Ranking based Multitask Learning of Scoring Functions

Ivan Stojkovic[1,2], Mohamed Ghalwash[1,3,4], and Zoran Obradovic[1]

[1] Center for Data Analytics and Biomedical Informatics, Temple University, Philadelphia PA 19122, USA
[2] School of Electrical Engineering, University of Belgrade, 11120 Belgrade, Serbia
[3] IBM T. J. Watson Research Center, Cambridge MA, USA
[4] Faculty of Science, Ain Shams University, 11566 Cairo, Egypt

Abstract. Scoring functions are an important tool for quantifying properties of interest in many domains; for example, in healthcare, a disease severity scores are used to diagnose the patient's condition and to decide its further treatment. Scoring functions might be obtained based on the domain knowledge or learned from data by using classification, regression or ranking techniques - depending on the type of supervised information. Although learning scoring functions from collected data is beneficial, it can be challenging when limited data are available. Therefore, learning multiple distinct, but related, scoring functions together can increase their quality as shared regularities may be easier to identify. We propose a multitask formulation for ranking-based learning of scoring functions, where the model is trained from pairwise comparisons. The approach uses mixed-norm regularization to impose structural regularities among the tasks. The proposed regularized objective function is convex; therefore, we developed an optimization approach based on alternating minimization and proximal gradient algorithms to solve the problem. The increased predictive accuracy of the presented approach, in comparison to several baselines, is demonstrated on synthetic data and two different real-world applications; predicting exam scores and predicting tolerance to infections score.

Keywords: sparse learning, multitask, mixed-norm regularization

1 Introduction

Quantifying the properties of interest is an integral part in many domains, e.g., assessing the condition of a patient [27], estimating the risk of an investment [1], or predicting binding affinity of a ligand [4] when developing new drugs. Various measuring technologies and sensors are devised to quantify such properties of interest, which would in turn be utilized for informing decisions and making appropriate actions. However, the properties of interest are often not easy to obtain, whether they are difficult to measure directly or completely unobservable. This is usually the case when the properties are conceptual, i.e. they are latent constructs, such as health, satisfaction, and even intelligence. Under these circumstances, other measurable characteristics, considered related and informative of the true target, are observed and used as surrogate variables. In clinical settings, variables like temperature, blood pressure and various biomarkers measured from tissues are commonly tracked and considered when determining the health of the patient.

Typically, some heuristic rules are decided to map these surrogate variables into the desired score. The process of deciding these heuristic rules (or scoring functions) is usually long and tedious. For example, disease severity scores that are needed in clinical practices for patient diagnostics require years of effort and consensus of the medical community before the scoring functions can become part of the protocols. Fortunately, developments in machine learning and increasing amounts of collected data allows an alternative and complementary way for engineering the scoring functions by extracting rules automatically from the data, which facilitates and complements traditional approaches.

Algorithms for learning scoring functions from data were previously proposed, mainly in the medical domain, with the objective to learn disease severity scores [11, 12, 21, 28, 31]. Initial approaches posed the problem as traditional supervised learning tasks of classification [21, 28] and regression [31]. However, classification and regression approaches require scores to be already accessible up front, which limits their applicability to problems with a good surrogate. The approach in [11, 12] suggests the very appealing idea that there is a more convenient alternative form of supervised information to learn the scoring function from. Namely, ranked pairs are much easier to obtain than direct score estimates, and moreover, learning from pairs of ranked examples may result in more reliable and robust scoring functions.

In this work, we extend the suggested ranking-based approach [11] for score learning in multitask settings. The efforts are motivated by the applications in which there are multiple related tasks, with a limited amount of data for each task. Related tasks commonly share underlying regularities which could be learned more accurately by modeling all tasks together. For example, in education, scores on different subjects (e.g. Math and English) are dependent on the same characteristics of a particular student and a particular school. In the medical domain, disease severity scores for related illnesses (e.g. various respiratory viral infections) are expected to share common underlying biological mechanisms.

Consequently, we propose a novel multitask formulation for learning scoring functions from pairwise comparisons, by enforcing structural regularities on joint parameter space, using a matrix norm regularizations. In addition, we provide another contribution by developing an optimization algorithm in the form of an alternate minimization scheme based on a proximal gradient method. We evaluated the proposed approach on a synthetic data and two real-world applications. The objective of the first application is learning exam scores of elementary school pupils, while the objective of the second application is learning the tolerance to respiratory viral infections in humans. The results showed increased prediction accuracy of the proposed approach over individual tasks.

2 Related Work

Early efforts to learn scoring functions were dependent on complete supervised information (e.g. classification and regression tasks). In the classification settings, where the discrete class labels are provided, the classification methods were used to estimate the probability of a sample belonging to a certain class; these probabilities were used as a scoring function. For example, the method in [28] uses sparsity inducing L_1 norm in

combination with a classical logistic loss function to learn the disease severity scoring function for assessing the abnormality of the skull in craniosynostosis cases.

Another similar approach is to learn the scoring function in a regression manner from the continuous outcome. In [31], Alzheimer's disease severity, as measured by cognitive scores, is modeled as a (temporal) multi-task regression problem using the fused sparse group lasso approach. The approach was more concerned with the progression of the disease; hence, the multi-task problem was formulated considering each time-step as a separate task. In contrast, we are interested in multiple score mapping from a single time-point set of measurements. There is also work on multitask learning to rank in the context of web search results ranking [6], where the ranking function is learned using the gradient boosted trees from the ranking scores provided by the human experts.

The problem with such completely-supervised methods is the necessity of providing direct values of scores for training purposes, which render the approaches as less powerful in settings where characteristics of interest are latent and not directly accessible. However, rather than giving direct estimates of the score, the easier task seems to be comparing two samples and asserting whether one has a higher score than the other. Ranking SVM [18] was the first approach that recognized the benefits of learning from ordered pairs of samples. This method was applied to learn an improved relevance function for documents retrieval from click-through data. Main insight was that clicked links are definitely more relevant for the search, as compared to non-clicked ones. And such kind of data is much more abundant than the user provided rankings. Recently, the ranking SVM-based method was adopted for Sepsis severity score learning [11] and extended for temporal applications by introducing a term that ensures gradual score change over consecutive time points.

Multitask learning is based on the idea that generalization (predictive performance) can be increased by accounting for the intrinsic relationships among multiple tasks. Multitask approach is found particularly effective when the number of samples per task is small. To the best of our knowledge, there are no published multitask formulations for ranking-based scoring functions, that is, for methods that learn from pairwise comparisons. The closest approaches are the previously mentioned multitask regression-based models for Alzheimer's disease progression [31] and search results ranking [6]. Other multitask regression methods exist that learn the structure among the tasks using norm regularization [30], or methods that utilize fixed relatedness structure [23] obtained from domain knowledge [25] or learned from a statistical correlation [24]. However, since they are not directly proposed for ranking-based learning of the scoring functions, we will not consider them, nor will compare with them in this work.

The main problem in multi-task learning is finding the most appropriate assumption on how the tasks are related and incorporating such assumption into the model. Typically, in linear models, such structural assumptions are imposed on the joint parameter matrix, where rows correspond to features and columns to different tasks. Kernel methods assume that all tasks are related and similar [13], but some methods enforce tasks to be grouped into clusters [16]. For example, "Dirty method" [17] encourages block-structured row-sparsity in the joint parameter matrix by $\|.\|_{1,1}$ norm, and element-wise sparsity with $\|.\|_{1,\infty}$. The robust approach [14] selects sparse rows of features for related tasks with $\|.\|_{2,1}$ and dense columns for outlier tasks with $\|.\|_{1,2}$, in order to discern

between related and unrelated tasks. Other approaches assume some shared common set of features [3] or shared common subspace [2,9]. The approach proposed in [10] attempts to learn such relatedness subspace with trace (nuclear) norm $\|.\|_*$ by encouraging the parameter matrix to have low rank, and finding outlier tasks with additional sparse group norm $\|.\|_{1,2}$.

In this work we use regularization composed of trace norm [10], and grouped Lasso penalty [3] to jointly learn multiple ranking based scoring tasks, from temporal data.

3 Model

Let us assume that we have N samples (examples), where each sample i is represented as $X_i \in \mathbb{R}^d$, and where X_{ij} is the value (measurement) of the feature $j = \{1, 2, \ldots, d\}$ for the sample $i = \{1, 2, \ldots, N\}$. Let us assume that $y_i \in \mathbb{R}$ represents the property of interest (outcome variable) for the sample i. Scoring function $score : \mathbb{R}^d \to \mathbb{R}$ is then a mapping $X_i \mapsto y'_i$ that provides a close estimate y'_i of the true score y_i.

However, in many cases the values of the true scoring function are difficult to obtain. In such situations, it is easier to assess the ranking between the scores of two samples p and q, i.e. to assert that one has perceived higher score than the other: $score(X_p) > score(X_q)$. Therefore, a set of multiple such ordered pairs can be used to find a projection in the space of measured features, that will preserve the orders in the best possible way, and that might be used as a scoring function.

Moreover, measurements collected on multiple occasions over time might belong to the same subject; In this case, the measurements at each time step will be considered as a sample. We assume that the outcome variable changes gradually (smoothly) over time for the same subject, e.g. the disease severity score changes smoothly over consecutive time points for the same patient. This assumption will lead to improving the quality of the scoring function. We assume that X_p represents the feature vector for the sample p (which could be one particular subject at one particular time point).

In this work, we constrain such functional mapping $score$ to the linear case, where the score estimate is computed as a weighted sum of the measured characteristics: $score(X) = w^T X$. Therefore, the problem of learning the scoring function becomes finding the appropriate weight (or parameter) vector $w \in \mathbb{R}^d$.

3.1 Single task model formulation

Maximizing the number of correctly ordered training pairs can be performed using the soft max-margin framework expressed in a Hinge loss form (1), as suggested in [18].

$$max(0, 1 - (X_p - X_q)w) \tag{1}$$

If sample p should have higher score compared to sample q, the formulation (1) will favor the weighted difference $(X_p - X_q)w$ that is positive and greater than 1, thus even achieving some margin in the score difference.

The L_2 norm on the weight vector $||w||^2$, is introduced to regularize the magnitude of the weights, and to turn the problem into simultaneous maximization of correct ordering and maximization of normalized margin.

Gradual (smooth) change of the scoring function over time can be obtained by penalizing high changes of the score (e.g. for two samples X^s_{i+1}, X^s_i of the same subject s), over short time intervals. In [12] such effect is achieved by using the temporal smoothness term:

$$\left(\frac{(X^s_{i+1} - X^s_i)w}{(t^s_{i+1} - t^s_i)} \right)^2 \tag{2}$$

, which essentially ensures that squared magnitude in difference, normalized with the time interval length, is kept low.

Therefore, for single task formulation of ranking-based scoring function learning, we adopted the Linear Disease Severity Score Learning formulation [11] which combines attractive properties of ranking SVM [18], with temporal smoothness term (2) that enforces the gradual change of the scoring function over time:

$$\begin{aligned} \hat{w} = \underset{w}{\operatorname{argmin}} \frac{1}{2}\|w\|_2^2 + c \sum_{\{p,q\} \in O} max(0, 1 - (X_p - X_q)w) \\ + b \sum_{\{i,i+1\}_s \in S} \left(\frac{(X^s_{i+1} - X^s_i)w}{(t^s_{i+1} - t^s_i)} \right)^2 \end{aligned} \tag{3}$$

Every measurement (row) vector X_i, $i = \{1, 2, \ldots, N\}$ has associated time-stamp t, while $\hat{w} \in \mathbb{R}^d$ denotes the solution of the objective 3.

Set O is composed of ordered pairs $\{p, q\}$, where p has a higher rank than q (p is perceived to have a higher score than q), and which corresponds to the measurement vectors X_p and X_q, respectively. Sum of the Hinge loss terms over all pairs from the O set, serves to reduce the extent of incorrectly ordered pairs.

Set of all consecutive pairs in all subjects is denoted S and the sum of the Temporal smoothness terms in eq. (3) penalizes high rates of change in score values in consecutive time steps t_i and t_{i+1} for all subjects $s \in S$. Scalar constants c and b are hyperparameters that determine the cost of the respective loss terms, the Hinge loss and the Temporal loss.

We aggregate the differences of measurements in the Hinge loss term into a single data matrix $D_{k \times d}$, where k is the number of pairs in the comparison set O. Similarly, measurement and temporal difference ratios in the Temporal loss term we write as matrix $R_{l \times d}$, where l is a number of pairs in the consecutive measurements set S. We aggregate the L_2 norm and temporal smoothness terms (they are essentially weighting the square of optimization parameters) into a single weighted quadratic term $\frac{1}{2} w^T Q w$, where Q is constant square matrix defined in eq. (4):

$$Q = I + 2bR^T R \tag{4}$$

, I being the d-dimensional identity matrix.

The formulation (3) can now be rewritten more concisely as (5):

$$\hat{w} = \underset{w}{\operatorname{argmin}} \frac{1}{2} w^T Q w + c \sum_i max(0, 1 - D^i w) \tag{5}$$

3.2 Multitask formulation

As mentioned before, in case of a limited amount of data for training the scoring function for a single task (5), it is beneficial to exploit the relatedness among the multiple similar tasks, by learning them together, as illustrated in Figure 1.

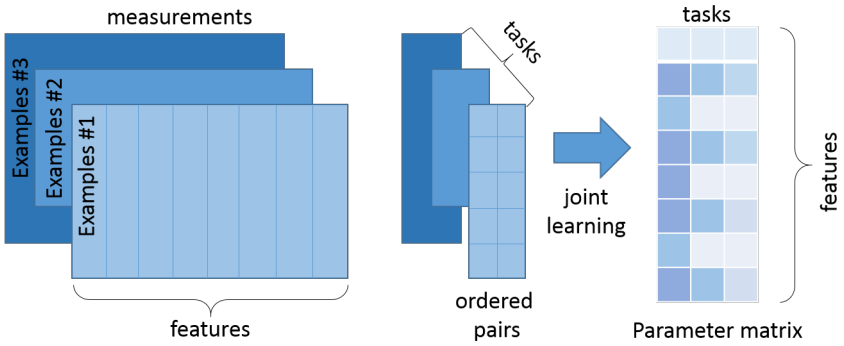

Fig. 1: Illustration of joint training of multiple ranking based score learning tasks. Three distinct task are depicted, where measured data in combination with supervision in form of ordered pairs, are jointly optimized to obtain the scoring function parameters, represented as parameter matrix. Parameter matrix is typically regularized to encode the structural assumptions regarding the task relatedness.

For m different tasks, individual parameter vectors w_i are aligned into a matrix $W_{d \times m}$, and a joint objective is obtained as a superposition of individual losses (eq. (5)) over the multiple tasks $i \in \{1, 2, ..., m\}$:

$$\underset{W}{\operatorname{argmin}} \sum_{i=1}^{m} \left(\frac{1}{2} W_i^T Q_i W_i + c \sum_j max(0, 1 - D_i^j W_i) \right) \tag{6}$$

Instead of the non-smooth Hinge loss $L(a) = max(0, a)$ in eq. (6), we work with the twice differentiable approximation in the form of Huber loss [11]:

$$L_h(a) = \begin{cases} 0 & \text{, if } a < -h \\ \frac{(a+h)^2}{4h} & \text{, if } |a| \leq h \\ a & \text{, if } a > h. \end{cases} \tag{7}$$

, where the approximation threshold h can be chosen arbitrarily small.

Further, we regularize the objective in eq. (6) with a joint norm on parameter matrix $\|W\|_{p,q} = (\sum_i((\sum_j (W_{ij}^q)^{\frac{1}{q}})^p)^{\frac{1}{p}}$. For $p = 2$ and $q = 1$, this approach is known as a group Lasso penalty on the row groups (of W), which forces sparsity in the parameter weights corresponding to certain features [3]. Additionally, we introduce the trace norm L_* in order to get the low rank component, or in other words, the parameter weight pattern common among all the tasks. To accommodate such a setup, which will be further clarified in the Optimization section, the parameter matrix W was split into two distinct matrices A and B, where $W = A + B$.

Multitask Ranking Based Scoring Function Learning (MultiRBSFL) objective is now given in eq. (8), and it takes as an input two matrices (per task i) obtained from the data: $Q^i_{d\times d}$ and $D^i_{k\times d}$; hyperparameters b, c, λ_1 and λ_2 weighting the influence of Temporal loss, Huber loss, trace norm and sparse group norm, respectively.

$$\underset{W=A+B}{\operatorname{argmin}} \ \mathcal{L}_1 + \lambda_1\|A\|_* + \lambda_2\|B\|_{2,1} \tag{8}$$

where

$$\mathcal{L}_1 = \frac{1}{m}\sum_{i=1}^{m}\left(\frac{1}{2}(A^i + B^i)^T Q^i (A^i + B^i) + c\sum_{j=1}^{k} L_h(1 - D^i_j(A^i + B^i))\right) \tag{9}$$

A^i and B^i are column vectors $\mathbb{R}^{d\times 1}$, and D^i_j is $\mathbb{R}^{1\times k}$ row-vector.

4 Optimization

The optimization (8) is composed of smooth and non-smooth terms. However, although the regularaization terms are separable in A and B, the loss term $\mathcal{L}_1$ is not separable. Therefore, we solve the problem by using the alternative minimization scheme, where, in each iteration, we fix A and minimize (8) with respect to B, and then fix B and minimize (8) w.r.t A. In this case, each subproblem can be decomposed into two different optimizations. This will be explained in the next section.

Fix A

$$\underset{B}{\operatorname{argmin}} \ \mathcal{L}_1 + \lambda_2 \left\|B\right\|_{2,1} \tag{10}$$

Fix B

$$\underset{A}{\operatorname{argmin}} \ \mathcal{L}_1 + \lambda_1 \left\|A\right\|_* \tag{11}$$

In general, problem (10) and (11) can be written as:

$$\underset{\boldsymbol{\Theta}}{\operatorname{argmin}} \ \mathcal{L}_1 + \gamma \left\|\boldsymbol{\Theta}\right\|_p \tag{12}$$

, where $\boldsymbol{\Theta} = \{A, B\}$ and $p = \{*, \{2, 1\}\}$.

The optimization (12) is convex. The expression $\mathcal{L}_1$ is smooth and the regulariation term (either group lasso or trace norm) is non-smooth. Therefore, we solve (12) using the proximal methods.

4.1 Proximal Algorithm

We solve (12) using the proximal gradient method [20].

$$\begin{aligned}\boldsymbol{\Theta}^{k+1} &:= \mathbf{prox}_{\lambda\|\boldsymbol{\Theta}\|_p}(\boldsymbol{\Theta}^k - \lambda\nabla\mathcal{L}_1(\boldsymbol{\Theta}^k)) \\ &= \underset{\boldsymbol{\Theta}}{\operatorname{argmin}}\left(\|\boldsymbol{\Theta}\|_p + \frac{1}{2\lambda}\left\|\boldsymbol{\Theta} - (\boldsymbol{\Theta}^k - \lambda\nabla\mathcal{L}_1(\boldsymbol{\Theta}^k))\right\|_2^2\right)\end{aligned} \tag{13}$$

, where $\mathbf{prox}_{\lambda\|\boldsymbol{\Theta}\|_p}$ is the proximal operator of the scaled function $\|\boldsymbol{\Theta}\|_p$, and $\lambda \in (0, 1/L]$ is a *constant* step size, and L is a Lipschitz constant of $\nabla\mathcal{L}_1$. Problem (12) can be solved analytically, where the proximal operator associated with the norm can be obtained as in [5].

Trace norm. Let us assume that $M = U\Sigma V$ is the singular value decompoistion of M, where Σ is a diagonal matrix and its entries σ_i are the singluar values of the matrix M. The proximal operator of the trace norm is defined as [8]:

$$\mathbf{prox}_{\lambda\|.\|_*}(M) = U\mathbf{diag}(\mathbf{prox}_{\lambda\|.\|_1}(\sigma(M)))V$$

i.e., the proximal operator of $\|.\|_*$ can be calculated by carrying out a singular value decomposition of Z and evaluating the proximal operator of the corresponding absolutely symmetric function at the singular values $\sigma(M)$. Therefore,

$$\mathbf{prox}_{\lambda\|.\|_*}(M) = U\mathbf{diag}(\overline{\sigma}_1, \overline{\sigma}_2, \ldots, \overline{\sigma}_n)V \tag{14}$$

, where:

$$\overline{\sigma}_i = \begin{cases}\sigma_i - \lambda & \sigma_i \geq \lambda \\ 0 & -\lambda \leq \sigma_i \leq \lambda \\ \sigma_i + \lambda & \sigma_i \leq -\lambda\end{cases}$$

Equation (14) is sometimes called the singular value thresholding operator.

Group lasso norm. The proximal operator associated with the group lasso norm is defined as:

$$\left[\mathbf{prox}_{\lambda\|.\|_{1,2}}(u)\right]_g = \begin{cases}(1 - \frac{\lambda}{\|u_g\|_2})u_g & \|u_g\|_2 > \lambda \\ 0 & \text{otherwise}\end{cases}$$

4.2 Step size

In order to find an adaptive step size λ^k in each iteration k, we employ the backtracking line search algorithm [7], which requires computing an upper bound for $\mathcal{L}_1$. Since $\mathcal{L}_1$ is convex and smooth, and $\nabla\mathcal{L}_1$ is L-Lipschitz continuous, it follows that:

$$\mathcal{L}_1(\boldsymbol{\Theta}) \leq \underbrace{\mathcal{L}_1(\boldsymbol{\Theta}^k) + \nabla\mathcal{L}_1(\boldsymbol{\Theta}^k)^T(\boldsymbol{\Theta} - \boldsymbol{\Theta}^k) + \frac{L}{2}\left\|\boldsymbol{\Theta} - \boldsymbol{\Theta}^k\right\|_2^2}_{\widehat{\mathcal{L}_1}_{\frac{1}{L}}(\boldsymbol{\Theta}, \boldsymbol{\Theta}^k)} \tag{15}$$

Algorithm 1 Fast Gradient Proximal Method with Backtracking Step Size

Input: $\boldsymbol{\Theta}^0$ (random), η (usually $1/2$), $L > 0$
$\lambda = \frac{1}{L}$, $\mathbf{z}^1 = \boldsymbol{\Theta}^0$, $t_1 = 1$, $k = 0$
repeat
  $k \longleftarrow k + 1$
  while true **do**
    $\mathbf{z} \longleftarrow$Solve (12) ▷ use λ and $\mathbf{z}^k$
    if $\mathcal{L}_1(\mathbf{z}) \leq \widehat{\mathcal{L}_1}(\mathbf{z}, \mathbf{z}^k)$ **then**
      break
    end if
    $\lambda \longleftarrow \eta\lambda$
  end while
  $\boldsymbol{\Theta}^k \longleftarrow \mathbf{z}$
  $t_{k+1} = \frac{1+\sqrt{1+4t_k^2}}{2}$
  $\mathbf{z}^{k+1} = \boldsymbol{\Theta}^k + (\frac{t_k - 1}{t_{k+1}})(\boldsymbol{\Theta}^k - \boldsymbol{\Theta}^{k-1})$
until Convergence

Algorithm 2 Alternative Minimization

Input: A^0, B^0 (random)
repeat
  Fix A, solve (10) using Algorithm (1).
  Fix B, solve (11) using Algorithm (1).
until Convergence

By utilizing (15), it can be shown that the optimization (13) is equivalent to [20]:

$$\boldsymbol{\Theta}^{k+1} := \underset{\boldsymbol{\Theta}}{\operatorname{argmin}}\ \widehat{\mathcal{L}_{1\lambda^k}}(\boldsymbol{\Theta}, \boldsymbol{\Theta}^k) + \|\boldsymbol{\Theta}\|_p \tag{16}$$

where $\lambda^k = \frac{1}{L}$. So at each iteration, the function $\mathcal{L}_1$ is linearized around the current point and the problem (16) is solved. The final fast proximal gradient method with backtracking is shown in Algorithm 1. The final alternative minimization algorithm is shown in Algorithm (2).

5 Empirical evaluation

The proposed approach for multitask learning of ranking-based scoring functions is tested on one synthetic and two real-world datasets. We compared our MultiRBSFL approach against the following baseline approaches:

1. L_2 - independently learning scoring functions for each task (objective (3));
2. L_1 - independently learning sparse (L_1 regularized) scoring functions for each task;
3. L_* - learning multiple scoring functions by imposing low rank regularization on their joint parameter matrix (L_* regularized objective (6));
4. $L_{2,1}$ - joint objective (6), regularized by mixed $\|.\|_{2,1}$ norm.

Our MultiRBSFL approach, which uses composite low rank and mixed norm regularized joint objective (8), we will denote as $L_* + L_{2,1}$ for consistency in naming the alternative approaches.

We measured the predictive performance in terms of accuracy, which is the number of correctly ordered test pairs. As the pairwise ranking relation is antisymmetric, it is sufficient to use only the positive training instances (i.e. where the first sample in a pair has the larger score). Test pairs are exclusively generated from examples not contained in the training set. Accuracy values that we report in this study are obtained by doing 5-fold cross-validation experiments.

5.1 Experiments on Synthetic Data

In this settings, a Gaussian processes model with an exponential kernel was used to generate the temporal data. We compiled 250 processes to mimic $d = 250$ measured variables (features) per subject. Each single process was used to generate a time series with 10 time points (10 samples). We followed the same principle to generate 10 different multivariate time series (subjects) for training and 10 subjects for test, resulting in 100 samples $X^{train}_{100\times250}$ for training, and 100 samples $X^{test}_{100\times250}$ for test.

Four different tasks were created by randomly generating the weight matrix $W_{250\times4}$, with only 5 nonzero rows, which corresponds to the $L_{2,1}$ assumption (row-sparsity). This row-wise sparse matrix was then superimposed with a dense rank-1 matrix, generated by multiplication of two random vectors, which suits the L_* trace norm part of the objective. True underlying scores on four tasks, for each of the 250-dimensional samples (one time point of one patient), are calculated as the weighted sum of the feature values $X * W$. Zero mean random vector was subsequently superimposed to input X data to model the measurement noise.

A training set is then obtained by making pairs out of samples whose scores are sufficiently different (in our case we set the threshold to 1). Pairs of examples were generated independently for each task based on their scores, totaling 14,187 pairs for all four tasks jointly. Test set pairs were generated in the same fashion, but with a smaller threshold and consisted out of 19,390 pairs. Training pairs were used to learn the weight matrix $\hat{W}$, which was used to estimate the testing scores from the test samples. The obtained estimates were used to infer the relative order of the testing pairs. The accuracy (percentage of correct guesses) is reported in the Table 1. It is no surprise that the proposed $L_* + L_{1,2}$ approach achieves the highest accuracy on all four tasks, as the underlying assumptions were explicitly built into the synthetic example.

5.2 School Exam Score

Intelligence as well as the capacity for understanding and using mathematics or languages are all examples of properties that are latent - yet important and often evaluated (estimated). We have tested the multitask score learning framework on data from an elementary school study [19], which contains longitudinal data on performance in Math and English language for pupils in 50 inner London schools [5]. In total there are scores

[5] http://www.bristol.ac.uk/cmm/media/migrated/jsp.zip

Table 1: Comparison of accuracy indicators (fraction of correctly ordered pairs) for alternative score learning methods on the synthetic data of four related tasks.

Task	L_2	L_1	L_*	$L_{1,2}$	$L_* + L_{1,2}$
TASK1	0.538	0.745	0.680	0.744	**0.757**
TASK2	0.556	0.707	0.763	0.782	**0.795**
TASK3	0.592	0.765	0.744	0.821	**0.837**
TASK4	0.466	0.864	0.700	0.874	**0.885**
AVG	0.538	0.770	0.722	0.805	**0.818**

for 3,236 exams (Math and English each), taken by 1,402 students over three consecutive school years. The goal is to rank the students' performances on Math and English test based on known score from Ravens ability test and additional information like demographics, social status, gender, class and school type. Distributions of scores for two tasks are given in the Figure 2.

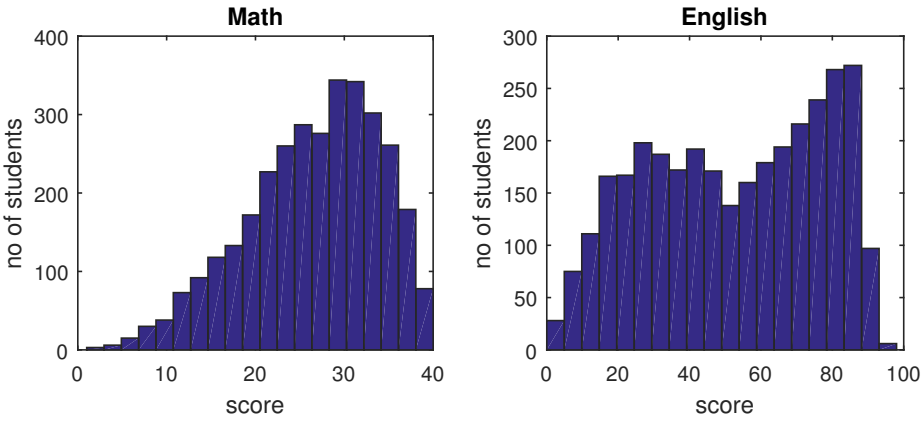

Fig. 2: Distributions of test scores for Math and English tasks, respectively.

According to results depicted in Table 2, our $L_* + L_{1,2}$ approach achieved the best predictive performance in both tasks.

5.3 Tolerance to Infections Score

Tolerance is the host's behavior that arises from interactions with a pathogen, which describes the ability of the host to preserve fitness despite the presence of a large amount of pathogen. Therefore, it is defined as changes in host fitness (health) with respect to changes in pathogen load [22]. However, tolerance is a very understudied topic, where there is no established scoring function, despite the necessity.

Table 2: Comparison of accuracy indicators (fraction of correctly ordered pairs) for alternative score learning methods on the task of learning the performance on Math and English tests.

task	L_2	L_1	L_*	$L_{1,2}$	$L_* + L_{1,2}$
MATH	0.780	0.794	0.725	0.789	**0.812**
ENGLISH	0.820	0.863	0.717	0.857	**0.870**
AVG	0.800	0.828	0.721	0.823	**0.841**

We analyzed three publicly available datasets [6] that allows characterization of the tolerance behavior in humans. The data comes from the human viral challenge studies [29] where human volunteers were infected with H3N2 influenza, rhinovirus (HRV) and respiratory syncytial virus (RSV), respectively. For all subjects in each dataset, symptoms were recorded twice a day and quantified by the modified Jackson Score [15]. Thereafter, subjects were classified based on the modified Jackson Score values into "symptomatic" and "asymptomatic" groups. In addition, viral load temporal measurements are available for 28 "symptomatic" subjects, given in Table 3. Gene expression measurements (for 12,023 genes) were collected temporally, starting at a baseline (24 hours prior to inoculation with virus) and measured at certain time points following the experimental procedure described in detail in [29], making a total of 16, 14 and 21 time-point measurements for H3N2, HRV and RSV datasets, respectively. Table 3 shows the viral shedding and symptom scores for subjects who developed clinically relevant symptoms from H3N2, HRV and RSV datasets.

Temporal measurements about symptoms (proxy for fitness) and viral (pathogen) load for each subject were used to derive tolerance scores according to the definition given in [22]. In particular, the tolerance score for each subject was calculated by dividing the maximum viral load with the maximum severity of symptoms observed for that subject (Table 3). Gene expression measurements were used as an explanatory variables in our ranking task.

Biological rationale behind the task relatedness is that the three infections are viruses that cause similar respiratory symptoms (runny nose, fever, cough) and are quantified by the same Jackson score, suggesting that some shared genetic mechanisms might be responsible for the disease manifestations. Consequently, we sought to learn the tolerance scoring functions jointly.

The tolerance scores were used to compile a set of ranked pairs, and the objective was to learn the scoring functions for tolerance to H3N2, HRV and RSV viruses (3 tasks), from high-dimensional gene expression data. Since 12,023 dimensions is very computationally expensive to optimize, we reduced the dimensionality of the data to the 100 most informative genes according to the correlation with the target. The results of learning the scoring functions with different approaches are summarized in the Table 4.

The results from the Table 4 show that the HRV task is the most difficult one in the described formulation. Although some alternative approaches achieved better accuracy

[6] http://people.ee.duke.edu/ lcarin/reproduce.html

Table 3: Tolerance scores (Ratio) derived by dividing maximum viral load (Max V) with maximum severity score (Max S).

H3N2				HRV				RSV			
Sub ID	Max S	Max V	Ratio	Sub ID	Max S	Max V	Ratio	Sub ID	Max S	Max V	Ratio
FLU05	12.00	5.45	0.45	HRV06	8.00	2.72	0.34	RSV01	11.00	0.00	0.00
FLU08	10.00	4.70	0.47	HRV19	2.00	0.95	0.47	RSV20	6.00	0.00	0.00
FLU01	9.00	4.25	0.47	HRV04	8.00	3.94	0.49	RSV07	20.00	4.46	0.22
FLU07	12.00	6.25	0.52	HRV15	7.00	3.45	0.49	RSV02	20.00	5.10	0.26
FLU06	7.00	5.00	0.71	HRV07	7.00	4.44	0.63	RSV12	4.00	2.50	0.62
FLU10	5.00	3.75	0.75	HRV20	6.00	4.44	0.74	RSV06	9.00	5.65	0.63
FLU12	4.00	5.00	1.25	HRV16	6.00	4.69	0.78	RSV14	6.00	4.54	0.76
FLU15	2.00	4.50	2.27	HRV09	3.00	2.46	0.82	RSV11	5.00	3.85	0.77
FLU13	2.00	5.45	2.70	HRV11	3.00	2.47	0.83	RSV03	6.00	4.70	0.78
				HRV03	4.00	3.45	0.86				

Table 4: Comparison of accuracy indicators (fraction of correctly ordered pairs) for alternative score learning methods on the tolerance to three viruses learning task.

task	L_2	L_1	L_*	$L_{1,2}$	$L_* + L_{1,2}$
FLU	0.766	0.980	0.809	0.988	**0.996**
HRV	0.344	0.122	0.389	**0.500**	0.400
RSV	0.806	**0.972**	0.861	0.306	0.861
AVG	0.638	0.692	0.686	0.598	**0.752**

in two of the tasks, the proposed approach achieved the best generalization trade-off as can be concluded from the highest average (overall) accuracy.

6 Discussion and Conclusions

We proposed the method that jointly learns multiple scoring functions from a set of ranked examples. The approach utilizes composite regularization consisting of the trace norm and row-wise grouped Lasso penalty, to impose the structural regularity among the model parameters of different tasks. We also provide optimization algorithm, based on the alternate minimization and proximal gradient techniques, for solving the proposed convex MultiRBSFL objective.

Presented empirical evaluations in one synthetic and two real world datasets suggest the benefits of utilizing the multitask approach for learning related ranking based scoring functions. According to the results, the model with only L_* performs worse than $L_{1,2}$, probably because sparsity in features seems to be the more dominant pattern in the data than the low-rank component. However, utilizing both L_* and $L_{1,2}$ in the same model turned out to be most beneficial for studied applications.

The proposed proximal gradient algorithm with alternating minimization for optimization of the multitask objective proved valuable for applications with low to moderate dimensionality of the feature space. However, as the contemporary applications have ever

increasing number of measured variables, more efficient optimization approaches and with better scalability would be required. One potential way to accelerate the proximal gradient algorithm is to adopt the approach proposed in [26].

Acknowledgments This research was supported in part by DARPA grant W911NF-16-C-0050 and in part by DARPA grant No. FA9550-12-1-0406 negotiated by AFOSR. Computations were performed on the OwlsNest HPC cluster at Temple University, which is supported in part by the National Science Foundation through NSF grant NSF-CNS-1625061 and Pennsylvania Department of Health CURE grant.

References

1. Anderson, R.: The credit scoring toolkit: theory and practice for retail credit risk management and decision automation. Oxford University Press (2007)
2. Ando, R.K., Zhang, T.: A framework for learning predictive structures from multiple tasks and unlabeled data. Journal of Machine Learning Research 6(Nov), 1817–1853 (2005)
3. Argyriou, A., Evgeniou, T., Pontil, M.: Convex multi-task feature learning. Machine Learning 73(3), 243–272 (2008)
4. Ashtawy, H.M., Mahapatra, N.R.: Machine-learning scoring functions for identifying native poses of ligands docked to known and novel proteins. BMC bioinformatics 16(6), S3 (2015)
5. Bach, F., Jenatton, R., Mairal, J., Obozinski, G., et al.: Convex optimization with sparsity-inducing norms. Optimization for Machine Learning 5 (2011)
6. Bai, J., Zhou, K., Xue, G., Zha, H., Sun, G., Tseng, B., Zheng, Z., Chang, Y.: Multi-task learning for learning to rank in web search. In: Proceedings of the 18th ACM conference on Information and knowledge management. pp. 1549–1552. ACM (2009)
7. Beck, A., Teboulle, M.: Gradient-based algorithms with applications to signal recovery. Convex optimization in signal processing and communications pp. 42–88 (2009)
8. Cai, J.F., Candès, E.J., Shen, Z.: A singular value thresholding algorithm for matrix completion. SIAM Journal on Optimization 20(4), 1956–1982 (2010)
9. Chen, J., Tang, L., Liu, J., Ye, J.: A convex formulation for learning shared structures from multiple tasks. In: Proceedings of the 26th Annual International Conference on Machine Learning. pp. 137–144. ACM (2009)
10. Chen, J., Zhou, J., Ye, J.: Integrating low-rank and group-sparse structures for robust multi-task learning. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 42–50. ACM (2011)
11. Dyagilev, K., Saria, S.: Learning (predictive) risk scores in the presence of censoring due to interventions. Machine Learning pp. 1–26 (2015)
12. Dyagilev, K., Saria, S.: Learning severity score for sepsis: a novel approach based on clinical comparisons. In: AMIA Annual Symposium Proceedings. pp. 1890–1898 (2015)
13. Evgeniou, T., Micchelli, C.A., Pontil, M.: Learning multiple tasks with kernel methods. Journal of Machine Learning Research 6(Apr), 615–637 (2005)
14. Gong, P., Ye, J., Zhang, C.: Robust multi-task feature learning. In: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 895–903. ACM (2012)
15. Jackson, G.G., Dowling, H.F., Spiesman, I.G., Boand, A.V.: Transmission of the common cold to volunteers under controlled conditions: I. the common cold as a clinical entity. AMA archives of internal medicine 101(2), 267–278 (1958)

16. Jacob, L., Vert, J.p., Bach, F.R.: Clustered multi-task learning: A convex formulation. In: Advances in neural information processing systems. pp. 745–752 (2009)
17. Jalali, A., Sanghavi, S., Ruan, C., Ravikumar, P.K.: A dirty model for multi-task learning. In: Advances in Neural Information Processing Systems. pp. 964–972 (2010)
18. Joachims, T.: Optimizing search engines using clickthrough data. In: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 133–142. ACM (2002)
19. Mortimore, P., Sammons, P., Stoll, L., Lewis, D., Ecob, R.: School matters: The junior years. Open Books (1988)
20. Parikh, N., Boyd, S.: Proximal algorithms. Foundations and Trends in optimization 1(3), 127–239 (2014)
21. Santolino, M., Boucher, J.P.: Modelling the disability severity score in motor insurance claims: an application to the spanish case. IREA–Working Papers, 2009, IR09/002 (2009)
22. Simms, E.L.: Defining tolerance as a norm of reaction. Evolutionary Ecology 14(4-6), 563–570 (2000)
23. Stojkovic, I., Jelisavcic, V., Milutinovic, V., Obradovic, Z.: Distance based modeling of interactions in structured regression. In: Procedeengs of the 25th International Joint Conference on Artificial Intelligence IJCAI-16. pp. 2032–2038 (2016)
24. Stojkovic, I., Jelisavcic, V., Milutinovic, V., Obradovic, Z.: Fast sparse gaussian markov random fields learning based on cholesky factorization. In: Procedeengs of the 26th International Joint Conference on Artificial Intelligence IJCAI-17 (2017)
25. Stojkovic, I., Obradovic, Z.: Predicting sepsis biomarker progression under therapy. In: Procedeengs of the 30th IEEE International Symposium on Computer-Based Medical Systems IEEE CBMS-17 (2017)
26. Toh, K.C., Yun, S.: An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems. Pacific Journal of optimization 6(615-640), 15 (2010)
27. Vincent, J.L., Moreno, R., Takala, J., Willatts, S., De Mendonça, A., Bruining, H., Reinhart, C., Suter, P., Thijs, L.: The sofa (sepsis-related organ failure assessment) score to describe organ dysfunction/failure. Intensive care medicine 22(7), 707–710 (1996)
28. Yang, S., Shapiro, L., Cunningham, M., Speltz, M., Birgfeld, C., Atmosukarto, I., Lee, S.I.: Skull retrieval for craniosynostosis using sparse logistic regression models. In: Medical Content-Based Retrieval for Clinical Decision Support, pp. 33–44. Springer (2012)
29. Zaas, A.K., Chen, M., Varkey, J., Veldman, T., Hero, A.O., Lucas, J., Huang, Y., Turner, R., Gilbert, A., Lambkin-Williams, R., et al.: Gene expression signatures diagnose influenza and other symptomatic respiratory viral infections in humans. Cell host & microbe 6(3), 207–217 (2009)
30. Zhou, J., Chen, J., Ye, J.: Malsar: Multi-task learning via structural regularization. Arizona State University 21 (2011)
31. Zhou, J., Liu, J., Narayan, V.A., Ye, J.: Modeling disease progression via fused sparse group lasso. In: Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 1095–1103. ACM (2012)