Exploiting Unrelated Tasks in Multi-Task Learning

Bernardino Romera-Paredes
Dept. of Computer Science
UCL Interactive Centre
Univ. College London, UK

Andreas Argyriou
Toyota Technology Institute
at Chicago, USA

Nadia Bianchi-Berthouze
UCL Interactive Centre
Division of Psychology
& Language Sciences

Massimiliano Pontil
Dept. of Computer Science
Univ. College London, UK

Abstract

We study the problem of learning a group of *principal* tasks using a group of *auxiliary* tasks, *unrelated* to the principal ones. In many applications, joint learning of unrelated tasks which use the same input data can be beneficial. The reason is that prior knowledge about which tasks are unrelated can lead to sparser and more informative representations for each task, essentially screening out idiosyncrasies of the data distribution. We propose a novel method which builds on a prior multitask methodology by favoring a shared *low dimensional representation* within each group of tasks. In addition, we impose a penalty on tasks from different groups which encourages the two representations to be *orthogonal*. We further discuss a condition which ensures convexity of the optimization problem and argue that it can be solved by alternating minimization. We present experiments on synthetic and real data, which indicate that incorporating unrelated tasks can improve significantly over standard multi-task learning methods.

1 Introduction

Multi-task learning [5, 8, 20] is a machine learning paradigm for learning a number of supervised learning tasks simultaneously, exploiting commonalities between them. It has been frequently observed in the recent literature that, when there are relations between the tasks to learn, it can be advantageous to learn all the tasks simultaneously instead of learning each task independently of the others – see, for example, [1, 2, 4, 5, 8, 9, 10, 17, 20] and references therein.

In this paper, we consider the scenario in which there are two groups of tasks which are known a priori to be *unrelated*, in the sense that the first group of tasks uses features which are not relevant for the second group of tasks and vice versa. In other words, the tasks that belong to the same group tend to share the same set of features while two tasks belonging to different groups tend not to share any features. One instance of the above scenario is the problem of identity/emotion recognition. Suppose that we have a data set of video clips of individuals expressing a set of emotions. We know from the literature that recognition of the identity of a person and recognition of the emotion expressed depend on different and uncorrelated features of the same image. Identity recognition is based on features describing rigid characteristics of the face (e.g., face width, hair color), whereas emotion recognition is based on features describing facial muscle configurations (e.g., eyes narrowed, corners of mouth raised) [7].

In this paper we propose to take advantage of the prior knowledge that these tasks are unrelated to improve the learning accuracy on one of the groups of tasks. We call this last group of tasks *principal tasks* (e.g., emotion recognition) and the other group *auxiliary tasks* (e.g., identity recognition). In the identity/emotion application described above, we are interested only in learning a good classifier for detecting emotions in images. If the training sample per task is small enough, a method which does not take into account the differentiation of groups can easily overfit, so that the facial features (idiosyncrasies) of a specific person can be mistaken as characteristics of a given emotion. To avoid this, our method exploits the identity labels of the instances in the training set, but does not use them for prediction of emotion on the test instances.

The approach we propose builds on the multi-task feature learning framework described in [2]. Specifically, we add a regularization term which penalizes the inner product between the predictor functions of any two tasks belonging to two different groups. In this way, our formulation can discriminate those features important for each group of tasks and can lead to improvements in statistical performance. We also present a simplified setting of our method which

Appearing in Proceedings of the 15^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2012, La Palma, Canary Islands. Volume XX of JMLR: W&CP XX. Copyright 2012 by the authors.

ensures that it is equivalent to a convex optimization problem.

Our methodology shares some aspects with some recent work in multi-task learning. For example, [3] and [11] extended the multi-task learning approach of [2] by assuming that there are a number of groups or clusters of tasks and that the weight vectors of the tasks belonging to the same group are similar to each other. In this case, the clusters are not known *a priori*. In addition, no constraint is imposed on tasks belonging to different clusters. The idea of exploiting unrelated groups of tasks to improve learning has been also addressed in [19, 21, 23]. These studies rely on multilinear models to describe the relations between different factors (e.g., emotion and identity). However, these studies present a number of limitations that make them not always suitable to applications in which the training sets are not equally distributed among the factors and the variability between instances belonging to the same factors is very high. Furthermore, their approach does not allow for addressing regression problems.

The paper is organized as follows. In Section 2, we review previous work on multi-task learning. In Section 3, we present our method for incorporating unrelated auxiliary tasks in a multi-task framework and an algorithm for solving the resulting optimization problem. In Section 4, we present our experiments with the proposed method. Finally, in Section 5 we discuss our findings and future questions.

2 Background on Multi-Task Learning

In this section we introduce our notation and describe a previous method for multi-task learning which forms the basis of our approach.

2.1 Notation

We are given a set of T supervised tasks. Each task $t = 1, \dots, T$ is identified by a function $f_t : \mathbb{R}^d \to \mathbb{R}$, which for simplicity we assume to be linear, that is $f_t(x) = w_t^\top x$. The vector of regression coefficients $w_t \in \mathbb{R}^d$ is unknown and we are provided with m data examples per task, $\{(x_{ti}, y_{ti}) : i = 1 \dots, m\} \subset \mathbb{R}^d \times \mathbb{R}$, such that $y_{ti} = w_t^\top x_{ti} + \eta_{ti}$, $i = 1, \dots, m$, $t = 1, \dots, T$, where η_{ti} is some zero mean i.i.d. noise process[1]. We call these the *principal tasks* and the goal is to learn them jointly under the assumption that they are related. We will focus only on multi-task learning in the following, but *transfer learning* (see e.g. [17]) – in which the goal is to learn a new task – is also straightforward within our framework.

2.2 Multi-Task Feature Learning

Our aim here is to review a learning algorithm which takes advantage of prior knowledge that the number of features used by the tasks is small. This is a well studied assumption in multi-task learning, see [2, 5, 6, 8, 17] and references therein. In the linear multi-task learning model, this assumption means that the vectors w_t lie on a *low dimensional subspace*. In other words, the matrix of tasks $W = [w_1, \dots, w_T]$ can be factorized as the product of a $d \times d$ orthogonal matrix U and a $d \times T$ coefficient matrix A, which has only *few nonzero rows*. Note that the rows of A are associated with the features while the columns with the tasks. To learn such a factorization, we define the average empirical error

$$\mathcal{E}_{\mathrm{pr}}(UA) = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{m}\sum_{i=1}^{m} L(y_{ti}, a_t^\top U^\top x_{ti}) \qquad (1)$$

and, following [2], minimize the regularized error

$$\mathcal{E}_{\mathrm{pr}}(UA) + \gamma \|A\|_{2,1}^2 \qquad (2)$$

over all matrices $A \in \mathbb{R}^{d\times T}$ and orthogonal matrices U, that is, $U^\top U = I$. The norm appearing in the regularization term in equation (2) is defined as

$$\|A\|_{2,1} \equiv \sum_{j=1}^{d}\sqrt{\sum_{t=1}^{T} a_{jt}^2}$$

namely, it is the sum of the ℓ_2 norms of the rows of matrix A. This choice is a special case of the regularization term used in the Group Lasso estimator [24] and it encourages matrices with many zero rows, under assumptions (e.g. Restricted Eigenvalue conditions) about the distribution of the data [12].

In [2] it is proved that the above problem is equivalent to the convex problem

$$\begin{aligned} &\inf\ \mathcal{E}_{\mathrm{pr}}(W) + \gamma\, \mathrm{tr}(W^\top D^{-1} W) \\ &\text{s.t. } W \in \mathbb{R}^{d\times T},\ D \succ 0,\ \mathrm{tr}\,(D) \le 1. \end{aligned} \qquad (3)$$

If $(\hat{A}, \hat{U})$ is an optimal solution of (2), then $\hat{W} = \hat{U}\hat{A}$ is an optimal solution of (3), see [2, Thm. 1]. Moreover, for a fixed W the optimal D is given by

$$D(W) = \frac{(WW^\top)^{\frac{1}{2}}}{\mathrm{tr}\,(WW^\top)^{\frac{1}{2}}}.$$

3 Exploiting Orthogonal Tasks

We now present our method, which uses an auxiliary group of tasks, assumed to be unrelated to the principal group, to improve the learning process. Here we use the term unrelated to signify that the two groups of tasks are defined by

[1] In practice, the number of examples per task may vary but we have kept it constant for simplicity of notation.

orthogonal set of features. The intuition is that, by exploiting this orthogonality – that will be formalized shortly – we will improve the estimation of the principal group of tasks (and possibly the auxiliary one as well).

We identify the auxiliary tasks by the column vectors $v_1, \dots, v_S$. We let V be the $d \times S$ matrix whose columns are given by the above vectors, in order. We also denote by $\{(x'_{si}, y'_{si}) : i = 1 \dots, m\} \subset \mathbb{R}^d \times \mathbb{R}$, $s = 1, \dots, S$ the examples for these additional tasks.

We make the following assumption about the two group of tasks:

- a *low dimensional* representation is shared by the tasks within each group, and
- the principal tasks w_t *share no features* with the auxiliary tasks v_s.

To formalize these requirements, we write $V = UB$, where B is a $d \times S$ matrix of coefficients and let $C = [A, B]$ so that $[W, V] = UC$. We require that

- the matrix C has *few nonzero rows* and
- each of these rows has nonzero values in *only one* group of columns.

A schematic example of a matrix which our method should favor is

$$C = \begin{bmatrix} a_{11} & a_{12} & a_{13} & 0 & 0 \\ a_{21} & a_{22} & a_{23} & 0 & 0 \\ 0 & 0 & 0 & b_{31} & b_{32} \\ 0 & 0 & 0 & b_{41} & b_{42} \\ 0 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

In this example, there are three principal tasks and two auxiliary tasks. Furthermore, there are two important features for each group of tasks, but these features are not shared across the groups. Finally, there is a large number of features which are not relevant to any of the tasks.

We incorporate the above constraints into our method as follows. We let

$$\mathcal{E}_{\text{aux}}(UB) = \frac{1}{S} \sum_{s=1}^{S} \frac{1}{m} \sum_{i=1}^{m} L(y'_{si}, b_s^\top U^\top x'_{si})$$

and minimize the regularized error

$$\mathcal{E}_{\text{pr}}(UA) + \mathcal{E}_{\text{aux}}(UB) + \gamma\, \Phi(A, B) + \lambda\, \Psi(A, B) \quad (4)$$

over all matrices $A \in \mathbb{R}^{d \times T}$, $B \in \mathbb{R}^{d \times S}$ and orthogonal matrices $U \in \mathbb{R}^{d \times d}$. There are two regularization parameters $\gamma, \lambda > 0$ which may be tuned by cross validation. The first parameter controls the number of features shared by the tasks – the larger γ, the smaller the number of shared features will be; the second parameter controls the degree of orthogonality between the two groups of tasks – the larger λ, the less "correlated" the tasks within the two groups will be. In particular, in the limit $\lambda \to \infty$, the two groups of tasks will be orthogonal to each other.

The regularization term in (4) consists of two parts. The term $\Phi(A, B)$ favors few nonzero rows in the matrix $[A, B]$ and the term $\Psi(A, B)$ penalizes features shared by the different groups of tasks. Regarding the first term, we may choose $\Phi(A, B) = \|[A, B]\|_{2,1}^2$ as in standard multi-task feature learning (Section 2.2). Regarding the second term, we want that $a_{jt} b_{js} = 0$, for every $t \in \{1 \dots T\}$, $s \in \{1 \dots S\}$ and $j \in \{1 \dots d\}$. A sufficient condition for this to hold is that $A^\top B = 0$, where 0 denotes the $T \times S$ matrix of zeros. At first sight this condition does not seem sufficient, since $a_t^\top b_s = 0$ imposes orthogonality only on a_t and b_s. However, since this condition holds for every choice of t and s in their range *and* the matrix U is orthogonal, it implies that the subspace spanned by the principal tasks is orthogonal to the subspace spanned by the auxiliary tasks. Consequently, it must be the case that there is an orthogonal matrix U' and matrices A', B' such that $W = U'A'$, $V = U'B'$ and $[A', B']$ has the desired structure. Thus, we can use the square of the Frobenius norm of $A^\top B$ as the second regularization term[2], that is,

$$\Psi(A, B) = \|A^\top B\|_{\text{F}}^2 . \quad (5)$$

We now make the change of variable $[W, V] = U[A, B]$ in a way similar to Section 2.2 and derive the equivalent problem

$$\begin{aligned} &\inf\ \mathcal{E}(W, V) + \mathcal{R}_0(W, V, D) \\ &\text{s.t. } W \in \mathbb{R}^{d \times T},\ V \in \mathbb{R}^{d \times S}, \\ &\quad D \succ 0,\ \text{tr}\,(D) \leq 1, \end{aligned} \quad (6)$$

where $\mathcal{E}(W, V) = \mathcal{E}_{\text{pr}}(W) + \mathcal{E}_{\text{aux}}(V)$ and

$$\mathcal{R}_0(W, V, D) = \gamma\, \text{tr}\left(D^{-1}(WW^\top + VV^\top)\right) + \lambda \|W^\top V\|_{\text{F}}^2 .$$

Note that unlike the standard multi-task optimization problem (3), problem (6) is *nonconvex* due to the $\|W^\top V\|_{\text{F}}^2$ term in the regularizer $\mathcal{R}$. To overcome this drawback, we add a strongly convex function to the regularizer. A natural choice, which we consider here, is to add a multiple of the squared Frobenius norm of the parameters. That is, we

[2] Another valid choice would be the ℓ_1-norm of the vector formed by the entries of matrix $A^\top B$, see [25]. However, the Frobenius norm, besides being differentiable and easier to deal with, seems more appropriate in our context, since it drives all the inner products towards zero, whereas the ℓ_1-norm does not prevent some of the inner products from being large.

consider the optimization problem

$$\begin{aligned}&\inf\ \mathcal{E}(W,V)+\mathcal{R}_0(W,V,D)+\rho(\|W\|_{\mathrm{F}}^2+\|V\|_{\mathrm{F}}^2)\\&\text{s.t. } W\in\mathbb{R}^{d\times T},\ V\in\mathbb{R}^{d\times S},\ D\succ 0,\ \mathrm{tr}\,(D)\leq 1\end{aligned}\tag{7}$$

where ρ is a positive parameter. The following result, whose proof can be found in the appendix, establishes a condition under which problem (7) is convex.

Theorem 3.1. *If $\rho > \sqrt{\frac{\mathcal{E}(0,0)\lambda}{2}}$ then problem* (7) *is convex.*

We solve problem (7) by alternating minimization, see Algorithm 1. For fixed W, V the optimal D is given by

$$D(W,V)=\frac{(WW^\top+VV^\top)^{\frac{1}{2}}}{\mathrm{tr}\,(WW^\top+VV^\top)^{\frac{1}{2}}}.\tag{8}$$

We note, in passing, that if we substitute the right hand side of this expression in the regularizer appearing in the objective function of problem (7), we obtain the following function of W and V

$$\gamma\|[W,V]\|_{\mathrm{tr}}^2+\rho(\|W\|_{\mathrm{F}}^2+\|V\|_{\mathrm{F}}^2)+\lambda\|W^\top V\|_{\mathrm{F}}^2$$

where $\|\cdot\|_{\mathrm{tr}}$ denotes the trace norm, that is the ℓ_1 norm of the vector of singular values. The first two terms in the right hand side of the above expression are similar to a matrix version of the elastic net regularizer [26]. For this reason, we will refer to the learning method solving problem (7) as orthogonal multi-task learning elastic-net (OrthoMTL-EN).

Returning to the Algorithm, we observe that, for fixed D, the regularizer separates across tasks. Indeed, using elementary properties of the trace of matrix products, it follows that

$$\begin{aligned}\mathcal{R}(W,V,D)&=\sum_{t=1}^{T}w_t^\top(\gamma D^{-1}+\rho I+\lambda VV^\top)w_t\\&\qquad+\mathrm{tr}((\gamma D^{-1}+\rho I)VV^\top)\\&=\sum_{s=1}^{S}v_s^\top(\gamma D^{-1}+\rho I+\lambda WW^\top)v_s\\&\qquad+\mathrm{tr}((\gamma D^{-1}+\rho I)WW^\top).\end{aligned}$$

Thus, the minimization over W (resp. V) can be carried out independently across the tasks since the regularizer decouples when D and V (resp. W) are fixed.

We remark that the alternating process decreases the objective function in problem (6) and hence is guaranteed to converge in objective value. One may modify the perturbation analysis in [2] to show that, under the hypothesis of Theorem 3.1, the iterates of the algorithm converge; a detailed discussion will be presented in a longer version. Note also that we may still apply Algorithm 1 to approximately solve Problem (7) for an arbitrary choice of the parameters γ, λ, ρ. In this case, however, the objective is not guarantee to be convex and, so, the algorithm is only guaranteed to converge to a stationary point.

Algorithm 1 Orthogonal Multi-Task Learning (OrthoMTL)

```
Input: training sets $\{(x_{ti},y_{ti})\}_{i=1}^m$, $\{(x'_{si},y'_{si})\}_{i=1}^m$, $t\in\{1,\ldots,T\}$, $s\in\{1,\ldots,S\}$.
Parameters: regularization parameters $\gamma$, $\lambda$, $\rho$, tolerance parameter $tol$
Output: regression matrices $W=[w_1,\ldots,w_T]$ and $V=[v_1,\ldots,v_S]$, $d\times d$ positive definite matrix $D$
Initialization: set $D=\frac{I}{d}$
while $\|W-W_{\mathrm{prev}}\|>tol$ or $\|V-V_{\mathrm{prev}}\|>tol$ do
  for $t=1\ldots T$ do
    compute the minimizer $w_t\in\mathbb{R}^d$ of the function $\sum_{i=1}^{m}L(y_{ti},w^\top x_{ti})+w^\top\left(\gamma D^{-1}+\rho I+\lambda VV^\top\right)w$
  end for
  for $s=1\ldots S$ do
    compute the minimizer $v_s\in\mathbb{R}^d$ of the function $\sum_{i=1}^{m}L(y'_{si},v^\top x'_{ti})+v^\top\left(\gamma D^{-1}+\rho I+\lambda WW^\top\right)v$
  end for
  set $D=\frac{(WW^\top+VV^\top)^{\frac{1}{2}}}{\mathrm{tr}(WW^\top+VV^\top)^{\frac{1}{2}}}$
end while
```

In practice our numerical experiments indicate that the algorithm converges in less than 20 iterations. Each W or V update can be executed very quickly by computing each column vector independently. For example, for the square loss this consists in solving a linear system of d equations. However if $d > m$, one may solve an equivalent dual problem, see e.g. [18]. Other loss functions, such as the hinge loss can be handled similarly. Finally, the D step requires the computation of a matrix square root, which we solve by singular value decomposition.

4 Experiments

In this section, we present numerical experiments to test our method on one synthetic and two real datasets. In all experiments we compare the following methods:

- OrthoMTL-EN: this is our method (cf. problem (7)).
- OrthoMTL-C: this is like OrthoMTL-EN but with parameter ρ set according to Theorem 3.1. This way problem (7) is guaranteed to be convex.
- OrthoMTL: this is like OrthoMTL-EN but with parameter $\rho = 0$.

- Ridge Regression: this standard method corresponds to the choice $\lambda = \gamma = 0$ and can be interpreted as learning the tasks independently.
- MTL: this is the multi-task feature learning method of [2] and corresponds to the choice of $\rho = \lambda = 0$.
- MTL-2G: this approach consists in applying the method of [2] to each of group of tasks separately.

In the figures below, to ease the visualization of the results, only the best five methods are reported. We use the same setting of parameters for all experiments and all algorithms: we perform 5-fold cross-validation to tune the value of the regularization parameters, whenever those were treated as free parameters. We considered the values of $\gamma = 10^k$ with $k \in \{-4, \ldots, 2\}$, $\lambda = 10^k$, with $k \in \{4, \ldots, 7\}$ and $\rho = 10^k$ with $k \in \{-2, \ldots, 2\}$.

Finally, in all experiments we have trained all learning methods using the square loss function $L(y, z) = (y - z)^2$, $y, z \in \mathbb{R}$.

4.1 Synthetic Data

We can use synthetic data to test whether Algorithm 1 finds the right solution on data that satisfy the prior orthogonality assumptions. To this end, we have created a dataset consisting of 20 tasks, 10 of them belonging to the first subset and the remaining ones to the second subset ($T = S = 10$). The data is in a $d = 100$ dimensional space. From these 100 dimensions, only the first 5 are useful for the first subset of tasks and the following 5 are useful for the second subset. Finally, the remaining 90 dimensions are not important at all. In this synthetic dataset, every task is represented as either $(w_{1t}, \ldots, w_{5t}, 0, \ldots, 0)$, $t = 1, \ldots, 10$ or $(0, \ldots, 0, w_{6s}, \ldots, w_{10s}, 0, \ldots, 0)$, $s = 1, \ldots, 10$, where each parameter w_{it} is chosen randomly from a uniform distribution, $\mathcal{U}(0, 0.1)$.

We build a set of $n = 1000$ instances, $Z \in \mathbb{R}^{d \times n}$, so that every element of matrix Z is sampled from the uniform distribution on the unit interval. The training set is composed of a random subset of m instances, for different values of the sample size $m = 10, 15, \ldots, 50$, and the test set is composed of the remaining instances. For every task t, we generate the output y_t as $y_t = Zw_t + \eta_t$, where $\eta_t \in \mathbb{R}^m$ and $\eta_{ti} \sim N(0, 1)$, $i = 1, \ldots n$. Finally we apply an orthogonal rotation to Z by sampling an orthogonal matrix U randomly from the Haar measure and set $X = UZ$.

We have repeated the described experiment 750 times for each value of m. The results can be seen in Figure 1. MTL-2G performed comparably to Ridge Regression and MTL. All of our methods performed better than both Ridge Regression and MTL. OrthoMTL-C gives the best results, followed by OrthoMTL-EN and OrthoMTL. We have applied a paired t-test to check whether the difference between OrthoMTL-C and OrthoMTL-EN and either Ridge Regression or MTL is equal to 0 and obtained a p-value below 10^{-7} for training set sizes below 45.

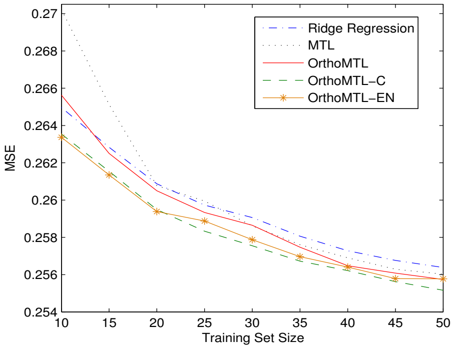

Figure 1: Synthetic data: Comparison between Ridge Regression, MTL [2], OrthoMTL, OrthoMTL-C and OrthoMTL-EN.

4.2 Real Data

Next, we tested the model with two real datasets. In both datasets we have two groups of supervised learning tasks so that the tasks belonging to one group are unrelated to the remaining ones.

4.2.1 JAFFE Dataset

The first experiment considered the Japanese Female Facial Expression (JAFFE) database [14]. It is composed of 213 images of 10 subjects displaying a range of expressions, like those shown in Figure 2 (top). There are 7, mutually exclusive emotion classes that need to be detected. The classes are: “happiness”, “sadness”, “surprise”, “anger”, “disgust”, “fear” and “neutral”. Given an unlabeled image, the objective is to predict the emotion expressed in it.

We represented an input image in the following manner. First we extracted the face from the background. To this end, we used the OpenCV implementation of Viola and Jones face detector [22] to detect the face and eyes in the image. After that, we rotated the face so that the eyes are horizontally aligned. Finally, we rescaled the face to a 200×200 size image. In order to obtain a descriptor of the textures of the image we used the Local Phase Quantization (LPQ) [16]. Specifically, we divided every image into 5×5 non overlapping regions. We computed the LPQ descriptor for each region and we created the image descriptor by concatenating all the LPQ descriptors. Finally we applied Principal Component Analysis to extract as many components as necessary to describe 99% of the data variance. After this process, we obtained a descriptor with 203 attributes for each image.

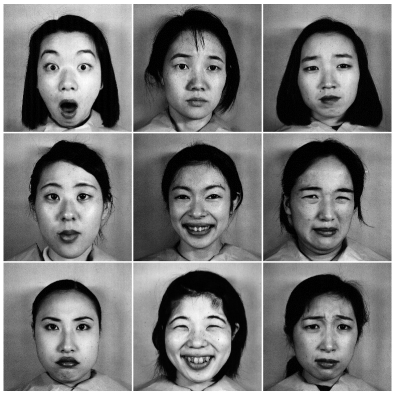

Figure 2: Sample images taken from the JAFFE dataset.

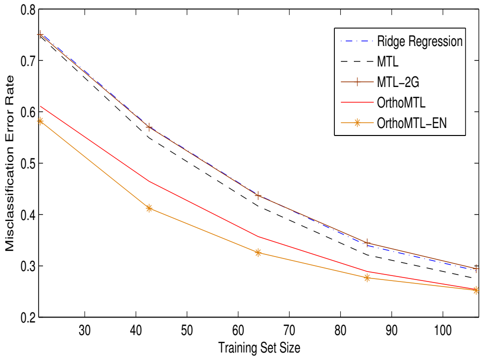

Figure 3: JAFFE dataset: Comparison between Ridge Regression, MTL, MTL-2G, OrthoMTL and OrthoMTL-EN.

As discussed in the introduction, we can assume that the features which are useful for recognizing the emotion are different from those which are useful for recognizing the identity of the subject. Therefore, it seems appropriate to apply our method when the principal tasks are those related to predicting the emotion and the auxiliary tasks are those related to the prediction of the identity. Each task discriminates one class from the others (one versus all), so that we have 7 tasks in the first group (one for each emotion) and 10 tasks in the second group (one for each actor).

We have carried out two experiments with this data set. In the first one we select randomly m instances as training set and use the remaining ones as test set. We run the experiments for different values of m so that we can plot the learning curve. The experiments were executed 200 times and the results are shown in Figure 3.

As we see, both OrthoMTL-EN and OrthoMTL outperform the other approaches, the improvement being more evident when the training set is small. This is reasonable since the prior information that we have (the emotion tasks are unrelated to the identity tasks) makes a significant difference when the training set size is smaller. We have applied a paired t-test between our methods and either MTL, MTL-2G and Ridge Regression, obtaining always a p-value below 10^{-3} for any value of m. This result supports the hypothesis that the differences between both approaches are significant. In this experiment, OrthoMTL-C (not shown in the plot) performed comparably to Ridge Regression.

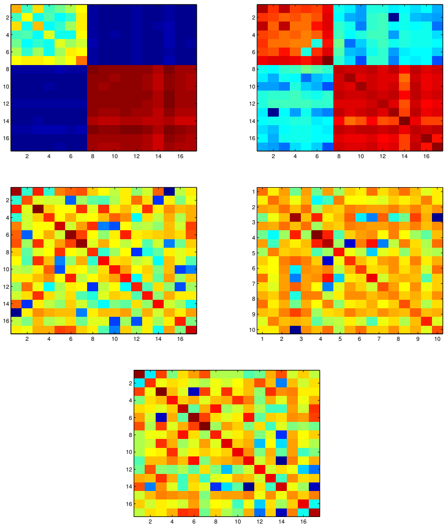

Figure 4: Tasks correlation matrix learned by different methods: OrthoMTL-EN (top left), OrthoMTL (top right), MTL-2G (middle left), MTL applied only to the emotion tasks (middle right) and Ridge Regression (bottom), Red (resp. blue) denotes high (resp. low) intensity values.

We also report in Figure 4 the task correlation matrix $[W, V]^\top [W, V]$ learned by the different methods. As it can be seen, the off-diagonal blocks of this matrix, which are formed by the inner products between tasks of different groups, are much smaller than the elements in the diagonal blocks, which correspond to inner products between tasks in the same group. This effect is more pronounced in the case of our methods, indicating that they can take advantage of the information contained in the auxiliary tasks.

In the second experiment, we have considered a transfer learning problem with the aim of comparing OrthoMTL-EN with the Bilinear Model proposed in [19]. A transfer learning problem requires test instances for identities which are not present in the training set. To do so, we have used a leave-one-subject-out strategy. To tune the parameters of the Bilinear Model we have also followed a cross-validation process. We have run 10 times the whole process (that is, each subject has been in the test set 10 times) and the results are shown in Figure 5. The results show that our approach clearly outperforms the Bilinear Model for this dataset. The resulting p-value is below 0.01 supporting our claim.

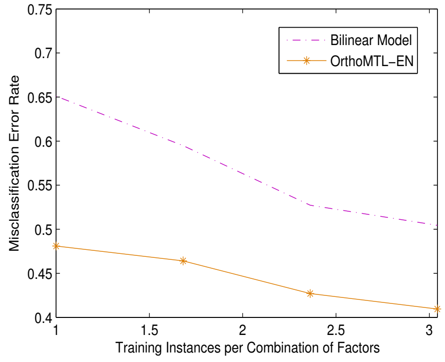

Figure 5: JAFFE dataset: Comparison between Bilinear Model and OrthoMTL-EN in a transfer learning experiment – see text for description.

4.2.2 UNBC-McMaster Shoulder Pain Expression Archive

As a final test, we apply our methods to the UNBC-McMaster Shoulder Pain Expression Archive [13]. Differently from the previous dataset, this data set contains spontaneous facial expressions, i.e. it presents higher variability than stereotypical acted expressions. It contains 200 video clips of facial expressions of 25 patients who suffer from shoulder pain. The facial expressions were captured while the subjects were performing a series of active and passive physical exercises. A label indicating the level of pain felt by the patient is provided for each video frame in each video clip. The dataset also provides 66 tracked landmarks points of the face for each frame of each clip. Our task here is to recognize if a frame of a clip shows a pain expression (i.e., pain value bigger than 0) or not. Instead of texture features, in this experiment, the attributes consist of distances between provided landmarks points as shown in Figure 6 (top).

Even though some people are more prone to feeling pain than others, we still can assume that the task of detecting pain is unrelated to the task of detecting a person's identity. To test the algorithm, the experiments have been carried out using a leave-one-subject-out protocol. At each step, the frames from one patient were used as test set and a percentage of $0.1\%, 0.125\%, \ldots, 0.325\%$ randomly selected frames from the remaining 24 patients were used as the training set. The process was repeated until all the subjects had been used as the testing set once. The whole protocol was executed 30 times. The mean results (using Area Under the Curve as a measure of accuracy) are reported in Figure 6 (bottom).

As it can be noted, all of OrthoMTL-EN, OrthoMTL-C and OrthoMTL perform significantly better than their competitors (MTL and Ridge Regression). The advantage of our methods is particularly clear in the case of OrthoMTL-EN which performs the best. OrthoMTL also performs well, especially as the training set decreases: By applying a paired t-test, we observe that when the training set is small, $m = 48$ (corresponding to 0.1% of the number of available frames), the difference between each of our methods and both MTL or Ridge Regression is significant ($p < 10^{-3}$) and it remains still significant as the training set increases to $m = 140$ ($p < 0.025$).

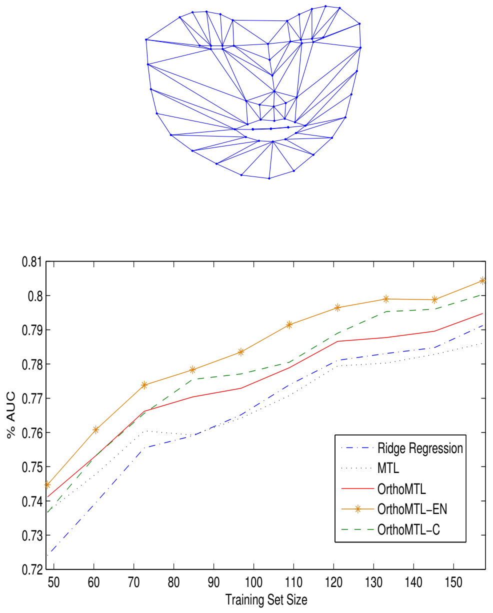

Figure 6: Top: Landmark points and edges used to build the attributes for the UNBC-McMaster Shoulder Pain Expression Archive (selected according to Figure shown in [13]). Bottom: Comparison between Ridge Regression, MTL, OrthoMTL-EN, OrthoMTL and OrthoMTL-C on the UNBC-McMaster Shoulder Pain Expression Archive Database.

5 Discussion

We have addressed the problem in which two or more groups of supervised learning tasks are unrelated in the sense that they involve different linear discriminative features of the input data. We have proposed a regularization formulation which incorporates this information in the learning method. The regularizer encourages both a low dimensional representation and penalizes the inner product between any pair of weight vectors of tasks from different groups. The implication of this constraint is that we look for common sparse representations within each group of tasks and also that tasks from different groups share as few features as possible. The method depends on three regularization parameters. For special choices of these parameters, the method reduces to the multi-task feature learning approach of [2] and to Ridge Regression (independent multi-task learning).

At first sight it seems surprising that we can exploit one group of tasks to improve learning of the other group. However, the fact that the two groups of tasks use different features provides an implicit constraint about which features could be used by each group, thereby helping the learning process. To validate this claim, we have presented experiments on a synthetic and on two well-characterized real datasets comparing our algorithm with Ridge Regression as a base line and with the linear multi-task feature learning method of [2]. The experimental results indicate that the proposed method consistently improves over the other methods, supporting our hypothesis that taking into account independence helps discriminate features for tasks in different groups.

Overall, our results indicate that our method performs best when all regularization parameters are tuned by cross validation. A simplified setting of the method, in which only two parameters are tuned, also provides improved results over the method of [2] and Ridge Regression. We have also discussed a special setting of our method, which leads to a convex optimization problem. Our experimental results in this setting are encouraging though not conclusive: We obtained good results on the synthetic dataset and one real dataset but no improvement was observed on the other real dataset.

The work presented here can be extended in different directions. On the theoretical side, it would be valuable to investigate whether the improved generalization performance of the method could be supported by a statistical analysis. When the auxiliary tasks are known *a priori* such a result would follow from the analysis in [15]. However when both the primary and auxiliary tasks need to be estimated from data, the above problem remains to be understood. On the practical side, it may be valuable to explore the application of our approach in the context of hierarchical classification where recent work has considered the incorporation of orthogonal constraints [25]. The ideas presented here could also be applied to matrix completion problems such as those arising in the context of collaborative filtering.

A Appendix

In this appendix we present the proof of Theorem 3.1. We define the function

$$\Omega(W,V) = \frac{1}{2}\left(\|W\|_{\mathrm{F}}^2 + \|V\|_{\mathrm{F}}^2 + \alpha\|W^\top V\|_{\mathrm{F}}^2\right).$$

The proof is based on the following lemma[3].

Lemma A.1. *Assume that $\|W\|_{\mathrm{F}}^2 + \|V\|_{\mathrm{F}}^2 < R^2$. Then the function Ω is convex on this domain provided that $\alpha < \frac{2}{R^2}$.*

Proof. We will compute the Hessian matrix H of function Ω and establish that it is positive semidefinite in the domain of interest, whenever $\alpha \leq \frac{2}{R^2}$. From calculus we find that

$$H(W,V) = \begin{bmatrix} A(W,V) & C(W,V) \\ C(W,V)^\top & B(W,V) \end{bmatrix}$$

where

$$A_{ti,\hat{t}j}(W,V) = \frac{\partial^2\Omega(W,V)}{\partial w_{ti}\partial w_{\hat{t}j}} = (\delta_{ij} + \alpha\sum_s v_{si}v_{sj})\delta_{t\hat{t}}$$

$$B_{si,\hat{s}j}(W,V) = \frac{\partial^2\Omega(W,V)}{\partial v_{si}\partial v_{\hat{s}i}} = (\delta_{ij} + \alpha\sum_t w_{ti}w_{tj})\delta_{s\hat{s}}$$

$$C_{ti,sj}(W,V) = \frac{\partial^2\Omega(W,V)}{\partial w_{ti}\partial v_{sj}} = \alpha(\langle w_t, v_s\rangle\delta_{ij} + v_{si}w_{tj}).$$

The matrix H is positive semidefinite if, for every $X \in \mathbb{R}^{d\times T}$ and $Z \in \mathbb{R}^{d\times S}$ it holds that

$$\sum_{tij} X_{ti}A_{ti,tj}X_{tj} + \sum_{sij} Z_{si}B_{si,sj}Z_{sj} + 2\sum_{stij} X_{ti}C_{tisj}Z_{sj} \geq 0$$

where $t \in \{1,\ldots,T\}$, $s \in \{1,\ldots,S\}$ and $i,j \in \{1,\ldots,d\}$. In matrix notation we obtain

$$\|X\|_{\mathrm{F}}^2 + \|Z\|_{\mathrm{F}}^2 + \alpha\|X^\top V + W^\top Z\|_{\mathrm{F}}^2 + 2\alpha\langle W^\top V, X^\top Z\rangle_{\mathrm{F}}.$$

Discarding the middle term and using Cauchy-Schwarz inequality, we bound from below the above quantity by

$$\|X\|_{\mathrm{F}}^2 + \|Z\|_{\mathrm{F}}^2 - 2\alpha\|W^\top V\|_{\mathrm{F}}\|X^\top Z\|_{\mathrm{F}}.$$

Next, using the inequality $2\|X^\top Z\|_{\mathrm{F}} \leq \|X\|_{\mathrm{F}}^2 + \|Z\|_{\mathrm{F}}^2$, we have the lower bound

$$(\|X\|_{\mathrm{F}}^2 + \|Z\|_{\mathrm{F}}^2)(1 - \alpha\|W^\top V\|_{\mathrm{F}}).$$

The result follows.

Proof of Theorem 3.1. We first use equation (8) and rewrite problem (7) as an optimization problem in W and V only. Specifically, we obtain the objective function

$$f(W,V) = \mathcal{E}(W,V) + \gamma\|[W,V]\|_{\mathrm{tr}}^2 + \lambda\|W^\top V\|_{\mathrm{F}}^2 + \rho(\|W\|_{\mathrm{F}}^2 + \|V\|_{\mathrm{F}}^2)$$

where $\|\cdot\|_{\mathrm{tr}}$ denotes the trace norm, that is the ℓ_1 norm of the vector of singular values.

Since the function f is continuous and grows at infinity, it has a minimum. Moreover, if the pair $(\hat{W},\hat{V})$ is a minimizer then $f(\hat{W},\hat{V}) \leq f(0,0)$, which readily implies that $\|W\|_{\mathrm{F}}^2 + \|V\|_{\mathrm{F}}^2 \leq \mathcal{E}(0,0)/\rho$. The result now follows by applying Lemma A.1 with $R^2 = \mathcal{E}(0,0)/\rho$ and $\alpha = \lambda/\rho$. □

[3] We also refer to [25] for a similar result for the regularizer $\Omega(W,V) = \|W\|_{\mathrm{F}}^2 + \|V\|_{\mathrm{F}}^2 + \alpha\|W^\top V\|_1$. See also our remarks preceding equation (5).

References

[1] R. K. Ando and T. Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6:1817–1853, 2005.

[2] A. Argyriou, T. Evgeniou, and M. Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.

[3] A. Argyriou, A. Maurer, and M. Pontil. An algorithm for transfer learning in a heterogeneous environment. In *ECML/PKDD*, pages 71–85, 2008.

[4] B. Bakker and T. Heskes. Task clustering and gating for bayesian multi–task learning. *Journal of Machine Learning Research*, 4:83–99, 2003.

[5] J. Baxter. A model for inductive bias learning. *Journal of Artificial Intelligence Research*, 12:149–198, 2000.

[6] S. Ben-David and R. Schuller. Exploiting task relatedness for multiple task learning. In *Proceedings of the Sixteenth Annual Conference on Learning Theory*, pages 567–580, 2003.

[7] A.J. Calder, A.M. Burton, P. Miller, A.W. Young, and S. Akamatsu. A principal component analysis of facial expressions. *Vision Research*, 41(9):1179–1208, 2001.

[8] R. Caruana. Multi–task learning. *Machine Learning*, 28:41–75, 1997.

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

[10] J. Guinney, Q. Wu, and S. Mukherjee. Estimating variable structure and dependence in multitask learning via gradients. *Machine Learning*, pages 1–23, 2011.

[11] L. Jacob, F. Bach, and J.P. Vert. Clustered multi-task learning: A convex formulation. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 745–752. 2009.

[12] K. Lounici, M. Pontil, A.B Tsybakov, and S. van de Geer. Taking advantage of sparsity in multi-task learning. In *Proc. of the 22nd Annual Conference on Learning Theory (COLT)*, 2009.

[13] P. Lucey, J.F. Cohn, K.M. Prkachin, P.E. Solomon, and I. Matthews. Painful data: The UNBC-McMaster shoulder pain expression archive database. In *Automatic Face & Gesture Recognition and Workshops (FG 2011)*, pages 57–64, 2011.

[14] M. Lyons and S. Akamatsu. Coding facial expressions with Gabor wavelets. *Computer*, pages 200–205, 1998.

[15] A. Maurer. Transfer bounds for linear feature learning. *Machine Learning*, 75(3):327–350, 2009.

[16] V. Ojansivu and J. Heikkilä. A method for blur and affine invariant object recognition using phase-only bispectrum. In *ICIAR*, pages 527–536, 2008.

[17] S.J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 10(22):1345–1359, 2009.

[18] J. Shawe-Taylor and N. Cristianini. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

[19] J.B. Tenenbaum and W.T. Freeman. Separating style and content with bilinear models. *Neural computation*, 12(6):1247–83, 2000.

[20] S. Thrun and J. O'Sullivan. Discovering structure in multiple learning tasks: The tc algorithm. In *ICML*, pages 489–497, 1996.

[21] M.A.O. Vasilescu and D. Terzopoulos. Multilinear image analysis for facial recognition. *Object recognition supported by user interaction for service robots*, pages 511–514, 2002.

[22] P. Viola and M.J. Jones. Robust real-time face detection. *International Journal of Computer Vision*, 57(2):137–154, 2004.

[23] H. Wang and N. Ahuja. Facial expression decomposition. *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 958–965 vol.2, 2003.

[24] M. Yuan and Y. Lin. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society, Series B*, 68(1):49–67, 2006.

[25] D. Zhou, L. Xiao, and M. Wu. Hierarchical classification via orthogonal transfer. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, 2011.

[26] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(2):301–320, 2005.