Biometrical Journal

WILEY-VCH

Biometrical Journal

RESEARCH ARTICLE OPEN ACCESS

A Network-Constrain Weibull AFT Model for Biomarkers Discovery

Claudia Angelini[1] | Daniela De Canditiis[2] | Italia De Feis[1] | Antonella Iuliano[3]

[1]Istituto per le Applicazioni del Calcolo "M.Picone" (CNR), Via Pietro Castellino, Napoli, Italy | [2]Istituto per le Applicazioni del Calcolo "M.Picone" (CNR), Via dei Taurini, Roma, Italy | [3]Department of Mathematics, Computer Science, and Economics, University of Basilicata, Viale dell'Ateneo Lucano, Potenza, Italy

Correspondence: Antonella Iuliano (antonella.iuliano@unibas.it)

Received: 5 October 2023 | **Revised:** 8 May 2024 | **Accepted:** 22 May 2024

Funding: C. Angelini was partially supported by PRIN 2022 PNRR P2022BLN38 project, "Compu- tational approaches for the integration of multi-omics data" funded by European Union - Next Generation EU, CUP B53D23027810001. A. Iuliano and C. Angelini acknowledge the INdAM - GNCS Project 2022 "Modelli di shock basati sul processo di conteggio geometrico e applicazioni alla sopravvivenza", CUP E55F22000270001. C. Angelini and I. De Feis were partially supported by the Project REGINA: "Rete di Genomica Integrata per Nuove Applicazioni in medicina di precisione" - Ministero della salute nell'ambito del Piano Operativo Salute. Traiettoria 3 "Medicina rigenerativa, predittiva e personalizzata". Linea di azione 3.1 "Creazione di un programma di medicina di precisione per la mappatura del genoma umano su scala nazionale", CUP B53C22002520006. I. De Feis and D. De Canditiis acknowledge the INdAM - GNCS Project 2023 "Metodi com- putazionali per la modellizzazione e la previsione di malattie neurodegenerative", CUP E53C22001930001. A. Iuliano was partially supported by PRIN 2022 PNRR P2022XSF5H project, "Stochastic Models in Biomathematics and Applications", funded by European Union - Next Generation EU, CUP C53D23008430001.

Keywords: accelerated failure time model | network regularization | proximal gradient descent method | survival analysis | Weibull model

ABSTRACT

We propose AFTNet, a novel network-constraint survival analysis method based on the Weibull accelerated failure time (AFT) model solved by a penalized likelihood approach for variable selection and estimation. When using the log-linear representation, the inference problem becomes a structured sparse regression problem for which we explicitly incorporate the correlation patterns among predictors using a double penalty that promotes both sparsity and grouping effect. Moreover, we establish the theoretical consistency for the AFTNet estimator and present an efficient iterative computational algorithm based on the proximal gradient descent method. Finally, we evaluate AFTNet performance both on synthetic and real data examples.

1 | Introduction

In the last 20 years, the development of high-throughput technology has produced a large amount of heterogeneous biomolecular data that, together with clinical information about patients, promises to identify stable and interpretable biomarkers to predict survival and characterize personalized therapy. Thereby, there has been a growing interest in developing methods for high-dimensional data that integrate genome-scale knowledge into regression models for survival data to create a comprehensive view of molecular mechanisms and disease progression. In this context, the Cox proportional hazard model (Cox 1972), which assumes a constant hazard ratio over time, has been naturally extended to deal with the so-called problem $p \gg n$, representing p the number of regressors and n the number of samples (see the pioneeristic papers by Antoniadis, Fryzlewicz, and Letué 2010; Benner et al. 2010; Du, Ma, and Liang 2010; Fan and Li 2002; Gui and Li 2005; Tibshirani 1997; H. H. Zhang and Lu 2007). At the same time, the availability of databases encoding information about genes/proteins regulatory mechanisms and pathways demanded more advanced techniques to integrate this knowledge into the models. Network-penalized Cox regression methods explicitly incorporated the relationships among the variables in the penalty term, improving the prediction capabilities and better addressing the inherent structure of omics data (Gong, Wu, and Clarke 2014; Huang et al. 2014, 2016; Iuliano et al. 2016,

2018; Iuliano et al. 2021; Jiang and Liang 2018; Kim et al. 2012; S. Wang et al. 2009; R. Li et al. 2021; Sun et al. 2014; Verissimo et al. 2016; Wu and Wang 2013; W. Zhang et al. 2013).

An attractive alternative to the Cox model is the accelerated failure time (AFT) model, where the covariates can accelerate or decelerate the life course of an event by some constant. Different papers extended the AFT model to the high-dimensional case, considering semiparametric estimators based on weighted least squares or rank-based losses. Huang, Ma, and Xie (2006) and Huang and Ma (2010) introduced the regularized Stute's weighted least squares estimator combining least absolute shrinkage and selection operator (LASSO) (Tibshirani 1996), the threshold-gradient-directed regularization method (Friedman and Popescu 2004), and the bridge penalty (Frank and Friedman 1993). Huang and Harrington (2005), Datta, Le-Rademacher, and Datta (2007), S. Wang et al. (2008), and Khan and Shaw (2013) presented the regularized Buckley–James estimator using LASSO, elastic net (Zou and Hastie 2005), and Dantzig selector (Candes and Tao 2007). Sha, Tadesse, and Vannucci (2006) developed a Bayesian variable selection approach; Engler and Li (2009) and Cai, Huang, and Tian (2009) developed the regularized Gehan's estimator considering LASSO and the elastic net penalties. Recently, Cheng et al. (2022) considered the Stute's weighted least squares criterion combined with the l_0-penalty extending the support detection and root finding (SDAR) algorithm (Huang et al. 2018) for the linear regression model to the AFT model.

In the context of network-penalized AFT models, Ren et al. (2019) developed a robust network-based variable selection method based on the regularized Stute's least absolute deviation estimator with a penalty of an $MCP + l_1$ form. MCP is the minimax concave penalty (C. Zhang 2010) encouraging sparsity, and the l_1 term promotes the network structure incorporating the network adjacency through the Pearson correlation coefficient. Recently, Suder and Molstad (2022) proposed a new alternating direction method of the multipliers algorithm based on proximal operators for fitting semiparametric AFT models. They minimize a penalized Gehan's estimator considering both the weighted elastic net and the weighted sparse group LASSO as penalties. The `R` package `penAFT` (https://cran.r-project.org/web/packages/penAFT/index.html) implements the proposed approach.

The literature on high-dimensional parametric AFT models is less extensive than semiparametric ones. However, these models offer an interesting alternative to weighted least squares or rank-based estimators since they are simple, relatively robust against the misspecification of the assumed distribution (Hutton and Monaghan 2002), and maximum likelihood estimation (MLE) can be used for inference. Park and Do (2018) proposed LASSO, adaptive LASSO (Zou 2006), and smoothly clipped absolute deviation (SCAD) (Fan and Li 2001) for both the log-normal and the Weibull AFT models. Barnwal, Cho, and Hocking (2022) presented an interesting implementation of the AFT model using XGBoost (Chen and Guestrin 2016), a widely used library for gradient boosting, considering the log-normal, log-logistic, and the Weibull AFT models. Alam, Rahman, and Bari (2022) penalized the log-likelihood function with Firth's penalty term (Firth 1993) to overcome the problems due to small sample or rare events for the log-normal, log-logistic, and the Weibull AFT models.

To the best of the authors' knowledge, the problem of incorporating biomolecular knowledge into high-dimensional parametric AFT models still needs to be addressed. In this paper, we fill this gap with AFTNet, a novel method for the Weibull AFT model based on a double penalty that combines LASSO and quadratic Laplacian penalties (C. Li and Li 2010) to promote both sparsity and grouping effect. When using the log-linear representation, the inference becomes a structured sparse regression problem for which we implement an efficient iterative computational algorithm based on the proximal gradient descent method and cross-validated linear predictors approach (CV-PL) (Dai and Breheny 2024). Moreover, we establish the proposed estimator's theoretical consistency and evaluate its performance on synthetic and real data examples.

The paper is organized as follows. Section 2 presents the mathematical background. Section 3 introduces the AFTNet estimator and its theoretical property. Section 4 discusses the numerical implementation. Results are shown in Sections 5 and 6, and conclusions are drawn in Section 7.

2 | Mathematical Background

The AFT model is a flexible mathematical framework that describes the relationship between a set of covariates and a time-to-event response. In this model, the nonnegative random variable T describing the time to event is related to covariates $\boldsymbol{x} = (x_1, \dots, x_p)^T \in \mathbb{R}^p$ through the hazard function

$$h(t) = h_0\left(te^{-\boldsymbol{x}^\top\boldsymbol{\beta}}\right)e^{-\boldsymbol{x}^\top\boldsymbol{\beta}}, \tag{1}$$

where $\boldsymbol{\beta} = (\beta_1, \dots, \beta_p)^T$ is the parameter vector, $\boldsymbol{x}^\top\boldsymbol{\beta}$ denotes the joint effect of covariates, and $h_0(\cdot)$ is the baseline hazard function. The baseline hazard function represents the hazard without the effects of covariates, that is, when $\boldsymbol{x} = (0, \dots, 0)^T$. The hazard function in Equation (1) explains the name AFT model, whereas in a proportional hazards (PH) model the covariates act multiplicatively on the hazard and in an AFT model the covariates act multiplicatively on time. Then in AFT models, the effect of covariates is such that if $e^{-\boldsymbol{x}^\top\boldsymbol{\beta}} > 1$ a deceleration of the survival (time) process ensues and if $e^{-\boldsymbol{x}^\top\boldsymbol{\beta}} < 1$ then an acceleration of the survival (time) process occurs. The term $e^{\boldsymbol{x}^\top\boldsymbol{\beta}}$ is known as the accelerated factor.

Let now consider a study with a number n of individuals from a homogeneous population and suppose that for each individual, the values of p explanatory variables $\boldsymbol{x}_i = (x_{1i}, \dots, x_{pi})^T$ have been recorded, $i = 1, \dots, n$. Let $T_i \geq 0$ be the survival time (failure time) and C_i the censoring time for each $i = 1, \dots, n$. We assume a right censoring mechanism and make the following hypotheses:

A1. Given covariates $\boldsymbol{x}_i$, T_is and C_is are conditionally independent and the pairs (T_i, C_i)s are also conditionally independent for $i = 1, \dots, n$;

A2. Given covariates $\boldsymbol{x}_i$, C_is are conditionally noninformative about T_is.

With the introduction of the censoring variable, the observations can be represented by the pairs (ξ_i, δ_i), where $\xi_i = \min(T_i, C_i)$ and $\delta_i = 0$ if the ith individual is censored, $\delta_i = 1$ if not.

AFT models are unified by adopting a log-linear representation of the model. This representation shows that the AFT model for survival data is closely related to the general linear model used in regression analysis. Moreover, most computer software packages adopt the log-linear AFT form when fitting the data. The log-linear representation of the AFT model describes a linear relationship between the logarithm of survival time and covariates given by

$$\log T_i = \boldsymbol{x}_i^\top \boldsymbol{\beta} + \sigma\varepsilon_i, \quad i = 1, \dots, n, \tag{2}$$

where $\varepsilon_i \in \mathbb{R}$ is a random error independent from $\boldsymbol{x}_i$, $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_p)^T$ is the parameter vector and σ is the scale parameter. Passing on the time scale from Equation (2), we get

$$T_i = e^{\boldsymbol{x}_i^\top \boldsymbol{\beta}} e^{\sigma\varepsilon_i}, \quad i = 1, \dots, n.$$

Hence, the survival function of variable T_i is given by

$$S_i(t) = P(T_i \geq t) = P(e^{\sigma\varepsilon_i} \geq te^{-\boldsymbol{x}_i^\top \boldsymbol{\beta}}) = S_0(te^{-\boldsymbol{x}_i^\top \boldsymbol{\beta}}), \tag{3}$$

where $S_0(t) = P(e^{\sigma\varepsilon_i} \geq t)$ is the survival function of an individual for whom $\boldsymbol{x} = 0$. Taking the logarithm of both sides of Equation (3), multiplying by -1, and differentiating with respect to t, we get

$$h_i(t) = -\frac{d \log S_i(t)}{dt} = h_0(t\, e^{-\boldsymbol{x}_i^\top \boldsymbol{\beta}}) e^{-\boldsymbol{x}_i^\top \boldsymbol{\beta}},$$

which is the general form of the hazard function for the ith individual in an AFT model; see Equation (1). For completeness, in the following, we give the survival, density, and hazard functions of the ith individual under the AFT model in terms of the random variable ε. These alternative formulations will shortly be used to write the likelihood. The survival function of the ith individual is

$$\begin{aligned} S_i(t) = P(T_i \geq t) = P(\log T_i \geq \log t) &= P\left(\varepsilon_i \geq \frac{\log t - \boldsymbol{x}_i^\top \boldsymbol{\beta}}{\sigma}\right) \\ &= S_{\varepsilon_i}\left(\frac{\log t - \boldsymbol{x}_i^\top \boldsymbol{\beta}}{\sigma}\right), \end{aligned} \tag{4}$$

the density function of the ith individual is

$$f_i(t) = -\frac{dS_i(t)}{dt} = \frac{1}{t\sigma} f_{\varepsilon_i}\left(\frac{\log t - \boldsymbol{x}_i^\top \boldsymbol{\beta}}{\sigma}\right), \tag{5}$$

the hazard function of the ith individual is

$$h_i(t) = \frac{1}{t\sigma} h_{\varepsilon_i}\left(\frac{\log t - \boldsymbol{x}_i^\top \boldsymbol{\beta}}{\sigma}\right). \tag{6}$$

Let denote $Y_i = \min(\log T_i, \log C_i)$. Then, under assumptions (A1) and (A2), from the expressions in Equations (4–6), we can express the AFT likelihood function as

$$\begin{aligned} &\prod_{i=1}^{n} \left[\frac{1}{t_i\sigma} f_{\varepsilon_i}\left(\frac{y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}}{\sigma}\right)\right]^{\delta_i} \left[S_{\varepsilon_i}\left(\frac{y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}}{\sigma}\right)\right]^{1-\delta_i} \\ &= \prod_{i=1}^{n} \left[\frac{1}{t_i\sigma} h_{\varepsilon_i}\left(\frac{y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}}{\sigma}\right)\right]^{\delta_i} S_{\varepsilon_i}\left(\frac{y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}}{\sigma}\right), \end{aligned}$$

and the AFT log-likelihood (a part of the term $\sum_{i=1}^{n} \delta_i \log t_i$) by

$$l(\boldsymbol{\beta}^\top, \sigma) = \sum_{i=1}^{n} \delta_i \left(\log(f_{\varepsilon_i}(e_i)) - \log(\sigma)\right) + (1 - \delta_i) \log(S_{\varepsilon_i}(e_i)), \tag{7}$$

with $e_i = \frac{y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta}}{\sigma}$ the standardized residuals.

Both $\boldsymbol{\beta}$ and σ are the unknown parameters we aim to estimate. Let us denote $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \sigma)^\top$. Then, the gradient vector of $l(\boldsymbol{\theta}) = l(\boldsymbol{\beta}^\top, \sigma)$ has the following expression:

$$\begin{cases} \dfrac{\partial l(\boldsymbol{\theta})}{\partial \beta_j} = \dfrac{1}{\sigma} \displaystyle\sum_{i=1}^{n} a_i x_{ij}, & \text{for } j = 1, \dots, p, \\ \dfrac{\partial l(\boldsymbol{\theta})}{\partial \sigma} = \dfrac{1}{\sigma} \displaystyle\sum_{i=1}^{n} (e_i a_i - \delta_i), \end{cases} \tag{8}$$

with

$$\begin{aligned} a_i &= -\delta_i \frac{d \log(f_{\varepsilon_i}(e_i))}{de_i} - (1 - \delta_i) \frac{d \log(S_{\varepsilon_i}(e_i))}{de_i} \\ &= -\delta_i \frac{d \log(f_{\varepsilon_i}(e_i))}{de_i} + (1 - \delta_i) h_{\varepsilon_i}(e_i). \end{aligned}$$

The observed information matrix $I(\boldsymbol{\theta}) = I(\boldsymbol{\beta}, \sigma)$ has entries

$$\begin{cases} -\dfrac{\partial l^2(\boldsymbol{\theta})}{\partial \beta_j \partial \beta_k} = \dfrac{1}{\sigma^2} \displaystyle\sum_{i=1}^{n} x_{ij} x_{ik} A_i, \\ -\dfrac{\partial l^2(\boldsymbol{\theta})}{\partial \beta_j \partial \sigma} = \dfrac{1}{\sigma^2} \displaystyle\sum_{i=1}^{n} x_{ij} e_i A_i + \dfrac{1}{\sigma} \dfrac{\partial l(\boldsymbol{\theta})}{\partial \beta_j}, \\ -\dfrac{\partial l^2(\boldsymbol{\theta})}{\partial^2 \sigma} = \dfrac{1}{\sigma^2} \displaystyle\sum_{i=1}^{n} (e_i^2 A_i + \delta_i) + \dfrac{2}{\sigma} \dfrac{\partial l(\boldsymbol{\theta})}{\partial \sigma}, \end{cases}$$

with $i, k = 1, \dots, p$ and $A_i = \frac{da_i}{de_i} = \delta_i \frac{d^2 \log(f_{\varepsilon_i}(e_i))}{de_i^2} + (1 - \delta_i)[h_{\varepsilon_i}(e_i) \times \frac{d \log(f_{\varepsilon_i}(e_i))}{de_i} + h_{\varepsilon_i}^2(e_i)]$.

This is a very general framework. We obtain different models depending on the distribution specified for ε_i in Equations (4–6). The members of the AFT model class include the exponential AFT model, Weibull AFT model, log-logistic AFT model, log-normal AFT model, and gamma AFT model. Here, we consider the Weibull AFT model presented in the following subsection.

2.1 | The Weibull AFT Model

The Weibull model is obtained when the variables ε_i in Equation (2) are independently and identically distributed as a standard Gumbel variable (extreme value variable) with density,

survival, and hazard functions given by

$$f_{\varepsilon_i}(x) = \exp(x - e^x); \quad S_{\varepsilon_i}(x) = \exp(-e^x);$$
$$h_{\varepsilon_i}(x) = e^x; \qquad x \in \mathbb{R}. \tag{9}$$

In this case, it is well known that the variable $T_i = \exp(\boldsymbol{x}_i^\top \boldsymbol{\beta} + \sigma\varepsilon_i)$, obtained from the log-linear model of Equation (2), is a Weibull distribution with scale parameter $\exp(\boldsymbol{x}^\top\boldsymbol{\beta})$ and shape parameter $\frac{1}{\sigma}$, (i.e., $T \sim W(\exp(\boldsymbol{x}^\top\boldsymbol{\beta}), \frac{1}{\sigma})$), the proof is given for example in Liu (2018). The Weibull distribution is the unique distribution (along with its special cases, like the exponential distribution) that satisfies both the PH and AFT assumptions. Moreover, in this case, we can specialize the general model given in the previous section since the standardized residual $e_i = \frac{y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta}}{\sigma}$ has density, survival, and hazard functions given in Equation (9). Hence, substituting $\log(f_{\varepsilon_i}(e_i)) = e_i - e^{e_i}$ and $\log(S_{\varepsilon_i}(e_i)) = -e^{e_i}$ into Equation (7), we get the log-likelihood function for the Weibull AFT model

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} \delta_i \left(-\log(\sigma) + \frac{y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta}}{\sigma} \right) - \exp\left(\frac{y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta}}{\sigma} \right) \tag{10}$$

as well as we get the expression of the gradient given in Equation (8) with parameter a_i given by

$$a_i = -\left[\delta_i - \exp\left(\frac{y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta}}{\sigma} \right) \right]. \tag{11}$$

More precisely, in the case of the Weibull AFT model, the log-likelihood gradient vector is given by

$$\begin{cases} \frac{\partial \ell(\boldsymbol{\theta})}{\partial \beta_j} = \frac{1}{\sigma} \sum_{i=1}^{n} \exp\left(\frac{y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta}}{\sigma} \right) x_{ij} - \delta_i x_{ij}, & \text{for } j = 1, \dots, p, \\ \frac{\partial \ell(\boldsymbol{\theta})}{\partial \sigma} = \frac{1}{\sigma} \sum_{i=1}^{n} \delta_i \left(-1 - \frac{y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta}}{\sigma} \right) + \exp\left(\frac{y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta}}{\sigma} \right) \frac{y_i - \boldsymbol{x}_i^\top\boldsymbol{\beta}}{\sigma}. \end{cases} \tag{12}$$

and the information matrix by

$$\begin{cases} -\dfrac{\partial \ell^2(\boldsymbol{\theta})}{\partial \beta_j \partial \beta_k} = \dfrac{1}{\sigma^2} \displaystyle\sum_{i=1}^{n} x_{ij} x_{ik} \exp(e_i), \\ -\dfrac{\partial \ell^2(\boldsymbol{\theta})}{\partial \beta_j \partial \sigma} = \dfrac{1}{\sigma^2} \displaystyle\sum_{i=1}^{n} x_{ij} e_i \exp(e_i) + \dfrac{1}{\sigma} \dfrac{\partial \ell(\boldsymbol{\theta})}{\partial \beta_j}, \\ -\dfrac{\partial \ell^2(\boldsymbol{\theta})}{\partial^2 \sigma} = \dfrac{1}{\sigma^2} \displaystyle\sum_{i=1}^{n} \left(e_i^2 \exp(e_i) + \delta_i \right) + \dfrac{2}{\sigma} \dfrac{\partial \ell(\boldsymbol{\theta})}{\partial \sigma}. \end{cases} \tag{13}$$

3 | Inference

Given a sample of n data satisfying the Weibull AFT model, we propose to estimate parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}^\top, \sigma)^\top \in \mathbb{R}^p \times \mathbb{R}^+$ using a penalized MLE. The penalty we impose on the coefficients has two motivations, which are very important in modern data analysis, especially for biomarker discovery. The first is the high dimension of the problems encountered in applications where often the number of data/individuals (sample size n) is smaller than the number of predictors/genes ($p \gg n$), so an MLE estimator is unfeasible without constraints. The second motivation is that in some applications (especially in genomic cancer applications), one has prior network-constrained information that is very important to exploit during the inference process.

Let (Y, δ), with $Y = \min(\log T, \log C)$ be the observable random variables from the censored Weibull AFT model. Given a data sample of size n $(y_i, \delta_i, \boldsymbol{x}_i)$, $i = 1, \dots, n$, under Assumptions (A1) and (A2), Equation (10) represents the log-likelihood function, with a_i given in Equation (11).

In these settings, we propose AFTNet as the following penalized estimator for parameter $\boldsymbol{\theta}$, that is,

$$\widehat{\boldsymbol{\theta}} = \underset{\theta}{\operatorname{argmin}} \; -\ell(\boldsymbol{\theta}) + n\,\mathcal{P}_{\lambda,\alpha}(\boldsymbol{\beta}), \tag{14}$$

where

$$\mathcal{P}_{\lambda,\alpha}(\boldsymbol{\beta}) = \lambda[\alpha||\boldsymbol{\beta}||_1 + (1-\alpha)\Omega(\boldsymbol{\beta})], \tag{15}$$

with $\lambda > 0$ the regularization parameter and $\alpha \in [0,1]$ a fixed parameter, balancing the convex combination of the two penalty terms. In the penalty in Equation (15), the first term is the classical ℓ_1-norm, that is, LASSO penalty, which forces the individual parameter estimate sparsity; the second term is a Laplacian matrix constraint $\Omega(\boldsymbol{\beta})$ that gives smoothness among connected variables in a prior known network. Following C. Li and Li (2010), Huang et al. (2011), Sun et al. (2014), this prior network is encoded by a weighted graph $\mathcal{G} = (V, E)$ with vertex set $V = \{1, \dots, p\}$ associated with covariates, edge set $E = \{(j,k) : (j,k) \in V \times V\}$ and zero diagonal adjacency matrix $\mathbf{A} \in \mathbb{R}^{p\times p}$. Defining the degree matrix $\mathbf{D} \in \mathbb{R}^{p\times p}$ as a diagonal matrix with $d_j = \sum_{i=1}^{p} |a_{ij}|$, we have $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{A} \in \mathbb{R}^{p\times p}$, so that the Laplacian constraint is

$$\Omega(\boldsymbol{\beta}) = \boldsymbol{\beta}^T \boldsymbol{L} \boldsymbol{\beta} = \boldsymbol{\beta}^T (\mathbf{D} - \mathbf{A}) \boldsymbol{\beta} = \sum_{1 \le j < k \le p} |a_{jk}| \left(\beta_j - s_{jk} \beta_k \right)^2, \tag{16}$$

where $s_{jk} = \operatorname{sign}(a_{jk})$. The edge (j,k) weighted by a_{jk} is labeled with a "+" or "−" sign to accommodate the case where two predictors can have a nonzero adjacent coefficient but are negatively correlated. Note that $\mathbf{L}$ is positive semidefinite since it is symmetric and diagonally dominant. The Laplacian matrix constraint, $\Omega(\boldsymbol{\beta})$, nearly sets all of the connected coefficients in the network to zero or nonzero values.

We stress that in Equation (15), we assume that the Laplacian matrix $\boldsymbol{L}$ is known and available from the literature, the regularization parameter $\lambda > 0$ can be estimated using some data-driven model selection procedure, the parameter $\alpha \in [0,1]$ is fixed and it is chosen by the user to balance the two penalties.

We provide theoretical properties of the AFTNet estimator in Section 3.1. A numerical solution to a problem in Equation (14) is presented in Section 4.

3.1 | Theoretical Properties

We can not write a closed-form expression for the solution to problem (14) using the Karush–Kuhn–Tucker conditions since the log-likelihood $\ell(\boldsymbol{\theta})$ is convex for variable $\boldsymbol{\beta}$ and not for variable σ. For this reason, it is not possible to extend asymptotic results for Cox given in Sun et al. (2014). Hence, we resort to the finite sample approach proposed in Loh and Wainwright (2015), which

establishes error bounds when both the loss and penalty are allowed to be non-convex, provided that the loss function satisfies a form of restricted strong convexity and the penalty satisfies suitable mild conditions.

Let us consider the likelihood in Equation (14), $\ell(\theta) = \sum_{i=1}^{n} \ell_i(\theta)$ with

$$\ell_i(\theta) = \delta_i \left(-\log(\sigma) + \frac{y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}}{\sigma} \right) - \exp\left(\frac{y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}}{\sigma} \right);$$

and let us define $\ell^{(n)}(\theta) = -\frac{1}{n}\ell(\theta)$. We have the following gradient vector and Hessian matrix:

- $\nabla \ell(\theta) = \sum_{i=1}^{n} \nabla \ell_i(\theta_i)$, where $\nabla \ell_i(\theta) \in \mathbb{R}^{p+1}$, by Equation (12), is

$$\nabla \ell_i(\theta_j) = \begin{cases} \frac{1}{\sigma} \exp\left(\frac{y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}}{\sigma} \right) x_{ij} - \delta_i \, x_{ij}, & j = 1, \dots, p, \\ \frac{1}{\sigma} \delta_i \left(-1 - \frac{y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}}{\sigma} \right) + \exp\left(\frac{y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}}{\sigma} \right) \frac{y_i - \boldsymbol{x}_i^T \boldsymbol{\beta}}{\sigma}, \end{cases} \tag{17}$$

- $\nabla^2 \ell(\theta) = \sum_{i=1}^{n} \nabla^2 \ell_i(\theta)$, where $\nabla^2 \ell_i(\theta) \in \mathbb{R}^{p+1} \times \mathbb{R}^{p+1}$, by Equation (13), is

$$\left[\nabla^2 \ell_i(\theta)\right]_{jk} = -\begin{cases} \frac{1}{\sigma^2} x_{ij} x_{ik} \exp(e_i), & j, k = 1, \dots, p, \\ \frac{1}{\sigma^2} x_{ij} e_i \exp(e_i) + \frac{1}{\sigma} \frac{\partial \ell_i}{\partial \beta_j}, & j = 1, \dots, p; \\ & k = p+1, \\ \frac{1}{\sigma^2}(e_i^2 \exp(e_i) + \delta_i) + \frac{2}{\sigma} \frac{\partial \ell_i}{\partial \sigma}, & j = k = p+1. \end{cases}$$

Inspired by the work of Reeder, Lu, and Haneuse (2023), we make the following assumptions:

Assumption 3.1 (Bounded data).

(i) $\exists \tau_y$: $|Y_i| \le \tau_y$, for all $i = 1, \dots, n$,

(ii) $\exists \tau_x$: $\|x_{ij}\| < \tau_x \ \forall i, j$.

Assumption 3.2 (Bounded true parameter). Let $\theta^* = (\boldsymbol{\beta}^{*T}, \sigma^*)^T \in \mathbb{R}^p \times \mathbb{R}^+$ be the true parameter vector, that is,

$$\theta^* = \underset{\theta}{\operatorname{argmin}} \, \mathbb{E}_{XY}[\ell^{(n)}(\theta)], \quad \mathbb{E}_{XY}[\nabla \ell^{(n)}(\theta^*)] = 0,$$

$$\mathbb{E}_{XY}[\nabla^2 \ell^{(n)}(\theta^*)] \succ 0,$$

then $\exists \, R$ and $s \ll p$, such that

$$\|\theta^*\|_1 \le R; \quad \|\boldsymbol{\beta}^*\|_0 = s = \left| \{ j \, : \, \beta_j^* \neq 0 \} \right|; \quad \sigma^* > 0. \tag{18}$$

Assumption 3.3 (Bounded minimum eigenvalue population Hessian).

$$\exists \gamma > 0 \, : \, \min_{\theta : \|\theta - \theta^*\|_2 \le 2R} \lambda_{min}\left(\mathbb{E}[\nabla^2 \ell^{(n)}(\theta)] \right) \ge \gamma,$$

with $\lambda_{min}(A)$ being the minimum eigenvalue of matrix A.

Since matrix $\boldsymbol{L} \succeq 0$, the solution $\hat{\theta}$ of the minimization problem in Equations (14) and (15) satisfies

$$\begin{aligned} \hat{\theta} &= \underset{\theta \in \mathbb{R}^p \times \mathbb{R}^+, \|\theta\|_1 \le R}{\operatorname{argmin}} \; -\ell(\theta) + n\lambda\alpha\|\boldsymbol{\beta}\|_1 + n\lambda(1-\alpha)\|\boldsymbol{L}^{\frac{1}{2}}\boldsymbol{\beta}\|_2^2 \\ &= \underset{\theta \in \mathbb{R}^p \times \mathbb{R}^+, \|\theta\|_1 \le R}{\operatorname{argmin}} \; -\frac{1}{n}\ell(\theta) + \lambda\alpha\|\boldsymbol{\beta}\|_1 + \lambda(1-\alpha)\|\boldsymbol{L}^{\frac{1}{2}}\boldsymbol{\beta}\|_2^2 \\ &= \underset{\theta \in \mathbb{R}^p \times \mathbb{R}^+, \|\theta\|_1 \le R}{\operatorname{argmin}} \; \ell^{(n)}(\theta) + \lambda\alpha\|\boldsymbol{\beta}\|_1 + \lambda(1-\alpha)\|\boldsymbol{L}^{\frac{1}{2}}\boldsymbol{\beta}\|_2^2, \end{aligned}$$

Here, we include the side condition $\|\theta\|_1 \le R$ to guarantee the existence of at least one local/global optima.

In the following, we state the main result.

Theorem 3.1. *Under Assumptions 3.1–3.3 for all $\hat{\theta}$ local minimizers that satisfy the first-order condition*

$$\langle \nabla \ell^{(n)}(\hat{\theta}) + \nabla \mathcal{P}_{\lambda,\alpha}(\hat{\theta}), \theta - \hat{\theta} \rangle \ge 0, \quad \forall \theta \, : \, \|\theta\|_1 \le R,$$

it holds

$$\|\hat{\theta} - \theta^*\|_2 \le \frac{\frac{3}{2}\lambda\alpha\sqrt{s+1} + 2\lambda(1-\alpha)R\lambda_{max}(\boldsymbol{L})}{\gamma}$$

with high probability.

The proofs of Theorem 3.1 and associated lemmas are given in the Appendix.

Remark 3.2. In high-dimensional statistics, beyond consistency in L_2 norm (as proved in Theorem 3.1), one should also consider consistency in terms of variable selection and the asymptotic normality of the estimator restricted to the support of the true vector β^*. Regarding variable selection and asymptotic normality, there are some results in the Cox and frailty models which use separable penalties like LASSO, SCAD, or MCP (see Fan and Li 2002, Huang et al. 2013, and D. Wang, Wu, and Zhao 2019). There are also results concerning the linear model with the sparse network penalty (see C. Li and Li 2010 and Huang et al. 2011). The first result on the consistency of the sparse Laplacian shrinkage estimator for the Cox model, both in terms of empirical risk for the coefficients and sparsity, was given by Sun et al. (2014). In the context of network-penalized MLE estimators for AFT models, to the best of the authors' knowledge, this is the first result of L_2 consistency.

4 | Numerical Implementation

We propose a two-step procedure to numerically solve Equation (14) in AFTNet. In the first step, we estimate $\hat{\sigma}$ using the `survreg` function in the `R` package `survival`, that is, regressing Y only on the intercept. In the second step, we plug this estimate into Equation (14) and solve the following:

$$\underset{\beta \in \mathbb{R}^p}{\operatorname{argmin}} \; -\frac{1}{n}\ell(\boldsymbol{\beta}) + \lambda\alpha \sum_{i=1}^{p} |\beta_i| + \lambda(1-\alpha)\boldsymbol{\beta}^T \mathbf{L} \boldsymbol{\beta}. \tag{19}$$

We use a proximal gradient technique to obtain the corresponding stationary point of Equation (19). In the following, we give a

brief description of this technique. Then, we specialize it for our problem. The proximal gradient technique is an iterative method that solves the composite model (Parikh and Boyd 2014)

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} f(\boldsymbol{\beta}) + g(\boldsymbol{\beta}),$$

under the following assumptions (Beck 2017, chapters 2 and 10):

i. f is proper, closed, differentiable, with the Lipschitz-continuous gradient, with convex domain;

ii. g is a proper, closed, and convex function.

At each iteration k, we linearize the function $f(\boldsymbol{\beta})$ around the current point and solve the problem

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} f\left(\boldsymbol{\beta}^{\boldsymbol{k}}\right) + \nabla f\left(\boldsymbol{\beta}^{\boldsymbol{k}}\right)^T\left(\boldsymbol{\beta} - \boldsymbol{\beta}^{\boldsymbol{k}}\right) + g(\boldsymbol{\beta}) + \frac{M}{2}\|\boldsymbol{\beta} - \boldsymbol{\beta}^{\boldsymbol{k}}\|_2^2.$$

The last term is the proximal term whose function is to keep the update in a neighborhood of the current iterate $\boldsymbol{\beta}^{\boldsymbol{k}}$, where $f(\boldsymbol{\beta})$ is close to its linear approximation. The constant $M > 0$ is the step parameter, an upper bound of the Lipschitz constant of the ∇f. We can rewrite the problem as

$$\min_{\boldsymbol{\beta}\in\mathbb{R}^p} \frac{1}{2}\left\|\boldsymbol{\beta} - \left(\boldsymbol{\beta}^k - \frac{1}{M}\nabla f\left(\boldsymbol{\beta}^k\right)\right)\right\|_2^2 + \frac{1}{M}g(\boldsymbol{\beta})$$

whose solution is

$$\boldsymbol{\beta}^{k+1} = \text{Prox}_{\frac{1}{M}g(\boldsymbol{\beta})}\left(\boldsymbol{\beta}^k - \frac{1}{M}\nabla f\left(\boldsymbol{\beta}^k\right)\right),$$

where $\text{Prox}_{\frac{1}{M}g(\boldsymbol{\beta})}()$ denotes the proximal operator associated with the $\frac{1}{M}g()$ function.

We can now specify such a technique for our case, defining

$$f(\boldsymbol{\beta}) = -\frac{1}{n}\ell(\boldsymbol{\beta}) + \lambda(1-\alpha)\boldsymbol{\beta}^T\mathbf{L}\boldsymbol{\beta} = -\frac{1}{n}\ell(\boldsymbol{\beta}) + \lambda(1-\alpha)\left\|\mathbf{L}^{\frac{1}{2}}\boldsymbol{\beta}\right\|_2^2,$$

and

$$g(\boldsymbol{\beta}) = \lambda\alpha\sum_{i=1}^{p}|\beta_i| = \lambda\alpha\|\boldsymbol{\beta}\|_1.$$

We recall that the proximal operator of the LASSO penalty $g(\boldsymbol{\beta})$ is the soft thresholding. Hence, we have

$$\boldsymbol{\beta}^{k+1} = \text{SOFT}\left(\boldsymbol{\beta}^k - \frac{1}{M}\nabla f\left(\boldsymbol{\beta}^k\right), \frac{\lambda\alpha}{M}\right),$$

which is applied component-wise, that is,

$$\beta_j^{k+1} = \text{SOFT}\left(u_j, \frac{\lambda\alpha}{M}\right) = \begin{cases} u_j - \frac{\lambda\alpha}{M}, & \text{if } u_j > \frac{\lambda\alpha}{M} \\ 0, & \text{if } -\frac{\lambda\alpha}{M} \le u_j \le \frac{\lambda\alpha}{M} \\ u_j + \frac{\lambda\alpha}{M}, & \text{if } u_j < -\frac{\lambda\alpha}{M} \end{cases}$$

with

$$u_j = \beta_j^k - \frac{1}{M}\left(\nabla\ell\left(\boldsymbol{\beta}^k\right)\right)_j - \frac{2\lambda(1-\alpha)}{M}\left(\mathbf{L}\boldsymbol{\beta}^k\right)_j,$$

for $j = 1, \ldots, p$.

We summarize in Algorithm 1 the numerical procedure proposed in AFTNet.a0~

ALGORITHM 1 | Proximal gradient algorithm for solving (19).

Require: $\lambda > 0$, $\alpha \in [0, 1]$

Input: $\boldsymbol{y}, \boldsymbol{X}, \boldsymbol{\delta}, \hat{\sigma}, \mathbf{L}$

Initialize $k \leftarrow 0$

Initialize $\boldsymbol{\beta}^{(k)}$

Initialize $M > 0$

while convergence not satisfied **do**

Backtracking line search for the step parameter M

evaluate $\nabla\ell\left(\boldsymbol{\beta}^k\right)$ by Equation (12) first formula

$\boldsymbol{u} \leftarrow \boldsymbol{\beta}^k - \frac{1}{M}\left(\nabla\ell\left(\boldsymbol{\beta}^k\right)\right) - \frac{2\lambda(1-\alpha)}{M}\left(\mathbf{L}\boldsymbol{\beta}^k\right)$

$\boldsymbol{\beta}^{k+1} \leftarrow \text{SOFT}\left(\boldsymbol{u}, \frac{\lambda\alpha}{M}\right)$

Check convergence

$k \leftarrow k + 1$

end while

In Algorithm 1, the *Backtracking line search* for the step parameter M is implemented as in Beck (2017). The *Check convergence* stops the algorithm when a maximum number of iterations has been reached, or the relative error norm between the estimates at two consecutive iterations is less than a given tolerance. Furthermore, Algorithm 1 assumes that the regularization parameter λ is known. In practice, we need to choose the regularization parameters by some data-driven selection procedure; in the following subsection, we describe our choice.

We implemented the proposed algorithm in the R package AFTNet and a set of auxiliary functions for fitting, cross-validation, prediction, and visualization. The source code to the AFTNet package and to data and codes to reproduce the simulations and real data analyses is available as the Supporting Information.

4.1 | Construction of Adjacency Matrix

In general, the adjacency matrix is known up to signs, so it is necessary to implement a method for estimating $s_{jk} = \text{sign}(a_{jk})$ (see Equation 16). We suggest the following two proposals:

- $s_{jk} = \text{sign}(\text{corr}(\mathbf{x}_{\cdot,j}, \mathbf{x}_{\cdot,k}))$, where corr is the Pearson correlation (see Huang et al. 2011).
- $s_{jk} = \text{sign}(\tilde{\beta}_j)\text{sign}(\tilde{\beta}_k)$, where $\tilde{\boldsymbol{\beta}}$ is given by a ridge estimate of the Weibul AFT model (see Sun et al. 2014).

4.2 | λ Parameter Selection

In our AFTNet implementation, we extend the CV-PL (Dai and Breheny 2024) to the Weibull AFT penalized model. More precisely, after dividing the training data into K nonoverlapping folds, for each $k = 1, \ldots, K$, we use the kth fold D^k as the test set, and the $K-1$ remaining folds T^k as the training set. We obtain

TABLE 1 | Simulation studies' dimensions: Weak and strong high dimensionality effects are tested in separate evaluations using both not-overlapping and overlapping cases. p is the number of potential explanatory variables (i.e., genes). n_T and n_D are the number of observations in the training set T and testing set D, respectively.

	Number of genes	Training set	Testing set
Effect	$\boldsymbol{p}$	$\boldsymbol{n_T}$	$\boldsymbol{n_D}$
Weak	220	110	55
Strong	1100	275	138

$\hat{\boldsymbol{\beta}}_\lambda^{(-k)}$ from T^k. Then, we evaluate the standardized residuals

$$\hat{e}_{i,\lambda}^{CV} = \frac{y_i - \boldsymbol{x}_i^\top \hat{\boldsymbol{\beta}}_\lambda^{(-k)}}{\hat{\sigma}}, \quad \forall i \in D^k.$$

Hence, after repeating this for all K folds, we combine all standardized residuals $\hat{e}_{i,\lambda}^{CV} \quad i = 1, \ldots, n$. We plug in $\hat{e}_{i,\lambda}^{CV}$ into the negative log-likelihood in Equation (10) and define $CV(\lambda)$ as

$$CV(\lambda) = -\sum_{i=1}^{n} \delta_i \left(-\log(\hat{\sigma}) + \hat{e}_{i,\lambda}^{CV} \right) - \exp\left(\hat{e}_{i,\lambda}^{CV} \right).$$

We evaluate $CV(\lambda)$ at each value of λ belonging to a predetermined grid of values within the interval $[\lambda_{min}, \lambda_{max}]$ and select the one for which $CV(\lambda)$ is the minimum.

5 | Simulation Results

To demonstrate the performance of AFTNet, we perform various numerical experiments on synthetic datasets and compare AFTNet with the elastic net regularized Gehan estimator under the AFT model described in Suder and Molstad (2022). In each synthetic dataset, we split n observations into n_T observations assigned to the training set and n_D observations to the test set, with $n = n_T + n_D$. The latter serves to evaluate the prediction performance. We consider scenarios that are likely to be encountered in genomic studies for biomarkers discovery, that is, the number of genes/variables $p >> n_T$ and availability of prior information about the regulatory networks among genes encoded by a known adjacency matrix $\mathbf{A}$. In all our experiments, the network consists of r regulatory modules. Similarly to Sun et al. (2014), each ith module, $i = 1, \ldots, r$, models one transcription factor (TF) that regulates a given number of genes p_i, such that the overall number of genes is $p = \sum_{i=1}^{r}(p_i + 1)$. We define the adjacency matrix $\mathbf{A}$ as a binary matrix of dimension $p \times p$, with $a_{ij} = 1$ between the TFs and their regulated genes, and $a_{ij} = 0$ otherwise. We consider two cases $r = 20$ and $r = 100$ according to Table 1, where $p/n_T = 2$ represents a *weak* effect of high dimensionality and $p/n_T = 4$ represents a *strong* effect of high dimensionality.

We consider two topological settings:

- *Not-overlapping regulatory modules*: The r regulatory modules are disjoint from each other. In each module, the TF regulates 10 genes. For each sample i, $i = 1, \ldots, n$, the ith row of the matrix $\mathbf{X}$ is given by the expression values of the p genes generated according to the following scheme: the expression value of each TF_j, $j = 1, \ldots, r$, is sampled from a standard normal distribution. The expression values of the 10 regulated genes are sampled from a conditional normal distribution with correlation ρ between their expressions and that of the corresponding TF. For each module, we randomly select υ genes to have a positive correlation $\rho = 0.7$ and the remaining $10 - \upsilon$ genes to have a negative correlation $\rho = -0.7$, mimicking the activation or repression of each gene under the effect of its corresponding TF.
- *Overlapping regulatory modules*: The first four regulatory modules overlap. In particular, the first two regulatory modules share 10 common genes (i.e., jointly regulated by TF_1 and TF_2), five genes are specific to the first regulatory module, and five genes are specific to the second regulatory module. Therefore, the first two regulatory modules contain 20 genes and two TFs. The third and fourth regulatory modules have six common genes (i.e., jointly regulated by TF_3 and TF_4), seven genes are specific to the third regulatory module, and seven are specific to the fourth module. Therefore, together, they are composed of 20 genes and two TFs. The remaining $r - 4$ regulatory modules do not overlap and the TF regulates 10 genes in each module, as in the not-overlapping case. This scenario mimics cases where some genes can belong to different pathways regulating different biological processes, as often observed in cancer. In this second setting, for each sample i, $i = 1, \ldots, n$, the ith row of the matrix $\mathbf{X}$ is given by the expression values of the p genes generated according to the following scheme: the expression value of each TF_j, $j = 1, \ldots, r$, is sampled from a standard normal distribution, the expression values of the module-specific regulated genes are sampled from a conditional normal distribution with correlation ρ between their expressions and that of the corresponding TF. In contrast, the expression values of the common regulated genes are sampled from a conditional normal distribution with correlation ρ between their expressions and that of the average of the two corresponding TFs. For both module-specific and common genes, we randomly select υ genes to have a positive correlation $\rho = 0.7$ and the remaining genes to have a negative correlation $\rho = -0.7$, mimicking the activation or repression of each gene under the effect of its corresponding TF.

In both dimensional cases, *weak* and *strong*, and topological settings, *not-overlapping* and *overlapping*, we consider $p_{active} = 88$ active genes, choosing the true parameter vector $\boldsymbol{\beta}^* \in \mathbb{R}^p$ with the last $p - 88$ components equal to 0, and the coefficients β_j^*, $j = 1, \ldots, 44$ are generated from the uniform distribution $\mathcal{U}(0.1, 0.5)$, while β_j^*, $j = 45, \ldots, 88$ are generated from $\mathcal{U}(-0.5, -0.1)$. In each dimensional case and setting, we generate times, T_i, $i = 1, \ldots, n$, sampling from a Weibull distribution with a shape parameter equal to $1/\sigma$ and scale parameter equal to $\exp(\boldsymbol{x}_i^T \boldsymbol{\beta}^*)$, with $\boldsymbol{x}_i^T$ being the ith row of matrix $\boldsymbol{X}$. We log-transform the times and consider moderate, medium, and high censoring rate (CR) scenarios with 30%, 50%, 80% CR, respectively.

We consider three values for the scale parameter σ: $\sigma = 0.5$ corresponding to increasing hazard, $\sigma = 1$ corresponding to an exponential distribution with constant hazard, and $\sigma = 1.5$ corresponding to decreasing hazard.

TABLE 2 | *Not-overlapping case*. Performance metrics with $n_T = 110$, $n_D = 55$, $p = 220$, and $\alpha = 0.5$ (weak effect) averaged over 100 independent replications. From top to bottom, results correspond to 30%, 50%, and 80% CR scenarios, respectively.

	$\sigma = 0.5$		$\sigma = 1$		$\sigma = 1.5$	
CR = 30%	**AFTNet**	**penAFT**	**AFTNet**	**penAFT**	**AFTNet**	**penAFT**
EMSE	2.609	2.802	2.810	2.915	2.927	2.966
PMSE	0.657	0.842	0.790	0.924	0.880	0.965
FNR	0.187	0.644	0.400	0.828	0.539	0.902
FPR	0.690	0.167	0.478	0.088	0.380	0.062
NSR	0.739	0.243	0.526	0.122	0.413	0.076
	$\sigma = 0.5$		$\sigma = 1$		$\sigma = 1.5$	
CR = 50%	**AFTNet**	**penAFT**	**AFTNet**	**penAFT**	**AFTNet**	**penAFT**
EMSE	2.623	2.818	2.796	2.938	2.925	2.982
PMSE	0.666	0.846	0.786	0.935	0.879	0.968
FNR	0.159	0.668	0.394	0.854	0.523	0,911
FPR	0.716	0.167	0.485	0.086	0.388	0.060
NSR	0.766	0.233	0.534	0.110	0.424	0.072
	$\sigma = 0.5$		$\sigma = 1$		$\sigma = 1.5$	
CR = 80%	**AFTNet**	**penAFT**	**AFTNet**	**penAFT**	**AFTNet**	**penAFT**
EMSE	2.565	2.901	2.800	2.959	2.913	2.996
PMSE	0.646	0.909	0.803	0.948	0.884	0.978
FNR	0.129	0.808	0.370	0.878	0.500	0.918
FPR	0.769	0.109	0.522	0.084	0.419	0.068
NSR	0.810	0.142	0.565	0.099	0.451	0.074

We use AFTNet as described in Section 4. In particular, in Algorithm 1, we initialize $\boldsymbol{\beta}^{(0)} = \mathbf{0}$, we fix $\alpha = 0.5$ and M is the largest eigenvalue of the likelihood Hessian matrix evaluated at $(\boldsymbol{\beta}^{(0)}, \hat{\sigma})$. We select the optimal parameter $\lambda_{opt} \in [\lambda_{min}, \lambda_{max}]$ by the CV-LP approach illustrated in Subsection 4.2 with $K = 5$ folds, $\lambda_{max} = \left\| \nabla \ell \left(\boldsymbol{\beta}^{(0)}, \hat{\sigma} \right) \right\|_\infty / \alpha$ and $\lambda_{min} = 0.01 \cdot \lambda_{max}$. For the λ's grid, we consider 50 equispaced points ζ_i in the interval $[\log 10(\lambda_{min}), \log 10(\lambda_{max})]$ and take $\lambda_i = 10^{\zeta_i}$.

For what concerns the implementation of the regularized Gehan estimator, we use the `penAFT` package with the following choice penalty = "EN," which stands for an elastic-net penalty, with $\alpha = 0.5$, $nlambda = 50$ different values of regularization parameter and $lambda.ratio.min = 0.01$.

To compare performance of the different approaches, we use the following indicators:

- the estimated mean square error (EMSE):

$$EMSE = \left\| \boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}} \right\|_2,$$

- the predictive mean square error (PMSE):

$$PMSE = \left\| \boldsymbol{X}_D^T \boldsymbol{\beta}^* - \boldsymbol{X}_D^T \hat{\boldsymbol{\beta}} \right\|_2^2 / n_D,$$

- the false negative rate (FNR):

$$FNR = \sum_{j=1}^{p} I\left[\hat{\boldsymbol{\beta}}_j = 0 \wedge \boldsymbol{\beta}_j^* \neq 0 \right] / p_{active},$$

  which is the relative number of not detected genes,

- the false positive rate (FPR):

$$FPR = \sum_{j=1}^{p} I\left[\hat{\boldsymbol{\beta}}_j \neq 0 \wedge \boldsymbol{\beta}_j^* = 0 \right] / (p - p_{active}),$$

  which is the relative number of (falsely) detected genes,

- the number of selected variables rate (NSR):

$$NSR = \sum_{j=1}^{p} I\left[\hat{\boldsymbol{\beta}}_j \neq 0 \right] / p,$$

  where $\hat{\boldsymbol{\beta}}$ denotes the estimated vector coefficients obtained by AFTnet or by `penAFT` package.

In the following, we show the average of the above indicators over 100 independent simulations.

Results for the not-overlapping case. We report simulation results for both models in Tables 2 and 3 for the weak and strong

TABLE 3 | *Not-overlapping case*. Performance metrics with $n_T = 275$, $n_D = 138$, $p = 1100$, and $\alpha = 0.5$ (strong effect) averaged over 100 independent replications. From top to bottom, results correspond to 30%, 50%, and 80% CR scenarios, respectively.

	$\sigma = 0.5$		$\sigma = 1$		$\sigma = 1.5$	
CR = 30%	**AFTNet**	**penAFT**	**AFTNet**	**penAFT**	**AFTNet**	**penAFT**
EMSE	2.701	2.838	2.837	2.962	2.944	3.005
PMSE	0.580	0.678	0.658	0.754	0.730	0.783
FNR	0.145	0.588	0.335	0.851	0.516	0.940
FPR	0.544	0.067	0.346	0.027	0.277	0.017
NSR	0.569	0.095	0.372	0.037	0.293	0.021
	$\sigma = 0.5$		$\sigma = 1$		$\sigma = 1.5$	
CR = 50%	**AFTNet**	**penAFT**	**AFTNet**	**penAFT**	**AFTNet**	**penAFT**
EMSE	2.750	2.875	2.869	2.985	2.955	3.014
PMSE	0.608	0.698	0.679	0.771	0.734	0.790
FNR	0.125	0.668	0.366	0.895	0.545	0.957
FPR	0.594	0.066	0.339	0.025	0.264	0.013
NSR	0.616	0.087	0.363	0.032	0.279	0.016
	$\sigma = 0.5$		$\sigma = 1$		$\sigma = 1.5$	
CR = 80%	**AFTNet**	**penAFT**	**AFTNet**	**penAFT**	**AFTNet**	**penAFT**
EMSE	2.715	2.949	2.873	3.015	2.962	3.039
PMSE	0.585	0.744	0.674	0.783	0.729	0.795
FNR	0.086	0.806	0.305	0.924	0.518	0.948
FPR	0.705	0.055	0.401	0.036	0.282	0.031
NSR	0.721	0.066	0.424	0.040	0.298	0.032

dimensional cases, respectively. Across all the considered topological settings and associated choices of σ (i.e., $\sigma = 0.5, 1, 1.5$), AFTNet outperforms penAFT in both the weak and strong effects in terms of EMSE, PMSE, and FNR. In contrast, penAFT tends to select fewer genes, resulting in a lower FPR. We note that both procedures AFTNet and penAFT are robust with respect to the CR scenarios.

Figures 1 and 2 provide analogous results using box plots. Here, a greater variability is visible in AFTNet for FNR and FPR indicators. Moreover, to assess the validity of our method, we plot the receiver operating characteristic (ROC) curve (for $\sigma = 1$ and $CR = 30\%$) in Figure 3. The figure shows that AFTnet performs better than penAFT in weak and strong dimensional scenarios.

Finally, the average running time of AFTNet for obtaining a single solution path over 100 independent replications, for $\sigma = 0.5, 1, 1.5$, and CR = 30% (50%; 80%), in the simulation settings with $(n, p) = (165, 220)$ is about 5.11, 7.52, 4.95 s, respectively, whereas, with $(n, p) = (413, 1100)$, is about 306.274, 195.753, 195.732 s, respectively. In the first case, the running time of AFTNet is slightly above the average running time for the penAFT; in the second one, it is larger. However, the comparison is only partially fair since penAFT implementation uses C++ functions to speed up the running time.

Results for the overlapping case. The simulation results for both methods are shown in Tables 4 and 5 for the weak and strong dimensional cases, respectively. Also, in this case, AFTNet performs better than `penAFT` in terms of EMSE, PMSE, and FNR when a weak or strong dimensional scenario is simulated. Instead, penAFT tends to detect fewer genes with lower FPR. We note that both procedures AFTNet and penAFT are robust with respect to the CR.

A visual summary of the metrics can be also found in Figures 4 and 5. Moreover, to assess the validity of our method, we show the ROC curve in Figure 6 where we can see that AFTnet performs better than penAFT in both weak and strong effects when $\sigma = 1$ and $CR = 30\%$.

Finally, similar to the previous results, the average running time of AFTNet for obtaining a single solution path over 100 independent replications increases compared to penAFT.

6 | Real Case Studies

In this section, we consider two examples of biomarkers discovery related to gene expression data of different types of cancer: breast and kidney.

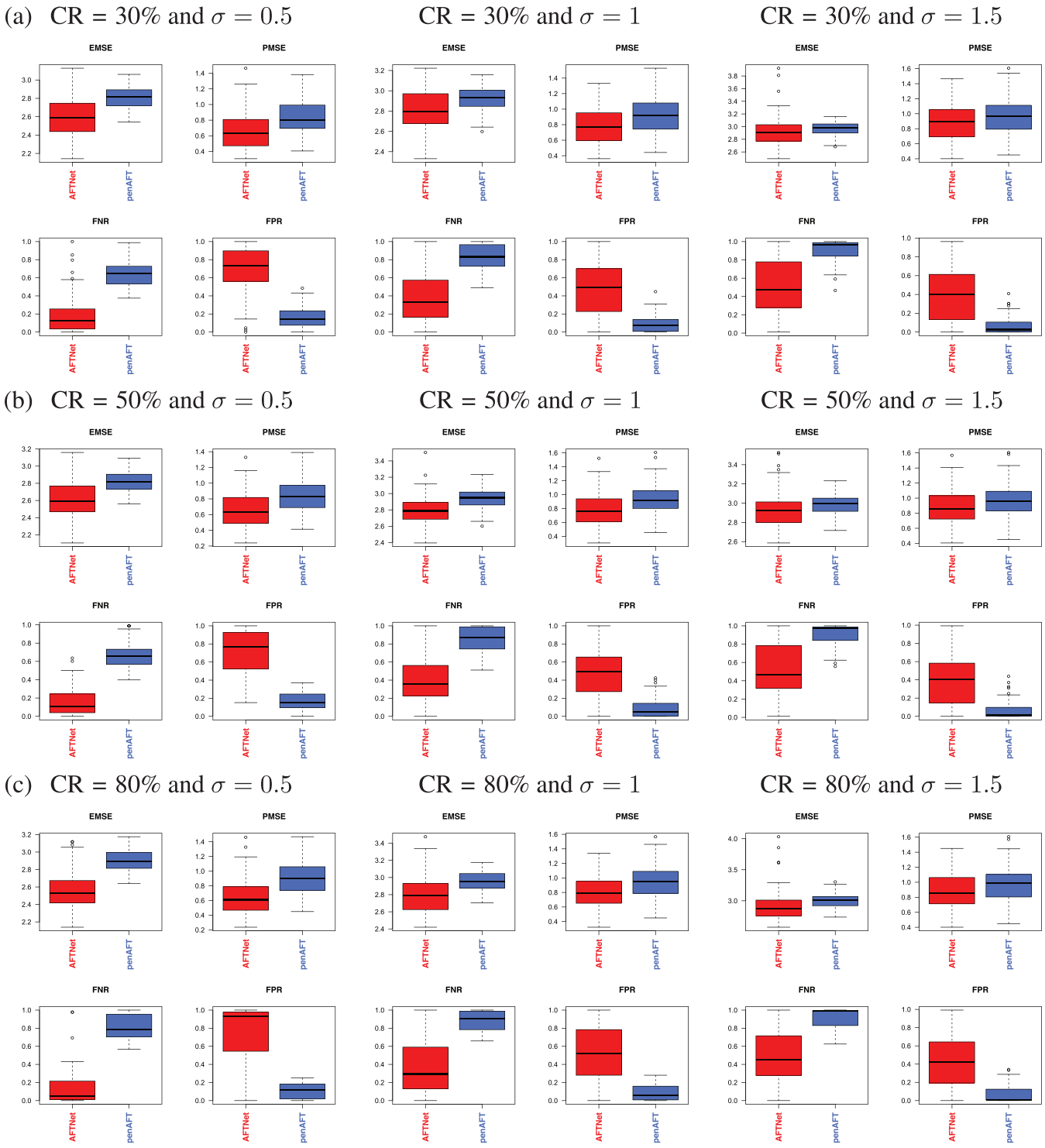

FIGURE 1 | *Not-overlapping case.* Box plots of the performance metrics results between AFTNet and penAFT with $n_T = 110$, $n_D = 55$, $p = 220$ (weak effect) and $\alpha = 0.5$ for $\sigma = 0.5, 1, 1.5$ (from the left side to the right side), respectively, averaged over 100 independent replications. From top to bottom, results correspond to (a) 30%, (b) 50%, and (c) 80% CR scenarios, respectively.

- *Breast cancer data (BC)*: We consider two independent gene expression datasets available from the GEO database (https://www.ncbi.nlm.nih.gov/geo/). We use GSE2034 as the training set and GSE2990 as the test set and select 13,229 genes common to both datasets. The first dataset contains $n_T = 286$ records of lymph-node-negative breast cancer patients. The second one contains $n_D = 189$ records of invasive breast carcinoma patients. The median survival time in the training dataset is 86 months with a censoring proportion of 62.59%, while in the test set, it is 77 months with a censoring proportion equal to 64.17%.
- *Kidney cancer data (TCGA-KIRC)*: We consider the TCGA-KIRC (Kidney Renal Clear Cell Carcinoma) gene expression dataset available in the GDC Data Portal

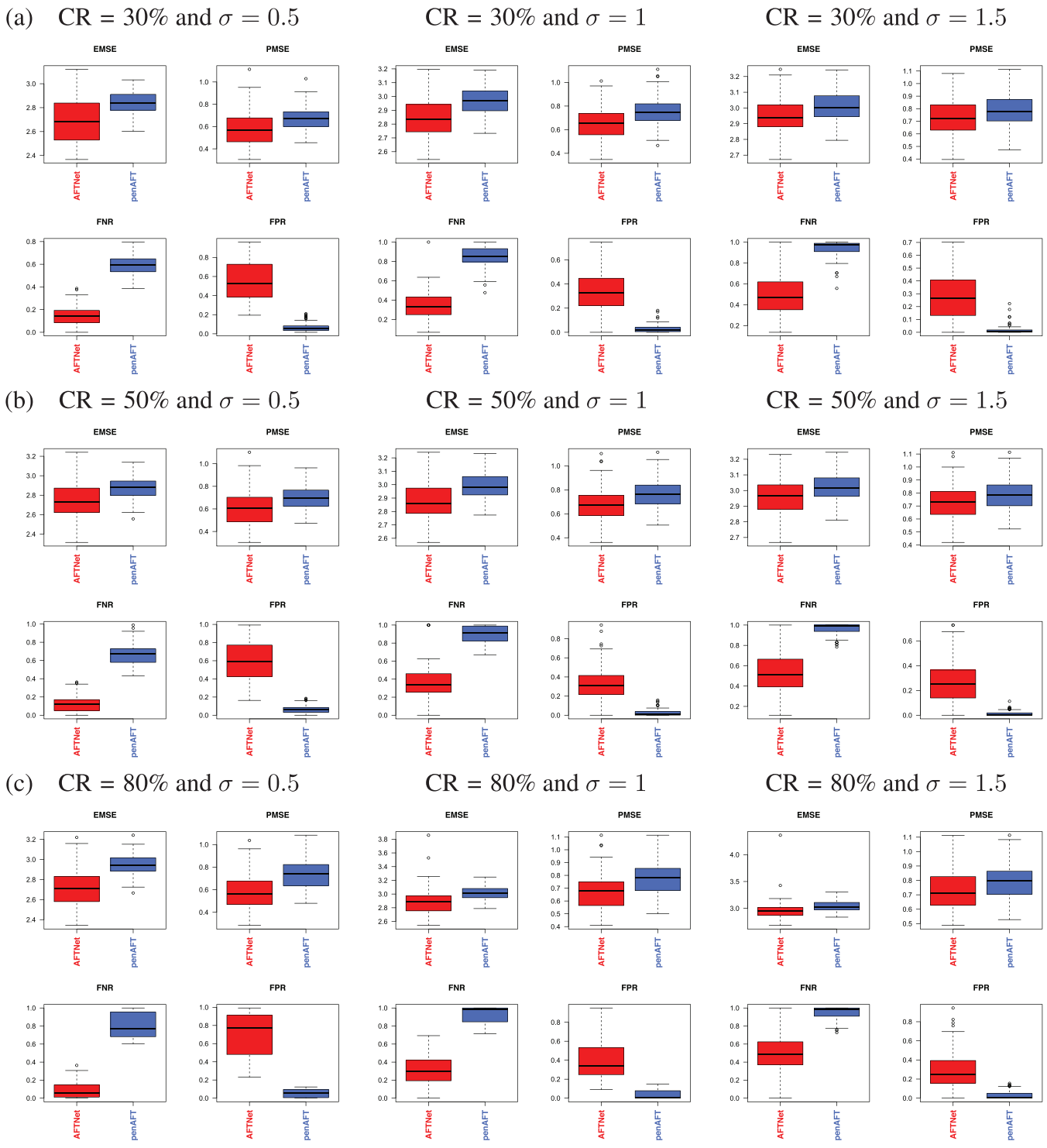

FIGURE 2 | *Not-overlapping case*. Box plots of the performance metrics results between AFTNet and penAFT with $n_T = 275$, $n_D = 138$, $p = 1100$ (strong effect) and $\alpha = 0.5$ for $\sigma = 0.5, 1, 1.5$ (from the left side to the right side), respectively, averaged over 100 independent replications. From top to bottom, results correspond to (a) 30%, (b) 50%, and (c) 80% CR scenarios, respectively.

(https://portal.gdc.cancer.gov). The data were obtained from the Illumina HiSeq platform, consisting of 518 kidney cancer patients and 20,159 genes. We download the preprocessed and normalized data (gene-level, RPKM) from the LinkedOmics portal (http://linkedomics.org/login.php). The median survival time is 1252 days with a censoring proportion of 66.60%. The data were randomly split into training T and test set D using a resampling method, which provides more accurate estimates of predictive accuracy (see, for instance, Simon et al. 2011, Iuliano et al. 2016).

As a preprocessing step, we first apply a screening procedure to reduce the dimension of the problem, selecting $p = 488$ genes for the datasets of BC and $p = 521$ genes for the

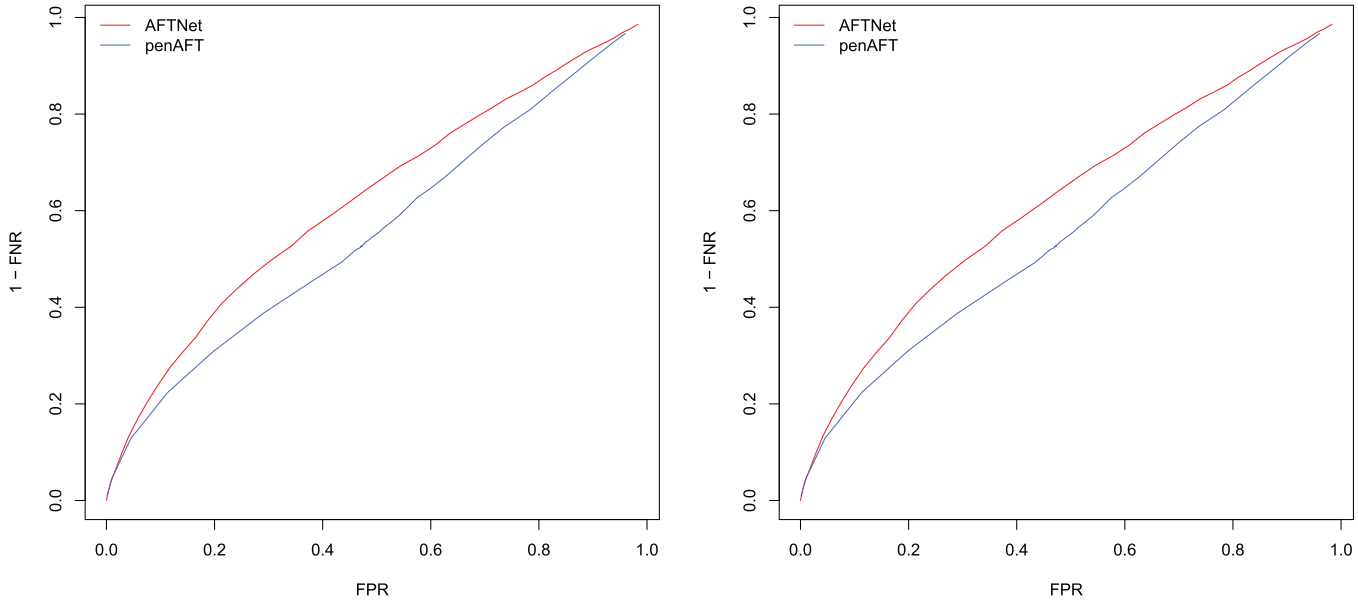

FIGURE 3 | *Not-overlapping case*. The ROC curve for $\sigma = 1$ and $CR = 30\%$ with a weak effect (on the left side) and a strong effect (on the right side).

TABLE 4 | *Overlapping case*. Performance metrics with $n_T = 110$, $n_D = 55$, $p = 220$ and $\alpha = 0.5$ (weak effect) averaged over 100 independent replications and a 30%, 50%, and 80% CR.

	$\sigma = 0.5$		$\sigma = 1$		$\sigma = 1.5$	
CR = 30%	**AFTNet**	**penAFT**	**AFTNet**	**penAFT**	**AFTNet**	**penAFT**
EMSE	2.659	2.838	2.846	2.963	2.929	2.997
PMSE	0.614	0.764	0.735	0.859	0.801	0.889
FNR	0.218	0.661	0.468	0.867	0.589	0.935
FPR	0.631	0.155	0.403	0.067	0.325	0.034
NSR	0.692	0.228	0.455	0.093	0.359	0.047
	$\sigma = 0.5$		$\sigma = 1$		$\sigma = 1.5$	
CR = 50%	**AFTNet**	**penAFT**	**AFTNet**	**penAFT**	**AFTNet**	**penAFT**
EMSE	2.690	2.874	2.856	2.971	2.928	3.003
PMSE	0.624	0.787	0.731	0.867	0.791	0.889
FNR	0.234	0.715	0.480	0.888	0.584	0.926
FPR	0.618	0.136	0.386	0.063	0.321	0.046
NSR	0.677	0.196	0.440	0.083	0.359	0.057
	$\sigma = 0.5$		$\sigma = 1$		$\sigma = 1.5$	
CR = 80%	**AFTNet**	**penAFT**	**AFTNet**	**penAFT**	**AFTNet**	**penAFT**
EMSE	2.608	2.941	2.845	2.990	2.936	3.029
PMSE	0.570	0.843	0.716	0.874	0.779	0.894
FNR	0.158	0.823	0.451	0.878	0.589	0.893
FPR	0.722	0.102	0.422	0.086	0.317	0.083
NSR	0.770	0.132	0.473	0.100	0.354	0.093

TCGA-KIRC dataset, respectively. The screening process is based on a combination of biomedical-driven (BMD) information and data-driven (DAD) knowledge (see, for instance, Iuliano et al. 2018). In particular, we perform this processing step using the `R` package `COSMONET` (Iuliano et al. 2021). Subsequently, we conduct inference by running AFTNet and penAFT according to the two real settings investigated (BC and TCGA-KIRC). To construct the prior network information, we map the screened genes from the

TABLE 5 | *Overlapping case.* Performance metrics with $n_T = 275$, $n_D = 138$, $p = 1100$, and $\alpha = 0.5$ (strong effect) averaged over 100 independent replications and a 30%, 50%, and 80% CR.

	$\sigma = 0.5$		$\sigma = 1$		$\sigma = 1.5$	
CR = 30%	**AFTNet**	**penAFT**	**AFTNet**	**penAFT**	**AFTNet**	**penAFT**
EMSE	2.755	2.868	2.880	2.975	2.953	3.006
PMSE	0.549	0.628	0.620	0.691	0.665	0.712
FNR	0.208	0.641	0.427	0.881	0.568	0.949
FPR	0.458	0.058	0.300	0.022	0.241	0.012
NSR	0.485	0.082	0.322	0.030	0.257	0.015
	$\sigma = 0.5$		$\sigma = 1$		$\sigma = 1.5$	
CR = 50%	**AFTNet**	**penAFT**	**AFTNet**	**penAFT**	**AFTNet**	**penAFT**
EMSE	2.765	2.895	2.881	2.993	2.955	3.015
PMSE	0.554	0.643	0.626	0.702	0.673	0.715
FNR	0.225	0.703	0.468	0.921	0.627	0.964
FPR	0.441	0.060	0.246	0.020	0.191	0.013
NSR	0.467	0.079	0.269	0.025	0.205	0.014
	$\sigma = 0.5$		$\sigma = 1$		$\sigma = 1.5$	
CR = 80%	**AFTNet**	**penAFT**	**AFTNet**	**penAFT**	**AFTNet**	**penAFT**
EMSE	2.728	2.962	2.882	3.012	2.959	3.027
PMSE	0.533	0.679	0.622	0.710	0.667	0.719
FNR	0.139	0.837	0.413	0.940	0.593	0.965
FPR	0.616	0.046	0.320	0.025	0.226	0.018
NSR	0.635	0.056	0.342	0.028	0.240	0.019

KEGG repository (https://www.genome.jp/kegg/pathway.html) and build a gene adjacency matrix (see Iuliano et al. 2021 for details). The sign of the adjacency matrix is chosen as $s_{ij} = \text{sign}(\text{corr}(\mathbf{x}_{\cdot i}, \mathbf{x}_{\cdot j}))$.

During the training phase in Algorithm 1, we initialize $\boldsymbol{\beta}^{(0)} = \mathbf{0}$, we fix $\alpha = 0.5$ and M the largest eigenvalue of the likelihood Hessian matrix evaluated at $(\boldsymbol{\beta}^{(0)}, \hat{\sigma})$. We select the optimal parameter $\lambda_{opt} \in [\lambda_{min}, \lambda_{max}]$ by the CV-LP approach illustrated in Subsection 4.2 with $K = 5$ folds, $\lambda_{max} = \left\| \nabla \ell \left(\boldsymbol{\beta}^{(0)}, \hat{\sigma} \right) \right\|_{\infty} / \alpha$ and $\lambda_{min} = 0.01 \cdot \lambda_{max}$. For the λ's grid, we consider 50 equispaced points ζ_i in the interval $[\log 10(\lambda_{min}), \log 10(\lambda_{max})]$ and take $\lambda_i = 10^{\zeta_i}$.

As regards the implementation of the regularized Gehan estimator, we use the `penAFT` package with the following choice *penalty* ="EN," which stands for the elastic-net penalty, with $\alpha = 0.5$, $nlambda = 50$ different values of regularization parameter and $lambda.ratio.min = 0.01$. In the testing phase, to assess the performance of both methods, we compute both concordance or Harrells C-index (Harrell et al. 1984) and the integrated area under the curve (AUC) measure (using the `survAUC` package in `R`) on the test set.

Quantitative results. The performance measures in the two examples of cancer data are displayed in Table 6. We observe that AFTNet performs better than penAFT in terms of concordance and integrated AUC in the BC dataset, while in the TCGA-KIRC dataset AFTNet performs better than penAFT only in terms of AUC. Interestingly, the running time of AFTNet is slightly higher than penAFT in the BC dataset, while it is much smaller in the TCGA-KIRC dataset compared to penAFT.

Pathway analysis. To investigate the gene signature, that is, the set of genes whose corresponding estimated coefficient in $\hat{\boldsymbol{\beta}}$ is nonzero, we perform a pathway analysis based on the information from the KEGG database (https://www.genome.jp/kegg/pathway.html). In particular, we use the `COSMONET` package (see Iuliano et al. 2021) to generate the subnetworks involved in the cancer mechanism and to identify the set of active pathways.

Before carrying out the KEGG pathway analysis, we first sort the list of the selected genes of each method (AFTnet and penAFT) in descending order, according to the regression coefficients. Then, we select the top-ranked and bottom-ranked genes into positive (top 20%) and negative (bottom 20%) genes. We perform the pathway analysis using not-isolated genes (i.e., connected in the adjacency matrix).

Specifically, for the BC dataset, we obtain that, for both methods, many of the selected genes belong to specific pathways as a group of the *KEGG pathways in cancer*, *KEGG*

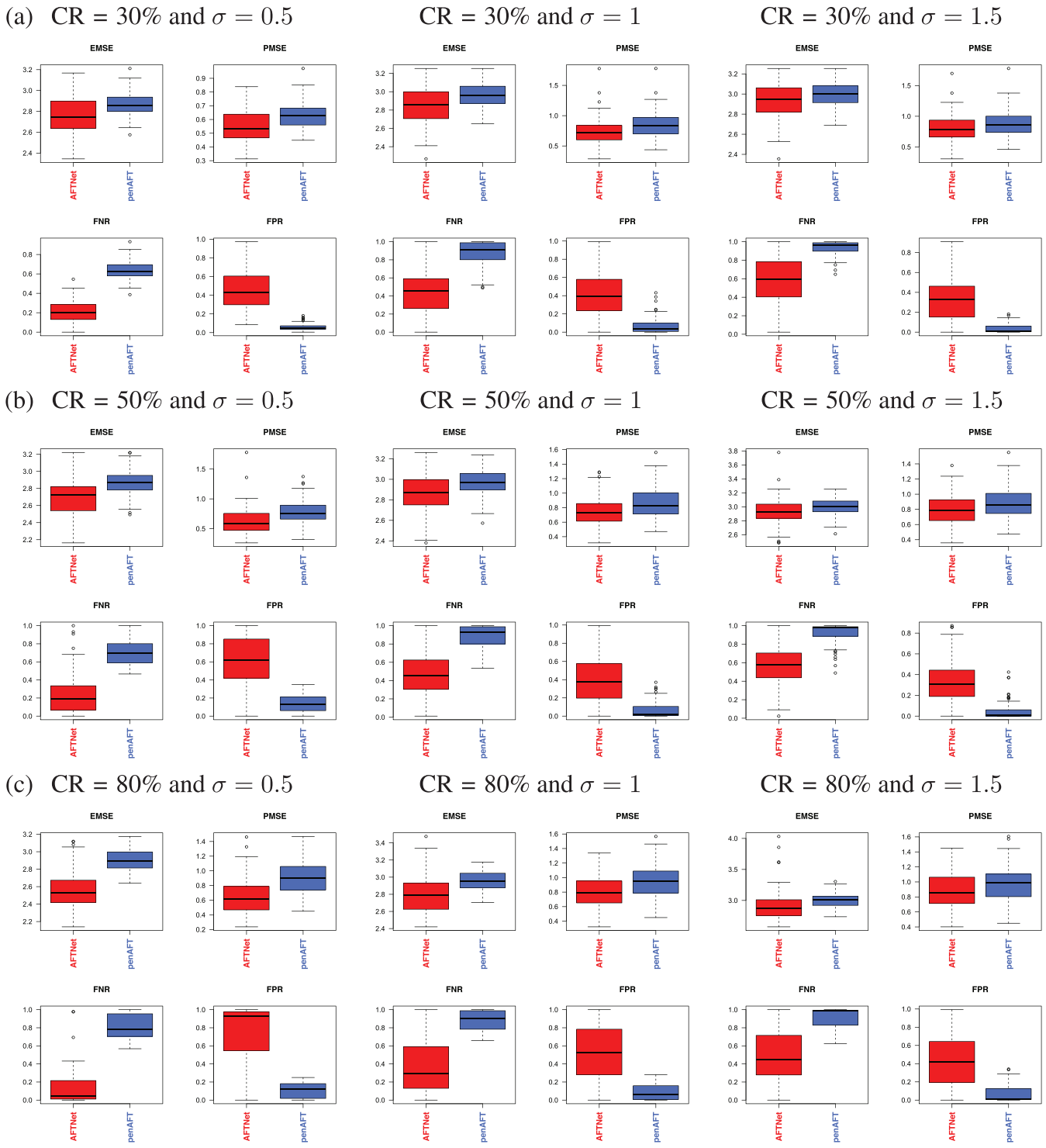

FIGURE 4 | *Overlapping case*. Box plots of the performance metrics results between AFTNet and penAFT with $n_T = 110$, $n_D = 55$, $p = 220$ (weak effect) and $\alpha = 0.5$ for $\sigma = 0.5, 1, 1.5$ (from the left side to the right side), respectively, averaged over 100 independent replications and (a) 30%, (b) 50%, and (c) 80% CR.

MAPK/ERBB/P53/chemokine signaling pathway and *KEGG cytokine–cytokine receptor interaction pathway*. However, AFTNet also includes the *KEGG VEGF signaling pathway*. Indeed, the family of the VEGF proteins has been identified as potential biomarkers of BC and indicators of treatment success and patient survival (Brogowska, Zajkowska, and Mroczko 2023).

In the TCGA-KIRC dataset, for both methods, the not-isolated genes detected are primarily included in *KEGG pathways in cancer*, *KEGG chemokine/insulin signaling pathway*, and *KEGG cycle cell*. Moreover, AFTNet contains the *KEGG NOTCH signaling pathway*, which plays an essential role in kidney development and disease treatments (Barak, Surendran, and Boyle 2012).

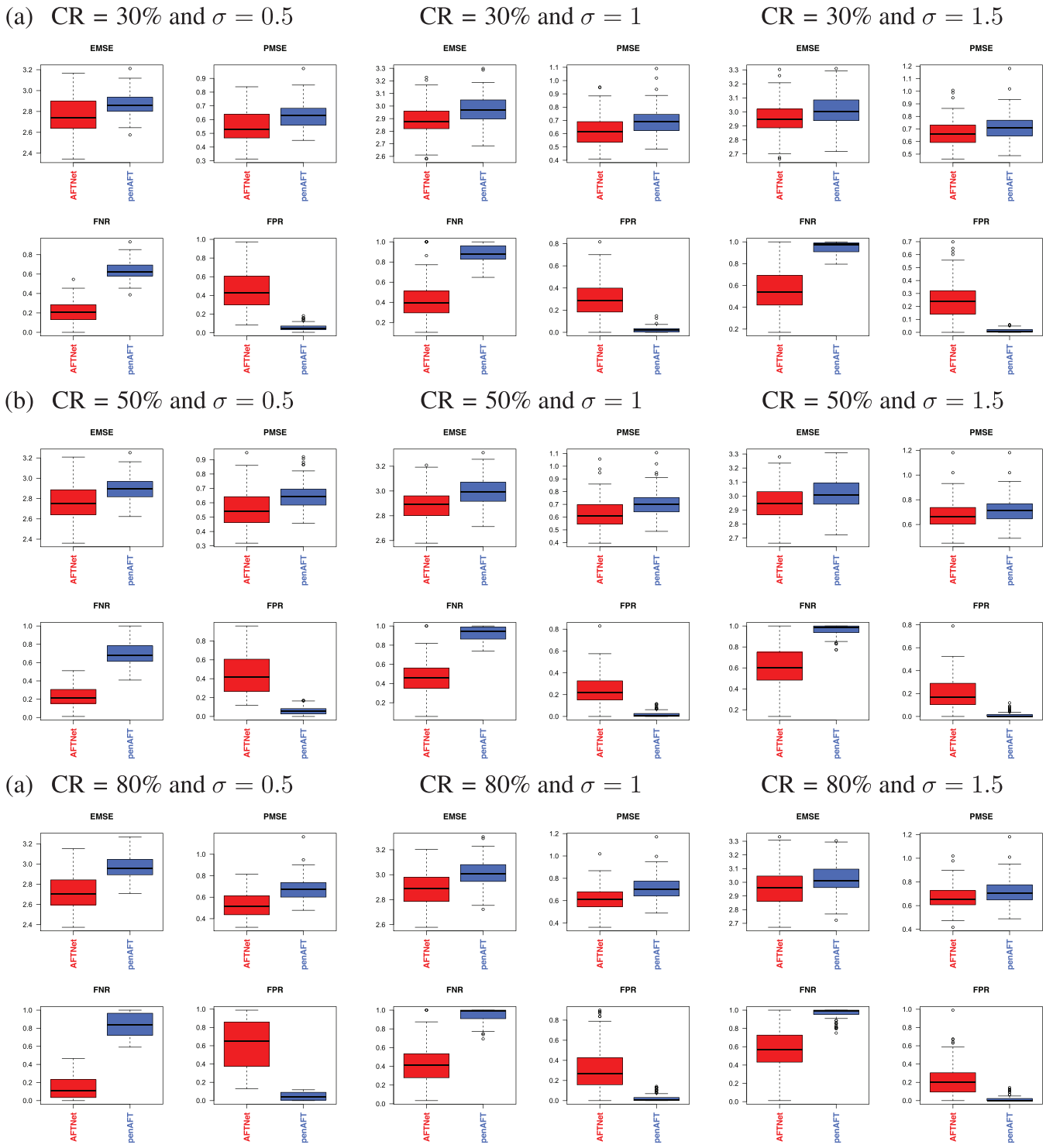

FIGURE 5 | *Overlapping case*. Box plots of the performance metrics results between AFTNet and penAFT with $n_T = 275$, $n_D = 138$, $p = 1100$ (strong effect) and $\alpha = 0.5$ for $\sigma = 0.5, 1, 1.5$ (from the left side to the right side), respectively, averaged over 100 independent replications and (a) 30%, (b) 50%, and (c) 80% CR.

7 | Conclusions

This article proposes AFTNet, a novel network-constraint survival analysis method considering the Weibull AFT model. AFTNet addresses the problem of high dimensionality (i.e., $p \gg n$) and the strong correlation among variables exploiting a double penalty that promotes both sparsity and grouping effect. We establish the finite sample consistency and present an efficient iterative computational algorithm based on the proximal gradient descent method.

We show AFTNet's effectiveness in simulations and on two real test cases, comparing its performance with the penAFT technique (Suder and Molstad 2022).

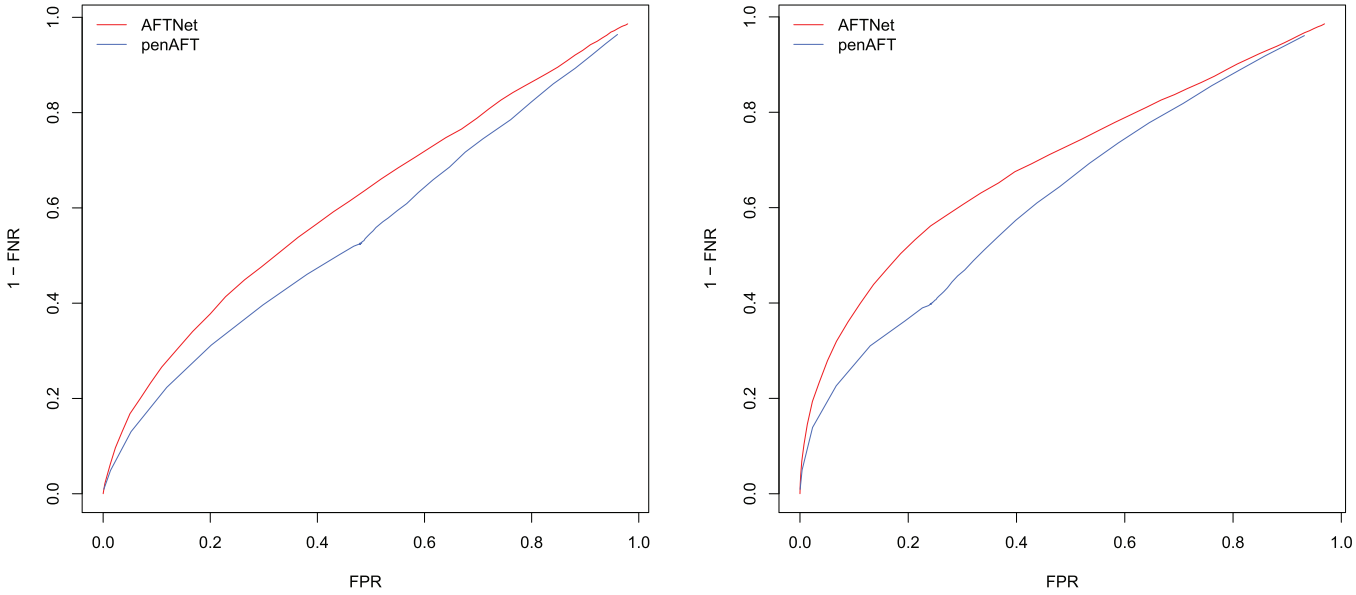

FIGURE 6 | *Overlapping case*. The ROC curve for $\sigma = 1$ and $CR = 30\%$ with a weak effect (on the left side) and a strong effect (on the right side).

TABLE 6 | Harrells C-index, integrated AUC, and computing time for AFNet and penAFT methods over the two cancer datasets. The running time includes the time taken for performing fivefold cross-validation (CV-LP) and model fitting to the complete training dataset with $\alpha = 0.5$.

	Harrells C-index		Integrated AUC		Running time (s)	
	AFTNet	penAFT	AFTNet	penAFT	AFTNet	penAFT
BC	0.671	0.639	0.396	0.325	35.174	37.757
TCGA-KIRC	0.682	0.722	0.279	0.251	52.737	297.515

Some interesting future research directions include extending the proposed framework to other parametric AFT models, that is, log-normal and log-logistic, and using other penalties, particularly non-convex ones such as SCAD or MCP.

Author Contributions

The authors contributed equally to this work.

Acknowledgments

I. De Feis and D. De Canditiis acknowledge the INdAM-GNCS Project 2023 "Metodi computazionali per la modellizzazione e la previsione di malattie neurodegenerative" (CUP E53C22001930001).

A. Iuliano and C. Angelini acknowledge the INdAM-GNCS Project 2022 "Modelli di shock basati sul processo di conteggio geometrico e applicazioni alla sopravvivenza" (CUP E55F22000270001).

C. Angelini e I. De Feis were partially supported by the Project REGINA: Rete di Genomica Integrata per Nuove Applicazioni in medicina di precisione - Ministero della salute nellambito del Piano Operativo Salute. Traiettoria 3 "Medicina rigenerativa, predittiva e personalizzata". Linea di azione 3.1 "Creazione di un programma di medicina di precisione per la mappatura del genoma umano su scala nazionale" (CUP B53C22002520006).

A. Iuliano was partially supported by Project Tech4You - Technologies for climate change adaptation and quality of life improvement, n. ECS0000009. This work was funded by the Next Generation EU - Italian NRRP, Mission 4, Component 2, Investment 1.5, call for the creation and strengthening of "Innovation Ecosystems", building "Territorial R&D Leaders" (Directorial Decree n. 2021/3277).

Open access publishing facilitated by Universita degli Studi della Basilicata, as part of the Wiley - CRUI-CARE agreement.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are openly available in Gene Expression Omnibus at https://www.ncbi.nlm.nih.gov/geo/ and in GDC Data Portal - National Cancer Institute at https://portal.gdc.cancer.gov.

Open Research Badges

This article has earned an Open Data badge for making publicly available the digitally-shareable data necessary to reproduce the reported results. The data is available in the Supporting Information section.

This article has earned an open data badge "**Reproducible Research**" for making publicly available the code necessary to reproduce the reported results. The results reported in this article could fully be reproduced.

References

Alam, T. F., M. S. Rahman, and W. Bari. 2022. "On Estimation for Accelerated Failure Time Models With Small or Rare Event Survival Data." *BMC Medical Research Methodology* 22: 169. https://doi.org/10.1186/s12874-022-01638-1.

Antoniadis, A., P. Fryzlewicz, and F. Letué. 2010. "The Dantzig Selector in Cox's Proportional Hazards Model." *Scandinavian Journal of Statistics* 37, no. 4: 531–552. https://doi.org/10.1111/j.1467-9892.2008.00586.x.

Barak, H., K. Surendran, and S. C. Boyle. 2012. "The Role of Notch Signaling in Kidney Development and Disease." *Advances in Experimental Medicine and Biology* 727: 99–113. https://doi.org/10.1007/978-1-4614-0899-4_8.

Barnwal, A., H. Cho, and T. Hocking. 2022. "Survival Regression With Accelerated Failure Time Model In XGBoost." *Journal of Computational and Graphical Statistics* 31, no. 4: 1292–1302. https://doi.org/10.1080/10618600.2022.2067548.

Beck, A. 2017. *First Order Methods in Optimization*. MOS-SIAM Series on Optimization. Philadelphia, PA: SIAM.

Benner, A., M. Zucknick, T. Hielscher, C. Ittrich, and U. Mansmann. 2010. "High-Dimensional Cox Models: The Choice of Penalty as Part of the Model Building Process." *Biometrical Journal* 52, no. 1: 50–69. https://doi.org/10.1002/bimj.200900064.

Brogowska, K. K., M. Zajkowska, and B. Mroczko. 2023. "Vascular Endothelial Growth Factor Ligands and Receptors in Breast Cancer." *Journal of Clinical Medicine* 12, no. 6: 2412. https://doi.org/10.3390/jcm12062412.

Cai, T., J. Huang, and L. Tian. 2009. "Regularized Estimation for the Accelerated Failure Time Model." *Biometrics* 65: 394–404. https://doi.org/10.1111/j.1541-0420.2008.01074.x.

Candes, E., and T. Tao. 2007. "The Dantzig Selector: Statistical Estimation When P is Much Larger Than N (With Discussion)." *The Annals of Statistics* 35: 2313–2404. https://doi.org/10.1214/009053606000001523.

Chen, T., and C. Guestrin. 2016. "XGBoost: A Scalable Tree Boosting System." In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. New York: ACM. https://doi.org/10.1145/2939672.2939785.

Cheng, C., X. Feng, J. Huang, Y. Jiao, and S. Zhang. 2022. "l_0-Regularized High-Dimensional Accelerated Failure Time Model." *Computational Statistics & Data Analysis* 170: 107430. https://doi.org/10.1016/j.csda.2022.107430.

Cox, D. R. 1972. "Regression Models and Life-Tables." *Journal of the Royal Statistical Society: Series B (Methodological)* 34, no. 2: 187–202. https://doi.org/10.1111/j.2517-6161.1972.tb00899.x.

Dai, B., and P. Breheny. 2024. "Cross-Validation Approaches for Penalized Cox Regression." *Statistical Methods in Medical Research* 33, no. 4: 702–715. https://doi.org/10.1177/0962280224123370.

Datta, S., J. Le-Rademacher, and S. Datta. 2007. "Predicting Patient Survival from Microarray Data by Accelerated Failure Time Modeling Using Partial Least Squares and LASSO." *Biometrics* 63, no. 1: 259–271. https://doi.org/10.1111/j.1541-0420.2006.00660.x.

Du, P., S. Ma, and H. Liang. 2010. "Penalized Variable Selection Procedure for Cox Models with Semiparametric Relative Risk." *The Annals of Statistics* 38, no. 4: 2092–2117. https://doi.org/10.1214/09-AOS780.

Engler, D., and Y. Li. 2009. "Survival Analysis With High-Dimensional Covariates: An Application in Microarray Studies." *Statistical Applications in Genetics and Molecular Biology* 8, no. 1: 1–22. (Article 14) https://doi.org/10.2202/1544-6115.1423.

Fan, J., and R. Li. 2001. "Variable Selection Via Nonconcave Penalized Likelihood and its Oracle Properties." *Journal of the American Statistical Association* 96, no. 456: 1348–1360. https://doi.org/10.1198/016214501753382273.

Fan, J., and R. Li. 2002. "Variable Selection for Cox's Proportional Hazards Model and Frailty Model." *The Annals of Statistics* 30, no. 1: 74–99. https://doi.org/10.1214/aos/1015362185.

Firth, D. 1993. "Bias Reduction of Maximum Likelihood Estimate." *Biometrika* 80, no. 1: 27–38. https://doi.org/10.1093/biomet/80.1.27.

Frank, I. E., and J. H. Friedman. 1993. "A Statistical View of Some Chemometrics Regression Tools." *Technometrics* 35: 109–148. https://doi.org/10.1080/00401706.1993.10485033.

Friedman, J. H., and B. E. Popescu. 2004. "Gradient Directed Regularization for Linear Regression and Classification." Technical Report, Department of Statistics, Stanford University, Stanford, CA.

Gong, H., T. T. Wu, and E. M. Clarke. 2014. "Pathway-Gene Identification for Pancreatic Cancer Survival Via Doubly Regularized Cox Regression." *BMC Systems Biology* 8: 1–9. https://doi.org/10.1186/1752-0509-8-S1-S3.

Gui, J., and H. Li. 2005. "Penalized Cox Regression Analysis in the High-Dimensional and Low-Sample Size Settings, With Applications to Microarray Gene Expression Data." *Bioinformatics* 21, no. 13: 3001–3008. https://doi.org/10.1093/bioinformatics/bti422.

Harrell Jr., F. E., K. L. Lee, R. M. Califf, D. B. Pryor, and R. A. Rosati. 1984. "Regression Modelling Strategies for Improved Prognostic Prediction." *Statistics in Medicine* 3, no. 2: 143–152. https://doi.org/10.1002/sim.4780030207.

Huang, J., and D. Harrington. 2005. "Iterative Partial Least Squares with Right-Censored Data Analysis: A Comparison to Other Dimension Reduction Techniques." *Biometrics* 61, no. 1: 17–24. https://doi.org/10.1111/j.0006-341X.2005.040304.x.

Huang, J., S. Ma, and H. Xie. 2006. "Regularized Estimation in the Accelerated Failure Time Model With High-Dimensional Covariates." *Biometrics* 62, no. 3: 813–820. https://doi.org/10.1111/j.1541-0420.2006.00562.x.

Huang, J., and S. Ma. 2010. "Variable Selection in the Accelerated Failure Time Model Via the Bridge Method." *Lifetime Data Analysis* 16: 176–195. https://doi.org/10.1111/10.1007/s10985-009-9144-2.

Huang, J., S. Ma, H. Li, and C.-H. Zhang. 2011. "The Sparse Laplacian Shrinkage Estimator for High-Dimensional Regression." *The Annals of Statistics* 39, no. 4: 2021–2046. https://doi.org/10.1214/11-AOS897.

Huang, J., T. Sun, Z. Ying, Y. Yu, and C.-H. Zhang. 2013. "Oracle Inequalities for the Lasso in the Cox Model." *The Annals of Statistics* 41, no. 3: 1142–1165. https://doi.org/10.1214/13-AOS1098.

Huang, J., L. Liu, Y. Liu, and X. Zhao. 2014. "Group Selection in the Cox Model With Diverging Number of Covariates." *Statistica Sinica* 24, no. 4: 1787–1810. https://doi.org/10.5705/ss.2013.061.

Huang, J., P. Breheny, S. Lee, S. Ma, and C.-H. Zhang. 2016. "The Mnet Method for Variable Selection." *Statistica Sinica* 26, no. 3: 903–923. https://doi.org/10.5705/ss.202014.0011.

Huang, J., Y. Jiao, Y. Liu, and X. Lu. 2018. "A Constructive Approach to l_0 Penalized Regression." *Journal of Machine Learning Research* 19, no. 1: 403–439.

Hutton, J. L., and P. F. Monaghan. 2002. "Choice of Parametric Accelerated Life and Proportional Hazards Models for Survival Data: Asymptotic Results." *Lifetime Data Analysis* 8: 375–393. https://doi.org/10.1023/A:1020570922072.

Iuliano, A., A. Occhipinti, C. Angelini, I. De Feis, and P. Lió. 2016. "Cancer Markers Selection Using Network-Based Cox Regression: A Methodological and Computational Practice." *Frontiers in Physiology* 7: 208. https://doi.org/10.3389/fphys.2016.00208.

Iuliano, A., A. Occhipinti, C. Angelini, I. De Feis, and P. Lió. 2018. "Combining Pathway Identification and Breast Cancer Survival Prediction via Screening-Network Methods." *Frontiers in Genetics* 9: 206. https://doi.org/10.3389/fgene.2018.00206.

Iuliano, A., A. Occhipinti, C. Angelini, I. De Feis, and P. Lió. 2021. "Cosmonet: An R Package for Survival Analysis Using Screening-Network Methods." *Mathematics* 9, no. 24: 3262. https://doi.org/10.3390/math9243262.

Jiang, H. K., and Y. Liang. 2018. "The L1/2 Regularization Network Cox Model for Analysis of Genomic Data." *Computers in Biology and Medicine* 100: 203–208. https://doi.org/10.1016/j.compbiomed.2018.07.009.

Khan, M. H. R., and J. E. H. Shaw. 2013. "Variable Selection With the Modified Buckley-James Method and the Dantzig Selector for High-Dimensional Survival Data." In *59th ISI World Statistics Congress Proceedings*, Hong Kong, August 25–30, 2013, 4239–4244. The Hague, The Netherlands: International Statistical Institute.

Kim, J., I. Sohn, S. H. Jung, S. Kim, and C. Park. 2012. "Analysis of Survival Data With Group Lasso." *Communications in Statistics - Simulation and Computation* 41, no. 9: 1593–1605. https://doi.org/10.1080/03610918.2011.611311.

Li, C., and H. Li. 2010. "Variable Selection and Regression Analysis for Covariates With Graphical Structure." *The Annals of Applied Statistics* 4: 1498–1516. https://doi.org/10.1214/10-AOAS332.

Li, R., Y. Tanigawa, Y. M. Justesen, et al. 2021. "Survival Analysis on Rare Events Using Group-Regularized Multi-Response Cox Regression." *Bioinformatics* 37, no. 23: 4437–4443. https://doi.org/10.1093/bioinformatics/btab095.

Liu, E. 2018. "Using Weibull Accelerated Failure Time Regression Model to Predict Survival Time and Life Expectancy." BioRxiv 362186. https://doi.org/10.1101/362186.

Loh, P. L., and M. J. Wainwright. 2015. "Regularized M-Estimators With Nonconvexity: Statistical and Algorithmic Theory for Local Optima." *Journal of Machine Learning Research* 16, no. 19: 559–616.

Parikh, N., and S. Boyd. 2014. "Proximal Algorithms." *Foundations and Trends® in Optimization* 1, no. 3: 127–239.

Park, E., and I. Do Ha. 2018. "Penalized Variable Selection for Accelerated Failure Time Models." *Communications for Statistical Applications and Methods* 25, no. 6: 591–604. https://doi.org/10.29220/CSAM.2018.25.6.591.

Reeder, H. T., J. Lu, and S. Haneuse. 2023. "Penalized Estimation of Frailty-Based Illness–Death Models for Semi-Competing Risks." *Biometrics* 79: 1657–1669. https://doi.org/10.1111/biom.13761.

Ren, J., Y. Du, S. Li, S. Ma, Y. Jiang, and C. Wu. 2019. "Robust Network-Based Regularization and Variable Selection for High-Dimensional Genomic Data in Cancer Prognosis." *Genetic Epidemiology* 43, no. 3: 276–291. https://doi.org/10.1002/gepi.22194.

Simon, R. M., J. Subramanian, M. C. Li, and S. Menezes. 2011. "Using Cross-Validation to Evaluate Predictive Accuracy of Survival Risk Classifiers Based on High-Dimensional Data." *Briefings in Bioinformatics* 12, no. 3: 203–214. https://doi.org/10.1093/bib/bbr001.

Sha, N., M. G. Tadesse, and M. Vannucci. 2006. "Bayesian Variable Selection for the Analysis of Microarray Data With Censored Outcome." *Bioinformatics* 22, no. 18: 2262–2268. https://doi.org/10.1093/bioinformatics/btl362.

Suder, P. M., and A. J. Molstad. 2022. "Scalable Algorithms for Semiparametric Accelerated Failure Time Models in High-Dimensions." *Statistics in Medicine* 41, no. 6: 933–949. https://doi.org/10.1002/sim.9264.

Sun, H., W. Lin, R. Feng, and H. Li. 2014. "Network-Regularized High-Dimensional Cox Regression for Analysis of Genomic Data." *Statistica Sinica* 24, no. 3: 1433. https://doi.org/10.5705/ss.2012.317.

Tibshirani, R. 1996. "Regression Shrinkage and Selection Via the Lasso." *Journal of the Royal Statistical Society. Series B (Methodological)* 58, no. 1: 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

Tibshirani, R. 1997. "The Lasso Method for Variable Selection in the Cox Model." *Statistics in Medicine* 16, no. 4: 385–395. https://doi.org/10.1002/(sici)1097-0258(19970228)16:4⟨385::aid-sim380⟩3.0.co;2-3.

Veríssimo, A., A. L. Oliveira, M. F. Sagot, and S. Vinga. 2016. "DegreeCox – A Network-Based Regularization Method for Survival Analysis." *BMC Bioinformatics* 17: 109–121. https://doi.org/10.1186/s12859-016-1310-4.

Wainwright, M. J. 2019. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge Series in Statistical and Probabilistic Mathematics. New York: Cambridge University Press.

Wang, S., B. Nan, J. Zhu, and D. G. Beer. 2008. "Doubly Penalized Buckley-James Method for Survival Data with High-Dimensional Covariates." *Biometrics* 64, no. 1: 132–140. https://doi.org/10.1111/j.1541-0420.2007.00877.x.

Wang, S., B. Nan, N. Zhou, and J. Zhu. 2009. "Hierarchically Penalized Cox Regression With Grouped Variables." *Biometrika* 96, no. 2: 307–322. https://doi.org/10.1093/biomet/asp016.

Wang, D., T. T. Wu, and Y. Zhao. 2019. "Penalized Empirical Likelihood for the Sparse Cox Regression Model." *Journal of Statistical Planning and Inference* 201: 71–85. https://doi.org/10.1016/j.jspi.2018.12.001.

Wu, T. T., and S. Wang. 2013. "Doubly Regularized Cox Regression for High-Dimensional Survival Data With Group Structures." *Statistics and Its Interface* 6: 175–186. https://doi.org/10.4310/SII.2013.v6.n2.a2.

Zhang, H. H., and W. Lu. 2007. "Adaptive Lasso for Cox's Proportional Hazards Model." *Biometrika* 94, no. 3: 691–703. https://doi.org/10.1093/biomet/asm037.

Zhang, C. 2010. "Nearly Unbiased Variable Selection Under Minimax Concave Penalty." *The Annals of Statistics* 38, no. 2: 894–942. https://doi.org/10.1214/09-AOS729.

Zhang, W., T. Ota, V. Shridhar, J. Chien, B. Wu, and R. Kuang. 2013. "Network-Based Survival Analysis Reveals Subnetwork Signatures for Predicting Outcomes of Ovarian Cancer Treatment." *PLoS Computational Biology* 9, no. 3: e1002975. https://doi.org/10.1371/journal.pcbi.1002975.

Zou, H., and T. Hastie. 2005. "Regularization and Variable Selection Via the Elastic Net." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67, no. 2: 301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x.

Zou, H. 2006. "The Adaptive Lasso and its Oracle Properties." *Journal of the American Statistical Association* 101, no. 476: 1418–1429. https://doi.org/10.1198/016214506000000735.

Supporting Information

Additional supporting information can be found online in the Supporting Information section.

Appendix

To prove Theorem 3.1, we need the following results.

Proposition A.1 (cf Proposition 2.5, p. 24 in Wainwright (2019) (Hoeffding bound)). *Assume X_i, $i = 1, \dots, n$, sub-Gaussian random variables with mean μ_i and parameters σ_i, $i = 1, \dots, n$. Then, for all $t \geq 0$, we have*

$$\mathbb{P}\left[\left|\sum_{i=1}^{n}(X_i - \mu_i)\right| \geq t\right] \leq 2\exp\left(-\frac{t^2}{\sum_{i=1}^{n}\sigma_i^2}\right).$$

In particular, for bounded random variables, $X_i \in [a, b]$ such that $\sigma = \frac{b-a}{2}$, we have

$$\mathbb{P}\left[\left|\sum_{i=1}^{n}(X_i - \mu_i)\right| \geq t\right] \leq 2\exp\left(-\frac{2t^2}{n(b-a)^2}\right). \tag{A1}$$

Lemma A.1 (Restricted strong convexity (RSC)). *Under Assumptions 3.1–3.3, it holds: $\exists\gamma > 0$ and $\exists\tau > 0$ such that*

$$\begin{aligned}&\langle\nabla\ell^{(n)}(\theta^* + \Delta\theta) - \nabla\ell^{(n)}(\theta^*), \Delta\theta\rangle\\ &\geq \gamma\|\Delta\theta\|_2^2\\ &- \tau\sqrt{\frac{\log n(p+1)}{n}}\|\Delta\theta\|_1^2, \quad \textit{for all } \Delta\theta : \ \|\Delta\theta\|_1 \leq 2R,\end{aligned}$$

with high probability.

Proof of Lemma A.1. For the integral definition of the mean value theorem generalized to vector-valued functions for some $u \in [0,1]$, we have

$$\langle \nabla \ell^{(n)}(\theta^* + \Delta\theta) - \nabla \ell^{(n)}(\theta^*), \Delta\theta \rangle = \Delta\theta^T \int_0^1 \nabla^2 \ell^{(n)}(\theta^* + u\,\Delta\theta) du\, \Delta\theta,$$

for $\|\Delta\theta\|_1 \leq 2R$ and where the matrix integral is element-wise. The right-hand side of the equality can be decomposed as TERM1 + TERM2, where

$$\text{TERM1} = \Delta\theta^T \int_0^1 \mathbb{E}_{XY}\left(\nabla^2 \ell^{(n)}(\theta^* + u\,\Delta\theta)\right) du \Delta\theta, \tag{A2}$$

$$\text{TERM2} = \Delta\theta^T \left\{ \int_0^1 \nabla^2 \ell^{(n)}(\theta^* + u\,\Delta\theta) - \mathbb{E}_{XY} \left(\nabla^2 \ell^{(n)}(\theta^* + u\,\Delta\theta)\right) du \right\} \Delta\theta. \tag{A3}$$

Let us consider first Equation (A2). Note that $\|\theta^* - (\theta^* + u\Delta\theta)\|_1 = u\|\Delta\theta\|_1 \leq 2R$. Hence, it holds Assumption 3.3 and we have

$$\text{TERM1} = \Delta\theta^T \int_0^1 \mathbb{E}_{XY}\left(\nabla^2 \ell^{(n)}(\theta^* + u\,\Delta\theta)\right) du \Delta\theta \geq \gamma \Delta\theta^T \Delta\theta = \gamma \|\Delta\theta\|_2^2.$$

Now consider Equation (A3) and define

$$G^{\Delta\theta}(u) = \nabla^2 \ell^{(n)}(\theta^* + u\,\Delta\theta) - \mathbb{E}_{XY}\left(\nabla^2 \ell^{(n)}(\theta^* + u\,\Delta\theta)\right),$$

then, Equation (A3) became

$$\text{TERM2} = \Delta\theta^T \int_0^1 G^{\Delta\theta}(u) du \Delta\theta.$$

Therefore, we have by Hölder's inequality

$$\begin{aligned}
|\text{TERM2}| &= \left| \Delta\theta^T \int_0^1 G^{\Delta\theta}(u) du \Delta\theta \right| \leq \|\Delta\theta\|_1 \left\| \int_0^1 G^{\Delta\theta}(u) du \Delta\theta \right\|_\infty \\
&= \|\Delta\theta\|_1 \max_{1\leq k\leq p+1} \left| \sum_{j=1}^{p+1} \left(\int_0^1 G^{\Delta\theta}(u) du \right)_{kj} \Delta\theta_j \right| \\
&\leq \|\Delta\theta\|_1 \max_{1\leq k\leq p+1} \left\| \left(\int_0^1 G^{\Delta\theta}(u) du \right)_{k\cdot} \right\|_\infty \|\Delta\theta\|_1 \\
&= \|\Delta\theta\|_1^2 \max_{1\leq k,j\leq p+1} \left| \left(\int_0^1 G^{\Delta\theta}(u) du \right)_{kj} \right| \\
&\leq \|\Delta\theta\|_1^2 \max_{1\leq k,j\leq p+1} \left(\int_0^1 \left| G^{\Delta\theta}(u) \right| du \right)_{kj} \\
&\leq \|\Delta\theta\|_1^2 \sup_{u\in[0,1]} \max_{1\leq k,j\leq p+1} \left| G_{kj}^{\Delta\theta}(u) \right|.
\end{aligned}$$

Using an ε-net argument, we define a grid of points for u, that is, $u_m = m/n$, $m = 1, \ldots, n$. Then, $\forall u \in [0,1]$, , $\exists m : |u - u_m| \leq 1/n$ and

$$\begin{aligned}
|\text{TERM2}| \leq & \underbrace{\|\Delta\theta\|_1^2 \max_{1\leq m\leq n} \max_{1\leq k,j\leq p+1} \left| G_{kj}^{\Delta\theta}(u_m) \right|}_{\text{TERM2}A} \\
& + \underbrace{\|\Delta\theta\|_1^2 \sup_{|u-u_m|\leq 1/n} \max_{1\leq k,j\leq p+1} \left| G_{kj}^{\Delta\theta}(u) - G_{kj}^{\Delta\theta}(u_m) \right|}_{\text{TERM2}B}
\end{aligned}$$

We consider first TERM2B. Note that $G_{kj}^{\Delta\theta}(u)$ is continuous in $u \in [0,1]$, $\forall k, j$ and therefore locally Lipschitz over [0,1]. So, we have

$$|\text{TERM2}B| \leq C \sup_{|u-u_m|\leq 1/n} |u - u_m|\, \|\Delta\theta\|_1^2 = \frac{C}{n} \|\Delta\theta\|_1^2,$$

where C is a Lipschitz constant. We use the Hoeffding inequality for TERM2A. Let us consider for all m, k, j fixed

$$\begin{aligned}
\left| G_{kj}^{\Delta\theta}(u_m) \right| &= \left| \left\{ \nabla^2 \ell^{(n)}(\theta^* + u_m\,\Delta\theta) - \mathbb{E}_{XY}\left(\nabla^2 \ell^{(n)}(\theta^* + u_m\,\Delta\theta)\right) \right\}_{kj} \right| \\
&= \left| -\frac{1}{n} \left\{ \sum_{i=1}^n \left[\nabla^2 \ell_i(\theta^* + u_m\,\Delta\theta) - \mathbb{E}_{XY}\left(\nabla^2 \ell_i(\theta^* + u_m\,\Delta\theta)\right) \right] \right\}_{kj} \right| \\
&= \left| \frac{1}{n} \sum_{i=1}^n [Z_i - \mathbb{E}(Z_i)] \right|,
\end{aligned}$$

with $Z_i = \left[\nabla^2 \ell_i(\theta^* + u_m\,\Delta\theta)\right]_{kj}$ bounded and independent random variables, implying Z_i sub-Gaussian variables. Now, we can use Equation (A1)

$$\mathbb{P}\left(\left| G_{kj}^{\Delta\theta}(u_m) \right| \geq t \right) = \mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^n [Z_i - \mathbb{E}(Z_i)] \right| \geq t \right) \leq 2\exp\left(-\frac{nt^2}{c} \right),$$

where c is the sub-Gaussian parameter. Therefore, from

$$\text{TERM2}A = \|\Delta\theta\|_1^2 \max_{1\leq m\leq n} \max_{1\leq k,j\leq p+1} \left| G_{kj}^{\Delta\theta}(u_m) \right|,$$

it follows

$$\begin{aligned}
\mathbb{P}\left(\max_{1\leq m\leq n} \max_{1\leq k,j\leq p+1} \left| G_{kj}^{\Delta\theta}(u_m) \right| \geq t \right) &= \mathbb{P}\left(\bigcup_{m=1}^n \bigcup_{k.j=1}^{p+1} \left| G_{kj}^{\Delta\theta}(u_m) \right| \geq t \right) \\
&\leq \sum_{m=1}^n \sum_{k.j=1}^{p+1} \mathbb{P}\left(\left| G_{kj}^{\Delta\theta}(u_m) \right| \geq t \right) \\
&\leq \sum_{m=1}^n \sum_{k.j=1}^{p+1} 2\exp\left(-\frac{nt^2}{c} \right) \\
&= 2n(p+1)^2 \exp\left(-\frac{nt^2}{c} \right).
\end{aligned}$$

We define $\varepsilon = 2n(p+1)^2 \exp\left(-\frac{nt^2}{c}\right)$. Thus, we get

$$\exp\left(-\frac{nt^2}{c} \right) = \frac{\varepsilon}{2n(p+1)^2} \quad \Rightarrow \quad t = \sqrt{\left[\log \frac{2n(p+1)^2}{\varepsilon} \right] \frac{c}{n}}.$$

Therefore, for all $\varepsilon > 0$,

$$\mathbb{P}\left(\text{TERM2}A \leq \sqrt{\left[\log \frac{2n(p+1)^2}{\varepsilon} \right] \frac{c}{n}} \right) \geq 1 - \varepsilon,$$

hence,

$$\begin{aligned}
|\text{TERM2}| &\leq \|\Delta\theta\|_1^2 \frac{C}{n} + \|\Delta\theta\|_1^2 \sqrt{\left[\log \frac{2n(p+1)^2}{\varepsilon} \right] \frac{c}{n}} \\
&\leq \|\Delta\theta\|_1^2 \left\{ \frac{C}{n} + \sqrt{\frac{c \log(2(p+1)/\varepsilon)}{n}} + \sqrt{\frac{c \log(n(p+1))}{n}} \right\}
\end{aligned}$$

with a probability of at least $1-\varepsilon$. Finally, combining all terms, it follows that it exists $\tau > 0$, such that

$$\langle \nabla \ell^{(n)}(\boldsymbol{\theta}^* + \Delta\boldsymbol{\theta}) - \nabla \ell^{(n)}(\boldsymbol{\theta}^*), \Delta\boldsymbol{\theta} \rangle = \text{TERM1} + \text{TERM2}$$
$$\geq \gamma \|\Delta\boldsymbol{\theta}\|_2^2 - \tau \|\Delta\boldsymbol{\theta}\|_1^2 \sqrt{\frac{\log n(p+1)}{n}},$$

with a probability of at least $1-\varepsilon$. □

Lemma A.2 (Bounded gradient). *Under Assumptions 3.1–3.3, it holds*

$$\max\left\{ \|\nabla \ell^{(n)}(\boldsymbol{\theta}^*)\|_\infty, \tau R \sqrt{\frac{\log n(p+1)}{n}} \right\} \leq \frac{\lambda\alpha}{4}, \tag{A4}$$

with high probability and suitable choice of $\lambda > 0$.

Proof of Lemma A.2. To obtain Equation (A4) is sufficient to prove that

$$\left\| \nabla \ell^{(n)}(\boldsymbol{\theta}^*) \right\|_\infty = O_p\left(\sqrt{\frac{\log(n(p+1))}{n}} \right). \tag{A5}$$

From Equation (17), for all $j = 1, \dots, p+1$ and $t \geq 0$, we have

$$\mathbb{P}\left(\left| \frac{\partial \ell^{(n)}(\boldsymbol{\theta}^*)}{\partial \theta_j} \right| \geq t \right) = \mathbb{P}\left(\left| -\frac{1}{n} \sum_{i=1}^n \frac{\partial \ell_i(\boldsymbol{\theta}^*)}{\partial \theta_j} \right| \geq t \right) = \mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^n Z_i \right| \geq t \right),$$

with $Z_i = \frac{\partial \ell_i(\boldsymbol{\theta}^*)}{\partial \theta_j}$ sub-Gaussian random variables with zero mean being $\mathbb{E}_{XY}(\nabla \ell^{(n)}(\boldsymbol{\theta}^*)) = 0$. Using Equation (A1), we have

$$\mathbb{P}\left(\left| \frac{\partial \ell^{(n)}(\boldsymbol{\theta}^*)}{\partial \theta_j} \right| \geq t \right) \leq 2\exp\left(-\frac{nt^2}{c} \right),$$

where c is the sub-Gaussian parameter. Hence, we get

$$\begin{aligned}
\mathbb{P}\left(\|\nabla \ell^{(n)}(\boldsymbol{\theta}^*)\|_\infty \geq t \right) &= \mathbb{P}\left(\max_{1\leq j\leq p+1} \left| \frac{\partial \ell^{(n)}(\boldsymbol{\theta}^*)}{\partial \theta_j} \right| \geq t \right) \\
&= \mathbb{P}\left(\bigcup_{1\leq j\leq p+1} \left| \frac{\partial \ell^{(n)}(\boldsymbol{\theta}^*)}{\partial \theta_j} \right| \geq t \right) \\
&\leq \sum_{j=1}^{p+1} \mathbb{P}\left(\left| \frac{\partial \ell^{(n)}(\boldsymbol{\theta}^*)}{\partial \theta_j} \right| \geq t \right) \leq \sum_{j=1}^{p+1} 2\exp\left(-\frac{nt^2}{c} \right) \\
&= 2(p+1)\exp\left(-\frac{nt^2}{c} \right).
\end{aligned}$$

We define $\varepsilon = 2(p+1)\exp\left(-\frac{nt^2}{c}\right)$, thus, we get

$$\exp\left(-\frac{nt^2}{c} \right) = \frac{\varepsilon}{2(p+1)} \quad \Rightarrow \quad t = \sqrt{\frac{c}{n} \log\left(\frac{2(p+1)}{\varepsilon} \right)}.$$

This implies that

$$\mathbb{P}\left(\|\nabla \ell^{(n)}(\boldsymbol{\theta}^*)\|_\infty \leq \sqrt{\frac{c}{n} \log\left(\frac{2(p+1)}{\varepsilon} \right)} \right) \geq 1-\varepsilon,$$

consequently, it holds Equation (A5). □

Now we are able to demonstrate Theorem 3.1.

Proof of Theorem 3.1. Define the error vector $\boldsymbol{\nu} = \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}^*$. Under Assumption 3.2, we have $\|\boldsymbol{\nu}\|_1 \leq \|\hat{\boldsymbol{\theta}}\|_1 + \|\boldsymbol{\theta}^*\|_1 \leq 2R$; hence, by Lemma A.1, it holds the RSC

$$\langle \nabla \ell^{(n)}(\hat{\boldsymbol{\theta}}) - \nabla \ell^{(n)}(\boldsymbol{\theta}^*), \boldsymbol{\nu} \rangle \geq \gamma \|\boldsymbol{\nu}\|_2^2 - \tau \sqrt{\frac{\log n(p+1)}{n}} \|\boldsymbol{\nu}\|_1^2,$$

from which follows that

$$\gamma \|\boldsymbol{\nu}\|_2^2 \leq \langle \nabla \ell^{(n)}(\hat{\boldsymbol{\theta}}) - \nabla \ell^{(n)}(\hat{\boldsymbol{\theta}}^*), \boldsymbol{\nu} \rangle + \tau \sqrt{\frac{\log n(p+1)}{n}} \|\boldsymbol{\nu}\|_1^2. \tag{A6}$$

Since $\|\boldsymbol{\theta}^*\|_1 \leq R$, by using the first-order condition for $\hat{\boldsymbol{\theta}}$, we have

$$\langle \nabla \ell^{(n)}(\hat{\boldsymbol{\theta}}) + \nabla \mathcal{P}_{\lambda,\alpha}(\hat{\boldsymbol{\theta}}), \boldsymbol{\theta}^* - \hat{\boldsymbol{\theta}} \rangle \geq 0$$

(note that the penalty is only on $\boldsymbol{\beta}$, so we have implicitly assumed that the last element of $\mathcal{P}_\lambda(\hat{\boldsymbol{\theta}})$ is zero), that is,

$$\langle \nabla \ell^{(n)}(\hat{\boldsymbol{\theta}}) + \nabla \mathcal{P}_{\lambda,\alpha}(\hat{\boldsymbol{\theta}}), -\boldsymbol{\nu} \rangle \geq 0, \quad \Rightarrow \quad \langle \nabla \ell^{(n)}(\hat{\boldsymbol{\theta}}), \boldsymbol{\nu} \rangle \leq \langle \nabla \mathcal{P}_{\lambda,\alpha}(\hat{\boldsymbol{\theta}}), -\boldsymbol{\nu} \rangle,$$

and using this inequality in Equation (A6), we have

$$\begin{aligned}
\gamma \|\boldsymbol{\nu}\|_2^2 &\leq \langle \nabla \ell^{(n)}(\hat{\boldsymbol{\theta}}), \boldsymbol{\nu} \rangle - \langle \nabla \ell^{(n)}(\boldsymbol{\theta}^*), \boldsymbol{\nu} \rangle + \tau \sqrt{\frac{\log n(p+1)}{n}} \|\boldsymbol{\nu}\|_1^2 \\
&\leq \langle \nabla \mathcal{P}_{\lambda,\alpha}(\hat{\boldsymbol{\theta}}), -\boldsymbol{\nu} \rangle - \langle \nabla \ell^{(n)}(\boldsymbol{\theta}^*), \boldsymbol{\nu} \rangle + \tau \sqrt{\frac{\log n(p+1)}{n}} \|\boldsymbol{\nu}\|_1^2 \\
&= \langle \lambda\alpha \nabla \|\hat{\boldsymbol{\beta}}\|_1 + \lambda(1-\alpha)\nabla \|\boldsymbol{L}^{\frac{1}{2}}\hat{\boldsymbol{\beta}}\|_2^2, \boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}} \rangle - \langle \nabla \ell^{(n)}(\boldsymbol{\theta}^*), \boldsymbol{\nu} \rangle \\
&\quad + \tau \sqrt{\frac{\log n(p+1)}{n}} \|\boldsymbol{\nu}\|_1^2 \\
&= \langle \lambda\alpha \nabla \|\hat{\boldsymbol{\beta}}\|_1, \boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}} \rangle + \langle 2\lambda(1-\alpha)\boldsymbol{L}\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}} \rangle - \langle \nabla \ell^{(n)}(\boldsymbol{\theta}^*), \boldsymbol{\nu} \rangle \\
&\quad + \tau \sqrt{\frac{\log n(p+1)}{n}} \|\boldsymbol{\nu}\|_1^2 \\
&\leq \langle \lambda\alpha \nabla \|\hat{\boldsymbol{\beta}}\|_1, \boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}} \rangle + |\langle 2\lambda(1-\alpha)\boldsymbol{L}\hat{\boldsymbol{\beta}}, \boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}} \rangle| - \langle \nabla \ell^{(n)}(\boldsymbol{\theta}^*), \boldsymbol{\nu} \rangle \\
&\quad + \tau \sqrt{\frac{\log n(p+1)}{n}} \|\boldsymbol{\nu}\|_1^2.
\end{aligned}$$

Now, since the LASSO penalty is convex, using Cauchy–Schwarz inequality and Hölder inequality, we have

$$\begin{aligned}
\gamma \|\boldsymbol{\nu}\|_2^2 &\leq \lambda\alpha \|\boldsymbol{\beta}^*\|_1 - \lambda\alpha \|\hat{\boldsymbol{\beta}}\|_1 + 2\lambda(1-\alpha)\|\boldsymbol{L}\hat{\boldsymbol{\beta}}\|_2 \|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}\|_2 \\
&\quad - \langle \nabla \ell^{(n)}(\boldsymbol{\theta}^*), \boldsymbol{\nu} \rangle + \tau \sqrt{\frac{\log n(p+1)}{n}} \|\boldsymbol{\nu}\|_1^2 \\
&\leq \lambda\alpha \|\boldsymbol{\beta}^*\|_1 - \lambda\alpha \|\hat{\boldsymbol{\beta}}\|_1 + 2\lambda(1-\alpha)\lambda_{max}(\boldsymbol{L})\|\hat{\boldsymbol{\beta}}\|_2 \|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}\|_2 \\
&\quad + |\langle \nabla \ell^{(n)}(\boldsymbol{\theta}^*), \boldsymbol{\nu} \rangle| + \tau \sqrt{\frac{\log n(p+1)}{n}} \|\boldsymbol{\nu}\|_1^2 \\
&\leq \lambda\alpha \|\boldsymbol{\beta}^*\|_1 - \lambda\alpha \|\hat{\boldsymbol{\beta}}\|_1 + 2\lambda(1-\alpha)\lambda_{max}(\boldsymbol{L})\, R \|\boldsymbol{\beta}^* - \hat{\boldsymbol{\beta}}\|_2 \\
&\quad + \|\nabla \ell^{(n)}(\boldsymbol{\theta}^*)\|_\infty \|\boldsymbol{\nu}\|_1 + \tau \sqrt{\frac{\log n(p+1)}{n}} \|\boldsymbol{\nu}\|_1^2, \\
&\leq \lambda\alpha \|\boldsymbol{\beta}^*\|_1 - \lambda\alpha \|\hat{\boldsymbol{\beta}}\|_1 + 2\lambda(1-\alpha)\lambda_{max}(\boldsymbol{L})\, R \|\boldsymbol{\nu}\|_2 \\
&\quad + \|\boldsymbol{\nu}\|_1 \left[\|\nabla \ell^{(n)}(\boldsymbol{\theta}^*)\|_\infty + \tau R \sqrt{\frac{\log n(p+1)}{n}} \right],
\end{aligned}$$

where $\lambda_{max}(L)$ is the maximum eigenvalue of $\boldsymbol{L}$. Now using Lemma A.2

$$\begin{aligned}
\gamma \|\boldsymbol{\nu}\|_2^2 &\leq \lambda\alpha \|\boldsymbol{\beta}^*\|_1 - \lambda\alpha \|\hat{\boldsymbol{\beta}}\|_1 + 2\lambda(1-\alpha)\lambda_{max}(\boldsymbol{L})\, R \|\boldsymbol{\nu}\|_2 + \frac{\lambda\alpha}{2}\|\boldsymbol{\nu}\|_1 \\
&\leq \frac{3}{2}\lambda\alpha \|\boldsymbol{\beta}^*\|_1 - \frac{\lambda\alpha}{2}\|\hat{\boldsymbol{\beta}}\|_1 + \frac{\lambda\alpha}{2}|\hat{\sigma} - \sigma^*| + 2\lambda(1-\alpha)\lambda_{max}(\boldsymbol{L})\, R \|\boldsymbol{\nu}\|_2.
\end{aligned} \tag{A7}$$

In case $\frac{3}{2}\lambda\alpha\|\boldsymbol{\beta}^*\|_1 - \frac{\lambda\alpha}{2}\|\hat{\boldsymbol{\beta}}\|_1 > 0$, by using Lemma 5 in Appendix A.1 of Loh and Wainwright (2015) with $\xi = 3$, we obtain

$$\frac{3}{2}\lambda\alpha\|\boldsymbol{\beta}^*\|_1 - \frac{\lambda\alpha}{2}\|\hat{\boldsymbol{\beta}}\|_1 \leq \frac{\lambda\alpha}{2}\left(3\|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_A\|_1 - \|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{A^c}\|_1\right),$$

where A is the set of the s largest elements of $\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*$ in magnitude and A_c is the complementary set, s, being defined in Equation (18).

Then, we have

$$\begin{aligned}
\gamma\|\boldsymbol{\nu}\|_2^2 &\leq \frac{\lambda\alpha}{2}\left[3\|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_A\|_1 - \|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_{A^c}\|_1,\right] + \frac{3}{2}\lambda\alpha|\hat{\sigma} - \sigma^*| \\
&\quad + 2\lambda(1-\alpha)\lambda_{max}(\boldsymbol{L})\,R\,\|\boldsymbol{\nu}\|_2 \\
&\leq \frac{3}{2}\lambda\alpha\|(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)_A\|_1 + \frac{3}{2}\lambda\alpha|\hat{\sigma} - \sigma^*| + 2\lambda(1-\alpha)\lambda_{max}(\boldsymbol{L})\,R\,\|\boldsymbol{\nu}\|_2 \\
&\leq \frac{3}{2}\lambda\alpha\,\|\boldsymbol{\nu}_A\|_1 + 2\lambda(1-\alpha)\lambda_{max}(\boldsymbol{L})\,R\,\|\boldsymbol{\nu}\|_2 \\
&\leq \frac{3}{2}\lambda\alpha\sqrt{s+1}\,\|\boldsymbol{\nu}\|_2 + 2\lambda(1-\alpha)\lambda_{max}(\boldsymbol{L})\,R\,\|\boldsymbol{\nu}\|_2
\end{aligned}$$

from which we finally obtain

$$\|\boldsymbol{\nu}\|_2 \leq \frac{\frac{3}{2}\lambda\alpha\sqrt{s+1} + 2\lambda(1-\alpha)\lambda_{max}(\boldsymbol{L})\,R}{\gamma}.$$

In case $\frac{3}{2}\lambda\alpha\|\boldsymbol{\beta}^*\|_1 - \frac{\lambda\alpha}{2}\|\hat{\boldsymbol{\beta}}\|_1 \leq 0$, the bound follows easily from Equation (A7). □