Research Article

A Flexible and Robust Approach to Analyze Survival Systems in the Presence of Extreme Observations

Maryam Sadiq ¹ and Tahir Mehmood ²

¹Department of Statistics, University of Azad Jammu and Kashmir, Muzaaffarabad, Pakistan

²School of Natural Sciences (SNS), National University of Sciences and Technology (NUST), Islamabad, Pakistan

Correspondence should be addressed to Maryam Sadiq; hussainulahmad@gmail.com

Received 26 March 2021; Accepted 7 May 2021; Published 19 May 2021

Academic Editor: Ishfaq Ahmad

Copyright © 2021 Maryam Sadiq and Tahir Mehmood. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Survival systems are difficult to analyze in the presence of extreme observations and multicollinearity. Finding appropriate models that provide a robust description of such survival systems and that address the smooth hazards in the context of covariates can be challenging given the sheer number of possibilities. Survival time algorithms that evaluate the efficiency of models in the presence of extreme observations over different datasets provide an effective tool to identify robust systems. However, the existing algorithms addressing the analysis of survival systems are limited in long-term evaluations. Therefore, an algorithm that can analyze survival time response on high-dimensional complex survival systems having extreme observations is developed which explores large margins dynamically. This algorithm is developed as a conjugate of flexible parametric models and partial least squares to estimate smooth, flexible, and robust functions to extrapolate the survival model in long-term evaluations in the presence of extreme observations. The algorithm is tested and validated using four distributions based on a simulated dataset generated from the Weibull distribution and compared with partial least squares-Cox regression. The comparison shows its flexibility and efficiency in handling different survival systems in the presence of extreme values. The algorithm is also used to analyze four real datasets of breast cancer survival time, each containing seven gene signatures. The coefficients of significant genes for each dataset are estimated. The flexibility in handling various distributions as parametric survival models supports the application of the algorithm to a large variety of different survival problems and represents a robust statistical framework for survival analysis in the presence of extreme observations.

1. Introduction

“Time-to-event” responses have become progressively relevant in the context of research studies, where the response of interest could be based on time to remission, overall survival [1], or progression-free survival [2] following treatment intervention. The Kaplan–Meier product-limit method [3] is the popular nonparametric approach to analyze time to event response. Cox’s proportional hazards (PH) regression model [4] is a widely used semiparametric survival approach when the proportional hazards assumption is valid [5].

Despite the popularity and advantages of nonparametric and semiparametric methods, parametric modeling procedures offer flexible, efficient, and robust alternatives. When

the proportional hazards assumption is violated and the distributional assumption on the survival times is valid, flexible parametric models (FPM) are an ample alternate of nonparametric and semiparametric methods resulting in more efficient estimates having smaller standard errors even in the presence of extreme observations. Also, statistical inferences are more precise as the full likelihood is used and results are easy to interpret. To approximate survival data, numerous theoretical distributions have been employed so far in FPM. The exponential distribution assists as the primary model for survival time. Other commonly used distributions to modeled big survival data with extreme observations are Weibull, gamma, Gompertz, log-normal, log-logistic, generalized gamma, and generalized F-distribution.

Also, the FPM can efficiently incorporate covariates to investigate the dependence of survival response in the context of estimated parameters, survival function, and cumulative hazard function [6–10]. Multivariate regression models assume that there is no multicollinearity among covariates of survival systems. This assumption, however, is often violated specifically in the case of the curse of dimensionality which analyses a large number of covariates with limited observations. Therefore, nonparametric and semiparametric survival methods are no longer appropriate to modeled large data with correlated covariates and extreme observations. Therefore, within the framework of partial least squares (PLS) regression, the partial least squares-Cox regression (PLS-CoxR) model was developed. The basic PLS algorithm transforms the original correlated covariates of the system into uncorrelated components [11, 12]. Hence, the PLS-CoxR model is used as a conjugate of Cox and PLS regression to model survival systems [13]. A major limitation of the nonparametric survival methods is that the determinants cannot be integrated into the analyses. Regarding the disadvantages of semiparametric survival methods, Cox regression is limited in the long-term evaluation as the cumulative hazard functions and survival functions cannot be extrapolated in the model as being incomplete. Because in long-term survival studies, the duration of the survey is often not sufficient to collect all the observations of every event that occurs. Another weakness of the Cox regression model is that it estimates survival function and cumulative hazards function as step functions. Hence, the PLS-CoxR model is limited and inflexible with unsmooth estimates. Alternatively, flexible parametric models (FPM) can be used to estimate smooth hazard ratios of predictors and corresponding cumulative hazard functions and hence to extrapolate the survival model. However, to use these models, the hazard function and survival function must follow a specific distribution. The FPM can estimate a continuous function instead of a step function due to its flexibility. Additionally, more complicated shapes can be modeled without sacrificing prediction accuracy and appropriate convergence is achieved. The generalization of several standard survival models is used as a series of models in FPM [14].

The motivation of this research was to develop a robust model that is specifically designed for the analysis of survival systems in the presence of extreme observations and is also able to handle various probability distributions. This robust algorithm, namely, Partial Least Squares Flexible parametric Models (PLS-FPM) supports the user defines its probability distribution, for estimating the survival, hazard, and cumulative hazard functions and returning a calculated model selection criterion.

After validating and testing the performance and efficiency of PLS-FPM using simulated data, as well as showing its flexibility to handle different distributions, the algorithm is applied to the analysis of four breast cancer survival datasets, and significant genes are estimated. The analyses based on different distributions using several datasets revealed the robustness of these models to estimate smooth survival and cumulative hazard functions in the presence of extreme values.

2. Materials and Methods

The Cox proportional hazards (PH) model is the most frequently used regression technique to address survival data. In the presence of multicollinearity, the Cox algorithm is integrated with PLS resulting in the PLS-Cox model.

2.1. The Partial Least Squares Cox (PLS-Cox) Regression Model. The PLS-Cox regression model is used as a reference model in this study. Let t represent survival time and let $z_j = z_{1j}, z_{2j}, \dots, z_{nj}$ be the vector of p correlated explanatory factors over n samples. The PLS model computes k latent components for p factors then the baseline hazard function takes the form of

$$\lambda(t) = \lambda_o(t) \exp(\beta_1 L_1 + \beta_2 L_2 + \dots + \beta_p L_k) = \lambda_o(t) \exp[\beta' L], \quad (1)$$

where $\lambda_o(t)$ is the unspecified baseline hazard function, β is the vector of coefficients, and L is a $(n \times k)$ matrix of components.

2.2. Flexible Parametric Survival Model (FPSM). Let T represent a survival time random factor and let Z denote the vector of covariates $z_1, \dots, z_p$ for a sample of size n . Let T have the probability density function $f(t)$ and cumulative distribution function $F(t) = \Pr(T \leq t)$. The survival function $S(t) = \Pr(T > t)$ is the probability of being alive at time t . The hazard function is $\lambda(t) = f(t)/S(t)$ and the risk function is

$$\Lambda(t) = \int_0^t \lambda(x) dx. \quad (2)$$

Any distribution defined for $t \in [0, \infty]$ can serve as a parametric function. In this study, the distributions integrated with PLS are as follows.

2.3. The Gamma Distribution. The inclusion of incomplete gamma integral in the survival and hazard functions of gamma distribution limited its use in survival analysis. A survival time random variable T is gamma distributed with shape parameter k and rate parameter β ; then, the survival function is

$$S(t) = 1 - I_k(\beta t), \quad (3)$$

where $I_k(t)$ is the incomplete gamma function and the gamma hazard function is

$$\Lambda(t) = \frac{\beta^k t^{k-1} e^{-\beta t}}{(1 - I_k(\beta t)) \Gamma(k)}. \quad (4)$$

Complicated numerical calculations are required to estimate parameters as maximum likelihood estimation is difficult to exercise due to incomplete gamma integrals. The gamma hazard function may be constant, monotonically increasing, or monotonically decreasing.

2.4. The Weibull Distribution. Let the survival time T be Weibull with scale parameter λ and shape parameter k ; then, the survival function is

$$S(t) = \exp(-\lambda t^k). \quad (5)$$

Transforming survival function to the log cumulative hazard scale, FPM is formulated as the linear function of log time:

$$\Lambda(t) = \ln\{H(t)\} = \ln[-\ln\{S(t)\}] = \ln(\lambda) + k \ln(t). \quad (6)$$

Adding covariates, the log cumulative hazard model becomes

$$\Lambda(t) = \ln(\lambda) + k \ln(t) + x_i \beta. \quad (7)$$

Thus, $\ln(\lambda) + k \ln(t)$ in the baseline log cumulative hazard function is combined with covariates. The Weibull distribution has constant, increasing, or decreasing hazard rates.

2.5. The Log-Logistic Distribution. For T following the log-logistic distribution with scale parameter α and shape parameter β , the survival function is

$$S(t) = (1 + \alpha t^\beta)^{-1}, \quad (8)$$

and the cumulative hazard function is

$$\Lambda(t) = \ln(1 + \alpha t^\beta). \quad (9)$$

The hazard function of log-logistic distribution may be increasing, decreasing, or hump-shaped.

2.6. Log-Normal Distribution. The natural logarithm of the lifetime T is assumed to be normally distributed with location parameters μ and scale parameter σ . The log-normal survival function is

$$S(t) = 1 - \Phi\left(\frac{\ln(t)}{\sigma}\right). \quad (10)$$

The cumulative hazard function of the lognormal distribution is

$$\Lambda(t) = -\ln\left(1 - \Phi\left(\frac{\ln(t)}{\sigma}\right)\right), \quad (11)$$

and Φ is the cumulative distribution function of the normal distribution. The log-normal hazard function is monotonically decreasing [15].

Various other standard and defined distributions including extreme value distributions can be used in FPM. In these models, proportional hazards imply proportional cumulative hazards; hence, covariates can be interpreted as hazard ratios similar to nonparametric models under PH assumption. The cumulative hazard, in FPM, is a more stable function compared to nonparametric models as being a function of a log time scale. For instance, the cumulative hazard function is a straight line in Weibull models. Thus, more stable-shaped functions are accurately captured. The PLSR model integrated with FPM using gamma, Weibull,

log-logistic, and log-normal distribution is introduced in this study for robust estimation in the context of high dimensional survival data in the presence of extreme values.

2.7. Partial Least-Squares Flexible Parametric (PLS-FP) Survival Regression Algorithm. Suppose Z denote the matrix of p correlated covariates $z_1, \dots, z_p$ for a sample of size n . The algorithm executes the FP model based on the K components (as $k \leq p$) of PLSR computed with time as a response variable and Z as a matrix of covariates. Figure 1 expresses the main steps of the proposed model.

The PLS-FP model works in two steps. It computes PLS components at the first step and executes FP distribution at the second step. This proposed model enhances model performance in terms of prediction and accuracy in the presence of extreme observations.

Generation of simulated data for survival response from standard parametric distributions.

The `simsurv` R package is used to generate simulated data to compare the performance of existing and proposed PLS-based models. The simulated dataset is generated from the Weibull distribution for scale parameter ($\lambda = 0.1$) and shape parameter ($k = 1.5$) over 5 years of censoring with extreme observations. The correlation structure between covariates is introduced as $R = [0.5, 0.5, 0.3, 0.5, 0, 0, 0, 0, 0]$ in the datasets over 100 samples with 100 covariates.

2.8. Breast Cancer Datasets. The 10-year censored survival time datasets of breast cancer patients used in this study contain the seven-gene signature innovated by [16]. The seven genes included in the data are AURKA representing proliferation; PLAU representing tumor metastasis; STAT1 representing immune response; VEGF representing angiogenesis; CASP3 representing apoptosis phenotypes; ESR1 representing estrogen receptor (ER) signaling; and ERBB2 representing human epidermal growth factor receptor 2 (HER2) signaling. The clinical data from four different datasets each with a subset of the gene expression and annotations in the presence of extreme observations are analyzed. The following four datasets were used to compare the models: mainz7g (200 observations) [17], transbig7g (198 observations) [18], vdx7g (344 observations) [19, 20], and upp7g (234 observations) [21]. The breast cancer datasets are freely available in an R package called `survcomp` and are available from <http://www.ulb.ac.be/di/map/bhaiibeka/software/survcomp/>. [22].

3. Results

The PLS-FP model parameterised with gamma, Weibull, log-logistic, and log-normal distributions are fitted over simulated dataset generated from Weibull distributions for comparison of five models with 100 correlated covariates. The results supported the application of proposed models over the traditional PLS-Cox model to deal with survival time response in the presence of collinear covariates and extreme observations. The model performances based on AIC and BIC for simulated data generated from the Weibull distribution are presented in Figure 2. Figure 2 shows a box

FIGURE 1: Flow diagram of survival analysis approaches.

FIGURE 2: The model comparison of standard PLS-Cox regression model with PLS-FPM parameterised over gamma, Weibull, log-logistic, and log-normal distribution by using simulated survival times data generated from Weibull distribution is presented.

plot comparison of models which is drawn in R package, namely, lattice. The Weibull, gamma, log-logistic, and log-normal distributions are fitted and compared to examine the difference in performance in the presence of correlated covariates and extreme observations. Figure 2 shows performances of five models based on AIC and BIC indicating that incorporated with PLS, the Weibull survival model has the highest performance for simulated data with known Weibull distribution having defined correlated covariates and extreme observations. More importantly, all proposed methods conjugated with PLSR performed more efficiently than PLSR conjugated with Cox regression. The results of simulated data demonstrate that the proposed methods are accurate and robust as showing the highest performance using the corresponding distribution. The application of simulated data generated from the Weibull distribution recommended the practical implementation of the proposed

methods to analyze high-dimensional survival time data in the presence of multicollinearity and extreme observations.

Before analyzing real datasets, multicollinearity among covariates and the presence of extreme observations is verified. For this purpose, correlations among covariates for all breast cancer datasets are plotted. The correlation maps for breast cancer datasets are presented in Figures 3 and 4 portray the high correlations between covariates for all datasets. Figure 3 presents the circle of correlations for mainz7g and transbig7g datasets. The correlation maps showed that gene symbols VEGFA and PLAUI represented by the probe names 211527_x_at and 211668_s_at, respectively, for the Affymetrix microarray are highly correlated for both datasets. This shows that the vascular endothelial growth factor A (VEGF-A) and Plasminogen Activator, Urokinase (PLAU) are highly correlated covariates in both datasets. Figure 4 visually displays the correlation strength

FIGURE 3: The circles of correlations for breast cancer dataset, namely, mainz7g and transbig7g, are presented.

FIGURE 4: The circles of correlations for breast cancer dataset, namely, vdx7g and upp7g, are presented.

among covariates of two datasets, namely, vdx7g and upp7g. The correlation plot shows that similar to mainz7g and transbig7g, the gene symbols Vegfa and Plau are highly correlated genes for the vdx7g dataset. The correlation maps showed that gene symbols AURKA and ERBB2 represented by the probe names 208079_s_at and 216836_s_at, respectively, for the Affymetrix microarray are highly correlated for the upp7g dataset. This means that the Aurora Kinase A (Aurka) and Erb-B2 Receptor Tyrosine Kinase 2 (ERBB2) have a high correlation for the upp7g dataset. Furthermore, gene symbol STAT1 and VEGFA represented by the probe names 209969_s_at and 211527_x_at, respectively, also showed correlation for upp7g dataset presenting association between signal transducer and activator of transcription 1 and vascular endothelial growth factor A (VEGF-A).

The correlation maps evidence the presence of multicollinearity. Moreover, for the identification of extreme observations, starburst graphs (also called bagplot) are plotted, presented in Figures 5 and 6. The starburst plot is usually used as a robust method to visualize two or three-

dimensional data. Detection of extreme observations is performed separately for each dataset. Figure 5 shows the bagplots for two breast cancer datasets, namely, mainz7g and transbig7g. The principal components are computed for the matrix of covariates and two-dimensional bagplots are constructed. In Figure 5, 16 outlying observations are detected for the mainz7g dataset and 26 outliers appeared for transbig7g data. Figure 6 represents the starburst plot for vdx7g and upp7g datasets. For vdx7g, 7 extreme observations are detected and 16 outliers are identified. The inner bags of the plot contain 50% of the observations. Hence, the presence of multicollinearity among covariates and extreme observations are detected visually. Regarding the parametric approach, the gamma, log-normal, and Weibull distributions are included addressing monotonically increasing, monotonically decreasing, arc-shaped, and bathtub hazards. Hence, incorporated with PLS, this family of distributions increases the efficiency of the model. The log-logistic distribution is included addressing only arc-shaped and decreasing hazards but still showed higher performance

FIGURE 5: The starburst plot for breast cancer dataset, namely, mainz7g and transbig7g, is presented.

FIGURE 6: The starburst plot for breast cancer dataset, namely, vdx7g and upp7g, is presented.

compared to PLS-Cox. For authentication of proposed methods, four gene signature breast cancer datasets are used to compare the model's accuracy presented in Figures 7 and 8. Each dataset is measured over seven genes and is randomly split into test and train components to increase reliability. The collinearity among genes is examined and verified. The existence of extreme observations is also detected. Then, the PLS-FPM parameterised with Weibull, gamma, log-logistic, and log-normal distribution is considered for all four datasets and compared with the PLS-Cox model. Figure 7 represents the comparison of performance based on AIC for mainz7g and transbig7g datasets. For the mainz7g dataset, the PLS-FPM showed nearly similar performance for gamma, log-

normal, Weibull, and log-logistic distribution. The graphical comparison evidence that PLS-FPM with all four distributions has higher efficiency compared to the standard PLS-Cox model. Figure 7 further displays the comparison of efficiency for transbig7g data. Regarding model comparability with PLS-Cox, the FPM-PLS shows higher performance integrated with all four distributions, while the comparison of PLS-FPMs demonstrates that log-normal distribution performs better among other distributions.

Figure 8 illustrates the model comparison for Vdx7g and upp7g datasets. The box plot for vdx7g shows that PLS-FPM integrated with log-normal distribution has a minimum value of AIC representing the higher efficiency of this

FIGURE 7: The model comparison of standard PLS-Cox regression model with PLS-FPM parameterised over gamma, Weibull, log-logistic, and log-normal distribution using breast cancer datasets, namely, mainz7g and transbig7g, is presented.

FIGURE 8: The model comparison of standard PLS-Cox regression model with PLS-FPM parameterised over gamma, Weibull, log-logistic, and log-normal distribution using breast cancer datasets, namely, vdx7g and upp7g, is presented.

distribution compared to other distribution. Regarding comparability with the PLS-Cox model, PLS-FPM with all four distributions has increased efficiency because these models have a lower value of AIC. The difference between the numerical measure of AIC among the PLS-Cox model and PLS-FPM integrated with all distributions is less for upp7g data than this value for vdx7g. The PLS-FPM shows

even higher efficiency than the PLS-Cox model for upp7g data. For this dataset, PLS-FPM shows nearly equal accuracy for all distributions.

Figure 9 shows the estimates of the baseline cumulative hazards from the PLS-Cox model and PLS-FPM based on gamma, Weibull, log-logistic, and log-normal distribution for mainz7g and transbig7g datasets of breast cancer. The

FIGURE 9: The estimates of the baseline cumulative hazard from standard PLS-Cox regression model and PLS-FPM parameterised over gamma Weibull, log-logistic, and log-normal distribution for breast cancer datasets, namely, mainz7g and transbig7g.

PLS-FPM integrated with all four distributions produce smooth estimates of the baseline cumulative hazards extrapolated to a time of 10 years showing a monotonically increasing linear trend for both datasets. The PLS-Cox model resulted in an unsmooth baseline cumulative hazards curve with unusual fluctuations at certain time points. All the parametric models exhibit monotonically increasing smooth curves of the baseline cumulative hazards due to parametric flexibility and the unsmooth pattern is observed for the PLS-Cox model for all datasets. Figure 10 illustrates the estimates of the baseline cumulative hazards from the PLS-Cox model and PLS-FPM based on gamma, Weibull, log-logistic, and log-normal distribution for vdx7g and upp7g datasets of breast cancer. The most efficient model for the vdx7g dataset is PLS-FPM integrated with log-normal distribution which has lower estimates of the baseline cumulative hazards compared to other distribution at initial survival time and it gradually increases with the increase of survival time. Consistent with the vdx7g dataset, the log-normal distribution shows decreased estimates of the baseline cumulative hazards compared to other distribution at the first two years of survival time and this estimate is minimum for gamma distribution for further years.

All seven genes are found to be significantly associated with breast cancer in six datasets. The parameters of all genes are estimated for each survival dataset and presented in Table 1. The coefficients are estimated for the optimal model corresponding to each dataset. The PLS-FPM based on log-normal distribution is used for parameter estimation of mainz7g, transbig7g, vdx7g, upp7g, and nki7g datasets based on performance. For unt7g data, Weibull distribution is integrated with PLS-FPM for the estimation of coefficients

having higher accuracy. Positive and negative associations are observed for different genes regarding various datasets of breast cancer.

4. Discussion

Efficient model selection with robust estimates remains a challenging and computationally intensive task, especially for survival systems in the presence of extreme observations. Therefore, the number of candidate models for evaluations and comparisons is usually limited in studies. However, nonparametric and semiparametric survival methods can misappropriate model structures without considering specific probability distribution. The PLS-Cox regression model is used to deal with multicollinearity among covariates in survival time analysis. The new approach is proposed mainly for two disadvantages of Cox regression. Firstly, Cox regression is a semiparametric approach; hence, it produces unsmooth estimates which are limited in the long-term evaluations. Secondly, the standard Cox regression model is not robust to extreme values. In this article, PLS-FPM is proposed as a fully parametric survival technique to examine hazard function and efficiently estimate the parameters. The PLS-FPM was particularly projected to address survival systems in the presence of multicollinearity by using various distributions to produce smooth estimates to extrapolate the survival model. Since this approach is flexible enough to combine different distributions, it can produce a robust model in the presence of extreme observations by using a suitable probability distribution.

Overall, PLS-FPM compares favorably with the benchmark method on both simulated and real datasets in the

FIGURE 10: The estimates of the baseline cumulative hazard from standard PLS-Cox regression model and PLS-FPM parameterised over gamma Weibull, log-logistic, and log-normal distribution for breast cancer datasets, namely, vdx7g and upp7g.

TABLE 1: PLS-FPM regression coefficients are presented where inflectional factors are extracted by generalized F distribution.

Genes	Coefficient estimates					
	mainz7g	transbig7g	upp7g	unt7g	vdx7g	nki7g
PLAU	0.65	0.57	0.37	0.81	-0.23	-0.27
CASP3	-0.54	0.22	-0.48	-0.05	-0.59	0.69
VEGFA	-0.39	-0.18	-0.18	-0.28	-0.15	0.34
STAT1	0.35	0.57	-0.25	0.12	0.10	-0.39
ESR1	0.06	0.37	0.18	-0.31	0.137	0.11
AURKA	-0.01	-0.35	0.71	-0.24	0.07	0.09
ERBB2	-0.01	-0.17	-0.04	-0.31	-0.74	0.40

presence of multicollinearity and extreme observations. The PLS-FPM using Weibull distribution turns out to be the best model to estimate cumulative hazards according to AIC and BIC over simulated data generated from Weibull distribution. More generally in the setting of simulated survival data, the fully parametric PLS-FPM had better performance than the semiparametric PLS-Cox regression model. The optimal model for each real dataset shows that seven genes are found significantly associated with each breast cancer survival dataset. Tumor cell proliferation is found to be one of the most significant predictors of breast cancer survival. Various previous studies investigated proliferation in tumor cells and found it a significant factor of breast cancer [23, 24]. Supporting previous studies, tumor metastasis is found to be an important factor of breast cancer survival [25, 26]. Previous literature reported that high immune scores are significantly associated with improved disease-free survival and overall survival in breast cancer patients [27]. The present study also finds the

association of immune response with breast cancer. Like previous studies, it is found that the apoptosis phenotype in a breast tumor seems to predict survival rate [28, 29]. Consistent with the present study, various previous studies have reported the association of angiogenesis and breast cancer to citelongatto2010angiogenesis, min2010tie2. The ESR1 is reported as the most attractive target for clinical therapy of ER-positive breast cancer [30]. The amplification of ERBB2 of clinical samples of breast cancers along with its association with aggressive disease and poor survival is discussed in literature [31]. Both the genes are found to be significantly associated with breast cancer survival for the present analysis supporting past evidence.

The overall accuracy of these algorithms enhances the model performance to a higher extend, considering collinear covariates and extreme observations. This efficiency suggests that survival function, hazard function, cumulative hazard function, and parameters of distribution for the survival time data with unknown distribution can be estimated more efficiently in terms of smooth lines.

5. Conclusion

PLS-FPM not only extrapolates survival outcomes beyond the available follow-up data but also supports a wide range of hazard shapes including monotonically increasing, monotonically decreasing, arc-shaped, and bathtub-shaped hazards. In a word, PLS-FPM is viewed as a useful fully parametric addition to the toolbox of robust estimation and prediction of survival time approaches for the widely used PLS-Cox model in the survival settings.

Data Availability

The breast cancer datasets are freely available in an R package called *survcomp* and are available at <http://www.ulb.ac.be/di/map/bhaibeka/software/survcomp/>.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Authors' Contributions

Maryam Sadiq and Tahir Mehmood contributed equally to this work.

References

- [1] A. Belot, A. Ndiaye, M.-A. Luque-Fernandez et al., "Summarizing and communicating on survival data according to the audience: a tutorial on different measures illustrated with population-based cancer registry data," *Clinical Epidemiology*, vol. 11, pp. 53–65, 2019.
- [2] C. Kelly, P. Majewska, S. Ioannidis, M. H. Raza, and M. Williams, "Estimating progression-free survival in patients with glioblastoma using routinely collected data," *Journal of Neuro-Oncology*, vol. 135, no. 3, pp. 621–627, 2017.
- [3] P. Deheuvels and J. H. Einmahl, "Functional limit laws for the increments of Kaplan-Meier product-limit processes and applications," *Annals of Probability*, vol. 28, pp. 1301–1335, 2000.
- [4] D. R. Cox, "Regression models and life-tables," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 34, no. 2, pp. 187–202, 1972.
- [5] A. Antoniadis, P. Fryzlewicz, and F. Letué, "The dantzig selector in cox's proportional hazards model," *Scandinavian Journal of Statistics*, vol. 37, no. 4, pp. 531–552, 2010.
- [6] B. Saraçoğlu and C. Tanış, "A new statistical distribution: cubic rank transmuted Kumaraswamy distribution and its properties," *Journal of the National Science Foundation of Sri Lanka*, vol. 46, no. 4, pp. 505–518, 2018.
- [7] C. Tanış and B. Saraçoğlu, "On the record-based transmuted model of balakrishnan and He based on weibull distribution," *Communications in Statistics-Simulation and Computation*, pp. 1–21, 2020.
- [8] C. Tanış, B. Saraçoğlu, C. Kuş, A. Pekgör, and K. Karakaya, "Transmuted lower record type fréchet distribution with lifetime regression analysis based on type I-censored data," *Journal of Statistical Theory and Applications*, vol. 20, no. 1, pp. 86–96, 2021.
- [9] C. Tanış, B. Saraçoğlu, C. Kuş, and A. Pekgör, "Transmuted complementary exponential power distribution: properties and applications," *Cumhuriyet Science Journal*, vol. 41, no. 2, pp. 419–432, 2020.
- [10] B. Saraçoğlu and C. Tanış, "A new lifetime distribution: transmuted exponential power distribution," *Communications Faculty of Sciences University of Ankara Series A1 Mathematics and Statistics*, vol. 70, no. 1, pp. 1–14, 2021.
- [11] T. Mehmood, M. Sadiq, and M. Aslam, "Filter-based factor selection methods in Partial Least Squares regression," *IEEE Access*, vol. 7, pp. 153499–153508, 2019.
- [12] M. Sadiq, T. Mehmood, and M. Aslam, "Identifying the factors associated with cesarean section modeled with categorical correlation coefficients in partial least squares," *PloS One*, vol. 14, no. 7, Article ID e0219427, 2019.
- [13] P. Bastien, V. E. Vinzi, and M. Tenenhaus, "PLS generalised linear regression," *Computational Statistics and data analysis*, vol. 48, no. 1, pp. 17–46, 2005.
- [14] P. Royston and M. K. B. Parmar, "Flexible parametric proportional-hazards and proportional-odds models for censored survival data, with application to prognostic modelling and estimation of treatment effects," *Statistics in Medicine*, vol. 21, no. 15, pp. 2175–2197, 2002.
- [15] G. Rodriguez, *Parametric Survival Models. Rapport Technique*, Princeton University, Princeton, NJ, USA, 2010.
- [16] C. Desmedt, B. Haibe-Kains, P. Wirapati et al., "Biological processes associated with breast cancer clinical outcome depend on the molecular subtypes," *Clinical Cancer Research*, vol. 14, no. 16, pp. 5158–5165, 2008.
- [17] M. Schmidt, D. Böhm, C. von Törne et al., "The humoral immune system has a key prognostic impact in node-negative breast cancer," *Cancer Research*, vol. 68, no. 13, pp. 5405–5413, 2008.
- [18] C. Desmedt, F. Piette, S. Loi et al., "Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series," *Clinical Cancer Research*, vol. 13, no. 11, pp. 3207–3214, 2007.
- [19] A. J. Minn, G. P. Gupta, D. Padua et al., "Lung metastasis genes couple breast tumor size and metastatic spread," *Proceedings of the National Academy of Sciences*, vol. 104, no. 16, pp. 6740–6745, 2007.
- [20] Y. Wang, J. G. Klijn, Y. Zhang et al., "Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer," *The Lancet*, vol. 365, no. 9460, pp. 671–679, 2005.
- [21] L. D. Miller, J. Smeds, J. George et al., "An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival," *Proceedings of the National Academy of Sciences*, vol. 102, no. 38, pp. 13550–13555, 2005.
- [22] B. Haibe-Kains, C. Desmedt, C. Sotiriou, and G. Bontempi, "A comparative study of survival models for breast cancer prognostication based on microarray data: does a single gene beat them all?" *Bioinformatics*, vol. 24, no. 19, pp. 2200–2208, 2008.
- [23] P. J. Van Diest, G. Brugal, and J. P. Baak, "Proliferation markers in tumours: interpretation and clinical value," *Journal of Clinical Pathology*, vol. 51, no. 10, pp. 716–724, 1998.
- [24] M. Dowsett, C. Archer, L. Assersohn et al., "Clinical studies of apoptosis and proliferation in breast cancer," *Endocrine-related Cancer*, vol. 6, no. 1, pp. 25–28, 1999.
- [25] J. Ma, D.-X. Luo, C. Huang et al., "AKR1B10 overexpression in breast cancer: association with tumor size, lymph node metastasis and patient survival and its potential as a novel serum marker," *International Journal of Cancer*, vol. 131, no. 6, pp. E862–E871, 2012.
- [26] D. Van Uden, M. Van Maaren, L. J. Strobbe et al., "Metastatic behavior and overall survival according to breast cancer subtypes in stage IV inflammatory breast cancer," *Breast Cancer Research*, vol. 21, no. 1, p. 113, 2019.
- [27] J. Wang, Y. Li, W. Fu et al., "Prognostic nomogram based on immune scores for breast cancer patients," *Cancer Medicine*, vol. 8, no. 11, pp. 5214–5222, 2019.
- [28] N. Krawczyk, A. Hartkopf, M. Banys et al., "Prognostic relevance of induced and spontaneous apoptosis of disseminated tumor cells in primary breast cancer patients," *BMC Cancer*, vol. 14, no. 1, p. 394, 2014.

- [29] M. Parton, M. Dowsett, and I. Smith, "Studies of apoptosis in breast cancer," *Bmj*, vol. 322, no. 7301, pp. 1528–1532, 2001.
- [30] Z. Zhou, J. X. Qiao, A. Shetty et al., "Regulation of estrogen receptor signaling in breast carcinogenesis and breast cancer therapy," *Cellular and Molecular Life Sciences*, vol. 71, no. 8, p. 1549, 2014.
- [31] P. Adusumilli, M. Konatam, S. Gundeti, S. Bala, and L. Maddali, "Treatment challenges and survival analysis of human epidermal growth factor receptor 2-positive breast cancer in real world," *Indian Journal of Medical and Paediatric Oncology*, vol. 38, no. 1, p. 22, 2017.