Estimating High-Dimensional Linear Models with Interaction Terms using Latent Variable Models

Mohammadreza Nemati[1]

Case Western Reserve University, Cleveland OH 44106**

Abstract. This study introduces a novel approach for estimating high-dimensional linear models using latent variable models, specifically applied to the domain of kidney transplantation. By integrating latent space models with Cox Proportional Hazards (Cox PH) models, we seek to improve the accuracy of predicting graft survival times through enhanced modeling of the compatibility between donor and recipient human leukocyte antigens (HLAs). Our methodology, designed to handle high-dimensional data with interaction terms, shows promise for broader applications in medical fields where complex interactions significantly impact outcomes. Preliminary results indicate that our approach significantly enhances kidney graft survial prediction. This research contributes to the fields of biomedical data analysis and transplant immunology by providing a more nuanced understanding of HLA interactions and their impact on transplant outcomes, with potential extensions to other complex medical scenarios.

Keywords: High-dimensional data, latent variable models, Cox proportional hazards model, kidney transplantation, HLA compatibility.

1 Introduction

Linear models, such as linear regression, logistic regression, and cox proportional hazard models, are foundational tools in data analysis, favored for their simplicity and effectiveness in handling high-dimensional data. However, their reliance on linear combinations of features limits their flexibility and adaptability, especially in complex biomedical applications like kidney transplantation. This limitation is critical because it affects the accuracy of predicting clinical outcomes based on the compatibility of human leukocyte antigens (HLAs) between donors and recipients, where non-linear interactions frequently occur.

The goal of this research is to enhance the predictive performance of linear models by incorporating latent variable models that capture the intricate interactions within high-dimensional biomedical data. This approach is particularly aimed at improving the understanding and prediction of graft survival times in kidney transplantation, an area where current models often fall short due to the complexity and high-dimensionality of the data.

Our research questions are as follows:

** PhD Advisor: Kevin S. Xu - email: ksx2@case.edu

1. Can integrating latent space models with Cox Proportional Hazards (Cox PH) models provide more accurate predictions of graft survival times compared to traditional methods?
2. How effectively can this approach handle the high-dimensional data characteristic of biomedical datasets, particularly those with significant interaction terms?

By addressing these questions, this study aims to contribute to the field of transplant immunology and potentially other medical areas where modeling complex interactions is crucial for outcome prediction.

2 Background

2.1 Linear Models

Linear models comprise a significant portion of the predictive modeling techniques used in machine learning, particularly within the biomedical field. These models include linear regression, logistic regression, and the Cox proportional hazards (Cox PH) model. Linear models typically represent the dependent variable y as a function of the independent variables $x_1, x_2, \ldots, x_p$ using the linear combination:

$$y = w_0 + w_1 x_1 + w_2 x_2 + \ldots + w_p x_p + \epsilon,$$

where $w_0, w_1, \ldots, w_p$ are the coefficients that quantify the influence of each independent variable, and ϵ represents the error term, capturing the deviation of the predictions from the actual observations.

Estimation In scenarios where the number of examples n significantly exceeds the number of covariates p, the coefficients can typically be estimated using the maximum likelihood estimation (MLE). This method provides parameter estimates that are both efficient (unbiased with minimum variance) and consistent as the sample size increases. However, in many biomedical datasets, n is often on the same order as p or even smaller, leading to a situation where MLE can result in severe overfitting. Such overfitting results in models that perform well on training data but poorly on unseen test data. To counteract this, regularization methods like the lasso [1] (employing an ℓ_1 penalty) and elastic net [2] (a combination of ℓ_1 and ℓ_2 penalties) are employed to enhance the generalization ability of the model.

Interaction Terms To capture non-linear relationships between variables, linear models can be extended by incorporating interaction terms and polynomial features. The most commonly augmented terms are the second-order interactions, denoted by products $x_i x_j$, which enable the model to account for the effects of interacting features. Terms with $i \neq j$ are known as interaction terms, capturing the interaction between different features. Conversely, when $i = j$, the model includes squared terms x_i^2, referred to as self-interactions. Although it is possible

to consider higher-order interactions involving three or more features, this paper limits its focus to second-order interactions without self interaction due to their prevalence and manageability in computational implementations.

2.2 Latent Variable Models

Latent variable models form a key category of statistical models utilized extensively in data analysis and machine learning [3]. These models aim to uncover "latent" variables—hidden factors not directly observable but inferred from available data. Employing a probabilistic approach, latent variable models can handle a diverse array of modeling scenarios [4]. They are particularly effective for simplifying complex data sets by reducing them to a smaller number of latent factors, as seen in techniques such as principal component analysis (PCA) [5]. This dimensionality reduction is crucial for making the visualization and interpretation of high-dimensional data more manageable.

A prominent example of a latent variable model is the latent space model proposed by [6], which is tailored for network analysis. This model posits that the likelihood of an edge between any two nodes in a network is influenced by their respective positions in an unseen latent space, usually conceptualized as a Euclidean space. Here, the adjacency matrix A signifies network connections, with $a_{ij} = 1$ indicating the presence of an edge, and $a_{ij} = 0$ indicating its absence. Under this model, the independence of node pairs is assumed conditional on their positions in the latent space. The log odds of an edge forming between two nodes, i and j, are calculated as:

$$\alpha + \beta x_{ij} - \|z_i - z_j\|,$$

where x_{ij} represents observed covariates, z_i and z_j are the latent positions of nodes i and j in a d-dimensional latent space, and α and β are parameters. This model implies a greater probability of connection between nodes that are closer together in the latent space.

3 Methods and Model Formulation

Let n and p denote the number of examples and covariates (features), respectively. We consider interaction terms between all p covariates $x_1, \ldots, x_p$, which include terms such as $x_1x_2, x_1x_3, \ldots, x_{p-1}x_p$. These additional covariates are selectively calculated only for those features of interest, excluding other covariates not considered for interaction.

To analyze the relationship between these covariates, including interaction terms, and a response variable such as survival time, we employ a Generalized Linear Model (GLM). The coefficients for both the primary covariates and interaction terms are typically estimated using regularization methods like the elastic net penalty. However, these methods might face challenges when n is smaller than the number of interaction terms p^2, particularly because they do not account for dependencies among the interaction terms.

3.1 Latent Variable Model for Interaction Terms

We suggest enhancing the estimation accuracy of GLMs with interaction terms by introducing a latent variable model that captures the low-dimensional structure of the interaction coefficients matrix V. Importantly, this does not imply a low-dimensional structure for the data matrix X itself.

Low Rank Model We hypothesize that the interaction matrix V may exhibit an approximately low-rank structure. This can be represented as:

$$V = ZZ^T + \epsilon,$$

where Z is a $p \times d$ matrix with $d \ll p$ representing the reduced dimensions, and *epsilon* is a zero-mean noise matrix that accounts for deviations from the ideal low-rank structure.

Latent Distance Model Alternatively, V might be described within a latent space model, where the distances in a low-dimensional space encapsulate the interaction strengths between covariates. The interaction coefficient for any pair i, j is defined as:

$$v_{ij} = \alpha_0 - \|z_i - z_j\|_2^2 + \epsilon_{ij},$$

where α_0 serves as the baseline interaction level, and ϵ_{ij} represents a zero-mean error term associated with the pair i, j.

3.2 Estimation Procedure

The comprehensive loss function devised for our model integrates three primary components. The initial term, $\mathcal{L}_{\text{GLM}}$, addresses the conventional loss associated with Generalized Linear Models (GLMs), including models such as Linear Regression, Logistic Regression, and the Cox Proportional Hazards model. This component ensures that the model adheres closely to the observed data, capturing the essential trends and patterns.

The second component, $\mathcal{L}_{\text{regularizer}}$, reflects the influence of elastic net regularization. This regularization technique, which combines both L1 and L2 penalties, is critical for promoting sparsity while controlling model complexity. This aspect is particularly advantageous when handling datasets characterized by high dimensionality.

The third component, $\mathcal{L}_{\text{lsm}}$, is associated with the loss from the latent space model representation of the interaction term weights. This component is pivotal for ensuring that the representation of these weights aligns with the hypothesized low-dimensional structure, facilitating both interpretability and robustness in predictions.

The total loss function is expressed as:

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{GLM}} + \lambda_r \mathcal{L}_{\text{regularizer}} + \lambda_n \mathcal{L}_{\text{lsm}},$$

where λ_r and λ_n denote the regularization and latent space model penalties, respectively. These hyperparameters are crucial as they require precise tuning to optimally balance the contributions from each component of the loss function to enhance the model's overall efficacy.

The primary theoretical contribution of our model, $\mathcal{L}_{\text{lsm}}$, distinguishes it by emphasizing the distance between the predicted GLM coefficients and their latent representations. This latent representation could be structured either as a low-rank matrix approximation:

$$\mathcal{L}_{\text{low-Rank}} = \|\hat{V} - Z_i Z_j^T\|_2^2,$$

or as a function of the distances within a latent space:

$$\mathcal{L}_{\text{latent\_distance}} = \|\hat{V} - (\alpha - \|Z_i - Z_j\|_2^2)\|_2^2,$$

where Z_i, Z_j are the latent position matrices for the features x_i, x_j. These representations facilitate the exploration of complex interaction dynamics within the data, enhancing both the interpretative power and predictive accuracy of the model.

4 Experiments

4.1 Data Description

This research utilized the dataset from the Scientific Registry of Transplant Recipients (SRTR), which collects comprehensive information on all donors, candidates on the waitlist, and transplant recipients in the United States, as reported by the Organ Procurement and Transplantation Network (OPTN).

Our analysis follows the inclusion criteria and data preprocessing guidelines as defined by Nemati et al. [7], summarized as follows: This study considers kidney transplants using deceased donors conducted from 2000 to 2016, including only recipients who are 18 years or older and receiving their first transplant. The outcome variable of interest, death-censored graft failure, defines cases where the patient died with a functioning graft as censored, with the censoring occurring at their last known alive date. The dataset analyzed comprises 106,372 kidney transplant instances, with 74.6% being censored cases. We include the same features that is used by Nemati et al. except the number of mismatches between donors and recipients into our study

4.2 Experiment Setting

We partition the dataset evenly, allocating 50% for training and 50% for testing purposes. Hyperparameter tuning is conducted through an exhustive search using the training subset. We employ a 5-fold cross-validation strategy to systematically explore and evaluate different combinations of hyperparameters. Specifically, we test the coefficients for lasso, ridge, and latent distance losses at predefined values: $[0.01, 0.1, 1, 10, 100]$. We choose a set of hyperparameters that maximizes the Cox's partial likeilihood.

Upon identifying the optimal hyperparameters, the models are trained using these parameters. Subsequently, the trained models are tested on the hold-out test dataset. We report Harrel's concordance index as the evaluation metric [8].

5 Results and Discussion

To conduct the experiments, we developed two distinct variants of the Cox Proportional Hazards (Cox PH) model. The first model, labeled as *cph*1, incorporates both lasso and ridge regularization techniques in its loss function. The second model, referred to as *cph*2, additionally includes a latent distance loss alongside the lasso and ridge penalties. For *cph*2, we explore different values for d, which represents the low-dimension latent representation for the HLA types.

The experimental outcomes indicate a significant enhancement in the predictive accuracy for graft survival times when integrating the latent distance model into the Cox Proportional Hazards (Cox PH) model. The introduction of latent distance loss in *cph*2 not only constrains the interaction weight estimates within the human leukocyte antigen (HLA) compatibility network but also stabilizes them against the variability caused by infrequently observed HLA interactions. This approach addresses a crucial challenge in transplantation immunology, where the rarity of certain HLA matches can lead to unreliable and noisy estimates, potentially compromising the decision-making process in donor-recipient matching. Moreover, the improvement in concordance indices, particularly for *cph*2 with varying dimensions of the latent space, underscores the utility of incorporating dimensional representations in modeling complex biological interactions.

These findings affirm the hypothesis that a well-structured latent space model, by encapsulating the underlying interactions within a compact, low-dimensional framework, can more accurately reflect the true biological processes influencing graft survival. This modeling approach not only offers a methodological advancement over traditional linear models with static penalty functions but also provides a more nuanced understanding of HLA interactions, which are often oversimplified in conventional models.

A key objective of this research is to evaluate how effectively our model can capture lower-dimensional representations of HLA types. Specifically, we apply this model to generate 2-dimensional representations for three HLA networks: HLA-A, HLA-B, and HLA-DR. Within the Cox Proportional Hazards (Cox PH) model framework, negative interaction weights correlate with extended survival times. Thus, if the interaction between two HLA types yields a significantly negative value, our proposed latent distance model suggests these HLA types should

Table 1. Comparison of test set concordance index between *cph*1 and *cph*2

Model	*cph*1	*cph*2($d = 2$)	*cph*2($d = 3$)	*cph*2($d = 6$)	*cph*2($d = 16$)
C-index	0.624 ± 0.001	0.628 ± 0.001	$\mathbf{0.629 \pm 0.001}$	0.628 ± 0.001	0.627 ± 0.001

be closely positioned within the latent space. To illustrate this, we visualize the 2D latent space representation of HLA-DR.

To validate our model's efficacy, we select the three smallest and three largest interaction term values. The smallest values should correspond to HLA types located proximally in the latent space, whereas the largest values should relate to distantly positioned HLA types. For clarity and focus in our visual analysis, we limit our detailed exploration to HLA-DR in this study. In Figure 1, uppercase letters denote donor HLA types, while lowercase letters represent recipient HLA types. The interaction terms with the largest values are associated with the HLA pairs (DR18, dr18, 2.41), (DR18, DR4, 2.37), and (DR8, dr13, 1.88), with each tuple comprising the donor's HLA type, recipient's HLA type, and the calculated weight. Consistent with the model's predictions, these pairs are positioned distantly from one another in the plot.

Conversely, the smallest values are linked to the pairs (DR1, dr1, -0.64), (DR9, dr5, -0.64), and (DR16, dr103, -0.63). As anticipated, these HLA types are located in close proximity within the latent space representation, validating the accuracy of the latent distance model in reflecting the strength of HLA compatibility through spatial closeness.

6 Conclusion

This research not only advances our understanding of HLA compatibility in kidney transplants but also paves the way for further studies to explore the application of similar latent space models in other areas of biomedical research where complex interaction networks are crucial. The methodology developed in this study, while initially applied to Cox Proportional Hazards (Cox PH) models for analyzing graft survival times, is not limited to this context. We aim to leverage this approach in other medical fields where linear models are utilized, demonstrating its applicability beyond Cox PH models.

Future work could focus on refining these models to incorporate additional biological insights and extend their applicability to other types of organ transplants, potentially improving outcomes across a broader spectrum of transplantation scenarios. Additionally, we plan to explore the utility of our latent variable modeling approach with other linear models, such as logistic regression and linear regression, in different medical fields. This would enable a broader application of our methodology, enhancing the predictive accuracy and interpretability of complex medical data and leading to improved clinical decision-making.

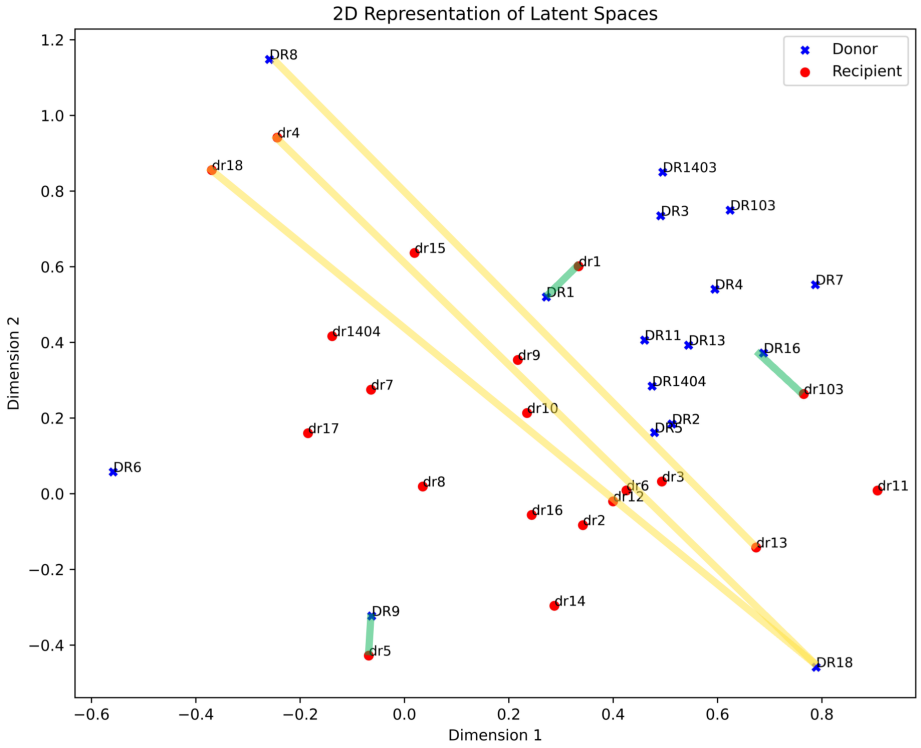

Fig. 1. The 2-D latent space plot for HLA-DR network. The 3 highest and lowest edge weights are shown with yellow and green lines, respectively. Donor-recipient pairs with the lowest edge weights tend to be placed closer together in the latent space compared to those with the highest edge weights

References

1. R. Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society: Series B (Methodological) 58 (1) (1996) 267–288.
2. H. Zou, T. Hastie, Regularization and variable selection via the elastic net, Journal of the royal statistical society: series B (statistical methodology) 67 (2) (2005) 301–320.
3. B. Everett, An introduction to latent variable models, Springer Science & Business Media, 2013.
4. A. Skrondal, S. Rabe-Hesketh, Latent variable modelling: A survey, Scandinavian Journal of Statistics 34 (4) (2007) 712–745.
5. S. Wold, K. Esbensen, P. Geladi, Principal component analysis, Chemometrics and intelligent laboratory systems 2 (1-3) (1987) 37–52.
6. P. D. Hoff, A. E. Raftery, M. S. Handcock, Latent space approaches to social network analysis, J. Am. Stats. Assoc. 97 (460) (2002) 1090–1098.
7. M. Nemati, H. Zhang, M. Sloma, D. Bekbolsynov, H. Wang, S. Stepkowski, K. S. Xu, Predicting kidney transplant survival using multiple feature representations for hlas, Artificial Intelligence in Medicine 145 (2023) 102675.
8. F. E. Harrell, R. M. Califf, D. B. Pryor, K. L. Lee, R. A. Rosati, Evaluating the yield of medical tests, Jama 247 (18) (1982) 2543–2546.