

Project Proposal: An Attempt to Reproduce the Findings of "Deep Cox Mixtures for Survival Regression"

Chris Yu¹, Jimmy Lee¹

¹University of Illinois Urbana-Champaign
hmyu2@illinois.edu, jl279@illinois.edu

Abstract

This project aims to reproduce the findings of "Deep Cox Mixtures for Survival Regression". The Deep Cox Mixtures (DCM) method addresses limitations of traditional survival models like the Cox Proportional Hazards model, which relies on restrictive assumptions. By leveraging neural networks and mixture models, DCM captures complex survival patterns for heterogeneous populations. Using public datasets including SUPPORT, we will evaluate DCM's reproducibility and compare its performance against established baselines.

Paper — <https://arxiv.org/abs/2101.06536>

Ref. Code — <https://github.com/autonlab/auton-survival>

Introduction

Survival analysis is used across fields to model and predict time-to-event outcomes while accounting for censored data, such as predicting patient survival times based on clinical data and treatments.

Estimating survival curves is challenging due to censored outcomes. Models such as Cox regression assume proportional hazards, which often fail and result in miscalibrated curves in critical cases. The DCM approach enhances survival curve calibration while preserving strong discriminative performance.

Problem Statements

The Cox model assumes that all individuals share the same baseline hazard, which simplifies the analysis but ignores potential differences in risk across diverse groups, such as racial or ethnic minorities. In reality, baseline risks can vary significantly between majority and minority populations due to systemic disparities in healthcare access, socioeconomic conditions, or biological factors. When the model fails to account for this heterogeneity, it tends to align more closely with the risk patterns of the majority group, often leading to less accurate predictions for minority populations. This can perpetuate biases in survival analysis, particularly for racial or ethnic minorities, and underscores the need for more flexible approaches that can better capture and address these variations in risk.

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Architecture of Deep Cox Mixtures (DCM)

The **Deep Cox Mixtures (DCM)** model is a neural network-based survival analysis method that combines the flexibility of deep learning with the interpretability of mixture models. The architecture consists of three main components:

- **Embedding Network:** The input features x are passed through a feedforward neural network to obtain a lower-dimensional representation \tilde{x} . This network is defined by a series of hidden layers (e.g., $[100, 50]$), which capture complex, non-linear relationships between the covariates and survival outcomes.
- **Gate Network:** A linear layer computes the log probabilities $\log P(Z = k | X = x)$ for latent cluster assignments, where Z represents the latent cluster and k is the number of clusters. These probabilities are normalized using a softmax function.
- **Expert Network:** Another linear layer computes the log hazard ratios $\log \lambda_k(t | x)$ for each latent cluster k . These hazard ratios are used to estimate the survival function for each cluster.

The final survival function $S(t | x)$ is computed as a weighted average of the cluster-specific survival functions, where the weights are the gate probabilities. This architecture allows the model to capture heterogeneity in the data by partitioning the population into latent subgroups, each with its own Cox proportional hazards model.

Training Process

The DCM model is trained using an **Expectation-Maximization (EM)** algorithm, which iteratively updates the model parameters and estimates the baseline survival functions. The training process consists of the following steps:

- **E-Step:** In the expectation step, the posterior probabilities $P(Z = k | X = x, T = t, \delta = e)$ for latent cluster assignments are computed using the current model parameters. These probabilities are derived from the likelihood of the data given the model and the Breslow estimates of the baseline survival functions.
- **M-Step:** In the maximization step, the model parameters are updated by minimizing the **Q-function**, which combines the partial log-likelihood loss for the Cox model

and the gate probabilities. This step is performed using gradient-based optimization (e.g., Adam).

- **Breslow Estimation:** Periodically, the Breslow estimator is used to compute the baseline survival functions for each latent cluster. These functions are smoothed using splines to ensure numerical stability.

The training process continues until convergence or a pre-defined number of epochs is reached. Early stopping is implemented based on the validation loss to prevent overfitting.

Usage of DCM

- **Survival Prediction:** Given a set of input features x , the model predicts the probability of survival beyond specified times t . This is achieved by computing the weighted average of the cluster-specific survival functions, where the weights are the gate probabilities. The output is a survival curve $S(t | x)$ that provides insights into the individual's risk over time.
- **Latent Cluster Assignment:** The model assigns each individual to one of the k latent clusters based on the gate probabilities. These clusters represent subgroups with distinct survival patterns, enabling personalized risk assessment and subgroup analysis.

Novelty of Deep Cox Mixtures (DCM)

The **Deep Cox Mixtures (DCM)** model introduces several key innovations that distinguish it from traditional and deep learning-based survival models:

- **Latent Clustering:** DCM partitions the population into k latent clusters, each with its own Cox proportional hazards model. This allows the model to capture heterogeneity in the data, which is often overlooked by traditional models like Cox Proportional Hazards (CPH) and Faraggi-Simon Net (FSN)/DeepSurv.
- **Flexible Baseline Hazards:** Instead of assuming a shared baseline hazard function (as in CPH), DCM estimates cluster-specific baseline hazards using non-parametric methods (e.g., Breslow estimator). This flexibility improves the model's ability to fit complex survival patterns.
- **Neural Network Integration:** DCM leverages neural networks to learn complex, non-linear relationships between covariates and survival outcomes. This is a significant improvement over linear models like CPH and AFT, which cannot capture such relationships.
- **Full Likelihood Optimization:** Unlike many deep learning models (e.g., DeepHit, DSM), DCM optimizes the full likelihood of the data, ensuring better calibration and discriminative performance.

Superiority Over Baselines

DCM outperforms other models in following measures:

Area Under the Curve (AUC):

AUC evaluates how well the model distinguishes between individuals who experience the event and those who don't at specific time points.

- Latent clustering captures subgroup differences, improving survival probability estimates.
- Full likelihood optimization ensures calibrated, reliable survival curves.

Time-Dependent Concordance Index (Ctd):

Ctd measures the model's ability to correctly rank individuals by their risk over time.

- Neural networks model complex, non-linear relationships between covariates and survival outcomes.
- Latent clustering enhances subgroup-specific risk ranking accuracy.

Data Access and Implementation Details

Dataset

- **SUPPORT dataset:** Data on 9,105 critically ill U.S. patients (1989-1994), available in the code repo.
- **FLCHAIN dataset:** Data on 7,874 Minnesota subjects aged 50+, linking serum free light chains to mortality, available in the code repo.
- **SEER Dataset:** A collection of data on cancer incidence, prevalence, and survival, covering approximately 47.9% of the U.S. population and could be accessed from the <https://seer.cancer.gov/> (SEER Program).

Planned Ablations

- **Regularization:** Add dropout or L2 regularization to the neural network to prevent overfitting. This helps the model generalize better to new data.

Feasibility of the computation

- The datasets are moderately sized, allowing training on standard hardware.
- The codebase is publicly available for replication. We designed to use the existing codebase as basics for further exploration as the original code didn't cover all the findings.
- Exact result replication may be challenging due to complex evaluation and implementation details.

Conclusion

This project aims to reproduce "*Deep Cox Mixtures for Survival Regression*" and evaluate it on public datasets like SUPPORT, FLCHAIN, and SEER. DCM improves upon traditional survival models by capturing complex patterns and heterogeneity using neural networks and mixture models. While datasets and code are accessible, reproduction may be challenging due to evaluation complexity. We will use the **auton-survival** library to compare DCM with the Cox Proportional Hazards (CPH) model, evaluating both using **AUC** and **Time-Dependent Concordance Index (Ctd)**. This study will provide insights into DCM's strengths and limitations.

References

- Chen, G. H. (2024). **An introduction to deep survival analysis models for predicting time-to-event outcomes.** *Foundations and Trends® in Machine Learning*, 17(6), 921–1100. <http://dx.doi.org/10.1561/22000000114>
- Nagpal, C., Jeanselme, V., & Dubrawski, A. (2021). **Deep parametric time-to-event regression with time-varying covariates.** In *Proceedings of AAAI Spring Symposium on Survival Prediction—Algorithms, Challenges, and Applications 2021* (pp. 1–15). PMLR.
- Nagpal, C., Li, X., & Dubrawski, A. (2021). **Deep survival machines: Fully parametric survival regression and representation learning for censored data with competing risks.** *IEEE Journal of Biomedical and Health Informatics*, 25(8), 3163–3175. <https://doi.org/10.1109/JBHI.2021.3068357>
- Nagpal, C., Yadlowsky, S., Rostamzadeh, N., & Heller, K. (2021). **Deep Cox mixtures for survival regression.** In *Machine Learning for Healthcare Conference* (pp. 674–708). PMLR.
- Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., & Kluger, Y. (2018). **DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural network.** *BMC Medical Research Methodology*, 18(1), 24. <https://doi.org/10.1186/s12874-018-0482-1>
- Fotso, S. et al. (2019). **PySurvival: Open source package for survival analysis modeling.** [Computer software]. <https://www.pysurvival.io/>
- Beil, M., Proft, I., van Heerden, D., Sviri, S., & van Heerden, P. V. (2019). **Ethical considerations about artificial intelligence for prognostication in intensive care.** *Intensive Care Medicine Experimental*, 7(1), 70. <https://doi.org/10.1186/s40635-019-0286-6>
- Lee, C., Zame, W., Yoon, J., & van der Schaar, M. (2018). **DeepHit: A deep learning approach to survival analysis with competing risks.** *Proceedings of the AAAI Conference on Artificial Intelligence*, 32(1). <https://doi.org/10.1609/aaai.v32i1.11842>
- Kingma, D. P., & Ba, J. (2014). **Adam: A method for stochastic optimization.** *arXiv preprint arXiv:1412.6980*.
- Gerds, T. A., Kattan, M. W., Schumacher, M., & Yu, C. (2013). **Estimating a time-dependent concordance index for survival prediction models with covariate dependent censoring.** *Statistics in Medicine*, 32(13), 2173–2184. <https://doi.org/10.1002/sim.5681>
- Hung, H., & Chiang, C.-T. (2010). **Optimal composite markers for time-dependent receiver operating characteristic curves with censored survival data.** *Scandinavian Journal of Statistics*, 37(4), 664–679. <https://doi.org/10.1111/j.1467-9469.2010.00708.x>
- Ishwaran, H., Kogalur, U. B., Blackstone, E. H., Lauer, M. S., et al. (2008). **Random survival forests.** *The Annals of Applied Statistics*, 2(3), 841–860. <https://doi.org/10.1214/08-AOAS169>
- Gerds, T. A., & Schumacher, M. (2006). **Consistent estimation of the expected Brier score in general survival models with right-censored event times.** *Biometrical Journal*, 48(6), 1029–1040. <https://doi.org/10.1002/bimj.200610232>
- Faraggi, D., & Simon, R. (1995). **A neural network model for survival data.** *Statistics in Medicine*, 14(1), 73–82. <https://doi.org/10.1002/sim.4780140108>
- Wei, L. J. (1992). **The accelerated failure time model: A useful alternative to the Cox regression model in survival analysis.** *Statistics in Medicine*, 11(14-15), 1871–1879. <https://doi.org/10.1002/sim.4780111409>
- Cox, D. R. (1972). **Regression models and life-tables.** *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2), 187–220. <https://doi.org/10.1111/j.2517-6161.1972.tb00899.x>