Y} \subset \mathbb{R}$) or classification ($\mathcal{Y} = \{+1, -1\}$) tasks. With respect to the general formulation (15) in this case the unknown operators are actually linear functionals $A : \mathcal{H}_1 \to \mathbb{R}$, $B : \mathcal{H}_2 \to \mathbb{R}$ and $\|\cdot\|$ boils down to the classical $2-$norm. By Theorem 2, the problem of finding A and B corresponds to finding single functions a and b which are fully identified by respectively $\alpha \in \mathbb{R}^{np}$ and $\beta \in \mathbb{R}^{np}$. On the other hand Theorem 1 ensures that the feature representation of the $l-$th point can be written as $\Phi_l = A^*_{U_l} Z_l B_{V_l}$ where Z_l is any solution of $K^U_{l,l} Z_l K^V_{l,l} = \Sigma_l$ and

$$(K^U_{l,m})_{ij} = \langle \phi_1(u^l_i), \phi_1(u^m_j) \rangle \ , \quad (K^V_{l,m})_{ij} = \langle \phi_2(v^l_i), \phi_2(v^m_j) \rangle \tag{19}$$

where u^l_i (resp. v^l_i) denotes the $i-$th left (resp. right) singular vector corresponding to the factorization of the $l-$th point $X_l = U_l \Sigma_l V_l^\top$. Relying on these facts the single task problem can be stated as

$$\min \left\{ \frac{1}{2} \sum_{l \in \mathbb{N}_n} \left(Y_l - \alpha^\top G^U_{:,l} Z_l G^V_{l,:} \beta \right)^2 + \frac{\lambda}{2} (\alpha^\top G^U \alpha)(\beta^\top G^V \beta) \ : \ \alpha \in \mathbb{R}^{np}, \ \beta \in \mathbb{R}^{np} \right\} \tag{20}$$

where $G^U, G^V \in \mathbb{R}^{np} \otimes \mathbb{R}^{np}$ are structured matrices defined block-wise as $[G^U]_{l,m} = K^U_{l,m}$ and $[G^V]_{l,m} = K^V_{l,m}$ and by $G^V_{l,:}$ and $G^U_{:,l}$ we mean respectively the $l-$th block row of G^V and the $l-$th block column of G^U. Define now the matrices $S_{\alpha,\beta}, S_{\beta,\alpha} \in \mathbb{R}^n \otimes \mathbb{R}^{np}$ row-wise as

$$(S_{\alpha,\beta})_{l,:} = \left(G^U_{:,l} Z^i G^V_{l,:} \beta \right)^\top \qquad \text{and} \qquad (S_{\beta,\alpha})_{l,:} = \alpha^\top G^U_{:,l} Z^i G^V_{l,:} \ .$$

A solution of (20) can be found iteratively solving the following systems of linear[9] equations dependent on each over

$$\left(S^\top_{\alpha,\beta} S_{\alpha,\beta} + \lambda_\beta G^U \right) \alpha = S^\top_{\alpha,\beta} y, \qquad \lambda_\beta := \lambda(\beta^\top G^V \beta) \tag{21}$$

$$\left(S^\top_{\beta,\alpha} S_{\beta,\alpha} + \lambda_\alpha G^V \right) \beta = S^\top_{\beta,\alpha} y, \qquad \lambda_\alpha := \lambda(\alpha^\top G^U \alpha) \ . \tag{22}$$

In practice, starting from a randomly generated $\beta \in \mathbb{R}^{np}$, we alternate between problems (21) and (22) until the value of the objective in (20) stabilizes. Once a solution has been found, the evaluation of the model on a test point $X_\star = U_\star \Sigma_\star V_\star^\top$ is given by $\alpha^\top G^U_{:,\star} Z^\star G^V_{\star,:} \beta$ where $Z^\star$ is any solution of $K^U_{\star,\star} Z^\star K^V_{\star,\star} = \Sigma_\star$, $(K^U_{\star,\star})_{ij} = \langle \phi_1(u^\star_i), \phi_1(u^\star_j) \rangle$ and

$$G^U_{:,\star} = \left[K^U_{1,\star} \ K^U_{2,\star} \dots K^U_{n,\star} \right]^\top, \qquad G^V_{\star,:} = \left[K^U_{\star,1} \ K^U_{\star,2} \dots K^U_{\star,n} \right] \ .$$

[9] The two systems are linear in the active unknown conditioned on the fixed value of the other.

5 Experimental Results

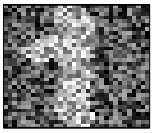
In linear tensor-based learning exploiting natural matrix representation has been shown to be particularly helpful when the number of training points is limited [7]. Hence in performing our preliminary experiments we focused on small scale problems. We compare a standard (vectorized) kernel approach versus our non-linear tensor method highlighted in Section 4.2. Both the type of kernel matrices in (19) were constructed upon the Gaussian RBF-kernel with the same value of width parameter. As standard kernel method we consider LS-SVM [12] also trained with Gaussian RBF-kernel. We do not consider a bias term as this is not present in problem (4.2) either. In both the cases we took a 20×20 grid of kernel width and regularization parameter (λ in problem (20)) and perform model selection via leave-one-out cross-validation (LOO-CV).

Robot Execution Failures [4]. Each input data point is here a 15×6 multivariate time-series where columns represent a force or a torque. The task we considered was to discriminate between two operating states of the robot, namely *normal* and *collision_ in_ part*. Within the 91 observations available, n were used for training and the remaining $n - 91$ for testing. We repeated the procedure over 20 random split of training and test set. Averages (with standard deviation in parenthesis) of Correct classification rates (CCR) of models selected via LOO-CV are reported on Table 1 for different number n of training points. Best performances are highlighted.

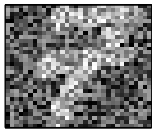
Table 1: Test performances for the Robot Execution Failures Data Set.

Correct Classification rates				
	n=5	n=10	n=15	n=20
RBF-LS-SVM	0.55(0.06)	0.64 (0.08)	0.66 (0.08)	0.70(0.06)
RBF-Tensor	**0.62(0.07)**	**0.66(0.08)**	**0.68(0.10)**	**0.71(0.11)**

Optical Recognition of Handwritten Digits [4]. Here we considered recognition of handwritten digits. We took 50 bitmaps of size 32×32 of handwritten 7s and the same number of 1s and add noise to make the task of discriminating between the two classes more difficult (Figure 3(a) and 3(b)). We followed the same procedure as for the previous example and report results on Table 2.



(a) A noisy 1



(b) A noisy 7

Correct Classification rates			
RBF-LS-SVM		RBF-Tensor	
n=5	n=10	n=5	n=10
0.71(0.20)	0.85 (0.14)	**0.84 (0.12)**	**0.88(0.09)**

Fig. 3 & Table 2: Instances of handwritten digits with high level of noise ((a) and (b)) and CCR on test for different number n of training points.

6 Conclusions

We focused on problems where input data have a natural 2-way representation. The proposed approach aims at combining the flexibility of kernel methods with the capability of exploiting structural information typical of tensor-based data analysis. We then presented a general class of supervised problems and gave explicitly an algorithm for the special case of regression and classification problems.

Acknowledgements

Research supported by Research Council KUL: GOA Ambiorics, GOA MaNet, CoE EF/05/006 Optimization in Engineering(OPTEC), IOF-SCORES4CHEM. Flemish Government: FWO: PhD/postdoc grants, projects: G0226.06 (cooperative systems and optimization), G0321.06 (Tensors), G.0302.07 (SVM/Kernel), G.0588.09 (Brain-machine) research communities (ICCoS, ANMMM, MLDM); IWT: PhD Grants, Eureka-Flite+, SBO LeCoPro, SBO Climaqs, SBO POM, O&O-Dsquare Belgian Federal Science Policy Office: IUAP P6/04 (DYSCO, Dynamical systems, control and optimization, 2007-2011); EU: ERNSI; FP7-HD-MPC (INFSO-ICT-223854), COST intelliCIS, FP7-EMBOCON (ICT-248940).

References

1. M. Aizerman, E.M. Braverman, and L.I. Rozonoer, *Theoretical foundations of the potential function method in pattern recognition learning.*, Automation and Remote Control **25** (1964), 821 – 837.
2. A. Argyriou, T. Evgeniou, and M. Pontil, *Multi-task feature learning*, Advances in Neural Information Processing Systems **19** (2007), 41.
3. N. Aronszajn, *Theory of reproducing kernels*, Transactions of the American Mathematical Society **68** (1950), 337 – 404.
4. A. Asuncion and D.J. Newman, *UCI machine learning repository http://www.ics.uci.edu/∼mlearn/MLRepository.html*, 2007.
5. A. Berlinet and C. Thomas-Agnan, *Reproducing Kernel Hilbert Spaces in Probability and Statistics*, Kluwer Academic Publishers, 2004.
6. G.H. Golub and C.F. Van Loan, *Matrix Computations*, third ed., Johns Hopkins University Press, 1996.
7. X. He, D. Cai, and P. Niyogi, *Tensor subspace analysis*, Advances in Neural Information Processing Systems **18** (2006), 499.
8. T.G. Kolda and B.W. Bader, *Tensor decompositions and applications*, SIAM review **51** (2009), no. 3, 455–500.
9. C.A. Micchelli and M. Pontil, *On learning vector-valued functions*, Neural Computation **17** (2005), no. 1, 177–204.
10. B. Recht, M. Fazel, and P.A. Parrilo, *Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization*, To appear in SIAM Review.
11. B. Schölkopf and A.J. Smola, *Learning with kernels: support vector machines, regularization, optimization, and beyond*, MIT Press, 2002.
12. J.A.K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor, and J. Vandewalle, *Least squares support vector machines*, World Scientific, 2002.
13. M. Vasilescu and D. Terzopoulos, *Multilinear subspace analysis of image ensembles*, IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2, 2003.
14. N.I.A. Vilenkin, *Special functions and the theory of group representations*, American Mathematical Society, 1968.
15. G. Wahba, *Spline models for observational data*, CBMS-NSF Regional Conference Series in Applied Mathematics, vol. 59, SIAM, Philadelphia, 1990.