Multi-task Survival Analysis

Lu Wang* †, Yan Li* ‡, Jiayu Zhou¶, Dongxiao Zhu† and Jieping Ye‡§
†Dept. of Computer Science, Wayne State University, Detroit, MI - 48202. Email: {lu.wang3, dzhu}@wayne.edu
‡Dept. of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI -48109.
Email: {yanliwl, jpye}@umich.edu
§Dept. of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI -48109.
¶Dept. of Computer Science, Michigan State University, East Lansing, MI - 48824. Email: jiayuz@msu.edu

***Abstract*—Collecting labeling information of time-to-event analysis is naturally very time consuming, i.e., one has to wait for the occurrence of the event of interest, which may not always be observed for every instance. By taking advantage of censored instances, survival analysis methods internally consider more samples than standard regression methods, which partially alleviates this data insufficiency problem. Whereas most existing survival analysis models merely focus on a single survival prediction task, when there are multiple related survival prediction tasks, we may benefit from the tasks relatedness. Simultaneously learning multiple related tasks, multi-task learning (MTL) provides a paradigm to alleviate data insufficiency by bridging data from all tasks and improves generalization performance of all tasks involved. Even though MTL has been extensively studied, there is no existing work investigating MTL for survival analysis. In this paper, we propose a novel multi-task survival analysis framework that takes advantage of both censored instances and task relatedness. Specifically, based on two common used task relatedness assumptions, i.e., low-rank assumption and cluster structure assumption, we formulate two concrete models, *COX-TRACE* and *COX-cCMTL*, under the proposed framework, respectively. We develop efficient algorithms and demonstrate the performance of the proposed multi-task survival analysis models on the The Cancer Genome Atlas (TCGA) dataset. Our results show that the proposed approaches can significantly improve the prediction performance in survival analysis and can also discover some inherent relationships among different cancer types.**

***Keywords*—Survival analysis; Multi-task learning; regularization; Cox model.**

I. Introduction

Accurately predicting the time to the event of interest is a critical and practical problem in many real-world applications. The event of interests can be various of things in different problem settings, e.g., patient death in healthcare [1], device failure in reliability engineering [2] and user clicking in customer behavior analysis [3], etc. One major challenge in this context is labeling sufficient number of training instances, which often incurs prohibitive cost of time, i.e., one has to wait for the occurrence of the event of interest and the latter may not always be observed for every instance. Survival analysis is an important branch of statistics which aims at solving the aforementioned time prediction problem. Survival analysis superiors to standard regression models because it not only takes the instances whose event of interests have been observed (uncensored instances) into account, but also considers the instances whose event of interests have not been observed (censored instances). Therefore, it leverages more information than standard models, which could alleviate data insufficiency and improve prediction performance.

In many real-world applications we often need to build multiple survival prediction models that are related. For example, predicting the time of occurrence of patients death in multiple cancer types, predicting the time of occurrence of defaults in multiple types of loans, and predicting the battery life of multiple types of electronic devices. These scenarios provide a chance to increase the sample size of time-to-event prediction from both internal, through handling censored instances, and external, learning multiple related survival tasks simultaneously. However, in the field of survival analysis, most of the prior works [4], [5], [6], [7], [8], [9] only focus on dealing with a single survival prediction task; they mainly concentrate on encoding the censored instances into the learning formulations but barely consider the task relatedness among multiple related survival prediction problems.

The concept of learning multiple related tasks in parallel was introduced in [10]. Over the past two decades multi-task learning (MTL) has been extensively studied to deal with classification, standard regression, and clustering problems. Learning multiple related tasks simultaneously could effectively increase the information for training each task and hence improve the prediction performance. Thus, MTL is especially beneficial when the training sample size is limited for each task. Such problems are especially prevalent in several domains such as healthcare and bioinformatics, where MTL has achieved significant success, e.g., predicting disease progression [11], HIV therapy screening [12], and genomic data analysis [13], etc. Survival analysis also plays an important role in healthcare analysis [6], [4], [1]. However, multi-task survival analysis has barely been studied so far, in spite of the clear practical needs.

The goal of this paper is to bridge these two active research fields of survival analysis and multi-task learning. In this paper, we propose a unified framework for multi-task survival analysis, where we employ the Cox proportional hazards model [4], one of the most popular survival analysis methods, to encode censored instances. The Cox model is a semi-parametric model such that the model coefficients can be learned without knowing (or assuming) the underlying distribution, and this property makes Cox model superior to parametric censored regressions [6] in most cases. The proposed multi-task survival analysis framework belongs to the regularized MTL approach [14], where the assumptions of task relatedness is encoded via regularization terms, and this approach has been extensively studied in the past decade [15], [16], [17], [18], [19], [14],

* These two authors contributed equally to this work

2374-8486/17 $31.00 © 2017 IEEE
DOI 10.1109/ICDM.2017.58

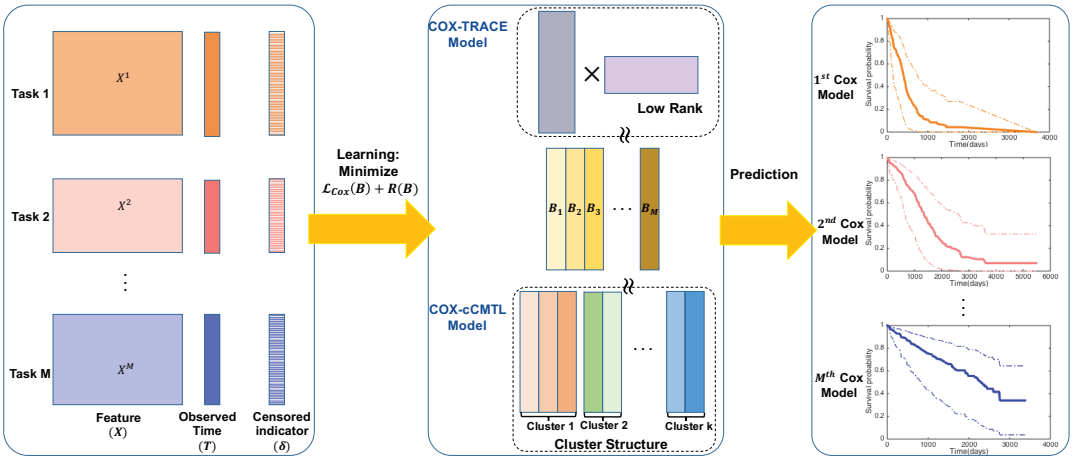


Fig. 1: Illustration of the proposed framework for multi-task survival analysis. We further develop two concrete models *COX-Trace* and *COX-cCMTL* based on this framework and illustrate the underlying assumptions. We note that the proposed framework can be used to develop more versatile multi-task survival models by introducing regularization terms to capture the task relatedness.

[13]. Fig. 1 illustrates the proposed frame work, and under this framework we study two concrete models according to different task relatedness assumptions: low-rank assumption and group structure assumption.

Based on the low-rank assumption, i.e., the coefficient vectors of tasks come from the same subspace, we employ the trace norm [17] and propose the *Cox-Trace* model. Based on the group structure assumption, i.e., all tasks can be clustered into different groups and the estimated coefficients of tasks from the same group are closer to each other than those from a different group [18], we propose *Cox-cCMTL* model. In this paper, to solve the proposed models, we employ the proximal gradient methods (PGM) [20] to achieve an effective learning process. In the experiment part of this paper, we demonstrate the prediction performance of the proposed multi-task survival analysis models using the well known The Cancer Genome Atlas (TCGA) dataset to predict the patient death from multiple cancer types.

The main contributions of this paper can be summarized as follows:

- Based on Cox proportional hazards model, we propose a unified framework for multi-task survival analysis, which extends the concept of multi-task learning to survival analysis.
- We proposed two novel concrete models under our framework to encode the task relatedness of multiple survival prediction problems under different assumptions.
- Our comprehensive empirical studies demonstrate the advantage of multi-task survival analysis over traditional single-task survival analysis and discover some inherent relationships among different cancer types.

The rest of this paper is organized as follows: Section II provides some relevant background regarding various MTL methods and regularized Cox regression models. The unified framework for multi-task survival analysis and two concrete multi-task survival analysis models is explained in details in Section III. In Section IV, the effectiveness of the MTL in survival analysis is demonstrated using the benchmark microarray gene expression datasets. Finally, Section V concludes our discussion and gives some future research directions for the proposed work.

II. Related Work

In this section, we present the related works in the areas of survival analysis and multi-task learning and discuss the relationships and primary distinctions of the proposed models compared to the existing methods that are available in the literature.

A. Survival Analysis

Survival analysis is the field of statistics, which aims at predicting the time to the event of interests. However, due to time limitation or some unexpected interruptions, the event of interests can not always be observed and this phenomenon is known as censoring [6]. The censored instances only provide the partially informative label information, which makes survival analysis more challenging compared to the standard regression. To deal with this problem, statistical approaches have been widely developed in the literature, and these approaches can be roughly categorized into three types: non-parametric, parametric, and semi-parametric.

The non-parametric methods, such as Kaplan-Meier (KM) method [21], provide a rough general description of the survival probability in a given group. This type of methods do not need any pre-assumption, but they ignore the individual differences within the group. The parametric methods [6] are more efficient and accurate for estimation when the time to event of interest follows a particular distribution; however, the model performance heavily relies on the choice of distribution, which makes the parametric methods impractical in most real-world scenarios. The semi-parametric methods, or more specifically, the Cox proportional hazards models [4] alleviate the weakness of aforementioned two types of models, to a considerable extent. This type of methods take individual differences into account and hence provide personalized prediction for each

individual; moreover, the parameter estimation does not require knowledge of the underlying distribution. Therefore, the Cox model is the most widely used model in survival analysis. Moreover, several useful variants of the basic Cox model have been extensively studied in the past two decades. For example, to deal with high-dimensional data and alleviate model over-fitting, some sparsity-inducing regularization have been integrated with the basic Cox model such as COX-LASSO [7] which employs the L_1 norm penalty, Elastic-Net Cox (COX-EN) [8] which uses the elastic net penalty term, and the group lasso penalized Cox regression [22]. In this paper, the proposed multi-task survival analysis approaches also belong to regularized Cox model, and in the experiment we will compare the proposed models with Cox model and corresponding related regularized Cox models to show the advantage of multi-task survival analysis.

Besides the statistical approaches, some machine learning based methods have been proposed in survival analysis. Recently, in [23] and [9] survival prediction problem has been viewed as a sequence of dependent classification tasks, and the tasks relatedness are encoded via some MTL approaches. Note that in these two papers the MTL approaches are used to solve a single survival prediction problem; however, in our paper we deal with multi-task survival analysis that learns multiple related survival analysis problems in parallel.

B. Multi-task learning

MTL is a machine learning paradigm that leverages relatedness among the tasks to improve the generalization performance of all machine learning models, by simultaneously learning all the related tasks and transferring knowledge among the tasks. The key building block of MTL algorithms is how task relatedness is assumed and encoded into the learning formulations. A conventional approach to achieve MTL is to couple the learning process by using multi-task regularizations [14]. The regularized MTL approach has a clear advantage over other MTL approaches, because it can leverage large-scale optimization algorithms such as proximal gradient techniques [20], [16], [17], [19], which can efficiently handle complicated constraints and/or non-smooth terms in the objective function.

In the past decade, there are many regularization terms designed to impose different assumptions about how the tasks are related. For example, multi-task feature learning [15], [16] assumes that all tasks share a subset of features and some group sparsity penalties are used to encode this assumption. Multi-task subspace learning [17] assumes the coefficient vectors of tasks come from the same subspace, which leads to a low-rank structure within coefficient matrix. Multi-task relationship learning assumes the task relatedness can be represented by some abstract structures such as cluster structure [18], [19], tree structure [24], and graph structure [14], [13]. In [25], the authors provide a comprehensive study and implementation of the commonly used multi-task regularizations. In addition to these commonly used assumptions, there are more regularized MTL formulations, which take domain specific knowledge into account [11], [26] and make regularized MTL more attractive.

In the all aforementioned methods and other related works (refer to [25]), the learning tasks are either classification or standard regression. However, in this paper, the learning tasks are the survival prediction problems. Recently, in [1] a transfer learning model, *Transfer-Cox*, has been proposed to enable knowledge transfer in survival analysis, which is a $L_{2,1}$-norm regularized Cox proportional hazards model. The *Transfer-Cox* can be viewed as a special case in multi-task survival analysis, i.e., it only has two tasks, source task and target task, and the model emphasizes more on the target task. In the experiment of our paper, we will generalize this model to equally learn arbitrary number of learning tasks simultaneously, and denote this model as *COX-*$L_{2,1}$ for the sake of naming convention.

III. Proposed Model

In this section, firstly, some basic concepts of survival analysis and Cox proportional hazards model are introduced. Then we will propose a unified framework and two concrete models to achieve the multi-task survival analysis.

A. Preliminaries

The primary goal in survival analysis is to model the relationship between the feature vector ($X_i \in \mathbb{R}^{1\times p}$) and its corresponding survival/failure time ($O_i \in \mathbb{R}^+$). The survival time can be observed from uncensored instances; however, as censoring happens, for a censored instance we can only obtain its last observed time, which is known as censored time ($C_i \in \mathbb{R}^+$). In practice, a censoring indicator ($\delta_i \in \{0, 1\}$) is introduced to incorporate these two types of instances in a same triplet format, (X_i, T_i, δ_i), where T_i is the *observed time*, which equals to O_i for uncensored instances ($\delta_i = 1$) and C_i for censored instances ($\delta_i = 0$). In addition, the most common form of censoring that occurs in real-world scenarios is *right censoring*, where the potential/unobserved survival time of a censored instance is longer than or equal to its corresponding censored time. In this paper, we deal with the survival prediction under the scenario of right censoring.

The *survival function* $S_i(t) = Pr(O_i \geq t)$ is an intuitive description of the survival prediction, which represents the probability that the survival time is not less than t. It can be easily found out that all of the survival functions have a same pattern, i.e., monotone decreasing and range from 1 to 0. Therefore, it is very difficult to model the slight difference among different survival functions. To overcome this drawback, the *hazards function*

$$h_i(t) = \lim_{\Delta t \to 0} \frac{Pr(t \leq O_i < t + \Delta t | O_i \geq t)}{\Delta t}$$

is introduced in survival analysis, which is the event rate at time t conditionally on survival until the time t or later. It is also a non-negative function, but it has a wider range of values and can take on a variety of shapes.

The hazard function for the Cox proportional hazards model has the form:

$$h(t, X_i) = h_0(t) \exp(X_i \beta), \text{ for } i = 1, 2, ..., N, \qquad (1)$$

where the $h_0(t)$ is the *baseline hazard function*, which can be an arbitrary non-negative function of time, and $\beta \in \mathbb{R}^{p\times 1}$ is the coefficient vector which needs to be learned in model training. The Cox proportional hazards model is named after the fact that in the model the hazard rate of any pair of instances

is a time-invariant constant number. Moreover, it is a semi-parametric model as all the instances share a same baseline hazard function, and the coefficient estimation is independent from the $h_0(t)$, which can be achieved via maximizing the partial likelihood [4]. In practice, to accommodate with tied failures, i.e., two or more failure events that occur at the same time, some methods such as the Efron's method [27] and the Breslow's method [28] have been proposed. In this paper, we employ the Breslow's method in our model, for N instances with a increasing list of unique failure times, $O_1 < O_2 < \cdots < O_q$, the partial likelihood is defined as follows:

$$L(\beta) = \prod_{i=1}^{q} \frac{\exp(\sum_{j \in D_i} X_j \beta)}{[\sum_{j \in R_i} \exp(X_j \beta)]^{d_i}}, \tag{2}$$

where R_i is the risk set at O_i, which consists of all instances whose observed times are equal to or greater than O_i, D_i contains all instances whose failure time is O_i and $d_i = |D_i|$ is the size of D_i. Therefore, the coefficient vector can be learned via minimizing the negative partial log-likelihood:

$$l(\beta) = -\sum_{i=1}^{q} \left\{ \sum_{j \in D_i} X_j \beta - d_i \log[\sum_{j \in R_i} \exp(X_j \beta)] \right\}. \tag{3}$$

B. A united framework for multi-task survival analysis

In data mining and machine learning, a common paradigm for MTL can be formulated as a regularized empirical loss:

$$\min_B \mathcal{L}(B) + R(B), \tag{4}$$

where $\mathcal{L}(B) = \sum_{m=1}^{M} \frac{1}{N_m} l(B_m)$; in addition, B_m, N_m, and $l(B_m)$ denote the parameters to be estimated, the number of training instances, and the empirical loss on the training set with respect to the m-th task, respectively. $R(B)$ is the regularization term that encodes task relatedness and $B = [B_1, B_2, \cdots, B_M] \in \mathbb{R}^{p \times M}$. In standard multi-task classification/regression problems, the logistic regression and least squares are commonly used empirical loss function. In survival analysis, Cox proportional hazards model is one of the most widely used prediction methods, and we employ its loss function (Eq. (3)) for all tasks. Therefore, the proposed united framework for multi-task survival analysis can be formulated as:

$$\min_B \sum_{m=1}^{M} \frac{-1}{N_m} \sum_{i=1}^{q_m} \left\{ \sum_{j \in D_i^m} X_j^m B_m - d_i^m \log \Big[\sum_{j \in R_i^m} \exp(X_j^m B_m) \Big] \right\} + R(B), \tag{5}$$

where X^m is the training dataset of m-th task, and q_m is the corresponding number of unique failure times. D_i^m, d_i^m, and R_i^m denote the index of set of failure instances, number of failure instances, and the risk set at i-th unique failure time of m-th task, respectively.

Different assumptions on task relatedness lead to different regularization terms. In the field of MTL, there are many prior works that model relationships among tasks using novel regularization terms, and most of them are non-smoothing. In this paper, we will provide several prediction models for multi-task survival analysis, based on two commonly used assumptions of task relatedness.

In this paper, we employ the proximal gradient methods (PGM) [20] as the workhorse to optimize the proposed learning problem and estimate the model coefficients. This type of methods only require $\mathcal{O}(\frac{1}{\sqrt{\varepsilon}})$ iterations to achieve an accuracy of ε, which is the optimal among first order methods. The Algorithm 1 in the Appendix outlines the learning procedure of PGM, and the key building block of PGM is to compute the proximal operator, which is a regularized Euclidean projection problem (Line 6 in Algorithm 1). In the subsequent sections, for each proposed multi-task survival analysis model we will provide an analytical solution for the proximal operator with corresponding regularization term.

C. Trace-norm regularized multi-task survival analysis: COX-TRACE model

The low rank assumption is a commonly and widely used constraint in MTL, which assumes the estimated coefficients from different tasks sharing a low-dimensional subspace. Intuitively, this assumption results in the following rank minimization:

$$\min_B \mathcal{L}(B) + \lambda \, \text{rank}(B), \tag{6}$$

which is a NP-hard problem, and λ is a positive scale. In practice, the trace norm (or nuclear norm) is a commonly-used convex relaxation of the rank function, which is defined as the sum of the singular values: $\| B \|_* = \sum_i \sigma_i(B)$. Therefore, the proposed trace-norm regularized multi-task survival analysis model, *COX-TRACE*, can be formulated as:

$$\min_B \sum_{m=1}^{M} \frac{-1}{N_m} \sum_{i=1}^{q_m} \left\{ \sum_{j \in D_i^m} X_j^m B_m - d_i^m \log \Big[\sum_{j \in R_i^m} \exp(X_j^m B_m) \Big] \right\} + \lambda \| B \|_* . \tag{7}$$

Optimization:

The trace norm regularization has been studied extensively in MTL, and proximal gradient based optimization method is proposed in [17]. The key subroutine of proximal gradient methods is to compute the proximal operator:

$$\hat{B} = \arg\min_B \frac{1}{2} \| B - G \|_F^2 + \lambda \| B \|_*, \tag{8}$$

where G is known as gradient step: $G = S - \frac{1}{\gamma} \Delta \mathcal{L}(S)$. In addition, γ is the step size, S is the current search point that is a combination of previous points, i.e., in the i-th iteration it is defined as:

$$S^{(i)} = B^{(i)} + \alpha_i (B^{(i)} - B^{(i-1)}), \tag{9}$$

and α_i is the combination scalar. $\Delta \mathcal{L}(S)$ is the gradient of empirical loss at search point S, specifically, for all M tasks we have:

$$\Delta \mathcal{L}(S) = \left[\frac{l'(S_1)}{N_1}, \frac{l'(S_2)}{N_2}, \cdots, \frac{l'(S_M)}{N_M} \right]. \tag{10}$$

In the Cox proportional hazards model based multi-task survival analysis, for all tasks the derivative of negative partial log-likelihood function share the same formulation:

$$l'(\beta) = -\sum_{i=1}^{q} \left\{ \sum_{j \in D_i} X_j - d_i \frac{\sum_{j \in R_i} X_j \exp(X_j \beta)}{\sum_{j \in R_i} \exp(X_j \beta)} \right\}. \tag{11}$$

For each task, to calculate the derivative we just need to plug into the corresponding training samples and search point. In [17], an analytical solution of the corresponding proximal operator has been proposed and can be summarized in the following theorem.

Theorem 1: Given the gradient step $G \in \mathbb{R}^{p\times M}$ defined in Eq.(8), and let $G = U\Sigma V^T$ be its singular value decomposition (SVD), where $U \in \mathbb{R}^{p\times r}$ and $V \in \mathbb{R}^{M\times r}$ have orthonormal columns, $\Sigma = diag(\sigma_1, \cdots, \sigma_r) \in \mathbb{R}^{r\times r}$, and $r = \text{rank}(G)$. Then the optimal solution of the proximal operator in Eq.(8) is given by $\hat{B} = U\Sigma^{(\lambda)}V^T$, where $\Sigma^{(\lambda)} = diag(\sigma_1^{(\lambda)}, \cdots, \sigma_r^{(\lambda)})$ and $\sigma_i^{(\lambda)} = \max\{0, \sigma_i - \lambda\}$.

D. Clustered multi-task survival analysis: COX-cCMTL model

Many MTL algorithms assume that all learning tasks are related. In practical applications, the tasks may exhibit a more sophisticated group structure, where the estimated coefficients of tasks from the same group are closer to each other than those from a different group. This type of approaches are known as clustered multi-task learning (CMTL), and intuitively we can employ the sum-of-square error (SSE) function in K-means clustering as the regularization term to encode the assumption of clustering structure among multiple learning tasks.

Suppose all M tasks can be clustered into $K < M$ clusters, and the index set of the k-th cluster is defined as $\mathcal{I}_k = \{v | v \in \text{cluster}k\}$. Let $\bar{B}_k = \frac{1}{n_k}\sum_{v\in\mathcal{I}_k} B_v$ be the mean of the coefficient vectors of the k-th cluster, the SSE [29] can be formulated as:

$$\sum_{k=1}^{K}\sum_{v\in\mathcal{I}_k} \parallel B_v - \bar{B}_k \parallel_2^2 = \text{tr}(BB^T) - \text{tr}(BFF^TB^T), \quad (12)$$

where $\text{tr}(\cdot)$ represents the trace of matrix and $F \in \mathbb{R}^{M\times K}$ is an orthogonal cluster indicator matrix:

$$F_{m,k} = \begin{cases} \frac{1}{\sqrt{n_k}} & \text{if} \quad m \in \mathcal{I}_k \\ 0 & \text{if} \quad m \notin \mathcal{I}_k \end{cases} \quad (13)$$

However, the SSE in Eq.(12) is not easy to minimize as it is not-convex due to F has the aforementioned special structure. To deal with these issues, a spectral relaxation [29] has been proposed, which ignores the special structure of F but keeps the orthogonality requirement only, i.e., $F^TF = I_K$; moreover, a convex relaxation [18] has been proposed, which relaxes the feasible domain of FF^T into a convex set $\mathcal{W} = \{W | \text{tr}(W) = M, W \preceq I, W \in \mathbb{S}_+^M\}$ and approximate FF^T via W. In summary, these two aforementioned relaxation results in the following convex relaxed CMTL:

$$\begin{aligned} &\min_{B,W} \mathcal{L}(B) + \rho_1[\text{tr}(BB^T) - \text{tr}(BWB^T)] + \rho_2\text{tr}(BB^T), \\ &\text{s. t. } \text{tr}(W) = K, W \preceq I, W \in \mathbb{S}_+^M \end{aligned} \quad (14)$$

where $\text{tr}(BB^T) = \parallel B \parallel_F^2$, the square of Frobenius norm of B, which is used to shrink the coefficients and alleviate multicollinearity. Let $\eta = \frac{\rho_2}{\rho_1} > 0$ and through some simple algebra calculations we can finally formulate the convex relaxed clustered multi-task survival analysis model, *COX-cCMTL*, as:

$$\begin{aligned} \min_{B,W} &\sum_{m=1}^{M}\frac{-1}{N_m}\sum_{i=1}^{q_m}\left\{\sum_{j\in D_i^m} X_j^m B_m - d_i^m \log\Big[\sum_{j\in R_i^m}\exp(X_j^m B_m)\Big]\right\} \\ &+ \rho_1\eta(1+\eta)\text{tr}(B(\eta I + W)^{-1}B^T). \\ &\text{s. t. } \text{tr}(W) = K, W \preceq I, W \in \mathbb{S}_+^M \end{aligned} \quad (15)$$

Optimization:

The model proposed in Eq.(15) is jointly convex with respect to B and W. Moreover, it is an convex unconstrained smooth optimization problem with respect to B, and its global optimal can be achieved via iteratively updating its gradient step:

$$G_B = S - \frac{1}{\gamma}[\Delta\mathcal{L}(S) + 2\rho_1\eta(1+\eta)(\eta I + W_S)^{-1}S^T],$$

where S is the search point of B that is defined in Eq.(9), W_S is the search point of W and in the i-th iteration it can be similarly calculated as $W_S^{(i)} = W^{(i)} + \alpha_i(W^{(i)} - W^{(i-1)})$, and $\Delta\mathcal{L}(S)$ is the gradient of loss function as shown in Eq.(10).

The optimization of W is a convex constrained minimization problem, and its corresponding proximal operator is formulated as:

$$\min_W \|W - G_W\|_F^2, \quad \text{s.t.} \quad \text{tr}(W) = K, W \preceq I, W \in \mathbb{S}_+^M, \quad (16)$$

where G_W is the gradient step of W at the search point W_S and can be calculated as:

$$G_W = W_S + \frac{\rho_1\eta(1+\eta)}{\gamma}S^TS(\eta I + W_S)^{-2}. \quad (17)$$

An analytical solution of Eq.(16) has been proposed [19], which can be summarized in the following theorem.

Theorem 2: Given the gradient step $G_W \in \mathbb{S}^{M\times M}$, and let $G_M = V\hat{\Sigma}V^T$ be its eigen-decomposition, where $V \in \mathbb{R}^{M\times M}$ is orthonormal, and $\hat{\Sigma} = diag(\hat{\sigma}_1, \cdots, \hat{\sigma}_M) \in \mathbb{R}^{M\times M}$. Let $\Sigma^* = diag(\sigma_1^*, \cdots, \sigma_M^*) \in \mathbb{R}^{M\times M}$, where $\{\sigma_1^*, \cdots, \sigma_M^*\}$ is the optimal solution to the following optimization problem:

$$\begin{aligned} &\min_{\{\sigma_m\}} \sum_{m=1}^{M}(\sigma_m - \hat{\sigma}_m)^2. \\ &\text{s. t.} \quad \sum_{m=1}^{M}\sigma_m = K, \ 0 \le \sigma_m \le 1, \ \forall m = 1, \cdots, M \end{aligned} \quad (18)$$

Then the optimal solution of the proximal operator in Eq.(16) is given by $\hat{M} = V\Sigma^*V^T$.

IV. Experimental Result

In this section, we will first describe the dataset used in our experiment and demonstrate the prediction performance of the proposed multi-task survival analysis models.

TABLE I: Basic statistics of the 21 selected cancer types.

Cancer name	Primary Site	Acronym	# Instances	# Uncensored
Adrenocortical Carcinoma	Adrenal Gland	ACC	80	29
Bladder Urothelial Carcinoma	Bladder	BLCA	407	178
Breast Invasive Carcinoma	Breast	BRCA	754	105
Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma	Cervix	CESC	307	72
Cholangiocarcinoma	Bile Duct	CHOL	36	18
Esophageal Carcinoma	Esophagus	ESCA	184	77
Head and Neck Squamous Cell Carcinoma	Head and Neck	HNSC	484	203
Kidney Renal Clear Cell Carcinoma	Kidney	KIRC	254	76
Kidney Renal Papillary Cell Carcinoma	Kidney	KIRP	290	44
Brain Lower Grade Glioma	Brain	LGG	510	124
Liver Hepatocellular Carcinoma	Liver	LIHC	371	128
Lung Adenocarcinoma	Lung	LUAD	441	157
Lung Squamous Cell Carcinoma	Lung	LUSC	338	137
Mesothelioma	Pleura	MESO	86	73
Prostate Adenocarcinoma	Prostate	PRAD	178	93
Sarcoma	Soft Tissue	SARC	259	98
Skin Cutaneous Melanoma	Skin	SKCM	97	26
Stomach Adenocarcinoma	Stomach	STAD	382	147
Uterine Corpus Endometrial Carcinoma	Uterus	UCEC	410	72
Uterine Carcinosarcoma	Uterus	UCS	56	34
Uveal Melanoma	Eye	UVM	80	23

A. Experimental Dataset

To evaluate the models and demonstrate the effectiveness of multi-task survival analysis, we used The Cancer Genome Atlas (TCGA) dataset in our experiment. TCGA is one of the most well-known cancer genome programs, which is supported by the National Cancer Institute's Genomic Data Commons. It includes both molecular profiles and clinical data for 33 types of tumors profiled with different high-throughput platforms. In our experiment, we focus on analyzing the relationship between micro-RNAs (miRNA) and the survival time of cancer patients, where the miRNA functions in RNA silencing and post-transcriptional regulation of gene expression. We use the R package "TCGA2STAT" [30] to query and download TCGA data directly into a unified data repository; more specifically, we download clinical data and the raw counts of miRNA sequencing via the *getTCGA* function.

In TCGA, the miRNA expression of 29 out of 33 tumor types was profiled using Illumina HiSeq 2000 miRNA sequencing and each miRNA sequence has a length of 1046. Among these 29 tumor types, the number of uncensored instances in 8 tumor types are too small (< 18) and hence these 8 tumor types were eliminated for our evaluation. Therefore, finally we get 21 tumor types which are shown in Table I with their basic statistics. In the table, the columns titled "# Uncensored" correspond to the number of uncensored instances in each selected cancer type, respectively. For these tumor types, the event of interest is patient death; therefore, an uncensored instance refers to the patient being dead during the study, while a censored instance refers to the corresponding patient is still alive at the last observed time (which will be the censored time).

B. Performance Comparison

We compare our proposed multi-task learning survival analysis models with several related single-task learning (STL) survival analysis models and the state-of-the-art multi-task learning survival analysis models. Since our proposed models are Cox-based models, we choose both Cox proportional hazards model and two other popular regularized Cox models: COX-LASSO and COX-EN as single-task learning comparison methods. In our experiments, these three survival analysis methods are applied under two settings: 1) Individual setting, i.e., a prediction model is trained for each tumor type; 2) Global setting, i.e., a prediction model is trained for all tumor types. In the individual setting the heterogeneity among tasks are fully considered but the task relatedness are totally ignored; on the contrary, in the global setting all heterogeneities have been ignored. In our experiment, the Cox model is trained by using the *coxph* function in the *survival* package [31], and the other two STL regularized Cox models are trained by using the *cocktail* function in the *fastcox* package [32]. We implement all the multi-task learning survival models, including both the proposed two models (*COX-TRACE* and *COX-cCMTL*) and the *COX-*$L_{2,1}$ model, via Matlab and the source code can be download at the following address [1].

[1]https://github.com/yanlirock/Multi-task_Survival_Analysis

TABLE II: Performance comparison of the multi-task survival analysis models and related single-task survival analysis models using C-index values (along with their standard deviations).

Tumor Type	Individual setting			Global setting			Multi-task survival analysis models		
	COX	**COX-LASSO**	**COX-EN**	**COX**	**COX-LASSO**	**COX-EN**	**COX-$L_{2,1}$**	**COX-TRACE**	**COX-cCMTL**
ACC	0.6912 (0.0301)	0.6737 (0.0324)	0.7186 (0.0326)	0.7116 (0.0307)	0.7662 (0.0565)	0.7679 (0.0366)	0.8128 (0.0761)	**0.8154 (0.0319)**	0.8008 (0.0328)
BLCA	0.5610 (0.4971)	0.5244 (0.4904)	0.5385 (0.4982)	0.5142 (0.0396)	0.5473 (0.0317)	0.5429 (0.0399)	0.6048 (0.0178)	0.6129 (0.0067)	**0.6206 (0.0132)**
BRCA	0.5478 (0.0351)	0.5881 (0.0502)	0.5641 (0.0418)	0.5456 (0.0634)	0.5383 (0.0599)	0.5475 (0.0702)	**0.6265 (0.0422)**	0.6024 (0.0295)	0.6232 (0.0380)
CESC	0.5715 (0.5784)	0.5756 (0.5495)	0.5539 (0.5615)	0.6019 (0.0833)	0.6403 (0.0141)	0.6274 (0.0425)	**0.6791 (0.0686)**	0.5915 (0.0321)	0.6299 (0.0663)
CHOL	0.4798 (0.0314)	0.5517 (0.0254)	0.5428 (0.0127)	0.5169 (0.0651)	0.5118 (0.1211)	0.5008 (0.0732)	**0.6453 (0.1824)**	0.5565 (0.1075)	0.5615 (0.1037)
ESCA	0.5600 (0.6938)	0.5176 (0.5908)	0.5166 (0.6510)	0.5382 (0.0857)	0.5365 (0.0876)	0.5640 (0.0694)	0.5935 (0.0674)	0.5770 (0.0723)	**0.5969 (0.0428)**
HNSC	0.5092 (0.0379)	0.5134 (0.0554)	0.5155 (0.0148)	0.5412 (0.0347)	0.5791 (0.0463)	0.5732 (0.0452)	**0.5839 (0.0159)**	0.5231 (0.0248)	0.5542 (0.0257)
KIRC	0.6006 (0.5283)	0.6294 (0.5625)	0.6083 (0.5671)	0.5308 (0.0175)	0.6069 (0.0151)	0.5899 (0.0013)	0.6704 (0.0608)	0.6913 (0.0348)	**0.7037 (0.0248)**
KIRP	0.7451 (0.0392)	0.7403 (0.0459)	0.7494 (0.0380)	0.6964 (0.0620)	0.7678 (0.0553)	0.7410 (0.0191)	0.8030 (0.0528)	0.7943 (0.0410)	**0.8042 (0.0539)**
LGG	0.6948 (0.5928)	0.6803 (0.5966)	0.6976 (0.5942)	0.6736 (0.0858)	0.7186 (0.0988)	0.7232 (0.1160)	0.7502 (0.0783)	0.7441 (0.0895)	**0.7661 (0.0934)**
LIHC	0.5341 (0.0234)	0.5517 (0.0355)	0.5454 (0.0257)	0.5477 (0.0472)	0.5903 (0.0469)	0.5800 (0.0513)	0.6496 (0.0438)	0.6055 (0.0181)	**0.6514 (0.0348)**
LUAD	0.4971 (0.5355)	0.4904 (0.5776)	0.4982 (0.5273)	0.5677 (0.0316)	0.6024 (0.0156)	**0.6081 (0.0415)**	0.5690 (0.0107)	0.5548 (0.0443)	0.5967 (0.0363)
LUSC	0.5784 (0.0423)	0.5495 (0.1503)	0.5615 (0.1214)	0.5434 (0.0627)	0.5340 (0.0380)	0.5547 (0.0533)	0.5714 (0.0195)	**0.6006 (0.0267)**	0.5998 (0.0406)
MESO	0.6938 (0.4625)	0.5908 (0.4986)	0.6510 (0.4855)	0.5882 (0.0556)	0.6646 (0.0529)	0.6534 (0.0246)	0.6793 (0.0438)	0.7091 (0.0751)	**0.7169 (0.0529)**
PAAD	0.5283 (0.0156)	0.5625 (0.0237)	0.5671 (0.0225)	0.5436 (0.0097)	0.5573 (0.0760)	0.5665 (0.0561)	**0.5796 (0.0638)**	0.5453 (0.0342)	0.5572 (0.0157)
SARC	0.5928 (0.6374)	0.5966 (0.5422)	0.5942 (0.5764)	0.5523 (0.0269)	0.5759 (0.0687)	0.5594 (0.0455)	0.6177 (0.0175)	0.6457 (0.0219)	**0.6573 (0.0244)**
SKCM	0.5355 (0.0162)	0.5776 (0.0227)	0.5273 (0.0266)	0.4918 (0.0960)	0.5726 (0.0614)	0.5621 (0.0987)	**0.6535 (0.0538)**	0.6160 (0.0387)	0.5960 (0.0380)
STAD	0.4625 (0.4859)	0.4986 (0.4468)	0.4855 (0.4492)	0.5431 (0.0375)	0.4852 (0.0383)	0.5144 (0.0356)	**0.5544 (0.0345)**	0.4850 (0.0257)	0.5237 (0.0337)
UCEC	0.6374 (0.0427)	0.5422 (0.0739)	0.5764 (0.0506)	0.4737 (0.0915)	0.5678 (0.0894)	0.5465 (0.0880)	0.6259 (0.0412)	0.6435 (0.0319)	**0.6554 (0.0126)**
UCS	0.4859 (0.7242)	0.4468 (0.6415)	0.4492 (0.7630)	0.4210 (0.0724)	0.3934 (0.0578)	0.4007 (0.0776)	**0.6764 (0.0489)**	0.4745 (0.0425)	0.5440 (0.1021)
UVM	0.7242 (0.0776)	0.6415 (0.0839)	0.7630 (0.0290)	0.5611 (0.1559)	0.5809 (0.0528)	0.5480 (0.1790)	0.8005 (0.0551)	0.8050 (0.0106)	**0.8176 (0.0109)**

The concordance index (C-index), or *concordance probability*, is a commonly used evaluation metric in survival analysis [33]. For a pair of bivariate observations $(T_1, \hat{T}_1)$ and $(T_2, \hat{T}_2)$, the concordance probability is defined as:

$$c = Pr(\hat{T}_1 > \hat{T}_2 | T_1 \geq T_2), \tag{19}$$

where T_i is the actual time, and $\hat{T}_i$ is the predicted one. In practice it can be calculated based on the proportion of corrected ordered comparable instance pairs among all comparable instance pairs. In the standard Cox and regularized Cox models, the hazard ratio is modeled to describe the time-to-event data. The instances with a low hazard rate should survive longer, so the C-index is calculated as follows:

$$c = \frac{1}{num} \sum_{i \in \{1 \cdots N | \delta_i = 1\}} \sum_{T_j > T_i} I[X_i \hat{\beta} > X_j \hat{\beta}], \tag{20}$$

where num denotes the number of comparable instance pairs and $I[\cdot]$ is the indicator function. We can observe that the C-index computation requires a certain number of comparable instance pairs, so the testing data should contain enough samples; therefore, we use 3-folds cross validation for model evaluation as the sample size of some tumor types are very small.

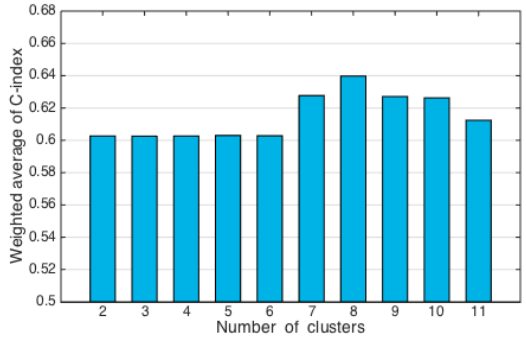


Fig. 2: The effect of cluster numbers in the *COX-cCMTL* model.

The number of clusters is an important parameter in the proposed *COX-cCMTL* model. In our experiment, to determine a suitable cluster number for our cancer patients survival analysis, we have conducted an exhaustive search that starts from 2 clusters and lasts until more than one cluster are observed with only one tumor type. In Fig. 2, we present the performance of *COX-cCMTL* under each setting of the number of clusters by taking the weighted average of C-index, i.e., $\frac{\sum_{m=1}^{M} N_m c_m}{\sum_{m=1}^{M} N_m}$ where c_m is the C-index value of the m-th task.

In Table II, we show the performance results of C-index values of different algorithms, and for *COX-cCMTL* we present its results under the best cluster number, which is 8 as the corresponding weighted average of C-index in Fig. 2 is the highest one. The results show that in general the multi-task survival analysis models performs better than traditional related single-task survival analysis models. Specifically, the *COX-cCMTL* model and *COX-*$L_{2,1}$ model perform very well in most tumor types; the *COX-cCMTL* model is suitable for cancer patients death prediction as its assumption agrees with the fact that not all tumor types are related under a same pattern, and the *COX-*$L_{2,1}$ model works well because $L_{2,1}$-norm can induce sparsity and our experimental dataset is high-dimensional miRNA sequencing data.

C. Tumor Group Discovery

In *COX-cCMTL* model, we assume that there is a underlying group structure among different tasks, e.g., in our cancer patients study we assume not all types of tumor are related under a same pattern and some of them can be clustered into different groups. The results in Table II show that this group structure assumption is suitable for survival analysis in cancer patients. The proposed *COX-cCMTL* model is able to discover the underlying group structure and leverage the structure information to improve prediction performance.

In this section, we would like to discuss our observations about the group structure. From Fig. 2, we observe that in our experimental dataset the *COX-cCMTL* model performs the best when the tumor types are grouped into 8 clusters. Therefore, in Table III, we present the corresponding discovered group structure.

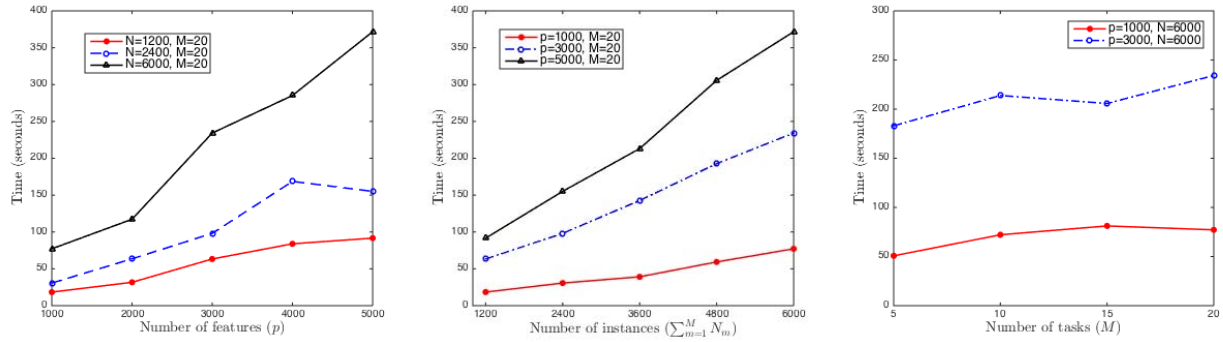
We observe that the clustering result of some groups are supported by domain knowledge and related literature. For example, UCEC and UCS belong to one group that meets the fact that their primary sites are both uterus, and similarly two kidney cancers, KIRC and KIRP, belong to the same group. However, some group results against our common sense, e.g., two types of lung cancers, LUAD and LUSC, belong to different groups. To explain this phenomenon we have consulted some clinical research papers. The LUAD usually originates in peripheral lung tissue (gland cell), while LUSC tends to be more centrally located and commonly originates in epithelial cells [34]; therefore, these two types of cancer have been grouped into different clusters. In addition, epithelial is one of the four basic types of animal tissue and lots of tumors can be viewed as epithelial tumors such as LIHC [35], LUSC and HNSC [36], and this supports our results of group 6 (G6 in Table III).

TABLE III: The group structure of tumor types under 8 clusters

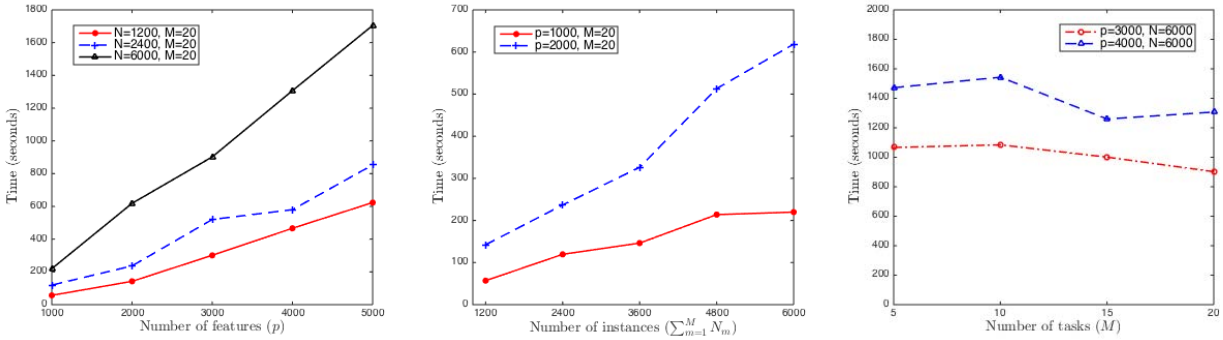
	Tumor Types		Tumor Types
G1	BRCA, CHOL, ESCA	G5	BLCA, CESC
G2	SKCM, UCEC, UCS	G6	HNSC, KIRC, KIRP, LIHC, LUSC
G3	LUAD, SARC, STAD	G7	LGG, MESO
G4	ACC, PAAD	G8	UVM

D. Scalability study of the proposed two models

The computational time of the proposed two models are mainly depended on the computational costs of the function value, gradient value, and proximal projection. Thanks to the risk set updating method proposed in [8], for the m-th task the computational costs of negative partial log-likelihood, $l(\beta)$, and its gradient, $l^{'}(\beta)$, are both $\mathcal{O}(N_m p)$. For *COX-TRACE* model, the computational cost of proximal projection is dominated by the SVD in Theorem 1, which is $\mathcal{O}(\min\{p^2 M, M^2 p\})$. In our case, usually $M < p$; therefore, the total computational cost of *COX-TRACE* model is $\mathcal{O}((\sum_{m=1}^{M} N_m + M^2)p)$. For *COX-cCMTL* model, it will take $\mathcal{O}(p^2 M)$ to calculate the penalty term and $\mathcal{O}(M^3)$ to calculate



(a) Scalability of *COX-TRACE* w.r.t. feature dimension (b) Scalability of *COX-TRACE* w.r.t. sample size (c) Scalability of *COX-TRACE* w.r.t. task number



(d) Scalability of *COX-cCMTL* w.r.t. feature dimension (e) Scalability of *COX-cCMTL* w.r.t. sample size (f) Scalability of *COX-cCMTL* w.r.t. task number

Fig. 3: Empirical scalability study of *COX-TRACE* model (upper panel) and *COX-cCMTL* model (lower panel) in terms of computational time. The y-axis in each sub-figure represents the total running time for 10 regularization scales, i.e., λ for *COX-TRACE* and $\{\rho_1, \eta\}$ for *COX-cCMTL*, averaged over five trials. Note that, we employ warm-start technology, i.e., the initial search point of the coefficient matrix is the optimal value learned in the previous training phase, which helps the model start with a searching point that is not far from the optimal solution. Therefore, the practical scalability of these two models are better than their corresponding theoretical upper bound.

eigen-decomposition in Theorem 2. Hence, in summary, the time complexity of *COX-cCMTL* model is $\mathcal{O}(\sum_{m=1}^{M} N_m p + p^2 M + M^3)$. In addition, in order to present the scalability of the proposed models in practice, in Fig.3, we demonstrate the scalability of the proposed two models with respect to the sample size, feature dimensionality, and task number, respectively.

V. Conclusion

In this paper, we proposed a unified framework for multi-task survival analysis, which extends the concept of multi-task learning to survival analysis. The proposed framework belongs to the regularized multi-task learning, where the Cox model is used to model the time to the event of interests and regularization terms are used to encode the assumption of task relatedness. Based on the proposed framework, we develop two concrete models, *COX-TRACE* and *COX-cCMTL*. These two models encode two commonly used task relatedness assumptions, i.e., low-rank assumption and group structure assumption, and the proximal gradient methods are employed to train them effectively. We demonstrate the performance of the proposed multi-task survival analysis models using the well known The Cancer Genome Atlas (TCGA) dataset to model the death time of cancer patients and discover the relationship among various of cancers. In the future, we plan to develop more advanced multi-task survival analysis methods, which can take the domain knowledge into account during the problem formulation.

Acknowledgments

This work was supported by the US National Institutes of Health grants 1RF1AG051710-01, National Science Foundation grants CNS-1637312, CCF-1451316, IIS-1615597 and IIS-1565596, and Office of Naval Research grants N00014-14-1-0631 and N00014-17-1-2265.

References

[1] Y. Li, L. Wang, J. Wang, J. Ye, and C. K. Reddy, "Transfer learning for survival analysis via efficient l2, 1-norm regularized cox regression," in *IEEE 16th International Conference on Data Mining, ICDM 2016, December 12-15, 2016, Barcelona, Spain*, 2016, pp. 231–240.

[2] M. Modarres, M. P. Kaminskiy, and V. Krivtsov, *Reliability engineering and risk analysis: a practical guide*. CRC press, 2009.

[3] N. Barbieri, F. Silvestri, and M. Lalmas, "Improving post-click user engagement on native ads via survival analysis," in *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 2016, pp. 761–770.

[4] D. R. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 187–220, 1972.

[5] L.-J. Wei, "The accelerated failure time model: a useful alternative to the cox regression model in survival analysis," *Statistics in medicine*, vol. 11, no. 14-15, pp. 1871–1879, 1992.

[6] E. T. Lee and J. Wang, *Statistical methods for survival data analysis*. John Wiley & Sons, 2003, vol. 476.

[7] R. Tibshirani *et al.*, "The lasso method for variable selection in the cox model," *Statistics in medicine*, vol. 16, no. 4, pp. 385–395, 1997.

[8] N. Simon, J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for cox's proportional hazards model via coordinate descent," *Journal of statistical software*, vol. 39, no. 5, pp. 1–13, 2011.

[9] Y. Li, J. Wang, J. Ye, and C. K. Reddy, "A multi-task learning formulation for survival analysis," in *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ser. KDD '16. ACM, 2016, pp. 1715–1724. [Online]. Available: http://doi.acm.org/10.1145/2939672.2939857

[10] R. Caruana, "Multitask learning," *Machine Learning*, vol. 28, no. 1, pp. 41–75, 1997.

[11] J. Zhou, L. Yuan, J. Liu, and J. Ye, "A multi-task learning formulation for predicting disease progression," in *Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2011, pp. 814–822.

[12] S. Bickel, J. Bogojeska, T. Lengauer, and T. Scheffer, "Multi-task learning for hiv therapy screening," in *Proceedings of the 25th international conference on Machine learning*. ACM, 2008, pp. 56–63.

[13] C. Li and H. Li, "Network-constrained regularization and variable selection for analysis of genomic data," *Bioinformatics*, vol. 24, no. 9, pp. 1175–1182, 2008.

[14] T. Evgeniou and M. Pontil, "Regularized multi–task learning," in *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2004, pp. 109–117.

[15] A. Argyriou, T. Evgeniou, and M. Pontil, "Convex multi-task feature learning," *Machine Learning*, vol. 73, no. 3, pp. 243–272, 2008.

[16] J. Liu, S. Ji, and J. Ye, "Multi-task feature learning via efficient l2, 1-norm minimization," in *Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*. AUAI Press, 2009, pp. 339–348.

[17] S. Ji and J. Ye, "An accelerated gradient method for trace norm minimization," in *Proceedings of the 26th annual international conference on machine learning*. ACM, 2009, pp. 457–464.

[18] L. Jacob, J.-p. Vert, and F. R. Bach, "Clustered multi-task learning: A convex formulation," in *Advances in neural information processing systems*, 2009, pp. 745–752.

[19] J. Zhou, J. Chen, and J. Ye, "Clustered multi-task learning via alternating structure optimization," in *Advances in neural information processing systems*, 2011, pp. 702–710.

[20] Y. Nesterov, *Introductory lectures on convex optimization: A basic course*. Springer Science & Business Media, 2013, vol. 87.

[21] E. L. Kaplan and P. Meier, "Nonparametric estimation from incomplete observations," *Journal of the American statistical association*, vol. 53, no. 282, pp. 457–481, 1958.

[22] S. Ma, X. Song, and J. Huang, "Supervised group lasso with applications to microarray data analysis," *BMC bioinformatics*, vol. 8, no. 1, p. 60, 2007.

[23] C.-N. Yu, R. Greiner, H.-C. Lin, and V. Baracos, "Learning patient-specific cancer survival distributions as a sequence of dependent regressors," in *Advances in Neural Information Processing Systems*, 2011, pp. 1845–1853.

[24] S. Kim and E. P. Xing, "Tree-guided group lasso for multi-task regression with structured sparsity," in *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 2010, pp. 543–550.

[25] J. Zhou, J. Chen, and J. Ye, *MALSAR: Multi-tAsk Learning via StructurAl Regularization*, Arizona State University, 2011. [Online]. Available: http://www.public.asu.edu/~jye02/Software/MALSAR

[26] J. Zhou, F. Wang, J. Hu, and J. Ye, "From micro to macro: data driven phenotyping by densification of longitudinal electronic medical records," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2014, pp. 135–144.

[27] B. Efron, "The efficiency of cox's likelihood function for censored data," *Journal of the American statistical Association*, vol. 72, no. 359, pp. 557–565, 1977.

[28] N. E. Breslow, "Contribution to the discussion of the paper by DR cox," *Journal of the Royal Statistical Society, Series B*, vol. 34, no. 2, pp. 216–217, 1972.

[29] H. Zha, X. He, C. Ding, M. Gu, and H. D. Simon, "Spectral relaxation for k-means clustering," in *Advances in Neural Information Processing Systems*, 2002, pp. 1057–1064.

[30] Y.-W. Wan, G. I. Allen, and Z. Liu, "Tcga2stat: simple tcga data access for integrated statistical analysis in r," *Bioinformatics*, p. btv677, 2015.

[31] T. Therneau, "A package for survival analysis in S. R package version 2.37-4," *URL http://CRAN. R-project. org/package= survival. Box*, vol. 980032, pp. 23 298–0032, 2013.

[32] Y. Yang and H. Zou, "A cocktail algorithm for solving the elastic net penalized cox's regression in high dimensions," *Stat and its Interface*, vol. 6, no. 2, pp. 167–173, 2012.

[33] F. E. Harrell, R. M. Califf, D. B. Pryor, K. L. Lee, and R. A. Rosati, "Evaluating the yield of medical tests," *JAMA*, vol. 247, no. 18, pp. 2543–2546, 1982.

[34] W. D. Travis, L. B. Travis, and S. S. Devesa, "Lung cancer," *Cancer*, vol. 75, no. S1, pp. 191–202, 1995. [Online]. Available: http://dx.doi.org/10.1002/1097-0142(19950101)75:1+<191::AID-CNCR2820751307>3.0.CO;2-Y

[35] T. K. Lee, R. T. Poon, A. P. Yuen, M. T. Ling, W. K. Kwok, X. H. Wang, Y. C. Wong, X. Y. Guan, K. Man, K. L. Chau *et al.*, "Twist overexpression correlates with hepatocellular carcinoma metastasis through induction of epithelial-mesenchymal transition," *Clinical cancer research*, vol. 12, no. 18, pp. 5369–5376, 2006.

[36] F. M. Johnson, B. Saigal, M. Talpaz, and N. J. Donato, "Dasatinib (bms-354825) tyrosine kinase inhibitor suppresses invasion and induces cell cycle arrest and apoptosis of head and neck squamous cell carcinoma and non–small cell lung cancer cells," *Clinical Cancer Research*, vol. 11, no. 19, pp. 6924–6932, 2005.

APPENDIX

Algorithm 1: proximal gradient algorithm used for model training

Input: Initial coefficient matrix $B^{(0)}$, and corresponding regularization scales
Output: $\hat{B}$

1 **Initialize**: $B^{(1)} = B^{(0)}$, $d_{-1} = 0$, $d_0 = 1$,$\gamma_0 = 1$,$i = 1$;
2 **repeat**
3 Set $\alpha_i = \frac{d_{i-2}-1}{d_{i-1}}$, $S^{(i)} = B^{(i)} + \alpha_i(B^{(i)} - B^{(i-1)})$;
4 **for** $j = 1, 2, \cdots$ **do**
5 Set $\gamma = 2^j\gamma_{i-1}$;
6 Calculate proximal operator: $B^{(i+1)} = \arg\min_{B\in\mathcal{C}} \Pi_{\gamma,S}(B)$;
7 **if** $f(B^{(i+1)}) \leq \Pi_{\gamma,S^{(i)}}(B^{(i+1)})$ **then**
8 $\gamma_i = \gamma$, **break** ;
9 **end**
10 **end**
11 $d_i = \frac{1+\sqrt{1+4d_{i-1}^2}}{2}$;
12 $i = i + 1$;
13 **until** *Convergence of* $B^{(i)}$;
14 $\hat{B} = B^{(i)}$;

Algorithm 1 outlines the general framework of proximal gradient methods. Its key building block is the calculation of proximal operator which shows in line 6, and it varies with respect to different problems. In line 1, the current search point is defined as a combination of previous two search points, and in lines 4-10, the optimal step size γ_i is chosen by the line search strategy.