# Multi-Task Kernel Null-Space for One-Class Classification

Shervin Rahimzadeh Arashloo and Josef Kittler, *Life Member*

**Abstract**—The one-class kernel spectral regression (OC-KSR), the regression-based formulation of the kernel null-space approach, has been shown to be an effective Fisher criterion-based methodology for one-class classification (OCC), achieving the state-of-the-art performance while providing a relatively high robustness against data corruption. This work extends the OC-KSR methodology to a multi-task setting where multiple one-class problems share information for improved performance. Accordingly, first, the OC-KSR method is extended to learn the multiple tasks structure *linearly* by posing it as an instance of the separable kernel learning problem in a vector-valued reproducing kernel Hilbert space where a linear output kernel encodes the tasks structure while another kernel captures input similarities. Next, by viewing the multi-task structure learning problem as one of composition function learning, a non-linear structure learning mechanism is proposed, which models the relationship between multiple tasks more effectively via a *non-linear* output kernel. The non-linear structure learning method is then reformulated for a sparse setting where different tasks compete in an output composition mechanism, leading to a sparse non-linear structure between the multiple problems. Through extensive experiments conducted on different data sets, the merits of the proposed multi-task kernel null-space techniques are validated and benchmarked against baseline and existing state-of-the-art techniques.

**Index Terms**—one-class classification, anomaly detection, multi-task learning, kernel null-space technique, reproducing kernel Hilbert space for vector valued functions (RKHSvv), regression.

## 1 Introduction

In the presence of large within-class variations, pattern classification techniques typically require a sufficiently large and representative set of training data to achieve a reasonable generalisation performance. With the growing complexity of the learning problems and the corresponding decision-making systems, the need for larger sets of training data has become indisputable. While there exist applications where the available data is abundant, there are other situations where the number of training observations is hard to increase. Such situations arise when the cost of collecting training samples is relatively high or samples are rare by nature. In other cases, where sufficient training observations are available, for effective training, multiple passes through the available samples may be required, increasing the computational complexity of learning. The problems associated with the training data are also pertinent to the settings where a large number of training observations might exist, but they fail to capture the real distribution of the underlying phenomena. In these situations, any deficiencies of the training data, on top of the limitations of the learning systems, may lead to a sub-optimal performance. Although other alternatives exist, in these circumstances, sharing knowledge among multiple tasks, facilitated by the multi-task learning (MTL) paradigm, has been found to be an effective strategy to improve the performance when individual problems are in some sense related [1]. Sharing knowledge among multiple problems may enhance the generalisation performance of individual learners, reduce the required number of training samples or the number of learning cycles needed to achieve a particular performance level by exploiting the commonalities/differences among different problems. As such, MTL is known to be an effective mechanism of inductive transfer, which enhances generalisation by exploiting the domain information available in the training signals of individual problems as an inductive bias [2]. This objective is typically achieved by learning multiple tasks in parallel, while using a shared representation.

Notwithstanding other strategies, the MTL approach may be cast within the framework of the reproducing kernel Hilbert space for vector-valued functions (RKHSvv) [3]. In this context, the problem may be viewed as one of learning vector-valued functions where each vector component is a real-valued function corresponding to a particular task. In the RKHSvv, the relationship between multiple inputs and the outputs is modelled through a positive definite multi-task kernel [4]. A plausible and computationally attractive simplification of this methodology is offered by the separable kernel learning paradigm, assuming a decomposition of the multi-task kernel in terms of a kernel on the inputs and another on task indices [4], [5], [6]. In this formalism, the input and outputs are decoupled in the sense that the input feature space does not vary by task while the structure of different problems is solely represented through the corresponding output kernel. Since a decomposition of the multi-task kernel facilitates the optimisation of the kernel on the task indices simultaneously with learning the predictive vector-valued function, it is widely applied as a kernel-based approach to model learning problems with multiple outputs.

The MTL strategy has been successfully applied to a wide spectrum of different problems [1]. Among others, a

- *S.R. Arashloo is with the centre for vision, speech and signal processing (CVSSP), university of Surrey, Guildford, GU2 7XH, UK. E-mail: s.rahimzadeh@surrey.ac.uk*
- *J. Kittler is with the centre for vision, speech and signal processing (CVSSP), university of Surrey, Guildford, GU2 7XH, UK. E-mail: j.kittler@surrey.ac.uk*

*Manuscript received ? ?, 2019; revised ? ?, 2019.*

relatively challenging classification problem is known to be one-class classification (OCC) [7]. OCC is defined as the problem of identifying patterns which conform to a specific behaviour, known as normal/target observations and distinguishing them from all other patterns, referred to as anomalies/novelties, etc. The interest in one-class learning is fuelled, in part, by the observation that very often a closed form definition of normality does exist whereas typically no such definition for an anomalous state is available. While one-class classification forms the backbone of a wide variety of applications [8], [9], [10], [11], [12], [13], [14], it usually suffers from a lack of representative training samples. The complexity of the problem may be attributed to the difficulty of obtaining non-target samples for training or their propensity to appear in unpredictable novel forms during the operational phase of the system.

These adversities suggest that the OCC problem may be a strong candidate to benefit from a multi-task learning strategy. While there exists some previous effort on utilising tasks' structures in designing one-class classification methods [15], [16], [17], [18], they typically rely on different flavours of the support vector machine paradigm. A plausible alternative to the SVM formulation is regularised regression [19]. By utilising the shared information across multiple relevant targets in a non-OCC setting, the performance of multi-target regression has been shown to improve [20], [21], [22], [23], [24]. Nevertheless, certain challenges exist in the context of multi-target regression that relate, for instance, to jointly modelling inter-target dependencies and non-linear input-output relations [22]. Very often in practice, multiple outputs represent higher level concepts which form highly complex relationships that call for powerful non-linear regression models, commonly formulated in the reproducing kernel Hilbert space. Despite the relative success achieved in the multi-target regression problem (in a general context beyond the OCC setting), the relationship among multiple tasks is typically modelled via a *linear* kernel, which limits the representational capacity of the existing methods.

In the current study, the kernel null-space technique for one-class classification [25], [26], [27], and in particular, its regression-based formulation, known as one-class kernel spectral regression (a.k.a. OC-KSR) [28], [29], is extended to a multi-task learning framework. The OC-KSR method, as compared to other alternatives, has been found to provide better performance and computational efficiency, while being more resilient to data corruption. In the context of the OC-KSR method, we show that the relationship among multiple related OCC problems may be encoded effectively by learning related tasks concurrently, based on the notion of separability of the multi-task kernel. To this end, multiple one-class learning problems are modelled as the components of a vector-valued function, while learning their structure corresponds to choosing suitable functional spaces.

### 1.1 Overview of the proposed approach

As noted earlier, in this work, the kernel regression-based formulation of the Fisher null-space technique for the one-class classification problem is reformulated to benefit from a multi-task learning strategy. For this purpose, first, it is shown that the kernel decomposition approach for learning vector-valued functions in the Hilbert space is directly applicable to the OC-KSR methodology, which in turn facilitates learning a predictive one-class vector-valued function and a linear structure among multiple tasks, concurrently. Next, as a second contribution, and in contrary to the common approach, which assumes a linear inter-target relation (modelled as a single output composition matrix), a new *non-linear* multi-task structure learning method is proposed, where the relationship among multiple OCC problems is encoded via a non-linear kernel function. The task-specific coefficients, as well as the output mixing parameters, are then learned concurrently via a new alternating direction block minimisation method. Finally, it is illustrated that the proposed non-linear approach for one-class vector-valued function learning may naturally be extended to a group-sparse representation, where different tasks interact in a sparse non-linear multi-task structure.

To summarise, the main contributions of the current study may be stated as:

- We extend the Fisher null-space one class classification approach to the multi-task case by means of a separable kernel learning, where the structure among multiple problems is captured *linearly* in terms of an output composition matrix;
- We generalise the multi-task Fisher null-space one-class learning approach to the non-linear case by modelling the composition function defining the task structure using a *non-linear* kernel function;
- We propose an extension of the non-linear multi-task structure learning mechanism to a sparse setting, where the structure between multiple problems is encoded in a group-sparse fashion;
- We validate the merit of the advocated methodology by its extensive experimental evaluation on a range of data sets and compare its different variants to existing approaches.

### 1.2 Outline of the paper

The rest of the paper is organised as follows: a summary of the existing work on multi-task one-class learning, as well as a brief overview of non-OCC multi-target regression approaches, most relevant to the current study, is provided in §2. In §3, once an overview of the one-class kernel spectral regression method for one-class learning [28], [29] is presented, the vector-valued function learning methodology, with an emphasis on separable kernel learning in the RKHSvv, is briefly reviewed. The discussion is then followed by its generalisation, formulated as a problem of composition function learning in the RKHSvv. The proposed multi-task one-class kernel null-space approach is introduced in §4, where the linear and non-linear structure learning mechanisms subject to Tikhonov, as well as sparse regularisation are presented. An experimental evaluation of the proposed multi-task structure learning methods on different data sets is carried out in §5, where a comparison with the baseline, as well as other existing approaches in the literature, is also discussed. Finally, §6 offers brief conclusions.

## 2 Related Work

In this section, a brief overview of the existing multi-task one-class learning approaches is presented. A number of non-OCC multi-target regression methods, relevant to the present study, shall be briefly reviewed too. For a detailed review on multi-task learning cf. [1].

As instances of the multi-task learning approaches for OCC, two multi-task learning formulations based on one-class support vector machines are presented in [15]. They are based on the assumption of closeness of the related tasks and the proximity of their corresponding models. Both multi-task learning problems are solved by optimising the objective function of a single one-class SVM. The work in [16], presents a multi-task approach, which incorporates additional new features in the one-class classification task. In [17], based on the one-class $\nu$-SVM, an MTL framework for one-class classification is presented, which constrains different problems to have similar solutions. Such a formulation is cast as a second-order cone programme to derive a global solution. In [18], the authors propose a method for anomaly detection, when collectively monitoring many complex systems. The proposed multi-task learning approach is based on a sparse mixture of Gaussian graphical models (GGM's), where each task is represented by a mixture of GGM's, providing the functionality to handle multiple modalities in the data. A new regularised formulation is then proposed with guaranteed sparsity in mixture weights. By introducing a vector-valued function subject to regularisation in the vector-valued reproducing kernel Hilbert space, an unsupervised classifier to detect the outliers and inliers simultaneously is proposed in [19], where preserving the local similarity of data in the input space is encouraged via manifold regularisation.

In the general context of multi-target regression and apart from the one-class classification paradigm, there exist a variety of different methods. These methods are not directly related to the present work in that they do not solve a one-class classification problem. Nevertheless, similar to the current study, in these methods, the multi-task learning problem is formulated as one of kernel regression. As an instance, in [20], an output kernel learning method, based on the solution of a suitable regularisation problem over a reproducing kernel Hilbert space of vector-valued functions, is proposed. A block-wise coordinate descent method is then derived that efficiently exploits the structure of the objective functional. The work in [21], addresses the MTL problem by illustrating that multiple tasks and their structure can be efficiently learned by formulating the problem as a convex optimisation problem, which is then solved by means of a block coordinate method. More recently, the authors in [22] propose a multi-target regression approach via robust low-rank learning. Their approach can encode inter-target correlations in a structure matrix by matrix elastic nets. Other method [23] models intrinsic inter-target correlations and complex non-linear input-output relationships via multi-target sparse latent regression, where inter-target correlations are captured via $L_{2,1}$-norm-based sparse learning. The work in [30] presents a two layer approach to jointly learn latent features shared by the tasks and a multi-task model based on Gaussian processes. In [24], in order to take into account the structure in the input data, while benefiting from kernels in the input space, the reproducing kernel Hilbert space theory for vector-valued functions is applied. In [31], the objective for multitask learning is formulated as a linear combination of two sets of eigen-functions such that the eigen-functions of one task provide additional information to the other and help to improve its performance. For a detailed review on multi-target regression one may consult [32].

## 3 Background

### 3.1 One-Class Kernel Spectral Regression

The Fisher criterion is a design objective commonly applied in the statistical pattern recognition, where a projection function from the input space into a feature space is inferred in a way that the between-class scatter of the data is maximised, while minimising the within-class scatter:

$$\boldsymbol{\varphi}^{\star} = \arg\max_{\boldsymbol{\varphi}} \frac{\boldsymbol{\varphi}^{\top}\mathbf{S}_b\boldsymbol{\varphi}}{\boldsymbol{\varphi}^{\top}\mathbf{S}_w\boldsymbol{\varphi}} \tag{1}$$

where $\mathbf{S}_b$ denotes the between-class scatter matrix, $\mathbf{S}_w$ stands for the within-class scatter matrix and $\boldsymbol{\varphi}$ is a basis corresponding to one axis of the subspace. A theoretically optimal projection that provides the best separability with respect to the Fisher criterion is the *null* projection [25], [26], [28], yielding a positive between-class scatter while providing a zero within-class scatter:

$$\begin{aligned} \boldsymbol{\varphi}^{\top}\mathbf{S}_w\boldsymbol{\varphi} &= 0 \\ \boldsymbol{\varphi}^{\top}\mathbf{S}_b\boldsymbol{\varphi} &> 0 \end{aligned} \tag{2}$$

In a one-class classification problem, the single optimiser for Eq. 1 is found as the eigenvector corresponding to the largest eigenvalue of the following generalised eigen-problem:

$$\mathbf{S}_b\boldsymbol{\varphi} = \lambda\mathbf{S}_w\boldsymbol{\varphi} \tag{3}$$

Once the null projection direction is determined, a sample $x$ is projected onto the null-space as

$$\mathbf{y} = \boldsymbol{\varphi}^{\top}\mathbf{x} \tag{4}$$

In order to handle data with an inherently non-linear structure, kernel extensions of this methodology are proposed [25], [26], [28]. While solving for the discriminant in a kernel space requires eigen-analysis of dense matrices, a computationally efficient method (one-class kernel spectral regression, a.k.a. OC-KSR) based on the spectral regression is proposed in [28], which poses the problem as one of solving a regularised regression problem in the Hilbert space:

$$\mathbf{a}^{opt} = \arg\min_{\mathbf{a}} \|\mathbf{K}\mathbf{a} - \mathbf{r}\|_2^2 + \gamma\mathbf{a}^{\top}\mathbf{K}\mathbf{a} \tag{5}$$

where $\gamma$ is a regularisation parameter, $\mathbf{r}$ denotes the desired responses and $\mathbf{K}$ stands for the kernel matrix. The optimal solution $\mathbf{a}^{opt}$ to the problem above is given as

$$\mathbf{a}^{opt} = (\mathbf{K} + \gamma\mathbf{I_n})^{-1}\mathbf{r} \tag{6}$$

where $\mathbf{I_n}$ denotes an identity matrix of size $n$ ($n$ being the number of training samples). Once $\mathbf{a}^{opt}$ is determined, the projections of samples onto the null feature space are found

as $\mathbf{y} = \mathbf{K}\mathbf{a}^{opt}$. For classification, the distance between the projection of a test sample and that of the mean of the target class is employed as a dissimilarity criterion.

In the conventional single-task OC-KSR approach, the procedure starts with building a separate kernel matrix for each one-class classification problem, followed by assigning optimal responses to each individual observation in each task. The optimal response vector $\mathbf{r}$ in the OC-KSR algorithm, when only positive instances are available for training, is shown to be a vector of ones (up to a scale factor). When negative training observations are also available, they are mapped onto the origin [28].

### 3.2 Vector-Valued Functions in the Hilbert Space

Let us assume there exist $T$ scalar learning problems (tasks), each associated with a training set $D_t$ of $n_t$ input-output observations $D_t = \{(x_i^t, r_i^t)\}_{i=1}^{n_t}$ with $x_i^t \in \mathcal{X}$ denoting the input space and $r_i^t \in \mathbb{R}$ denoting the output space data, and $t \in \{1, \ldots, T\}$ indexing a task. Given a loss function $\mathcal{L} : \mathbb{R}\times\mathbb{R} \rightarrow \mathbb{R}_+$ that measures the per-task prediction errors, the goal in the problem of learning vector-valued functions in the Hilbert space is to estimate a function $\mathbf{f}(.)$ which jointly minimises the regularised errors corresponding to multiple learning problems, i.e. $\mathbf{f}^*(.) = \arg\min_{\mathbf{f}\in\mathcal{H}} Q_L$, where $Q_L$ is defined as

$$Q_L = \sum_{t=1}^{T} \frac{1}{n_t} \sum_{i=1}^{n_t} \mathcal{L}(r_i^t, f_t(x_i^t)) + \mathcal{R}(\mathbf{f}) \tag{7}$$

$\mathcal{R}(\mathbf{f})$ denotes a regularisation imposed on the function $\mathbf{f}(.)$, with scalar components $f_t$, in the Hilbert space.

The multi-task learning approach is applicable to different classification problems. One possible example is that of object classification, where multiple hypotheses are considered in parallel, each hypothesis being a different OCC task and the goal is to exploit the correlation between the predictions of different objects through a multi-task learning approach. A different example is that of detecting a presentation attack (spoofing) in biometrics, where an unauthorised user attempts to gain access as a legitimate subject. The detection problem may be cast as an OCC problem (task) [33]. The goal may then be to discover any dependencies among multiple spoofing detection problems through the use of a multi-task learning strategy for improved performance. A further example may be that of multi-modal classification, where multiple sensing modalities are fused in order to exploit the complementary information they provide. In this case, the classification based on each modality may be considered as a separate task and the goal would be to combine the information coming from multiple sources, taking into account possible dependencies between them through a multi-task learning strategy.

A popular sub-class of vector-valued function learning methods is that of multi-target kernel *regression*, where the loss function $\mathcal{L}$ encodes a least squares loss in the Hilbert space. A commonly applied simplifying assumption in this case is that of the separability of input-output relations, which leads to expressing $\mathbf{f}(.)$ in terms of a separable kernel. Separable kernels are functions of the form $\mathbf{\Gamma}(x, x') = \kappa(x, x')\mathbf{B}$, where $\kappa : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ is a scalar reproducing kernel that captures similarities between the inputs and $\mathbf{B}$ is a symmetric positive semi-definite $T \times T$ matrix encoding dependencies among the outputs. In this case, $\mathbf{f}(.)$ is represented as

$$\mathbf{f}(.) = \sum_{i=1}^{n} \kappa(x_i, .)\mathbf{B}\mathbf{a}_i \tag{8}$$

where $\mathbf{a}_i$ stands for the coefficients. The output on the training data shall then be derived as $\mathbf{KAB}$ and the regularised loss given in Eq. 7 may be expressed in a matrix form as

$$Q_L = \|\mathbf{KAB} - \mathbf{R}\|_2^2 + \mathcal{R}(\mathbf{K}, \mathbf{A}, \mathbf{B}) \tag{9}$$

where $\mathbf{K}$ denotes the kernel matrix for the inputs, $\mathbf{A}^{n\times T}$ ($n = \sum_{t=1}^{T} n_t$) stands for a matrix the coefficient vectors $\mathbf{a}_i$'s denote its transposed rows and $\mathbf{R}$ denotes a matrix collection of the expected responses, while $\|.\|_2^2$ denotes the Frobenius norm. For this class of kernels, if $\mathbf{B}$ is the identity matrix, all outputs would be treated as being unrelated and the solution to the multi-task problem will be simplified to that of solving each task independently.

When the output structure matrix $\mathbf{B}$ is presumed to be other than the identity matrix, the tasks are regarded as being related and finding the optimal function $\mathbf{f}(.)$ is posed as the problem of learning the matrices $\mathbf{A}$ and $\mathbf{B}$, concurrently, subject to suitable regularisation constraints. The generic form of $Q_L$ in Eq. 9 may be considered as the common formulation to the multi-target regression problem in the Hilbert space, where a specific choice for $\mathcal{R}$ may be based on different apriori assumptions, leading to different instances of the problem. With reference to the separable kernel learning formulation for multi-task learning, one may interpret the output of a multi-task learning approach as finding the intermediate responses corresponding to each individual task via $\mathbf{KA}$ (similar to the OC-KSR approach) and then mixing them via a structure encoding mechanism to produce the final responses. From this standpoint, the final responses may be considered as the output of a composition function $\mathbf{f}(.) = \mathbf{g}(\mathbf{h}(.))$, where $\mathbf{h}(.)$ produces intermediate responses, while $\mathbf{g}(.)$ performs a composition on the intermediate responses to form the final outputs. In this regard, the relations in Eqs. 8 and 9 correspond to a non-linear mapping function $\mathbf{h}(.)$ expressed in terms of a non-linear kernel function $\kappa(.,.)$ and $\mathbf{A}$, while the linear function $\mathbf{g}(.)$ is defined as a linear mixing function, characterised via $\mathbf{B}$. The majority of the existing work on the multi-target regression problem is focused on the case where $\mathbf{g}(.)$ is a linear function.

In this work, we study the problem of jointly learning multiple one-class classification problems by modelling individual task-predictors as the components of a vector-valued function. In doing so, the utility of the composition function is in learning structures among multiple OCC problems. For this purpose, two cases are considered: 1-when the function $\mathbf{g}(.)$ is a linear function, we refer to the structure among multiple problems as a linear structure; and 2-when $\mathbf{g}(.)$ is a non-linear function, the structure shall be referred to as a non-linear structure. Note that for both alternative scenarios above, $\mathbf{h}(.)$ is assumed to be a non-linear function, defined in a Hilbert space. For a general Representer Theorem regarding composition functions in the Hilbert space see [34].

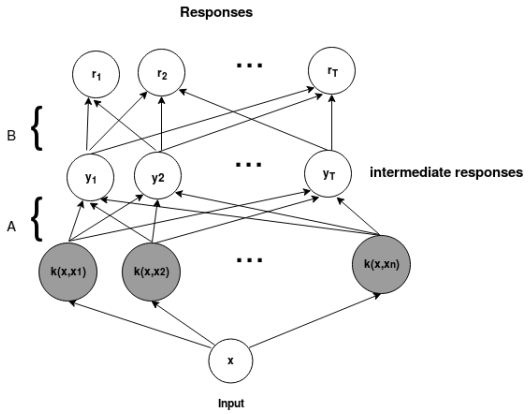


Fig. 1. Linear multi-task structure learning in the proposed multi-task one-class framework.

## 4 Multi-Task One-Class Kernel Null-Space

In this section, first, the proposed multi-task one-class learning method for linear structure learning is introduced. The discussion is then followed by introducing a non-linear multi-task OCC learning approach based on Tikhonov regularisation which is then modified to learn sparse non-linear multi-task structures.

### 4.1 Linear Structure Learning

Consider a linear form of the proposed multi-task one-class learning method (i.e. when $\mathbf{g}(.)$ is a linear function) defined in Eq. 9. Once the intermediate responses corresponding to different tasks are determined, they are mixed via an output matrix to produce the final responses. The key to the deployment of the cost function in Eq. 9, in the context of one-class classification based on the OC-KSR approach, is that the responses $\mathbf{R}$ should be real numbers. No other restrictions are imposed. This in turn offers the flexibility required to convert the problem into one of OCC. With reference to the OC-KSR formulation, in order for $Q_L$ to characterise a kernel null-space one-class classifier, the only requirement is to select suitable targets for the responses $\mathbf{R}$. In order to be consistent with the OC-KSR setting, a suitable choice for $\mathbf{R}$ is the one which forces all normal observations to be mapped onto a single point distinct from the projection of any possible anomalous sample. Accordingly, we set the expected responses for normal observations to 1, and any anomalies are mapped onto the origin. Choosing $\mathbf{R}$ as such would then lead to a zero within-class scatter, while providing a positive between-class scatter, i.e. a null projection function.

The learning machine induced by Eq. 9 admits a multi-layer structure where the second layer parameter $\mathbf{B}$ encodes a linear structure among multiple tasks, whereas the first layer coefficients $\mathbf{A}$ represent a collection of task-specific parameters, Fig. 1. The goal is then to concurrently learn the coefficient matrix $\mathbf{A}$ and the structure encoding matrix $\mathbf{B}$, subject to suitable regularisations on $\mathbf{A}$ and $\mathbf{B}$. While there exists different methods which differ from one another in terms of the regularisations imposed on the solution, recently, an effective approach has been advocated in [22], which controls the rank and shrinkage of $\mathbf{B}$, while penalising the norm of $\mathbf{A}$ in the Hilbert space. The advocated cost function in [22] is defined as

$$\begin{aligned} Q_L = &\|\mathbf{KAB} - \mathbf{R}\|_2^2 + \gamma_{L1}\text{trace}(\mathbf{A}^\top\mathbf{KA}) \\ &+\gamma_{L2}\text{trace}(\mathbf{B}^\top\mathbf{B}) + \gamma_{L3}\text{trace}(\sqrt{\mathbf{B}^\top\mathbf{B}}) \end{aligned} \tag{10}$$

For the optimisation of the objective function, a block coordinate descent method is suggested in [22], alternating between optimisation w.r.t. the parameters of the first layer and those of the second layer.

#### 4.1.1 Sub-problem w.r.t. $\mathbf{A}$

The first block of variables for the minimisation of the objective function $Q_L$ is that of $\mathbf{A}$. In order to optimise $Q_L$ with respect to $\mathbf{A}$, we set the partial derivatives to zero:

$$\frac{\partial Q_L}{\partial \mathbf{A}} = 2\mathbf{K}(\mathbf{KAB} - \mathbf{R})\mathbf{B} + 2\gamma_{L1}\mathbf{KA} = 0 \tag{11}$$

A sufficient condition for the above equality to hold is

$$\mathbf{KABB} - \mathbf{RB} + \gamma_{L1}\mathbf{A} = 0 \tag{12}$$

The linear matrix equation above is known as the discrete-time Sylvester equation, commonly arising in control theory [35]. The solution to $\mathbf{A}$ is given as

$$vec(\mathbf{A}) = (\gamma_{L1}\mathbf{I} \otimes \mathbf{K}^{-1} + \mathbf{BB} \otimes \mathbf{I})^{-1} vec(\mathbf{K}^{-1}\mathbf{RB}) \tag{13}$$

where $\otimes$ stands for the Kronecker product and $vec(.)$ denotes a concatenation of the columns of a matrix onto a vector. For large-scale problems, the solution above may be inefficient. In these cases, by utilising the structure of the problem, more efficient techniques have been developed [1].

#### 4.1.2 Sub-problem w.r.t. $\mathbf{B}$

For the minimisation of the error function $Q_L$ with respect to $\mathbf{B}$, a gradient descent approach may be applied [22]:

$$\mathbf{B} = \mathbf{B} - \eta_B \partial Q_L / \partial \mathbf{B} \tag{14}$$

where $\eta_B$ is the step size parameter and $\partial Q_L/\partial \mathbf{B}$ is derived as

$$\begin{aligned} \frac{\partial Q_L}{\partial \mathbf{B}} = &\frac{-2}{n}(\mathbf{KAB} - \mathbf{R})(\mathbf{AK})^\top \\ &+\beta \mathbf{U}\mathbf{\Sigma}^{-1}|\mathbf{\Sigma}|\mathbf{V}^\top + 2\gamma\mathbf{B} \end{aligned} \tag{15}$$

where $\mathbf{B} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ is an eigen-decomposition of the structure matrix and $|\mathbf{\Sigma}|$ is the matrix of element-wise absolute values of $\mathbf{\Sigma}$. For a detailed derivation of $\partial Q_L/\partial \mathbf{B}$ see [22].

The optimisation of the objective function $Q_L$ with respect to the unknown parameters $\mathbf{B}$ and $\mathbf{A}$ is realised via an alternating direction minimisation approach, summarised in Algorithm 1, where during the initialisation step, all tasks are deemed to be independent. That is, the structure matrix $\mathbf{B}$ is initialised to the identity matrix. Among others, one desirable property of the formulation in Eq. 10 over some other alternatives lies in the convexity of the objective function, which facilitates reaching the global optimum.

A number of observations regarding the proposed one-class multi-task linear structure learning approach are in order. First, it should be noted that the structure in Fig. 1

1. www.slicot.org

**Algorithm 1** Linear multi-task OC-KSR

1: $\mathbf{B} = \mathbf{I}_T$
2: **repeat**
3: $\quad vec(\mathbf{A}) = (\gamma_{L1}\mathbf{I} \otimes \mathbf{K}^{-1} + \mathbf{B}\mathbf{B} \otimes \mathbf{I})^{-1} vec(\mathbf{K}^{-1}\mathbf{R}\mathbf{B})$
4: $\quad \mathbf{B} = \mathbf{B} - \eta_B \frac{\partial Q_L}{\partial \mathbf{B}}$
5: **until** $|Q_L^{t+1} - Q_L^t| < \epsilon$

depicts the *learning* stage of the proposed one-class model. In the operational (test) phase, however, the parameter sets $\mathbf{A}$ and $\mathbf{B}$ may be combined to produce a model with a single layer of discriminants in the Hilbert space as $\mathbf{C} = \mathbf{A}\mathbf{B}$. Second, as noted earlier, the structure considered in Fig. 1 is not new and has been previously explored in the context of multi-target regression. The novelty, here, lies in enabling the kernel null-space one-class classification approach to benefit from the same learning structure, thanks to a kernel regression-based formulation of the OC-KSR approach.

### 4.2 Non-linear Structure Learning

In the proposed non-linear structure learning scheme and in contrast to the linear set-up, the relations between multiple tasks is modelled through a non-linear (kernel) function. The structure of the proposed learning machine in this case is depicted in Fig. 2. In this setting, once the intermediate responses corresponding to different problems ($y_t$'s, for $t = 1, \ldots, T$) for a given input $\mathbf{x}$ are produced, they collectively serve as a single input (i.e. $\mathbf{y}$) to the second layer. In the second layer, $\mathbf{y}$ is non-linearly mapped into a new space, induced by a kernel function (RBF kernel) and ultimately mixed via the coefficients $\mathbf{B}$ to produce the final responses corresponding to different tasks. The training/test data for the second layer thus consists of $T-$dimensional intermediate responses.

In the proposed non-linear structure learning method, the unknown matrices $\mathbf{A}$ and $\mathbf{B}$ are found by optimising an objective function $Q_N$ defined as a regularised kernel regression based on a kernel matrix $\mathbf{J}$, which captures the similarities between outputs of different tasks. The superiority of the non-linear model, as compared to the conventional linear structure of Fig. 1 (as will be demonstrated in the experimental evaluation section), may be justified from the perspective that the structure in Fig. 1 acts as a linear regression over the intermediate responses while that of Fig. 2 corresponds to a non-linear (kernel) regression. Different regularisations in the proposed non-linear setting of Fig. 2, namely Tikhonov and sparsity are examined in this work.

#### 4.2.1 Tikhonov Regularisation

A Tikhonov regularisation in the non-linear multi-task formulation favours models that provide predictions, that are as smooth functions of the intermediate responses as possible, by penalising parameters of larger magnitude and thereby producing a more parsimonious solution. Following a Tikhonov regularisation, the objective function for the model in Fig. 2 is defined as

$$Q_N = \|\mathbf{J}\mathbf{B} - \mathbf{R}\|_2^2 + \gamma_{N1}\text{trace}(\mathbf{A}^\top\mathbf{K}\mathbf{A}) + \gamma_{N2}\text{trace}(\mathbf{B}^\top\mathbf{J}\mathbf{B}) \quad (16)$$

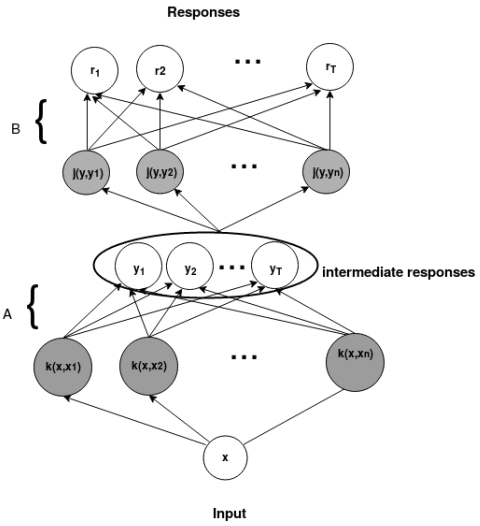


Fig. 2. Non-linear multi-task structure learning in the proposed multi-task one-class approach.

where $\mathbf{K}$ and $\mathbf{J}$ denote the kernel matrices associated with the first (the one closer to the input) and the second layer, respectively. The optimisation of the objective function associated with the non-linear model is realised via a block coordinate descent scheme, alternating between optimisation w.r.t. the parameters of the first layer and those of the second layer.

*Sub-problem w.r.t.* $\mathbf{A}$*:* The first direction of minimisation for the objective function $Q_N$ is that of $\mathbf{A}$. The partial derivatives of the term trace$(\mathbf{A}^\top\mathbf{K}\mathbf{A})$ w.r.t. $\mathbf{A}$ are readily obtained as

$$\frac{\partial \text{trace}(\mathbf{A}^\top\mathbf{K}\mathbf{A})}{\partial \mathbf{A}} = 2\mathbf{K}\mathbf{A} \quad (17)$$

Denoting the remaining terms of $Q_N$ as $Q_{N1} = \|\mathbf{J}\mathbf{B} - \mathbf{R}\|_2^2 + \gamma_{N2}\text{trace}(\mathbf{B}^\top\mathbf{J}\mathbf{B})$, we shall proceed with computing its partial derivative w.r.t. $\mathbf{A}$. The parameters involved in $Q_{N1}$ are independent of $\mathbf{A}$, except for the kernel matrix $\mathbf{J}$ (recall that the kernel matrix $\mathbf{J}$ models the similarities between $T$-dimensional intermediate responses $\mathbf{y}$'s). The dependency of the kernel matrix $\mathbf{J}$ on $\mathbf{A}$ is due to its dependency on the intermediate responses $\mathbf{Y}$, which are a function of $\mathbf{A}$, as $\mathbf{Y} = \mathbf{K}\mathbf{A}$. In order to compute the partial derivatives of $Q_{N1}$ w.r.t. $\mathbf{A}$, first, the following matrices are defined:

$$\begin{aligned} \mathbf{F} &= \mathbf{Y}\mathbf{Y}^\top \\ \mathbf{E} &= (\mathbf{I} \circ \mathbf{F})\mathbf{1} + \mathbf{1}^\top(\mathbf{I} \circ \mathbf{F})^\top - 2\mathbf{F} \end{aligned} \quad (18)$$

where $\circ$ stands for the Hadamard (component-wise) product and $\mathbf{1}$ denotes a matrix of ones. The kernel matrix $\mathbf{J}$ associated with the second layer may then be expressed as

$$\mathbf{J} = \exp[-\theta\mathbf{E}] \quad (19)$$

where the scalar parameter $\theta$ controls the RBF kernel width associated with the second layer. The partial derivatives of $Q_{N1}$ with respect to the kernel matrix $\mathbf{J}$ are

$$\frac{\partial Q_{N1}}{\partial \mathbf{J}} = 2(\mathbf{J}\mathbf{B} - \mathbf{R})\mathbf{B}^\top + \gamma_{N2}\mathbf{B}\mathbf{B}^\top \quad (20)$$

The partial derivatives $\partial Q_{N1}/\partial \mathbf{E}$, $\partial Q_{N1}/\partial \mathbf{F}$, $\partial Q_{N1}/\partial \mathbf{Y}$ are derived as [36]

$$\begin{aligned}\frac{\partial Q_{N1}}{\partial \mathbf{E}} &= (-\theta \mathbf{J}) \circ \frac{\partial Q_{N1}}{\partial \mathbf{J}} \\ \frac{\partial Q_{N1}}{\partial \mathbf{F}} &= \mathbf{I}_n \circ \Big( (\frac{\partial Q_{N1}}{\partial \mathbf{E}} + \frac{\partial Q_{N1}}{\partial \mathbf{E}}^\top) \mathbf{1}^\top \Big) - 2\frac{\partial Q_{N1}}{\partial \mathbf{E}} \\ \frac{\partial Q_{N1}}{\partial \mathbf{Y}} &= (\frac{\partial Q_{N1}}{\partial \mathbf{F}} + \frac{\partial Q_{N1}}{\partial \mathbf{F}}^\top) \mathbf{Y} \end{aligned} \tag{21}$$

For the computation of $\partial Q_{N1}/\partial \mathbf{A}$ we note

$$\delta Q_{N1} = \text{trace}(\frac{\partial Q_{N1}}{\partial \mathbf{Y}}^\top \delta \mathbf{Y}) = \text{trace}(\frac{\partial Q_{N1}}{\partial \mathbf{A}}^\top \delta \mathbf{A}) \tag{22}$$

Since $\mathbf{Y} = \mathbf{K}\mathbf{A}$, it holds that $\delta \mathbf{Y} = \mathbf{K}\delta \mathbf{A}$. Replacing $\delta \mathbf{Y}$ by $\mathbf{K}\delta \mathbf{A}$ in Eq. 22 yields

$$\delta Q_{N1} = \text{trace}(\frac{\partial Q_{N1}}{\partial \mathbf{Y}}^\top \mathbf{K}\delta \mathbf{A}) = \text{trace}(\frac{\partial Q_{N1}}{\partial \mathbf{A}}^\top \delta \mathbf{A}) \tag{23}$$

and hence

$$\frac{\partial Q_{N1}}{\partial \mathbf{A}} = \mathbf{K}\frac{\partial Q_{N1}}{\partial \mathbf{Y}} \tag{24}$$

Summing up, in order to compute $\partial Q_{N1}/\partial \mathbf{A}$, one first computes $\partial Q_{N1}/\partial \mathbf{J}$ and then $\partial Q_{N1}/\partial \mathbf{E}$, $\partial Q_N/\partial \mathbf{F}$ and $\partial Q_{N1}/\partial \mathbf{Y}$, respectively, followed by $\partial Q_{N1}/\partial \mathbf{A}$. Finally, $\partial Q_N/\partial \mathbf{A} = \partial Q_{N1}/\partial \mathbf{A} + 2\gamma_{N1}\mathbf{K}\mathbf{A}$.

*Sub-problem w.r.t.* $\mathbf{B}$: Minimising the regularised error over multiple tasks, represented by $Q_N$ w.r.t $\mathbf{B}$, may be performed by setting the partial derivative $\partial Q_N/\partial \mathbf{B}$ to zero:

$$\frac{\partial Q_N}{\partial \mathbf{B}} = 2\mathbf{J}^\top(\mathbf{J}\mathbf{B} - \mathbf{R}) + 2\gamma_{N2}\mathbf{J}\mathbf{B} = 0$$

which yields

$$\mathbf{B} = (\mathbf{J} + \gamma_{N2}\mathbf{I}_n)^{-1}\mathbf{R} \tag{25}$$

Finally, the partial derivatives of the objective function $Q_N$ w.r.t. $\theta$ are given as

$$\frac{\partial Q_N}{\partial \theta} = \text{trace}\Big(\frac{\partial Q_N}{\partial \mathbf{J}}^\top (-\mathbf{J} \circ \mathbf{E})\Big) \tag{26}$$

*Initialisation*: The initialisation step of the proposed non-linear structure learning approach is similar in spirit to that of the linear case. During the initialisation step, each task is presumed to be independent from all the others. Consequently, the kernel matrix $\mathbf{J}$ encoding inter-task relationships takes the form of a block-diagonal matrix, the diagonal elements of which are $\mathbf{1}^{n_t \times n_t}$ sub-matrices which leads to the initialisation of $\mathbf{B}$ as $\mathbf{B} = (\mathbf{J}_{init} + \gamma_{N2}\mathbf{I}_n)^{-1}\mathbf{R}$. The parameter controlling the width of the RBF kernel in the second layer ($\theta$) is initialised to the reciprocal of the average of $\mathbf{E}$, i.e. $\theta = 1/m_{\mathbf{E}}$ where $m_{\mathbf{E}}$ denotes the mean of $\mathbf{E}$. For the initialisation of $\mathbf{A}$, the problems are solved independently with the intermediate responses set to $\mathbf{R}$. Once all the parameters are initialised, the approximate optimisation of the objective function with respect to the parameters of the first and the second layer is performed via an alternating direction minimisation scheme, where for optimisation with respect to $\mathbf{A}$ and $\theta$, a gradient descent method is applied. The algorithm for the non-linear multi-task one-class learning is summarised in Algorithm 2, where

**Algorithm 2** Non-linear multi-task OC-KSR (Tikhonov regularisation)

1: $\mathbf{A} = (\mathbf{K} + \gamma_{N1}\mathbf{I}_n)^{-1}\mathbf{R}$
2: $\mathbf{B} = (\mathbf{J}_{init} + \gamma_{N2}\mathbf{I}_n)^{-1}\mathbf{R}$
3: $\theta = 1/\text{m}_{\mathbf{E}}$
4: **repeat**
5: $\quad \mathbf{A} = \mathbf{A} - \eta_A \frac{\partial Q_N}{\partial \mathbf{A}}$
6: $\quad \theta = \theta - \eta_\theta \frac{\partial Q_N}{\partial \theta}$
7: $\quad \mathbf{J} = \mathbf{J}(\mathbf{A}, \theta)$
8: $\quad \mathbf{B} = (\mathbf{J} + \gamma_{N2}\mathbf{I}_n)^{-1}\mathbf{R}$
9: **until** $|Q_N^{t+1} - Q_N^t| < \epsilon$

$\eta_A$ and $\eta_\theta$ denote the gradient descent step sizes for $\mathbf{A}$ and $\theta$, respectively. Note that Step 7 of the algorithm corresponds to updating the kernel matrix $\mathbf{J}$ associated with the second layer based on the most recent values for $\mathbf{A}$ and $\theta$.

#### 4.2.2 Sparse Regularisation

Besides the widely used Tikhonov regularisation, other regularisation schemes, encouraging sparseness on the solution, are widely applied as a guideline for inference. The underlying motivation in this case is to provide the simplest possible explanation of an observation as a combination of as few as possible atoms from a given dictionary. Typically, a more compact model is expected to provide better generalisation performance as compared with its non-sparse counterpart, especially in the presence of corruption in data or missing relations between problems. This is in contrast to the Tikhonov regularisation, which forces all problems to be related to one another. The sparsity in the proposed non-linear structure learning approach may be imposed at two different levels. The first level of sparsity is that of the task level. That is, a task either contributes in forming the discriminant of another task (the two tasks related) or not. The second level of sparsity is that of the within-task sparsity, where the response for a particular problem is derived using a sparse set of the corresponding training data. The two objectives above may be achieved via a group-sparse lasso formulation [37], [38] by enforcing an $L_1$-norm penalty on $\mathbf{B}$ in addition to an $L_2$-norm task-wise penalty on $\mathbf{B}$. Consequently, the objective function $Q_{NS}$ for the sparse non-linear setting is defined as

$$\begin{aligned} Q_{NS} = \|\mathbf{J}\mathbf{B} - \mathbf{R}\|_2^2 + \gamma_{N1}\text{trace}(\mathbf{A}^\top \mathbf{K}\mathbf{A}) \\ +\gamma_{N2}\|\mathbf{B}\|_1^1 + \gamma_{N3}\sum_{t=1}^{T}\|\mathbf{b}_t\|_2^2 \end{aligned} \tag{27}$$

where $\gamma_{N2}$ controls the within-task sparsity while $\gamma_{N3}$ governs task-wise sparseness. Accordingly, in the proposed sparse multi-task one-class learning approach, each response $\mathbf{r}_t$ may be generated using only a few tasks from among the pool of multiple problems.

The algorithm for the sparse non-linear multi-task one-class learning approach is similar to Algorithm 2 except for two differences. First, when optimising w.r.t. $\mathbf{A}$, the partial derivatives $\partial Q_{NS}/\partial \mathbf{J}$ would be

$$\frac{\partial Q_{NS}}{\partial \mathbf{J}} = 2(\mathbf{J}\mathbf{B} - \mathbf{R})\mathbf{B}^\top \tag{28}$$

**Algorithm 3** Non-linear multi-task OC-KSR (Sparse regularisation)

1: $\mathbf{A} = (\mathbf{K} + \gamma_{N1}\mathbf{I}_n)^{-1}\mathbf{R}$
2: $\mathbf{B} = (\mathbf{J}_{init} + \gamma_{N2}\mathbf{I}_n)^{-1}\mathbf{R}$
3: $\theta = 1/\mathrm{m}_{\mathbf{E}}$
4: **repeat**
5: $\quad \mathbf{A} = \mathbf{A} - \eta_A \frac{\partial Q_{NS}}{\partial \mathbf{A}}$
6: $\quad \theta = \theta - \eta_\theta \frac{\partial Q_{NS}}{\partial \theta}$
7: $\quad \mathbf{J} = \mathbf{J}(\mathbf{A}, \theta)$
8: $\quad \mathbf{B} = \mathrm{SLEP}(Q_{NS})$
9: **until** $|Q_{NS}^{t+1} - Q_{NS}^{t}| < \epsilon$

Second, in order to optimise the group-sparse problem in Eq. 27 w.r.t. $\mathbf{B}$, the Sparse Learning with Efficient Projections (SLEP) algorithm [37] is used in this work. Using the SLEP algorithm and by varying the regularisation parameters $\gamma_{N2}$ and $\gamma_{N3}$, solutions with different possible cardinalities of $\mathbf{B}$ may be obtained. The proposed sparse non-linear structure learning algorithm is summarised in Algorithm 3.

### 4.3 Analysis of the Algorithms

A few comments regarding the dynamics of the proposed non-linear (Tikhonov/sparse) multi-task learning approaches are in order.

While in the linear structure learning method (Algorithm 1) the impact of changing one block of parameters on the other (the impact of $\mathbf{A}$ on $\mathbf{B}$ or vice versa) is explicit, in the non-linear setting (Algorithms 2 and 3), the two sets of parameters $\mathbf{A}$ and $\mathbf{B}$ interact indirectly via the kernel matrix associated with the second layer, i.e. via $\mathbf{J}$ (see Step 7 of the Algorithms 2 and 3). In this respect, once $\mathbf{A}$ is updated, the intermediate responses are produced as $\mathbf{Y} = \mathbf{K}\mathbf{A}$. The new kernel matrix associated with the second layer may then be computed using the updated $\mathbf{Y}$ and the new parameter $\theta$. $\mathbf{B}$ is then derived based on the updated kernel matrix $\mathbf{J}$. Any modification to $\mathbf{B}$ would then affect parameter set $\mathbf{A}$ (see Eqs. 20 and 28).

In the operational phase of the proposed non-linear structure learning methods, upon the arrival of a new test sample $\mathbf{x}$, the corresponding intermediate outputs ($y_t$'s for $t = 1, \ldots, T$) for different problems are produced by the first layer. Treating the intermediate responses as the components of a single vector $\mathbf{y} = [y_1, \ldots, y_T]^\top$, its similarity is measured to those of training samples associated with the second layer (i.e. $\mathbf{y}_i$'s for $i = 1, \ldots, n$) using a non-linear (RBF) kernel function and subsequently combined via the corresponding mixing matrix $\mathbf{B}$ to produce the final responses.

## 5 Experimental Evaluation

In this section, an experimental evaluation of the proposed approaches for multi-task one-class classification is carried out. The efficacy of the proposed techniques is evaluated on four different data sets.

### 5.1 Data sets

#### 5.1.1 Face

This data set is created to perform a toy experiment in face recognition. The data set contains face images of different individuals and the task is to recognise a subject among others. For each subject, a one-class classifier is built using the training data associated with the subject under consideration, while all other subjects are considered as outliers with respect to the model. The experiment is repeated in turn for each subject in the data set. The features used for face image representation are obtained via the frontal-pose PAM deep CNN model [39] applied to face bounding boxes. The images of this data set are collected from the real-access videos of the Replay-Mobile data set [40], which is accompanied with face bounding boxes. In this work, ten subjects are used to form the data set and each task is to recognise a single subject. The number of positive training samples for each subject is set to 4. The number of positive and negative test samples for each subject are 40 and 160, respectively, where the negative test observations for each subject are selected from subjects other than the subject under consideration.

#### 5.1.2 MNIST

is a collection of $28 \times 28$ pixel images of handwritten digits 0-9 [41]. In our experiments, a single digit is considered as the target class while all the others correspond to non-target observations. The experiment is repeated in turn for all the digits. Similar to the face data set, each task is to recognise one digit among others. The number of positive training samples for each digit is set to 15. The number of positive and negative test samples for each class is set to 150 and 1350, respectively.

#### 5.1.3 Coil-100

The Coil-100 data set [42] contains 7,200 images of 100 different objects. Each object has 72 images taken at pose intervals of 5 degrees, with the images being of size $32 \times 32$ pixels. In the experiments conducted on this data set, 50 classes are selected randomly. A one-class classifier is then trained to recognise an object of interest among others and considered as a single task. The experiment is then repeated in turn for each of the 50 categories. Raw pixel intensities are used as feature representations in this data set. The number of positive train instances for each target class is 7. 65 positive and 585 negative test observations for each class are included in the experiments on this data set.

#### 5.1.4 Caltech256

is a challenging set of 256 object categories containing 30607 images in total [43]. Each class of images has a minimum of 80 images representing a diverse set of backgrounds, poses, lighting conditions and image sizes. In this experiment, 10 random classes are considered. The Bag-of-visual-words histograms from densely sampled SIFT features are used to represent images [2]. The setting for training and test is similar to the previous data sets. The number of positive train samples corresponding to each class is set to 12 (on average) while the number of positive test and negative test samples is 110 and 990, respectively.

2. http://homes.esat.kuleuven.be/ t̃uytelaa/unsup _ features.html

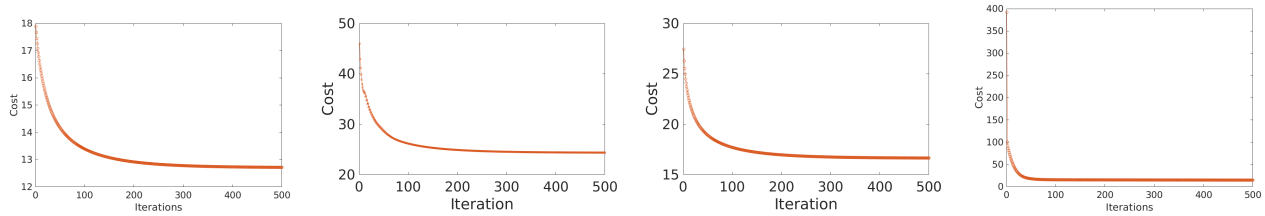


Fig. 3. Sample optimisation curves for the non-linear structure learning approach. From left to right: the face, the MNIST, the Coil-100 and Caltech256 data sets.

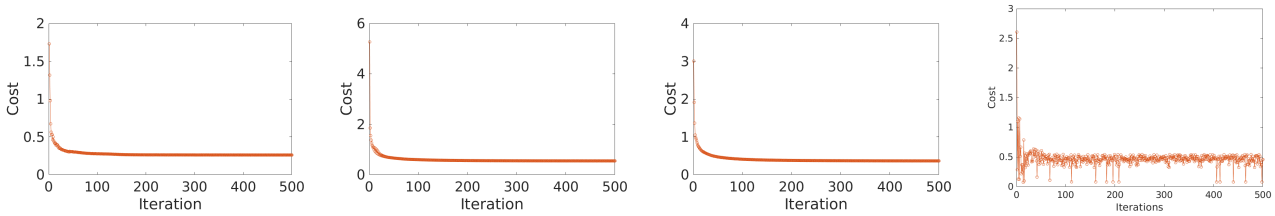


Fig. 4. Sample optimisation curves for the non-linear sparse structure learning approach. From left to right: the face, the MNIST, the Coil-100 and Caltech256 data sets.

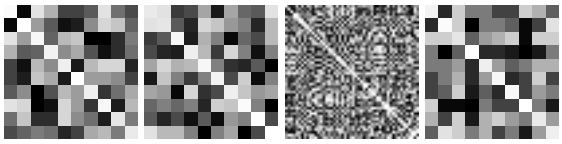

Fig. 5. Sample structural matrices for the linear structure learning approach. From left to right: the face, the MNIST, Coil-100 and Caltech256 data sets.

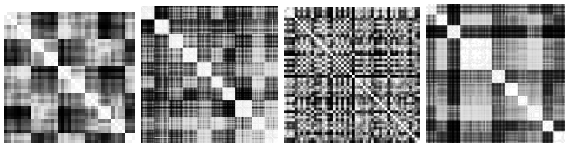

Fig. 6. Sample kernels for the non-linear structure learning approach. From left to right: the face, the MNIST, the Coil-100 and Caltech256 data sets.

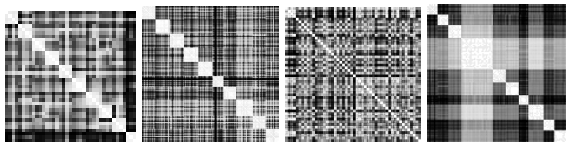

Fig. 7. Sample kernels for the non-linear sparse structure learning approach. From left to right: the face, the MNIST, the Coil-100 and Caltech256 data sets.

### 5.2 Methods

For the conversion of the OC-KSR method from a single-task to a multi-task setting, first, all training observations are combined to form a joint kernel matrix ($\mathbf{K}$) associated with the first layer. In this case, the positive training instances of one problem (target class) would serve as negative training observations for all the remaining tasks. As such, the number of training samples for all tasks would be similar. However, the number positive/negative training observations for each problem may be different. The optimal response would then be an $n \times T$ matrix where each row of the matrix is a vector of zeros except for a single element of one denoting the true class of an observation.

As previously demonstrated in [28], utilisation of negative training samples may boost the performance of the OC-KSR approach. In order to make a distinction between different variants of the OCKSR methodology, in this section, OCKSR would correspond to the algorithm when negative instances are not used for training while C-OCKSR shall be used to refer to the case when both positive and negative samples are used for training. This distinction is necessary to accurately gauge the benefits offered by a multi-task learning scheme independent of the effects of using non-target samples for training. Once a joint kernel matrix is built and the optimal responses are set, the rest of the procedure is performed according to either one of the structure learning mechanisms discussed earlier. Note that throughout the paper we have made the assumption that multiple problems use a shared representation, which is a commonly applied assumption. Nevertheless, if different representations (or modalities) are to be employed, the first layer of the proposed structure learning mechanisms needs to be modified to reflect the usage of multiple representations. More specifically, in this case, multiple kernels, each associated with a single task may be considered in the first layer.

A thorough evaluation and comparison between different one-class classification algorithms has been conducted in [28] and [29], with the conclusion of the OCKSR approach performing the best among other competitors. As such, the different methods included in these experiments are:

- OCKSR is the original single-task OC-KSR method presented in [28]. This method is used to learn an OCC classifier independently for each task and will serve as a baseline.
- C-OCKSR corresponds to the single-task OCKSR approach where negative observations are utilised for

training.

- OCKSR-L is the proposed multi-task OCKSR approach where a linear structure between different tasks is learned.
- OCKSR-N is the proposed multi-task OCKSR approach where a non-linear structure between different tasks subject to Tikhonov regularisation is learned.
- OCKR-NS is the proposed multi-task OCKSR approach where a non-linear structure between different tasks, subject to sparse group regularisation, is learned.
- SVDD is the Support Vector Data Description approach to solve the one class classification problem [44]. As a widely used method, it is used to learn an OCC classifier independently for each task to serve as a second baseline for comparison.
- MORVR is the multi-output relevance vector regression [45], which uses the Bayes theorem and the kernel trick to perform regression. The algorithm uses the matrix normal distribution to model correlated outputs.

### 5.3 Behaviour of the Optimisation Algorithms

In this section, the effectiveness of the proposed alternating direction minimisation scheme for the optimisation of the non-linear objective function (for both Tikhonov and sparse regularisation) is analysed. For an analysis of the convergence behaviour of the linear structure learning method one may consult [22]. The optimisation curves depicting the cost function vs. iterations for the Tikhonov and sparse regularisation are preented in Fig. 3 and 4, respectively. From Figs. 3 and 4, one may observe that the proposed alternating direction approach converges within a few hundred iterations, irrespective of the nature of the observations. Interestingly, the convergence of the sparse approach seems to be relatively faster than its non-sparse counterpart.

### 5.4 Visualisation of the Structure Matrices

The structures learned using different linear and non-linear multi-task approaches are illustrated in Figs. 5, 6 and 7, for the linear, non-linear and sparse non-linear setting, respectively. For the linear learning scheme, matrix $\mathbf{B}$ is depicted while for the non-linear setting, the kernel matrix associated with the second layer ($\mathbf{J}$) is visualised. Note that for the Coil-100 data set, as a relatively larger number of training sample is used, the kernel matrix is bigger in dimension compared to the other data sets. In the figure, the kernel matrix for this data set is rescaled to a similar size as those of others for the visualisation purposes. As noted earlier, at initialisation, the structural matrices are set to (block)-diagonal matrices. As may be observed from the figures, for all the data sets, the linear and non-linear multi-task learning approaches have discovered inter-task relations. This is manifest in all structural matrices exhibiting non-zero off (block)-diagonal elements.

### 5.5 Performance Comparison

In order to gauge the efficacy of the proposed multi-task OCC learning methods, multiple experiments are conducted on the face, MNIST, Coil-100 and Caltech256 data sets. In order to minimise the effect of any bias associated with the partitioning of the data, each data set is partitioned randomly into training and test sets, and each experiment is repeated 10 times and the average AUC measures results reported in Table 1. A number of observations from Table 1 are in order. First, the proposed multi-task Fisher null-space approaches are effective in improving the performance compared to other alternatives. Second, from among the proposed multi-task learning schemes, the non-linear learning methods perform better than their linear counterpart, which demonstrates the effectiveness of the proposed non-linear multi-task structure learning mechanism. Third, the proposed OCKSR-N method performs slightly better than its sparse counterpart. Nevertheless, the sparse method may provide an edge over the non-sparse variant when the data is corrupted or when some tasks are not related, as will be demonstrated in the subsequent sections.

### 5.6 The Effect of Initialisation

During the initialisation of the proposed multi-task learning methods, the structure matrices were initialised to (block)diagonal matrices. In this experiment, the effect of initialisation is analysed on the behaviour of the algorithms. For this purpose, the structural matrices were initialised randomly and the same set of experiments were conducted. Experimentally, it was observed that the initialisation had negligible effect on the performances of the proposed methods. As the performance was very similar to that of the previous experiment, the results are omitted. Nevertheless, the random initialisation of the structure matrices slighly improved the convergence speed of the algorithms.

### 5.7 The Effect of Regularisation

In the experiments conducted thus far, the regularisation parameter corresponding to the first layer is set to 1 for all the OCKSR-based methods, while the other parameters were optimised on the training set via cross validation. Typically a stronger regularisation reduces the flexibility of the model, but may provide relatively more robustness against data corruption. In a final set of experiments, the effect of changing the regularisation parameter of the first layer is analysed. For this purpose, the first layer regularisation parameter is chosen from $\{1, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}$ and the same set of experiments is repeated. The performances of different methods in terms of AUC for this experiment are reported in Fig. 8. Plots corresponding to the sum of squared errors for different variants of the OCKSR method are provided in Fig. 9. Note that since the sum of squared error measures corresponding to other methods were typically higher than those of the OCKSR variants, they are excluded from the figure in order to better visualise the effects of the proposed multi-target learning schemes.

A number of observations from the figures are in order. First, the proposed non-linear structure learning methods typically perform better than other alternatives irrespective of the degree of regularisation, confirming the efficacy of a non-linear structure learning mechanism. Second, while the linear structure learning approach OCKSR-L provides an edge over the single-task C-OCKSR method for stronger

TABLE 1
Average performance (in terms of AUC (%)) of different methods in a one-class classification scenario on different data sets

| Method | OCKSR | C-OCKSR | OCKSR-L | OCKSR-N | OCKSR-NS | SVDD | MORVR |
|---|---|---|---|---|---|---|---|
| Face | 97.70 | 99.55 | 99.71 | 99.78 | 99.76 | 97.69 | 97.63 |
| MNIST | 89.55 | 96.91 | 97.23 | 97.74 | 97.39 | 89.51 | 95.43 |
| Coil-100 | 92.08 | 97.32 | 97.40 | 98.87 | 97.95 | 93.18 | 78.27 |
| Caltech256 | 97.19 | 99.81 | 99.87 | 99.92 | 99.89 | 93.14 | 99.21 |

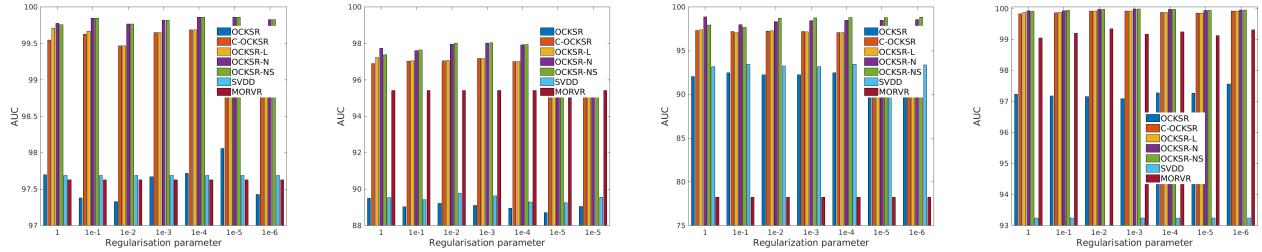


Fig. 8. The effect of regularisation on performance. From left to right: the face, MNIST, Coil-100 and Caltech256 data sets.

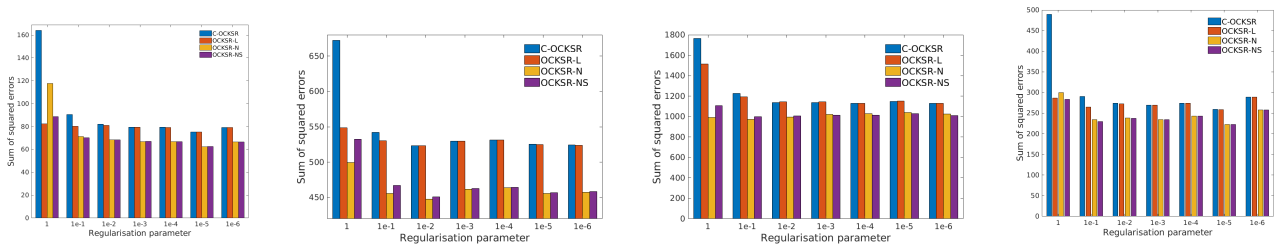


Fig. 9. The effect of regularisation on the sum of squared errors in predictions. From left to right: the face, MNIST, Coil-100 and Caltech256 data sets.

Fig. 10. Three different modalities used for recognition on the AR face data set.

regularisations (near 1), the advantage of learning a linear structure among multiple one-class problems vanishes towards lower regularisations levels, where the performances of the two methods nearly match. This may be observed from both the AUC as well as the sum of squared error plots. Third, although the proposed sparse non-linear structure learning approach performs slightly worse, compared to the non-sparse alternative for stronger levels of regularisation. Nevertheless, towards lower regularisation levels, it performs better than the non-sparse counterpart. A similar behaviour is observed both in terms of the AUCs as well as the sum of squared error measures.

### 5.8 Multimodal Learning

In the final set of experiments, we illustrate the utility of the proposed methodology for multimodal fusion for object classification in a toy experiment on face recognition. 100 subjects from the AR face data set [46], [47] are considered in this experiments for each of which 7 training samples from the first session of the data set are utilised. Similarly, 7 test images per subject, from the second session are included in this experiment. Three different modalities are considered: the whole face, a rectangular area around the left eye and another rectangular area around the right eye, Fig. 10. Raw zero-mean pixel intensities are used as image representations after normalising them to have unit $L_2$-norms. In order to perform a multimodal fusion, each task is defined as a recognition task, using a one-class classifier on one of the aforementioned modalities. Note that the multi-task structures considered previously assume a shared representation. In order to apply the proposed MTL approaches to a multimodal setting, the first layer of the previous structures need to be modified to reflect the existence of multiple representations. This may be realised by constructing multiple kernel matrices, each associated with a specific modality, in the first layer of the MTL structures. The rest of the learning mechanism remains similar to that previously discussed.

Similarly, in the operational phase, an object is represented through multiple modalities, the intermediate response associated with each of these corresponds to a single element of the intermediate response vector $\mathbf{y}$. Once the

TABLE 2
Average performance (in terms of AUC (%)) of different methods in a one-class classification scenario on the AR data set.

| Method | Average AUC (%) over 100 subjects |
|---|---|
| **C-OCKSR**+whole face | 93.84 |
| **C-OCKSR**+left eye | 92.52 |
| **C-OCKSR**+right eye | 92.86 |
| **OCKSR-N**-fusion | 97.02 |
| **OCKSR-NS**-fusion | 97.01 |

intermediate response vector (a vector of three elements in the current experiment) is constructed, the rest of the procedure is similar to that of the earlier experiments. For the purpose of this experiment, an OCC classifier for each of the three modalities is built separately for each subject using 7 positive samples of the corresponding subject and seven randomly chosen samples of subjects other than the subject under consideration. The number of positive test samples is 7, while there are 93 negative test samples. Driven by the earlier experiments, the two non-linear structure learning methods of OCKSR-N and OCKSR-NS, which perform better as compared to other approaches, are considered in this experiment. As a baseline for comparison, the single-task C-OCKSR approach is chosen. The performances in terms of the average AUC measures over all subjects are reported in Table 2. From the table the following observations may be made. Although the whole face seems to be more discriminative as compared with the eyes, nevertheless, the performance based on each eye does not fall far behind. Consequently, a fusion of the three different modalities is expected to improve the performance. This can be observed in both non-linear MT learning schemes, where an improvement over $3\%$ compared to the single-task method of C-OCKSR operating on the whole face is observed. A further point noting is the similarity of the performance of the two alternative non-linear MT learning methods.

## 6 Conclusion

We studied the one-class classification problem based on the kernel regression Fisher null-space technique (OCKSR) in a multi-task learning framework. For this purpose, first, it was shown that the OCKSR approach may be readily cast within a multi-target learning approach, where the dependencies among multiple problems are modelled linearly. Next, a non-linear structure learning mechanism was proposed, where the correlations among different problems were encoded more effectively via a non-linear kernel function. The non-linear multi-task learning approach was then extended to a sparse setting to account for any missing relationships among different problems. Different experiments conducted on multiple data sets verified the merits of multi-task learning for the OCC problem based on the OCKSR method. Moreover, it was observed that in certain cases, when the linear structure learning approach failed to provide an advantage over the single-task variant, the proposed non-linear multi-task learning methods maintained an edge over other alternatives.

## References


[1] Y. Zhang and Q. Yang, "A survey on multi-task learning," *CoRR*, vol. abs/1707.08114, 2017. [Online]. Available: http://arxiv.org/abs/1707.08114

[2] R. Caruana, "Multitask learning," *Machine Learning*, vol. 28, no. 1, pp. 41–75, Jul 1997. [Online]. Available: https://doi.org/10.1023/A:1007379606734

[3] C. A. Micchelli and M. Pontil, "Kernels for multi-task learning," in *Adv. Neur. Inf. Proc. Sys. (NIPS)*, 2004.

[4] T. Evgeniou, C. A. Micchelli, and M. Pontil, "Learning multiple tasks with kernel methods," *Journal of Machine Learning Research*, vol. 6, pp. 615–637, 2005.

[5] A. Caponnetto, C. A. Micchelli, M. Pontil, and Y. Ying, "Universal multi-task kernels," *J. Mach. Learn. Res.*, vol. 9, pp. 1615–1646, Jun. 2008. [Online]. Available: http://dl.acm.org/citation.cfm?id=1390681.1442785

[6] L. Baldassarre, L. Rosasco, A. Barla, and A. Verri, "Vector field learning via spectral filtering," in *Machine Learning and Knowledge Discovery in Databases*, J. L. Balcázar, F. Bonchi, A. Gionis, and M. Sebag, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2010, pp. 56–71.

[7] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, pp. 15:1–15:58, Jul. 2009.

[8] P. Nader, P. Honeine, and P. Beauseroy, "$l_p$-norms in one-class classification for intrusion detection in scada systems," *IEEE Transactions on Industrial Informatics*, vol. 10, no. 4, pp. 2308–2317, Nov 2014.

[9] A. Beghi, L. Cecchinato, C. Corazzol, M. Rampazzo, F. Simmini, and G. Susto, "A one-class svm based tool for machine learning novelty detection in hvac chiller systems," *IFAC Proceedings Volumes*, vol. 47, no. 3, pp. 1953 – 1958, 2014, 19th IFAC World Congress. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1474667016418999

[10] S. Budalakoti, A. N. Srivastava, and M. E. Otey, "Anomaly detection and diagnosis algorithms for discrete symbol sequences with applications to airline safety," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 39, no. 1, pp. 101–113, Jan 2009.

[11] S. Kamaruddin and V. Ravi, "Credit card fraud detection using big data analytics: Use of psoaann based one-class classification," in *Proceedings of the International Conference on Informatics and Analytics*, ser. ICIA-16. New York, NY, USA: ACM, 2016, pp. 33:1–33:8. [Online]. Available: http://doi.acm.org/10.1145/2980258.2980319

[12] G. G. Sundarkumar, V. Ravi, and V. Siddeshwar, "One-class support vector machine based undersampling: Application to churn prediction and insurance fraud detection," in *2015 IEEE International Conference on Computational Intelligence and Computing Research (ICCIC)*, Dec 2015, pp. 1–7.

[13] M. Yu, Y. Yu, A. Rhuma, S. M. R. Naqvi, L. Wang, and J. A. Chambers, "An online one class support vector machine-based person-specific fall detection system for monitoring an elderly individual in a room environment," *IEEE Journal of Biomedical and Health Informatics*, vol. 17, no. 6, pp. 1002–1014, Nov 2013.

[14] A. Rabaoui, M. Davy, S. Rossignol, and N. Ellouze, "Using one-class svms and wavelets for audio surveillance," *IEEE Transactions on Information Forensics and Security*, vol. 3, no. 4, pp. 763–775, Dec 2008.

[15] X. He, G. Mourot, D. Maquin, J. Ragot, P. Beauseroy, A. Smolarz, and E. Grall-Mas, "Multi-task learning with one-class svm," *Neurocomputing*, vol. 133, pp. 416 – 426, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925231214000356

[16] Yongjian Xue and P. Beauseroy, "Multi-task learning for one-class svm with additional new features," in *2016 23rd International Conference on Pattern Recognition (ICPR)*, Dec 2016, pp. 1571–1576.

[17] H. Yang, I. King, and M. R. Lyu, "Multi-task learning for one-class classification," in *The 2010 International Joint Conference on Neural Networks (IJCNN)*, July 2010, pp. 1–8.

[18] T. Id, D. T. Phan, and J. Kalagnanam, "Multi-task multi-modal models for collective anomaly detection," in *2017 IEEE International Conference on Data Mining (ICDM)*, Nov 2017, pp. 177–186.

[19] S. Dang, X. Cai, Y. Wang, J. Zhang, and F. Chen, "Unsupervised matrix-valued kernel learning for one class classification," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017, Singapore, November*

*06 - 10, 2017*, 2017, pp. 2031–2034. [Online]. Available: https://doi.org/10.1145/3132847.3133114
[20] F. Dinuzzo, C. S. Ong, P. Gehler, and G. Pillonetto, "Learning output kernels with block coordinate descent," in *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, ser. ICML '11. New York, NY, USA: ACM, Jun. 2011, pp. 49–56.
[21] C. Ciliberto, Y. Mroueh, T. Poggio, and L. Rosasco, "Convex learning of multiple tasks and their structure," in *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ser. ICML'15. JMLR.org, 2015, pp. 1548–1557. [Online]. Available: http://dl.acm.org/citation.cfm?id=3045118.3045283
[22] X. Zhen, M. Yu, X. He, and S. Li, "Multi-target regression via robust low-rank learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 2, pp. 497–504, Feb 2018.
[23] X. Zhen, M. Yu, F. Zheng, I. B. Nachum, M. Bhaduri, D. Laidley, and S. Li, "Multitarget sparse latent regression," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 5, pp. 1575–1586, May 2018.
[24] C. Brouard, M. Szafranski, and F. d'Alché Buc, "Input output kernel regression: Supervised and semi-supervised structured output prediction with operator-valued kernels," *Journal of Machine Learning Research*, vol. 17, no. 176, pp. 1–48, 2016. [Online]. Available: http://jmlr.org/papers/v17/15-602.html
[25] P. Bodesheim, A. Freytag, E. Rodner, M. Kemmler, and J. Denzler, "Kernel null space methods for novelty detection," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, June 2013, pp. 3374–3381.
[26] J. Liu, Z. Lian, Y. Wang, and J. Xiao, "Incremental kernel null space discriminant analysis for novelty detection," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017, pp. 4123–4131.
[27] P. Bodesheim, A. Freytag, E. Rodner, and J. Denzler, "Local novelty detection in multi-class recognition problems," in *2015 IEEE Winter Conference on Applications of Computer Vision*, Jan 2015, pp. 813–820.
[28] S. R. Arashloo and J. Kittler, "One-class kernel spectral regression for outlier detection," *CoRR*, vol. abs/1807.01085, 2018. [Online]. Available: http://arxiv.org/abs/1807.01085
[29] ——, "Robust one-class kernel spectral regression," *CoRR*, vol. abs/1902.02208, 2019. [Online]. Available: http://arxiv.org/abs/1902.02208
[30] P. Li and S. Chen, "Hierarchical gaussian processes model for multi-task learning," *Pattern Recognition*, vol. 74, pp. 134 – 144, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320317303746
[31] X. Tian, Y. Li, T. Liu, X. Wang, and D. Tao, "Eigenfunction-based multitask learning in a reproducing kernel hilbert space," *IEEE Transactions on Neural Networks and Learning Systems*, pp. 1–13, 2018.
[32] H. Borchani, G. Varando, C. Bielza, and P. Larrañaga, "A survey on multi-output regression," *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.*, vol. 5, pp. 216–233, 2015.
[33] S. R. Arashloo, J. Kittler, and W. Christmas, "An anomaly detection approach to face spoofing detection: A new formulation and evaluation protocol," *IEEE Access*, vol. 5, pp. 13 868–13 882, 2017.
[34] B. Bohn, M. Griebel, and C. Rieger, "A representer theorem for deep kernel learning," *Journal of Machine Learning Research*, vol. 20, no. 64, pp. 1–32, 2019. [Online]. Available: http://jmlr.org/papers/v20/17-621.html
[35] V. Sima, *Algorithms for linear-quadratic optimization*. New York : M. Dekker, 1996.
[36] M. Engin, L. Wang, L. Zhou, and X. Liu, "Deepkspd: Learning kernel-matrix-based spd representation for fine-grained image recognition," in *Computer Vision – ECCV 2018*, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham: Springer International Publishing, 2018, pp. 629–645.
[37] J. Liu and J. Ye, "Moreau-yosida regularization for grouped tree structure learning," in *Advances in Neural Information Processing Systems 23*, J. D. Lafferty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, Eds. Curran Associates, Inc., 2010, pp. 1459–1467. [Online]. Available: http://papers.nips.cc/paper/3931-moreau-yosida-regularization-for-grouped-tree-structure-learning.pdf
[38] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *JOURNAL OF THE ROYAL STATISTICAL SOCIETY, SERIES B*, vol. 68, pp. 49–67, 2006.
[39] I. Masi, F. Chang, J. Choi, S. Harel, J. Kim, K. Kim, J. Leksut, S. Rawls, Y. Wu, T. Hassner, W. AbdAlmageed, G. Medioni, L. Morency, P. Natarajan, and R. Nevatia, "Learning pose-aware models for pose-invariant face recognition in the wild," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 2, pp. 379–393, Feb 2019.
[40] A. Costa-Pazo, S. Bhattacharjee, E. Vazquez-Fernandez, and S. Marcel, "The replay-mobile face presentation-attack database," in *Proceedings of the International Conference on Biometrics Special Interests Group (BioSIG)*, Sep. 2016.
[41] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov 1998.
[42] S. A. Nene, S. K. Nayar, and H. Murase, "Columbia object image library (coil-100)," 1996. [Online]. Available: http://www1.cs.columbia.edu/CAVE/software/softlib/coil-100.php
[43] G. Griffin, A. Holub, and P. Perona, "Caltech-256 object category dataset," California Institute of Technology, Tech. Rep. 7694, 2007. [Online]. Available: http://authors.library.caltech.edu/7694
[44] D. M. Tax and R. P. Duin, "Support vector data description," *Machine Learning*, vol. 54, no. 1, pp. 45–66, Jan 2004. [Online]. Available: https://doi.org/10.1023/B:MACH.0000008084.60811.49
[45] Y. Ha, "Fast multi-output relevance vector regression," *CoRR*, vol. abs/1704.05041, 2017. [Online]. Available: http://arxiv.org/abs/1704.05041
[46] A. Martinez and R. Benavente, "The ar face database," CVC Tech. Report 24,1998.
[47] A. M. Martinez and A. C. Kak, "Pca versus lda," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 23, no. 2, pp. 228–233, Feb 2001.

**Shervin Rahimzadeh Arashloo** received the Ph.D. degree from the centre for vision, speech and signal processing, university of Surrey, UK. He is a visiting research fellow with the centre for vision, speech and signal processing, university of Surrey, UK and also holds an assistant professor position with the department of computer engineering, Bilkent university, Ankara, Turkey. His research interests includes secured biometrics, anomaly detection and graphical models with applications to image and video analysis.

**Josef Kittler** received the B.A., Ph.D., and D.Sc. degrees from the University of Cambridge, in 1971, 1974, and 1991, respectively. He is Professor of Machine Intelligence at the Centre for Vision, Speech and Signal Processing, Department of Electronic Engineering, University of Surrey, Guildford, U.K. He conducts research in biometrics, video and image database retrieval, medical image analysis, and cognitive vision. He published the textbook Pattern Recognition: A Statistical Approach (Englewood Cliffs, NJ, USA: Prentice-Hall, 1982) and over 700 scientific papers. He serves on the Editorial Board of several scientific journals in pattern recognition and computer vision.