Preference Models assume Proportional Hazards of Utilities

Chirag Nagpal
Meta Superintelligence Labs (MSL)

July 2025

Abstract

Approaches for estimating preferences from human annotated data typically involves inducing a distribution over a ranked list of choices such as the Plackett-Luce model. Indeed, modern AI alignment tools such as Reward Modelling and Direct Preference Optimization are based on the statistical assumptions posed by the Plackett-Luce model. In this paper, I will connect the Plackett-Luce model to another classical and well known statistical model, the Cox Proportional Hazards model and attempt to shed some light on the implications of the connection therein.

1 Introduction

Modelling of human preferences is an important step in modern post-training pipelines for AI alignment. One popular approach of building such models of human preference is assuming that human preference rankings assume a Plackett-Luce (Plackett, 1975; Luce et al., 1959) distribution. In this monograph, I draw a somewhat remarkable connection of the popular statistical model for estimating lifetimes, the Cox Proportional Hazard model (Cox, 1972) to the Plackett-Luce model and then consequently to algorithms such as Direct Preference Optimization, a popular algorithm for aligning modern Artifical Intelligence (Ouyang et al., 2022).

To the best of my knowledge, at the time of writing the connection between the Proportional Hazards model and the Plackett-Luce is relatively little known, and the subsequent connections to the AI alignment algorithms such as '*Direct Preference Optimization*' (Rafailov et al., 2023) are not well appreciated. I believe that explcitly stating this connection will help the AI research community build on existing research in semi-parametric statistics to build better models of human preference.

2 The Plackett-Luce and Bradley-Terry Model

The Plackett-Luce model was independently introduced by Plackett and Luce as a probabilistic framework to model ranked data. Consider a list of n observations $\{(\boldsymbol{x}_i, \boldsymbol{r}_i)_{i=1}^n\}$ where each example i consists of features $\boldsymbol{x}_i$ and $r_i \in [n]$ represents its observed ranking or position in the data. The central idea behind the Plackett-Luce model is to express the probability of observing a particular ranking of these items as a function of the features associated with each item.

Specifically, suppose $\sigma(\cdot)$ is a permutation on the set $[n]$ that defines the order of the items such that $\sigma(i)$ is the index of the item with the i^{th} rank. The Plackett-Luce model assigns a probability to the event that the ranking is exactly $\{\sigma(1) \succ \ldots \succ \sigma(n)\}$ based on their relative scores. Formally, this probability is given by the product:

$$\mathbf{P}(\sigma(1) \succ \ldots \succ \sigma(n) | \boldsymbol{f}, \{\boldsymbol{x}_i\}_{i=1}^n) = \prod_{i=1}^{n} \frac{\exp\Big(f(\boldsymbol{x}_{\sigma(i)})\Big)}{\sum\limits_{j=i}^{n} \exp\Big(f(\boldsymbol{x}_{\sigma(j)})\Big)}. \tag{1}$$

Here $\sigma(j)$ represents the j^{th} ranked item in the datum. $\boldsymbol{f} : \mathbb{R}^d \to \mathbb{R}$ is some function that operates on the covariates and scores the item. I use the terms Bradley-Terry and Plackett-Luce interchangeably, since the Bradley-Terry model is a special case of the model above when the number of choices $(k = 2)$.

Replacing $\boldsymbol{f}$ with $\log \frac{\pi(\cdot)}{\pi^{\text{ref}}(\cdot)}$ in Equation 1 recovers the popular AI alignment algorithm Direct Preference Optimization (DPO) (Rafailov et al., 2023). In the case of DPO $\pi(\cdot|\texttt{Input})$ refers to the probability of a response given an input to a language model.

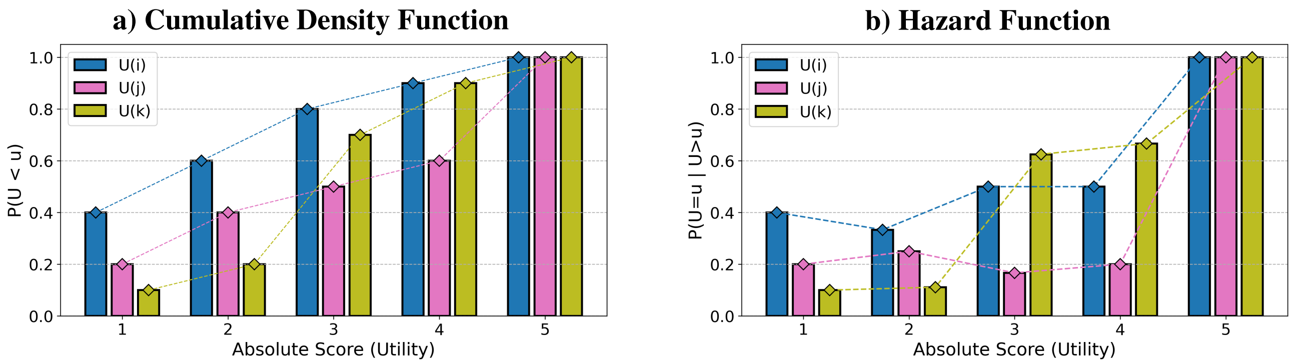

Figure 1: **The violation of the Proportional Hazards assumption in the Likert scale**. The cdfs of the utilities stochastically dominate each other ie. $\Pr(U(i) < u) > \Pr(U(j) < u)$, $\forall u \in \mathcal{U}$; while this condition is violated for $U(k)$. **Plackett-Luce can recover preferences corresponding to** (i, j) **but not when PH is violated as in** (i, k).

3 The Cox Proportional Hazards (PH) Model

The Cox model is popular in bio-statistics, reliability engineerring and acturial sciences for the estimation of *time-to-event* outcomes. With slight modification instead of modeling times let us consider a dataset of n examples $\{(\boldsymbol{x}_i, u_i)\}_{i=1}^{n}$. Each example i consists of features $\boldsymbol{x}_i \in \mathbb{R}^d$, a scalar *utility* $u_i \in \mathbb{R}_+$. In the human preference setting utility is akin to a 5-point rating scale or time for a snippet of AI generated code to run.

Given such a dataset, the Cox model's goal is to estimate a *hazard rate* for the utilities u conditioned on the features $\boldsymbol{x}$. Here, the hazard rate is a function that describes the instantaneous rate of change of the utility function at a certain value of utility. More formally the hazard rate is:

$$\boldsymbol{\lambda}(u) := \lim_{\Delta u \to 0} \frac{\mathbf{P}(u < U \leq u + \Delta u \mid U > u)}{\Delta u}. \tag{2}$$

Estimation in terms of the hazard rate $\boldsymbol{\lambda}(u)$ is natural in survival analysis and reliability engineering and can be used to estimate other quantities of interest such as the survival function.[1] This model assumes that the ratio between the hazard rate for a point with features $\boldsymbol{x}$ at a utility level u changes with respect to the *baseline hazard rate* $\boldsymbol{\lambda}_0(u)$[2] and that the rate of change is determined by the some function f operating on the covariates $\boldsymbol{x}$. We assume that the conditional hazard rate function follows a *proportional hazards* (PH) model:

$$\boldsymbol{\lambda}(U = u \mid X = \boldsymbol{x}) := \boldsymbol{\lambda}_0(u) \exp\left(f(\boldsymbol{x})\right). \tag{3}$$

The *proportionality of hazards* follows naturally from Equation 3 above since it implies that for any (i, j) pair:

$$\forall u \in \mathcal{U}, \quad \boldsymbol{\lambda}(u|X = \boldsymbol{x}_i) \Big/ \boldsymbol{\lambda}(u|X = \boldsymbol{x}_j) = \text{constant}\,.$$

The parameters of the Cox Proportional Hazards model $\boldsymbol{f}$ are estimated by minimizing the *partial likelihood*. Given a dataset we define the partial likelihood $\mathcal{L}(\boldsymbol{f})$ imposed by the Cox Proportional Hazards model as:

$$\mathcal{L}(f) = \prod_{i=1}^{n} \frac{\cancel{\boldsymbol{\lambda}_0(u)} \exp\left(f(\boldsymbol{x}_i)\right)}{\sum\limits_{j:\, u_j \geq u_i} \cancel{\boldsymbol{\lambda}_0(u)} \exp\left(f(\boldsymbol{x}_j)\right)} = \prod_{i=1}^{n} \frac{\exp\left(f(\boldsymbol{x}_i)\right)}{\sum\limits_{j:\, u_j \geq u_i} \exp\left(f(\boldsymbol{x}_j)\right)}. \tag{4}$$

The partial likelihood is called such as it is independent of $\boldsymbol{\lambda}_0(\cdot)$ representing the base hazard. Typically $\boldsymbol{\lambda}_0(\cdot)$ is treated a *nuisance* parameter and not estimated. One can however, estimate the cumulative density of the utilities using the following non-parametric estimator (Breslow, 1972):

$$\widehat{\boldsymbol{S}}_0(u) = \exp\left(- \sum_{i:\, u_i < u} \frac{1}{\sum\limits_{j:\, u_j \geq u_i} \exp\left(\widehat{\boldsymbol{f}}(\boldsymbol{x}_j)\right)} \right) \quad \text{and} \quad \widehat{\mathbf{P}}(U < u | X = \boldsymbol{x}, \widehat{\boldsymbol{f}}) = 1 - \widehat{\boldsymbol{S}}_0(u)^{\exp\left(\widehat{\boldsymbol{f}}(x)\right)}. \tag{5}$$

[1] The survival rate is the negative exponent of the cumulative hazard, ie. $\boldsymbol{S}(u) = \exp\left(-\int_0^u \boldsymbol{\lambda}(u)\right)$.

[2] The base hazard rate $\boldsymbol{\lambda}$ is an infinite dimensional functional parameter of the model, which is estimated non-parametrically.

4 The Connection and implication for AI Alignment

When typically working with Plackett-Luce models we only observe the relative ranking order. Let us now introduce some explicit notion of utility $\mathcal{U}(\cdot)$, that determines the intrinsic (absolute) quality or value of an associated choice $i \succ j \iff \mathcal{U}(i) > \mathcal{U}(j)$. Comparing Equation 4 to Equation 1 should make the connection abundantly clear. Assuming the rankings are associated with an observed scalar utility u, somewhat remarkably **the Plackett-Luce assumptions recovers exactly the same model as the Cox PH model.**

Plackett-Luce and its derivatives are sensitive to the assumption of *Proportional Hazards*. Models based on the Plackett-Luce assumptions (such as Bradley-Terry Reward Models or Direct Preference Optimization) are restricted to modelling preferences whose underlying utility functions are stochastically dominated (Figure 1). When fitted to data arising from utilities that violate PH, the Plackett-Luce model likely will mis-estimate human preferences. **This is more than just a theoretical insight and has practical consequences**. **Real world annotations on polarizing concepts manifest population level heterogeneity, in such situations preferences will likely be mis-estimated.**

Estimating the conditional distribution of utilties with preference data. I mentioned that in the Cox model $\boldsymbol{\lambda}_0(\cdot)$ is effectively a nuisance parameter that is not involved directly in the estimation. Now let us consider situations in the presence of data representing absolute utilities, such as point-wise feedback, the baseline hazard can simultaneously be estimated non-parametrically (Breslow, 1972; Lin, 2007). This can be used to recover an estimate of the true distribution over the absolute value of utility, leading to better estimation of relative utility.

5 Conclusion

In this paper, I have attempted to bring about a relatively little known connection between Plackett and Luce's statistical model of choice and the Cox's classic Proportional Hazards model. Although some work (Chen et al., 2024; Maystre and Russo, 2022) has similar flavor, they do not explicitly mention this connection[3]. I have further shed some light on the implications for the design of better approaches to align AI with human preferences.

References

Breslow, N. E. (1972). Discussion of the paper by D.R. Cox. *J R Statist Soc B*.

Chen, A., Malladi, S., Zhang, L., Chen, X., Zhang, Q. R., Ranganath, R., and Cho, K. (2024). Preference learning algorithms do not learn preference rankings. *Advances in Neural Information Processing Systems*.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*.

Lin, D. (2007). On the breslow estimator. *Lifetime data analysis*, 13(4).

Luce, R. D. et al. (1959). *Individual choice behavior*, volume 4. Wiley New York.

Maystre, L. (2019). *Efficient learning from comparisons*. Gesellschaft für Informatik eV.

Maystre, L. and Russo, D. (2022). Temporally-consistent survival analysis. *Advances in Neural Information Processing Systems*, 35:10671–10683.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35.

Plackett, R. L. (1975). The analysis of permutations. *Journal of the Royal Statistical Society Series C: Applied Statistics*.

Rafailov, R., Sharma, A., Mitchell, E., Manning, C. D., Ermon, S., and Finn, C. (2023). Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*.

Ranganath, R., Perotte, A., Elhadad, N., and Blei, D. (2016). Deep survival analysis. In *Machine Learning for Healthcare Conference*. PMLR.

[3] Although I have not had a correspondence with either, I am inclined to believe that Ranganath et al. and Maystre maybe privy to the connection with '*Proportional Hazards*' due to their prior art in the area.