ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

Tumor Grading via Decorrelated Sparse Survival Regression

Benjamin Paaßen[1], Nadine Gaisa[2,3], Michael Rose[2,3], Mark-Sebastian Bösherz[2]

1- Faculty of Technology
Bielefeld University, Bielefeld, Germany
bpaassen@techfak.uni-bielefeld.de

2- Institute of Pathology
University Hospital RWTH Aachen University, Aachen, Germany

3- Institute of Pathology
University Hospital Ulm, University of Ulm, Ulm, Germany

Abstract. In medical pathology, tumor grading is concerned with estimating the risk posed by a tumor, based on its pathological features. One way to infer risk scores is survival regression, i.e. using machine learning to infer a score that predicts the remaining survival time of a patient. Unfortunately, if applied naively, such a score is a mix of the intrinsic risk posed by the tumor and other risk factors, like the progression of the tumor or patient gender and age. We provide the first survival regression model that disentangles tumor grading from undesired correlations, while retaining a high degree of model interpretability, thanks to convex optimization, non-negativity constraints, sparsity, and linearity. We evaluate the proposed approach both on simulated and real-world data from $N = 114$ patients at the University Clinic Aachen.

1 Introduction

To make a prognosis for a tumor patient, the tumor needs to be *graded* and *staged*. *Staging* refers to estimating how far the tumor has already progressed in its growth, whereas *grading* refers to an estimate of the risk posed by the tumor due to its pathological characteristics, such as tumor cell counts or the differentiation between tumor cells and healthy cells [1]. Formally, the density of time until death of a patient can be described with a decreasing curve. Staging refers to the time that has already passed, grading to the slope of the curve (refer to Fig. 1). Typically, grading is performed by pathologists based on decision trees that have been derived manually from empirical data. However, manually deriving the most predictive tumor grading scheme can be challenging, especially in case of multiple relevant pathological chracteristics that have to be combined. Hence, it would be desirable to infer tumor grading schemes automatically, while retaining interpretability for medical experts, conforming to bio-medical domain knowledge, and ensuring that grading and staging are disentangled.

The most related branch of research is survival regression, which typically estimate a linear score from the input features that predicts the remaining survival time of patients [2, 3, 4, 5]. Because the scoring itself is linear, the models

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

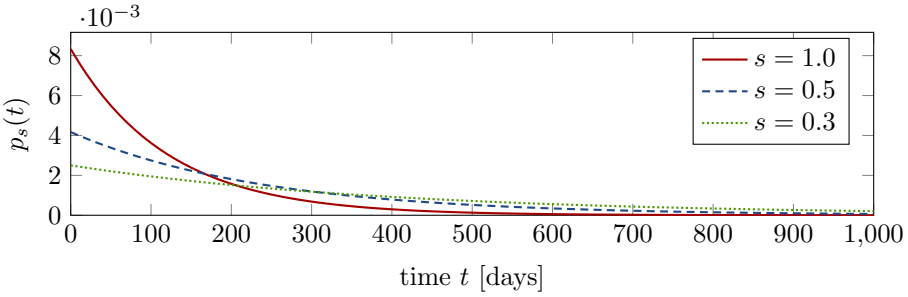

Figure 1: Illustration of tumor grading versus staging: The grading s corresponds to the slope of the probability density until time of death, whereas staging corresponds to the position on the time axis.

are interpretable to some degree. However, the parameters may be bio-medically implausible and grading and staging remain entangled. In fact, the inferred risk scores are likely to include many undesired correlations, such as with age or gender. Hence, we propose a new algorithm for survival regression that takes plausibility constraints and undesired correlations into account, which we name *decorrelated sparse survival regression* (DSSR).

In detail, our contributions are three-fold: 1) we propose DSSR as a new approach to automatically learn a sparse, linear tumor grading. We provide three algorithms based on linear and quadratic programming as well as evolutionary optimization. 2) In a simulation experiment, we demonstrate that DSSR is able to disentangle tumor grading from tumor staging, whereas baselines from the literature fail in that regard. 3) On real-world data from $N = 114$ patients, we demonstrate that DSSR achieves sparser and more biologically plausible models that generalize (slightly) better to new patients.

2 Method

Our goal is to derive a tumor grading scheme that is linear in input features $\vec{x}$, sparse, biologically plausible, predicts the remaining survival time of patients accurately, but avoids undesired correlations with features $\vec{z}$, such as staging indicators, age, or gender. More precisely, let $\boldsymbol{X} \in \mathbb{R}^{N \times n}$ be a matrix of n features for N patients and let $\boldsymbol{Z} \in \mathbb{R}^{N \times m}$ be the matrix of m indicators that ought to be decorrelated from the grading. We wish to find sparse weights $\vec{w} \in \mathbb{R}^n$ that minimize correlations between the vector of risk scores $\vec{s} = \boldsymbol{X} \cdot \vec{w}$ and any column of $\boldsymbol{Z}$. We express this target via a linear program: minimize the slack variable r under the side constraints $r > \vec{s}^T \cdot \vec{z}_k$ and $r > -\vec{s}^T \cdot \vec{z}_k$, where $\vec{z}_k$ is the z-normalized kth column of $\boldsymbol{Z}$.

At the same time, we wish to maximize the *concordance index* [2] between the risk scores $s_1, \ldots, s_N$ and the actual times until death $t_1, \ldots, t_N$, meaning the fraction of pairs (i, j) such that $s_i > s_j$ and $t_i > t_j$. Following the scheme of support vector machines for survival regression [3], we translate the concordance

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

target to a linear program, as well: we minimize the slack variables $\epsilon_{i,j}$ under the side constraints $s_i - s_j + \epsilon_{i,j} \geq 0$ and $\epsilon_{i,j} \geq 0$ for all (i, j) where $t_i < t_j$.

Our remaining targets are sparsity and biological plausibility. For the former, we apply L1 regularization to $\vec{w}$. For the latter, we impose the side constraint $\vec{w} \geq 0$. In other words, we assume that the features have been pre-processed such that higher values correspond to higher assumed risk from a bio-medical standpoint. In our real data, both categorial and ordinal features occur. For categorial features, we perform a one-hot coding. For ordinal features, we perform a multi-hot coding, such that the kth level of an ordinal scale with K values is encoded as the vector of k ones followed by $K - k$ zeros.

Putting all components together, we obtain the overall linear program:

$$\begin{aligned} \min_{\vec{w}\in\mathbb{R}^n, \boldsymbol{E}\in\mathbb{R}^{N\times N}, r\in\mathbb{R}} \quad & \sum_{i=1}^{N}\sum_{j=1}^{N} \epsilon_{i,j} + \lambda_1 \cdot \sum_{\ell=1}^{n} w_\ell + \lambda_2 \cdot r & (1) \\ \text{such that} \quad & (\vec{x}_i - \vec{x}_j)^T \cdot \vec{w} + \epsilon_{i,j} \geq 1 & \forall i, j : t_i < t_j \\ & \vec{z}_k^T \cdot \boldsymbol{X} \cdot \vec{w} \leq r \wedge -\vec{z}_k^T \cdot \boldsymbol{X} \cdot \vec{w} \leq r & \forall k \\ & \epsilon_{i,j} \geq 0 & \forall i, j \\ & w_\ell \geq 0 & \forall \ell \end{aligned}$$

where λ_1 and λ_2 are hyper-parameters controlling the L1 regularization and decorrelation, respectively. We call this the *linprog* variant of DSSR.

Quadprog variant: We also consider a quadratic programming variant where the slack variables are punished quadratically and the correlation term is replaced with $\vec{w}^T \cdot \boldsymbol{X}^T \cdot \left(\sum_{k=1}^{K} \hat{z}_k \cdot \hat{z}_k^T \right) \cdot \boldsymbol{X} \cdot \vec{w}$, i.e. the sum of squared correlations to all columns of $\boldsymbol{Z}$.

CMA-ES variant: Finally, we consider an evolutionary optimization scheme via CMA-ES [6], where the fitness function is the concordance index minus the regularization terms in (1).

3 Experiments

In the following, we compare our three variants of DSSR (linprog, quadprog, CMA-ES) against several baselines from the literature. In detail, we consider several accelerated failure time (AFT) regression models, namely Weibull, Log-logistic, and Log-normal [4], as well as Cox regression [5], all implemented in the *lifelines* software package [2]. For linprog, we use the scipy-linprog solver [7], for quadprog the OSQP-quadprog solver, and for CMA-ES the reference implementation of [6]. In all experiments, we set the L1 regularization strength λ_1 to 0.01 and the correlation regularization strength λ_2 to 1. All experiments were executed on a desktop machine with an Intel i9-10900 CPU and 32 GB RAM. The source code can be found at `https://gitlab.com/bpaassen/dssr`.

Simulation experiment: To investigate whether DSSR is able to recover a ground truth grading from observed times until death, we generated simulated data following Fig. 1 to achieve a simple scenario where input features contain

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

Table 1: Mean (± stdev.) concordance index and correlation between predicted and ground truth risk scores for all models in the simulation experiment.

model	concordance index	score correlation
Weibull AFT	0.83 ± 0.02	0.54 ± 0.08
Log-Logistic AFT	0.83 ± 0.03	0.53 ± 0.08
Log-normal AFT	0.82 ± 0.03	0.57 ± 0.08
Cox	$\mathbf{0.84 \pm 0.02}$	0.52 ± 0.08
linprog	0.65 ± 0.04	$\mathbf{0.97 \pm 0.01}$
quadprog	0.65 ± 0.04	0.96 ± 0.01
CMA-ES	0.70 ± 0.04	0.92 ± 0.04

entangled information about grading and staging. For each patient i, 1) we sampled a ground truth grading score s_i uniformly from the interval $[0.1, 1]$. 2) We sampled days until death t_i from the exponential distribution with rate parameter $\lambda = \frac{s_i}{120}$. 3) We computed a staging score δ_i as the actual time until death minus the expected time until death in years, i.e. $\frac{1}{365} \cdot (t_i - \frac{120}{s_i})$. 4) We computed features $\vec{x}_i$ as two noisy copies of the ground truth grading score s_i (grading features) and three noisy copies of the staging score δ_i (staging features). The noise was Gaussian with std. 0.1. 5) We set $\vec{z}_i$ to the staging features. We performed 10 repeats, simulating 200 patients as training data and 100 as test data each time.

Table 1 displays the results for all models on the simulated data in terms of concordance index (center column) and correlation between predicted risk score and ground truth grading score (right column). The baseline regression approaches perform better in terms of concordance index, meaning they predict the time until death more accurately. However, as expected, grading and staging remains entangled, such that the predicted score correlate only moderately with the ground truth. By contrast, all DSSR variants achieve lower concordance index but much better correlation with the ground truth grading risk score.

Real-world experiment: The real-world data was obtained from $N = 114$ patients with bladder cancer at the University Hospital of RWTH Aachen. For each patient, a biopsy of the tumor was performed and evaluated by a pathology expert, yielding the feature matrix $\boldsymbol{X}$. Further, staging indicators as well as age and gender were recorded for each patient, yielding the matrix $\boldsymbol{Z}$. Note that follow-up data was only available for 87 patients, and time-of-death data only for 30 patients, meaning that the data set was right-censored for most patients.

Table 2 shows the concordance index in five-fold crossvalidation. We observe that no model manages to generalize well but DSSR variants perform slightly better, especially CMA-ES. This is likely due to a low number of patients with recorded time of death, leading to overfitting. In terms of correlation, DSSR variants successfully avoid significant undesired correlations, whereas the base-

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

Table 2: Mean (± stdev.) concordance index on the Aachen data in 5-fold crossvalidation.

model	concordance index
Weibull AFT	0.55 ± 0.09
Log-logistic AFT	0.52 ± 0.08
Log-normal AFT	0.53 ± 0.09
Cox	0.51 ± 0.10
linprog	0.56 ± 0.07
quadprog	0.58 ± 0.08
CMA-ES	$\mathbf{0.64 \pm 0.07}$

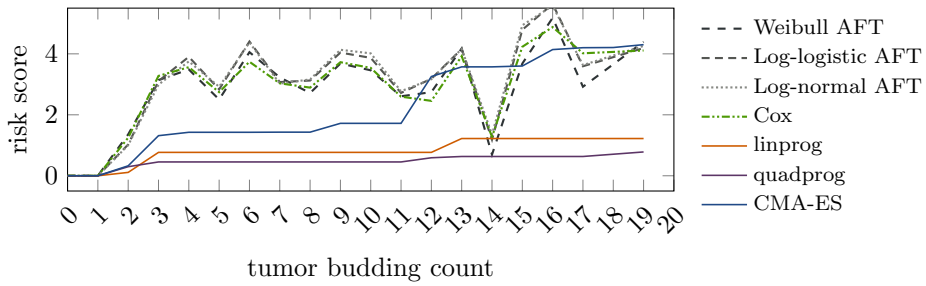

Figure 2: Weights for the tumor budding count indicator of all models.

lines tended to correlate with one of the staging features (F-Zahl), as well as patients' age. The most striking differences can be observed in terms of the learned parameters. Fig. 2 exemplarily shows the weights for the tumor budding count feature. From a bio-medical perspective, higher levels are deemed more dangerous. However, because the parameters of the baseline models are unrestricted, the risk score strongly fluctuates, sometimes yielding lower risk for higher feature values. By contrast, the non-negativity constraints of DSSR yield monotonously increasing risk scores, as desired.

In terms of training time (Fig. 3), we observe that the literature baselines scale better in terms of data points N and comparably in terms of features n but that CMA-ES has a particularly high constant factor (around 10s runtime). The linprog variant of DSSR scales particularly badly with higher N, whereas the quadprog variant remains competitive with literature baselines in terms of training time.

4 Conclusion

We proposed a novel survival regression approach which avoids undesired correlations and promotes interpretability via a sparse, non-negative linear model,

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

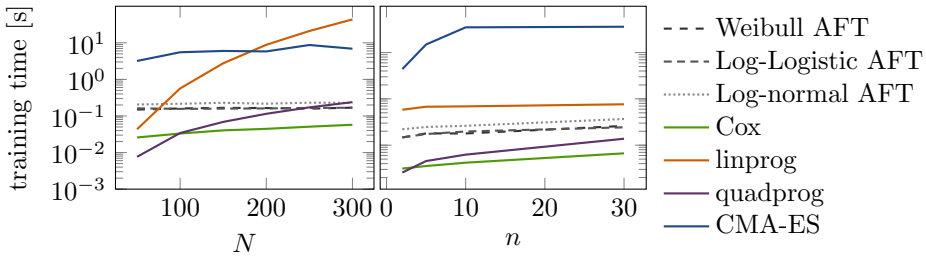

Figure 3: Training times with respect to number of patients (left) and number of features (right) for all algorithms.

which we name decorrelated sparse survival regression (DSSR). In a simulation experiment, we found that DSSR is better able to disentangle tumor grading from staging, compared to prior methods. On real-world data from $N = 114$ patients, we observed slightly better generalization performance, but mainly sparser and more plausible models that avoided undesired correlations. We tested three variants of DSSR, a linear programming one, a quadratic programming one, and an evolutionary optimization. The CMA-ES variant tends to achieve the best concordance but is less consistent across runs and tends to be slowest. Hence, in practice, the linear or quadratic variants are recommended.

The main limitation of our study is that the generalization performance on the real-world data set is only marginally better than random, indicating the challenges in survival regression on smaller data sets. Future work should evaluate DSSR variants on a bigger data set with undesired correlation data, which does not yet exist.

References

[1] S.M. Telloni. *Tumor Staging and Grading: A Primer*, pages 1–17. Springer New York, New York, NY, 2017.

[2] C. Davidson-Pilon. lifelines: survival analysis in python. *Journal of Open Source Software*, 4(40):1317, 2019.

[3] V. Van Belle, K. Pelckmans, S. Van Huffel, and J.A.K. Suykens. Support vector methods for survival analysis: a comparison between ranking and regression approaches. *Artificial Intelligence in Medicine*, 53(2):107–118, 2011.

[4] L.J. Wei. The accelerated failure time model: A useful alternative to the cox regression model in survival analysis. *Statistics in Medicine*, 11(14-15):1871–1879, 1992.

[5] H.H. Zhang and W. Lu. Adaptive Lasso for Cox's proportional hazards model. *Biometrika*, 94(3):691–703, 2007.

[6] N. Hansen and A. Ostermeier. Completely Derandomized Self-Adaptation in Evolution Strategies. *Evolutionary Computation*, 9(2):159–195, 2001.

[7] P. Virtanen, R. Gommers, T.E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature methods*, 17(3):261–272, 2020.