DRAFT: Enhancing Cox Model Interpretability Using Ghost Variables

Santiago Isaza Cadavid
C.C. 1007239660
sisazac@eafit.edu.co

Advisor
Santiago Ortiz
sortiza2@eafit.edu.co

MSc in Applied Mathematics
School of Applied Sciences and Engineering,
Universidad EAFIT, Medellín, Colombia

Abstract

XXXXXXX

Palabras Clave: XXXXXXX, XXXXXXX, XXXXXXX, XXXXXXX.

1 Introduction

Survival analysis investigates the time until an event of interest occurs. A cornerstone of this field is the Cox Proportional Hazards (CoxPH) model (Cox, 1972), which elegantly relates covariates to the hazard rate without specifying the baseline hazard. However, the standard Cox model relies on two key assumptions: the proportional hazards assumption (hazard ratios are constant over time) and, critically for interpretation of covariate effects, the assumption of a linear relationship between the covariates (or their chosen transformations) and the log-hazard rate.

While powerful and widely used, the linearity assumption can be overly restrictive when the true underlying relationships are complex and non-linear. This limitation can lead to model misspecification or the need for manual, often data-driven, variable transformations. In recent years, more flexible models capable of capturing non-linear effects, such as those based on machine learning (ML) techniques (Sundrani and Lu, 2021) or Generalized Additive Models (GAMs) (Hastie and Tibshirani, 1986a), have gained traction in survival analysis due to their potential for higher predictive accuracy. However, a major drawback of these complex models is their reduced interpretability; understanding why a prediction is made or how much each variable contributes becomes significantly challenging (Sundrani and Lu, 2021; Delicado and Peña, 2023). This "black box" nature hinders clinical trust and scientific understanding. While methods like SHAP have been adapted to derive Hazard Ratio (HR)-like interpretations from some ML survival models (Sundrani and Lu, 2021), there is still a need for robust methods to assess the fundamental relevance of variables within these non-linear frameworks.

This work proposes a novel methodology, termed "GhostCox," to enhance the interpretability of survival models allowing for non-linear predictor effects, while potentially maintaining the proportional hazards structure. We leverage the concept of Ghost Variables, introduced by Delicado and Peña (2023) in the context of general complex predictive models, to assess variable relevance within these potentially non-linear hazard structures. The core idea is to quantify the unique contribution of each variable to the model's predictive output (the non-linear component $f(\mathbf{x})$). This approach allows us not only to identify which predictors are important drivers of the model but also to quantify the magnitude of their unique impact, even when the relationship between predictors and the hazard rate is non-linear. It offers a complementary perspective to methods focused solely on effect-size estimation like HRs, by first addressing the fundamental question of variable relevance in the context of the specific model fitted.

2 Methodology: GhostCox for Interpreting Non-Linear Survival Models

This section details the proposed GhostCox methodology, designed to enhance the interpretability of survival models that accommodate non-linear covariate effects while potentially maintaining a proportional hazards structure. The approach integrates the established framework of proportional hazards models with the Ghost Variable technique developed by Delicado and

Peña (2023) for assessing variable relevance in complex predictive models.

2.1 The Proportional Hazards Framework and Linearity Limitations

The cornerstone of much survival analysis is the Cox Proportional Hazards (CoxPH) model (Cox, 1972), which relates the hazard rate $h(t, \mathbf{x}_i)$ for an individual i with a p-dimensional covariate vector $\mathbf{x}_i$ at time t to a baseline hazard function $h_0(t)$:

$$h(t, \mathbf{x}_i) = h_0(t) \exp(\boldsymbol{\beta}' \mathbf{x}_i) = h_0(t) \exp\left(\sum_{j=1}^{p} \beta_j x_{ij}\right)$$

Here, $h_0(t)$ is an unspecified non-negative function representing the hazard for an individual with all covariates equal to zero, and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is the vector of regression coefficients. A key feature is the estimation of $\boldsymbol{\beta}$ via maximization of the partial likelihood function $L(\boldsymbol{\beta})$, which circumvents the need to estimate $h_0(t)$ (Cox, 1975):

$$L(\boldsymbol{\beta}) = \prod_{i:\delta_i=1} \frac{\exp(\boldsymbol{\beta}' \mathbf{x}_i)}{\sum_{l \in R(t_i)} \exp(\boldsymbol{\beta}' \mathbf{x}_l)}$$

where the product is over individuals i experiencing the event ($\delta_i = 1$) at distinct event times t_i, and $R(t_i)$ denotes the risk set (individuals alive and uncensored just prior to t_i). The model assumes proportional hazards (the ratio of hazards for any two individuals is constant over time) and, critically for interpreting $\boldsymbol{\beta}$, assumes a linear relationship between the covariates and the log-hazard rate.

This linearity assumption, however, can be restrictive if the true underlying covariate effects are complex. While transformations can be applied, their selection is often non-trivial. This motivates the use of more flexible models that relax the linearity constraint.

2.2 Non-Linear Proportional Hazards Models

To accommodate potentially complex relationships, we consider a generalization of the Cox model that retains the proportional hazards structure but allows for a non-linear function of the covariates:

$$h(t, \mathbf{x}_i) = h_0(t) \exp(f(\mathbf{x}_i))$$

In this formulation, $f : \mathbb{R}^p \to \mathbb{R}$ is an unknown, potentially non-linear function that maps the covariate vector $\mathbf{x}_i$ to a risk score, which is proportional to the log-hazard relative to the baseline. The function $f(\mathbf{x})$ can be estimated using various flexible techniques suitable for censored survival data. Examples include:

- **Generalized Additive Models (GAMs):** Using spline functions or other smoothers for continuous predictors within the Cox framework Hastie and Tibshirani (1986b)
- **Machine Learning Ensembles:** Methods like Random Survival Forests (RSF) or Gradient Boosted Survival Trees, which inherently capture non-linearities and interactions (Sundrani and Lu, 2021; Ishwaran et al., 2008).

- **Neural Networks:** Architectures specifically designed for survival prediction Katzman et al. (2018)

The choice of method for estimating $f(\mathbf{x})$ depends on the data characteristics and the anticipated complexity of the covariate effects. In this work, we primarily utilize Random Survival Forests (RSF) as implemented in the `scikit-survival` library Pölsterl (2020) for estimating $\hat{f}(\mathbf{x})$, due to their demonstrated ability to capture complex non-linearities and interaction effects without pre-specification Ishwaran et al. (2008) and their availability in robust software packages.

While these models offer increased predictive power, the non-linear nature of $\hat{f}(\mathbf{x})$ makes direct interpretation of individual variable effects challenging, motivating the need for post-hoc interpretability methods.

2.3 The Ghost Variable Methodology

Delicado and Peña (2023) proposed the Ghost Variable methodology as a model-agnostic approach to quantify the relevance of individual predictor variables in complex predictive models $g(\mathbf{X})$. For a set of predictors $\mathbf{X} = (X_1, \ldots, X_p)$, the ghost variable $\hat{X}_j$ corresponding to predictor X_j is defined as its conditional expectation given all other predictors $\mathbf{X}_{-j} = (X_1, \ldots, X_{j-1}, X_{j+1}, \ldots, X_p)$:

$$\hat{X}_j = \mathbb{E}(X_j | \mathbf{X}_{-j})$$

The ghost variable $\hat{X}_j$ represents the best prediction of X_j using only the information available in the other predictors, effectively capturing the component of X_j's variability that is redundant given $\mathbf{X}_{-j}$.

The relevance of X_j to the model $g(\mathbf{X})$ is then assessed by measuring the impact of substituting X_j with its ghost $\hat{X}_j$ on the model's output. Let $g(\mathbf{X}_{\hat{j}})$ denote the model prediction when X_j is replaced by $\hat{X}_j$. A large difference between $g(\mathbf{X})$ and $g(\mathbf{X}_{\hat{j}})$ indicates that X_j provides unique, non-redundant information crucial for the prediction. Conversely, a small difference suggests X_j's contribution is largely captured by the other variables Delicado and Peña (2023).

The estimation of the conditional expectation $\mathbb{E}(X_j|\mathbf{X}_{-j})$ is flexible. While Linear Regression can be employed for computational efficiency, especially if inter-covariate relationships are approximately linear, the framework readily accommodates non-linear regression techniques such as Random Forests, Gradient Boosting, or GAMs to estimate $\hat{X}_j$. Using such flexible methods may provide more accurate estimates of the conditional expectation when inter-covariate relationships are complex Delicado and Peña (2023).

2.4 The GhostCox Interpretation Procedure

We adapt the Ghost Variable methodology to interpret the fitted non-linear component $\hat{f}(\mathbf{x})$ of the survival model $h(t, \mathbf{x}_i) = h_0(t) \exp(\hat{f}(\mathbf{x}_i))$. This involves a two-stage process:

Stage 1: Non-Linear Survival Model Fitting A non-linear survival model, such as a Random Survival Forest (RSF), is fitted to a designated training dataset $(X_{\text{train}}, y_{\text{train}})$, where y_{train} contains the time-to-event and event status information. The chosen non-linear survival model (e.g., RSF) inherently handles right-censored data during this fitting process. This stage yields the estimated risk prediction function $\hat{f}(\mathbf{x})$.

Stage 2: Interpretation using Ghost Variables on a Test Set To ensure the assessment of variable relevance reflects the model's generalization performance and avoids overfitting the interpretation to the training data, the Ghost Variable analysis is performed on an independent test set $(X_{\text{test}}, y_{\text{test}})$ using the survival model $\hat{f}$ fitted during Stage 1. The steps are as follows:

1. **Ghost Estimation:** For each variable X_j $(j = 1, \ldots, p)$, estimate its ghost variable values $\hat{\mathbf{x}}_j$ across the test set individuals using an appropriate regression model $\hat{\mathbb{E}}(X_j|\mathbf{X}_{-j,\text{test}})$. This model (e.g., Linear Regression for efficiency, or Random Forest Regressor, used primarily in our implementation, for flexibility) is fitted using only the test set covariates X_{test}. This results in a ghost matrix $X_{\text{ghost, test}}$.

2. **Original Prediction:** Compute the predicted risk scores from the fitted survival model $\hat{f}$ using the original test data: $\hat{f}_{\text{test}} = \hat{f}(X_{\text{test}})$.

3. **Ghost Predictions:** For each variable j, create a modified test set $X_{\text{test},\hat{j}}$ by replacing the j-th column of X_{test} with the corresponding j-th column of $X_{\text{ghost, test}}$. Compute the predictions on this modified set: $\hat{f}_{\text{test},\hat{j}} = \hat{f}(X_{\text{test},\hat{j}})$.

4. **Relevance Calculation:** Calculate the relevance of variable X_j, denoted $RV_{gh}(X_j)$, based on the average squared difference between the original and ghost predictions, normalized by a factor representing the variability of the model's output on the test set. Following Delicado and Peña (2023), we define the relevance numerator as:

$$R_j^{\text{num}} = \frac{1}{n_{\text{test}}} \sum_{i \in \text{test}} (\hat{f}(x_i) - \hat{f}(x_{i,\hat{j}}))^2$$

   As a normalization factor suitable for the risk score output $\hat{f}$, we use the variance of the predicted scores on the test set, $\text{Var}(\hat{f}_{\text{test}})$. The final relevance measure is:

$$RV_{gh}(X_j) = \frac{R_j^{\text{num}}}{\text{Var}(\hat{f}_{\text{test}})} = \frac{\frac{1}{n_{\text{test}}} \sum_{i \in \text{test}} (\hat{f}(x_i) - \hat{f}(x_{i,\hat{j}}))^2}{\text{Var}(\hat{f}(X_{\text{test}}))}$$

   This $RV_{gh}(X_j)$ quantifies the unique contribution of X_j to the variance of the predicted risk score $\hat{f}$ on the test set. Higher values indicate greater relevance. Ranking variables by $RV_{gh}(X_j)$ provides a measure of their relative importance to the non-linear survival model.

5. **Relevance Matrix (Optional):** To explore joint effects, construct the $n_{\text{test}} \times p$ matrix $\mathbf{A}$ of prediction changes, where the j-th column is the vector $(\hat{f}_{\text{test}} - \hat{f}_{\text{test},\hat{j}})$. The normalized Relevance Matrix $\mathbf{V}$ is then calculated as:

$$\mathbf{V} = \frac{1}{n_{\text{test}}} \frac{\mathbf{A}'\mathbf{A}}{\text{Var}(\hat{f}_{\text{test}})}$$

The diagonal elements V_{jj} correspond to $RV_{gh}(X_j)$. The off-diagonal elements V_{jk} measure the covariance between the prediction changes induced by substituting X_j and X_k with their respective ghosts, providing insight into the interplay of their unique contributions to $\hat{f}$. Eigenanalysis of $\mathbf{V}$ can reveal dominant axes of variable relevance and identify groups of variables with correlated importance Delicado and Peña (2023).

This two-stage GhostCox approach thus provides a principled framework for fitting flexible survival models and subsequently interpreting the contribution of each covariate to the potentially non-linear risk function $\hat{f}$. The validity of the relevance measures obtained through GhostCox relies on the assumption that the fitted survival model $\hat{f}(\mathbf{x})$ adequately captures the systematic covariate effects on hazard, and that the estimation of the ghost variables $\hat{\mathbb{E}}(X_j|\mathbf{X}_{-j,\text{test}})$ is reasonably accurate.

3 Experimentation

3.1 Validating Variable Ranking

3.1.1 Objective

The primary objective of this simulation study was to assess the ability of the proposed GhostCox methodology to correctly identify and rank the relevance of predictor variables under different controlled conditions. Specifically, we aimed to verify whether the calculated relevance measure, $RV_{gh}(X_j)$, consistently assigns higher importance (lower rank) to variables truly associated with the survival outcome compared to noise variables, even in the presence of non-linear predictor effects and correlations among features.

3.1.2 Data Generation

For each replication, datasets were generated with $N = 500$ samples, subsequently split into training (70%) and test (30%) sets. Five predictor variables ($\mathbf{X} = (X_1, X_2, X_3, X_4, X_5)'$) were generated. The true relationship between covariates and the log-hazard was defined via a function $f(\mathbf{x})$, incorporated into the hazard function $h(t|\mathbf{x}) = h_0(t)\exp(f(\mathbf{x}))$. Survival times were generated assuming a constant baseline hazard $h_0(t) = \lambda_0$ (with λ_0 implicitly defined by the simulation process, though not explicitly set as a parameter in the final generator code provided) and inverse transform sampling based on the resulting hazard. Censoring times were generated independently from an exponential distribution with rate $\lambda_c = 0.005$ to achieve moderate censoring levels (observed around 30-40

1. **Linear Uncorrelated (Linear_Uncorr):**

$$f(\mathbf{x}) = \boldsymbol{\beta}'\mathbf{x} = 1.0X_1 + 1.0X_2 - 1.0X_3 + 0.0X_4 + 0.0X_5$$

Features $\mathbf{X}$ were generated as independent standard normal variables, i.e., $X_j \sim N(0,1)$ for $j = 1, ..., 5$, and $Cov(X_j, X_k) = 0$ for $j \neq k$. Variables X_1, X_2, X_3 are relevant, while X_4, X_5 are noise variables.

2. **Linear Correlated (Linear_Corr):**

$$f(\mathbf{x}) = 1.0X_1 + 1.0X_2 - 1.0X_3 + 0.0X_4 + 0.0X_5$$

Features $\mathbf{X}$ were generated from a multivariate normal distribution $\mathbf{X} \sim N_5(\mathbf{0}, \Sigma)$ with the following correlation structure, designed to introduce dependencies primarily between X_1, X_2 and X_3, X_4:

$$\Sigma = \begin{pmatrix} 1.0 & 0.7 & 0.0 & 0.0 & 0.0 \\ 0.7 & 1.0 & 0.0 & 0.0 & 0.0 \\ 0.0 & 0.0 & 1.0 & 0.5 & 0.0 \\ 0.0 & 0.0 & 0.5 & 1.0 & 0.0 \\ 0.0 & 0.0 & 0.0 & 0.0 & 1.0 \end{pmatrix}$$

Variables X_1, X_2, X_3 are relevant, while X_4, X_5 are noise variables.

3. **Non-Linear Uncorrelated (NonLinear_Uncorr):** The predictor function $f(\mathbf{x})$ was defined as:

$$f(\mathbf{x}) = 1.0X_1 + \sin(\pi X_2) + 0.5X_3^2 + 0.0X_4 + 0.0X_5$$

Features $\mathbf{X}$ were generated as independent standard normal variables, as in Scenario 1. Variables X_1, X_2, X_3 are relevant (through linear, sinusoidal, and quadratic terms, respectively), while X_4, X_5 are noise variables.

4. **Non-Linear Correlated (NonLinear_Corr):** The predictor function $f(\mathbf{x})$ was the same non-linear function as in Scenario 3:

$$f(\mathbf{x}) = 1.0X_1 + \sin(\pi X_2) + 0.5X_3^2 + 0.0X_4 + 0.0X_5$$

Features $\mathbf{X}$ were generated from the correlated multivariate normal distribution $N_5(\mathbf{0}, \Sigma)$ used in Scenario 2. Variables X_1, X_2, X_3 are relevant, while X_4, X_5 are noise variables.

3.1.3 Model Fitting and Interpretation

For each scenario and replication ($N_{rep} = 100$), the following steps were performed:

1. A Random Survival Forest (RSF) model (from `scikit-survival`, using parameters `n_estimators=100`, `min_samples_leaf=15`, `random_state=42`, `n_jobs=-1`) was trained on the training set (X_{train}, y_{train}) to obtain the estimated risk function $\hat{f}(\mathbf{x})$.

2. On the test set covariates X_{test}, ghost variables $\hat{x}_j = \hat{\mathbb{E}}(X_j|\mathbf{X}_{-j,test})$ were estimated for each $j = 1, \ldots, 5$ using a Random Forest Regressor (RFR) (from `scikit-learn`, using default parameters including `n_estimators=100`, `max_depth=10`, `min_samples_leaf=5`, `random_state=42`, `n_jobs=1`).

3. The relevance $RV_{gh}(X_j)$ was calculated for each variable using the GhostCox Interpreter, based on the formula:

$$RV_{gh}(X_j) = \frac{\frac{1}{n_{test}} \sum_{i \in test} (\hat{f}(\mathbf{x}_i) - \hat{f}(\mathbf{x}_{i,\hat{j}}))^2}{Var(\hat{f}(X_{test}))}$$

4. Variables were ranked based on their $RV_{gh}(X_j)$ values (Rank 1 = highest relevance).

5. The concordance index (C-index) was calculated on the test set using the predicted risk scores $\hat{f}(X_{test})$ to assess the predictive performance of the RSF model.

3.1.4 Results Summary

The simulation study successfully validated the ranking capability of the GhostCox method. Table 1 summarizes the average ranks obtained across the $N_{rep} = 100$ replicates for each scenario.

Scenario Name	X_1	X_2	X_3	X_4	X_5
Linear_Uncorr_RSF_RFGhost	1.93	1.77	2.30	4.43	4.57
Linear_Corr_RSF_RFGhost	1.60	1.77	2.63	4.13	4.87
NonLinear_Uncorr_RSF_RFGhost	1.03	2.30	2.67	4.57	4.43
NonLinear_Corr_RSF_RFGhost	1.60	1.40	3.00	4.40	4.60

Table 1: Average Relevance Rank per Variable across Scenarios (Simulation 1)

As shown in Table 1, the relevant variables (X_1, X_2, X_3) consistently received substantially lower average ranks (indicating higher importance) than the noise variables (X_4, X_5) across all four scenarios. This demonstrates the robustness of the method in identifying influential predictors irrespective of the linearity of their effect or the presence of correlation among them. The average C-index values (Linear Uncorr: 0.795, Linear Corr: 0.823, NonLinear Uncorr: 0.728, NonLinear Corr: 0.740) indicated reasonable predictive performance of the underlying RSF models in all settings.

References

Cox, D. R. (1972), "Regression models and life-tables," *Journal of the Royal Statistical Society: Series B (Methodological)*, 34, 187–202.

Delicado, P. and Peña, D. (2023), "Understanding complex predictive models with ghost variables," *TEST*, 32, 107–145.

Hastie, T. and Tibshirani, R. (1986a), "Generalized additive models," *Statistical science*, 1, 297–310.

— (1986b), "Generalized additive models," *Statistical science*, 1, 297–310.

Ishwaran, H., Kogalur, U. B., Blackstone, E. H., and Lauer, M. S. (2008), "Random survival forests," *The Annals of Applied Statistics*.

Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., and Kluger, Y. (2018), "DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network," *BMC medical research methodology*, 18, 1–12.

Pölsterl, S. (2020), "scikit-survival: A Library for Time-to-Event Analysis Built on Top of scikit-learn," *Journal of Machine Learning Research*, 21, 1–6.

Sundrani, S. and Lu, J. (2021), "Computing the hazard ratios associated with explanatory variables using machine learning models of survival data," *JCO Clinical Cancer Informatics*, 5, 364–378.