Latent Variable Dyadic Regression Models for Predicting Over/Under Bets in Sports Betting

A thesis submitted in partial fulfillment
of the requirements for the degree of
Master of Science in Statistics and Analytics

by

Alexcia Trejo
Henderson State University
Bachelor of Science in Mathematics, 2023

May 2025
University of Arkansas

This thesis is approved for recommendation to the Graduate Council

Sean Plummer, Ph.D.
Thesis Director

Avishek Chakraborty, Ph.D.
Committee Member

Qingyang Zhang, Ph.D.
Committee Member

ABSTRACT

This thesis explores the use of latent factor models to uncover hidden structures in pairwise outcomes derived from Over/Under betting markets in sports betting. Specifically, we implement and evaluate the Eigen model, a latent space model that represents dyadic data using node-specific vectors whose inner product govern edge probabilities. By modeling relationships between teams as adjacency matrices of binary outcomes, we investigate the extent to which the Eigen model captures both homophily, the tendency of similar teams to yield consistent betting results, and stochastic equivalence, where different teams exhibit indistinguishable patterns of Over/Under outcomes. A Bayesian formulation of the model allows for posterior inference on team-level latent traits and model parameters, while posterior predictive checks are used to assess model fit. We further discuss the potential for such models to detect systematic biases in bookmaker lines, highlighting how latent structures in match-ups may inform profitable betting strategies under market inefficiencies. Simulated and real-world betting data are used to illustrate the model's capabilities and limitations.

ACKNOWLEDGEMENTS

First and foremost, I extend my deepest gratitude to my advisor, Dr. Sean Plummer. Thank you for giving me the freedom to explore my own ideas and the guidance to transform them into research applications. Your patient mentorship, incisive feedback, and encouragement pushed me to think more rigorously and to communicate more clearly.

I am equally indebted to the remarkable women in my life—my mother and sisters—who have provided daily examples of resilience, compassion, and quiet strength. They taught me that academic pursuits are not solitary endeavors but journeys nurtured by the people who stand beside me.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1 Introduction

1.1 Sports Betting

In 2018, the Supreme Court officially invalidated the Professional and Amateur Sports Protection Act (PASPA), opening the door for legalized sports betting across the United States [1]. In turn, the popularity of sports betting has risen dramatically, introducing a variety of wagering options. In this thesis, we will specifically address the wagering option of Over/Under bets.

In Over/Under sports betting, a bookmaker—commonly referred to as the "book" or, in other gambling contexts, the "house"—posts a predicted total score for a single game. This prediction, known as the *betting line*, represents the sum of both teams' final scores. Once the betting line is released to the public, bettors wager on the Over, meaning they believe the total score between both teams will exceed the betting line, or the Under, meaning they believe it will fall short.

The significance of Over/Under bets in our research lies in the way these bets encapsulate interactions between teams. Unlike traditional models that focus solely on win/loss or point spread outcomes, Over/Under bets aggregate the performance of both teams, making them uniquely suited for network analysis. By treating each team as a node in a network and defining edges based on the Over outcome, we can capture the dyadic relationships and systemic patterns that arise from team interactions. This approach allows us to reveal latent factors—such as team dynamics, rivalries, and strategic tendencies—that conventional models may overlook.

The rapid growth of the sports betting industry, combined with the nuanced information embedded in Over/Under bets, prompts new questions about how to model and draw inference on betting outcomes. We aim to leverage network structures in this space to capture team dynamics and underlying relational structures in a more systemic way than conventional models, which often consider team match-ups or game outcomes independently.

1.2 Network Data

Network data is prevalent in many fields and is used to quantify *dyadic relationships* among actors in a system. Dyadic relationships refer to the interactions between pairs of nodes and are measured over all permutations of nodes within the network. We illustrate the components of network models with an example based on Over/Under bets.

Consider all teams in the NBA (National Basketball Association) as forming a network, where each team is treated as a node. In this setting, edges indicate a relationship between nodes. For binary game outcomes, an edge is determined by whether a game between two teams goes over the bet line; if the game goes under, no edge is recorded. Thus, the Over outcome serves as an indication of a specific type of relationship or interaction between team i and team j. This network structure allows us to capture systemic tendencies and patterns, such as particular match-ups exceeding posted expectations due to style of play, rivalries, or other latent factors.

There are two kinds of edge characterizations: directed and undirected. In our work, we model the data as undirected, meaning that the presence of an edge indicates a dyadic relationship between nodes in both directions. In our example, an edge between team i and team j implies that the total sum both teams final scores in a single game went over

the betting line for the relationship between team i and team j (and, by symmetry, for the relationship between team j and team i).

In contrast, a directed network would be appropriate in scenarios where the direction of the relationship matters. For instance, if we were to model home and away teams along with Over/Under outcomes, the relationship from team i to team j would be treated as distinct from the relationship from team j to team i.

The presence or absence of an edge between pairs of nodes is modeled by an $N \times N$ square adjacency matrix, which can be either weighted or binary. In our study, we explore binary networks, meaning that the components of our adjacency matrix can only take on the values 0 or 1 and represent the presence of a dyadic relationship between teams.

In addition to the basic network structure, various network metrics can provide further insights into the data. Metrics such as the degree distribution, which quantifies the number of edges per node, can highlight teams with consistently high or low interactions. Clustering coefficients measure the extent to which nodes tend to cluster together, potentially revealing local groupings or rivalries. Centrality measures, including betweenness and eigenvector centrality, help identify influential nodes that may play key roles in the dynamics of Over/Under outcomes. These metrics set the stage for deeper analysis of systemic patterns and latent factors within sports betting networks.

1.3 Latent Space Models

Latent space models assume that relationships between nodes can be described as functions of node-specific variables and linear predictors. In these models, each node is embedded in an unobserved latent space, and the relationship between any two nodes—meaning,

the dyadic relationship—is modeled as a function of the node specific latent variables, for the model we will utilize in our simulation and experiment, this is the weighted inner product between the corresponding nodes latent vectors. The latent factors, which represent unobserved characteristics such as team dynamics or market influence, are estimated from the data. For instance, in the context of sports betting, a latent factor could be the size of a team's fan base or its underlying competitive strength—attributes that are difficult to measure directly but may explain why a team consistently out performs expectations in Over/Under bets.

Since it is impractical to measure all nodal attributes directly, latent space models aim to capture the key patterns of dyadic relationships among nodes. Earlier models, such as the latent class model introduced by [2], group nodes based on their similarity in relationship patterns, effectively capturing stochastic equivalence. Later, the latent distance model proposed by [3] focused on homophily, suggesting that nodes with similar characteristics tend to have stronger connections. Building on these ideas, the eigenmodel introduced by [4], also referred to as the bilinear mixed-effects model, generalizes both approaches by using eigendecomposition to derive dyadic relationships from the weighted inner product of latent factors, modeling both homophily and stochastic equivalence.

In the eigenmodel, each node is assigned a unique vector of latent factors, and the strength and directionality of the dyadic relationship between any two nodes are inferred from the weighted inner product of these vectors. This approach not only captures both stochastic equivalence and homophily but also offers flexibility in modeling complex network structures. Furthermore, the latent space can be visualized, providing an intuitive understanding of how teams cluster together or diverge based on their latent attributes. Such visualizations,

combined with quantitative metrics, set the stage for a deeper analysis of systemic patterns in Over/Under betting outcomes.

1.4 Bayesian Framework

We adopt a Bayesian framework to perform probabilistic inference on both the systems and the relationships within them. In this framework, latent factors are treated as parameters with prior distributions, and their uncertainty is quantified by the posterior distribution after observing the data. One key challenge in Bayesian inference is the intractability of the marginal likelihood (the denominator of the posterior), which makes direct evaluation of the posterior distribution difficult. To overcome this challenge, we use approximation methods, specifically a kind of Markov Chain Monte Carlo (MCMC) sampling, to estimate the posterior distribution.

Our implementation employs Stan to sample from the posterior distributions in both our simulation studies and experimental settings, through Hamiltonian Monte Carlo sampling. This approach allows us to evaluate the predictive performance of our model and to assess whether we can accurately recover the true latent factors underlying the network relationships, ultimately determining whether these latent factors provide significant information about the observed dynamics.

1.5 Thesis Organization

The remainder of this thesis is organized as follows. In Chapter 2, we introduce the derivations behind our modeling framework and discuss the key features of the model. Chapter 3 presents simulation results that validate the model and highlight its performance

under controlled conditions. In Chapter 4, we apply the methods developed in the first three sections to real sports betting data, providing an experimental evaluation of our approach. Finally, Chapter 5 concludes with a discussion of our key findings and outlines potential avenues for future research.

2 Methodology & Model Adaptation

2.1 Betting Advantage

Let's start with a simple example to understand why we believe quantifying a betting advantage is possible. Suppose the Oklahoma City Thunder are playing the Boston Celtics, and the book predicts that the total score should be 50 points. Assuming this prediction is as accurate as possible, we would expect half of the bets to go on the Over and half on the Under. For example, if $10,000 is wagered on this Over/Under market, then ideally $5,000 would be bet on the Over and $5,000 on the Under. The book sets odds, typically around -110, meaning that if you bet $100 and win, you receive $90.90 in winnings along with your initial $100 bet.

Now, assume the book sets the line at 50 and the true total score goes Over. In a balanced market, with $5,000 wagered on Over and $5,000 on Under, the book's odds at -110 mean that winning bets yield a payout of approximately $90.90 on every $100 wagered. In this scenario, if the Over wins, the book would have to pay out roughly $4,545 to the winning Over betters. Meanwhile, the Under side, where bets lose, contributes the full $5,000 as profit. This ensures that the book makes a profit because the payout on the winning side is less than the amount collected from the losing side. This balanced outcome is why the book strives for a 50/50 split of bets between Over and Under, any deviation from this balance risks either reducing their profit margin or, in extreme cases, causing a loss.

However, gamblers do not always bet solely based on statistical accuracy. If the book sets a line at 50 but, for example, Boston natives automatically bet in favor of their home

team regardless of the true predicted score, this would skew the betting proportions. Suppose that instead of a 50/50 split, Boston fans contribute only $2,500 to the Under bets, resulting in $7,500 wagered on the Over. In this case, if the Over wins, the book would have to pay out $6,817.50 to the winners while collecting only $2,500 from the losers, leading to a profit loss. Table 2.1 shows the bookmaker's profit as the betting split changes in this scenario, providing further insight into why a 50/50 split is optimal:

Table 2.1: Bookmaker Profit Under Different Betting Splits (Odds: -110)

Scenario	Over Wagered	Under Wagered	Payout to Winners	Net Profit
50/50 Split	$5,000	$5,000	$4,545	$5,000
60/40 Split	$6,000	$4,000	$5,454	$4,000
70/30 Split	$7,000	$3,000	$6,364	$3,000
75/25 Split	$7,500	$2,500	$6,818	$2,500
80/20 Split	$8,000	$2,000	$7,273	$2,000

Assuming that the book has access to the best possible statistics, this leads one to believe they will adjust the line to encourage a balanced 50/50 distribution of bets, therefore ensuring a profit regardless of the outcome. However, this adjustment creates a discrepancy between the true score and the posted line. In effect, the bookmaker's strategy shifts the problem from predicting the true score to predicting the match-ups that will yield Over or Under outcomes according to their profit-maximizing design. Since these outcomes are intentionally structured so that fifty percent of bettors win, the posted statistics are skewed to favor the book's margins rather than reflecting the true game dynamics.

Therefore, rather than attempting to beat the book's true score predictions, which inherently has a statistical advantage for the bookmaker, our goal is to capitalize on the misalignment introduced by their line adjustments. By uncovering the latent structure of team match-ups that drives these dyadic relationships, we aim to identify the side (Over or

Under) that the book is positioning to win. In doing so, we attempt to exploit the systematic bias in the posted line to inform more profitable betting decisions.

2.2 Network Data Models

In this work, we evaluate undirected binary networks by modeling the dyadic relationships between nodes through eigendecomposition. Our network is comprised of n labeled $N = \{1, \ldots, n\}$ nodes. Each node represents an actor in the system, in our context each node represents a team, and the potential relationship between any two distinct nodes i and j is captured by and edge indicator $y_{i,j}$. Specifically,

$$y_{ij} = \begin{cases} 1 & \text{if an edge is present between nodes } i \text{ and } j, \\ 0 & \text{otherwise,} \end{cases}$$

for $1 \leq i < j \leq n$.

These indicators form the entries of an $n \times n$ adjacency matrix, denoted $\mathbf{Y}$. Since our network is undirected, $\mathbf{Y}$ is symmetric, which means $y_{i,j} = y_{j,i}$ for all $1 \leq i \leq\leq n$. The diagonal entries are undefined (or set to 0), since a node cannot form an edge with itself. The formation of this adjacency matrix follows:

$$\mathbf{Y} = \begin{pmatrix} 0 & y_{12} & \cdots & y_{1n} \\ y_{21} & 0 & \cdots & y_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ y_{n1} & y_{n2} & \cdots & 0 \end{pmatrix}.$$

This binary representation simplifies our analysis by focusing on the mere presence or absence of a connection, rather than on weighted interactions. It allows us to efficiently capture and analyze the fundamental structure of the network. By modeling the network in this way, we can apply eigendecomposition techniques to extract latent factors that underlie the observed dyadic relationships. These latent factors can reveal important systemic patterns, such as clusters of nodes that exhibit similar behavior or strategic interactions between teams. The adjacency matrix representation therefore, provides a clear framework for investigating the relationships that drive Over/Under outcomes in sports betting.

2.2.1 Homophily and Stochastic Equivalence

A fundamental advantage of modeling our network data with the eigenmodel is the ability to capture critical structural tendencies, homophily, and stochastic equivalence [4].

Homophily refers to the idea that nodes with similar attributes are more likely to be connected, meaning nodes with similar unobserved characteristics form edges with one another similarly. In context of Over/Under betting, this structure will manifest if pairs of teams with comparable playing tendencies consistently push the total score above the bookmakers posted line. These latent similarities are quantified through the use of latent space. This quantification can be explained as, if the latent position of two teams are close in space, our model infers a higher probability of forming an edge, an Over result.

Stochastic Equivalence captures the idea that nodes can share similar probabilities of forming multiple edges. Meaning, if nodes exhibit stochastic equivalence, they appear in distinct clusters and are considered exchangeable, since they form edges with others in the system with very similar probabilities. Two teams are stochastically equivalent, if they

exhibit similar patterns of Over/Under outcomes across their match-ups. Within our latent space representation, stochastic equivalence is said to occur when different teams occupy similar positions in latent space respective to their edges. The quantification assigns similar interaction probabilities for those with comparable Over/Under records.

The ability of our eigenmodel to unify homophily and stochastic equivalence into a single latent space embedding allows us to reveal teams that collectively deviate from the bookmakers posted line in a consistent manner. These latent relationships provides a systematic way of anticipating lines that are set too high or too low for particular match-ups, guiding more informed betting strategies.

2.3 Network Eigenmodel

To model the probability of an edge between nodes i and j, assuming the edges appear independently for each pairing of the teams, we introduce θ_{ij}, representative of the equation for our latent and global parameters. $\theta_{i,j}$ is then transformed into a probability through the logistic link function,

$$\text{expit}(\theta_{i,j}) = \frac{e^{\theta_{ij}}}{1 + e^{\theta_{ij}}},$$

which maps each θ_{ij} to the interval $[0, 1]$.

Given the symmetric nature of $\mathbf{Y}$ and the properties of eigendecomposition, we parameterize $\theta_{i,j}$ in terms of latent factors and a global baseline parameter. Let $\mathbf{U}$ be an $n \times k$ matrix where each row u_i represents the k-dimensional latent factor vector for node i. Let Λ be a $k \times k$ diagonal matrix with eigenvalues $\lambda_1, \ldots, \lambda_k$ on the diagonal, and let ζ denote a

global additive parameter that captures a baseline level of connectivity. We then define this component of our model as,

$$\theta_{i,j} = \zeta + u_i \Lambda u_j^T, \quad \text{for } 1 \leq i \leq j \leq n.$$

Since our relational data is binary, we model each dyadic outcome as a Bernoulli random variable. The sampling model is therefore given by,

$$y_{i,j} \mid \zeta, \Lambda, \mathbf{U} \sim \text{Bernoulli}\Big(\text{expit}[\zeta + u_i \Lambda u_j^T]\Big). \tag{2.1}$$

This formulation is consistent with the Eigen framework introduced by Hoff, in which each dyadic relationship is modeled as the weighted inner product of node-specific latent factors, conditioned on the global and structural parameters.

2.3.1 Betting Adaptation and Over/Under Betting Model

Under the assumption that bookmakers possess the most accurate statistics, we can place belief in the true combined score for a game between teams i and j, denoted by T_{ij}. However, rather than posting T_{ij} directly, the bookmaker sets the Over/Under line L to maximize profit by balancing the betting dollars on both sides. In this framework, the posted line is given by $L = T_{ij} + b_{ij}$, where b_{ij} is a bias introduced by the bookmaker to achieve a roughly 50/50 split between Over and Under bets.

We posit that this bias is driven by unobserved (latent) team characteristics. To capture these latent factors, we assign each team a k-dimensional latent vector u_i (and

similarly u_j for team j). In addition, we introduce a diagonal scaling matrix Λ (with diagonal elements $\lambda_1, \ldots, \lambda_k$) that determines the strength of the interaction between the teams' latent characteristics. We then re-parameterize the bias as $b_{i,j} = u_i \Lambda u_j^T$. Substituting this into the equation for L gives $L = T_{i,j} + u_i \Lambda u_j^T$.

Since the true score T_{ij} is not directly measured, we instead observe a binary outcome y_{ij} defined by whether the combined score exceeds the posted line:

$$y_{i,j} = \begin{cases} 1, & \text{if } T_{i,j} > L \quad \text{(Over outcome)}, \\ 0, & \text{otherwise (Under outcome)}. \end{cases}$$

Because the discrepancy between T_{ij} and L is considered to be fully captured by the latent interaction $u_i \Lambda u_j^T$, we define a latent linear predictor for the probability of an Over outcome as

$$\theta_{i,j} = \zeta + u_i \Lambda u_j^T,$$

where ζ is a global baseline parameter representing overall market tendencies. This predictor is then transformed into a probability via the logistic link (expit) function:

$$\Pr(y_{i,j} = 1 \mid \zeta, \Lambda, u_i, u_j) = \frac{e^{\theta_{i,j}}}{1 + e^{\theta_{i,j}}}.$$

Thus, the observed binary outcome is modeled as

$$y_{i,j} \sim \text{Bernoulli}\Big(\text{expit}[\zeta + u_i \Lambda u_j^T]\Big).$$

To model Over/Under betting within the Eigen framework, we transform our data into a network representation. Let $\mathbf{Y}$ denote the $n \times n$ adjacency matrix for our network of $N = 30$ NBA teams, with each team defined as a node in latent space. Each component $y_{i,j}$ of $\mathbf{Y}$ represents the observed outcome for the game between teams i and j; since the game between i and j is identical to that between j and i, the matrix is symmetric. In the context of Over/Under betting, the binary outcome is formally defined as

$$y_{i,j} = \begin{cases} 1, & \text{if the combined score of teams } i \text{ and } j \text{ exceeds the posted line (Over),} \\ 0, & \text{otherwise (Under).} \end{cases}$$

Figure 2.1 illustrates a social network representation of our adjacency matrix for the 30 NBA teams. An edge (i.e., $y_{i,j} = 1$) indicates that the combined score in a game exceeded the posted Over/Under line.

It is important to note that although our initial network representation is based on teams whose game outcomes exceed the bookmaker's posted prediction line, our modeling question requires further refinement. Specifically, the relationship defined as a team match-up going Over should be reinterpreted in terms of aligning with the winning side of the bettors. Given that the sports book operates as a business with the primary goal of splitting betting dollars evenly, the book is inherently indifferent to which side wins. Consequently, we consider the presence of an edge between nodes to indicate that the cumulative betting behavior—driven by the true score exceeding the posted line—provides insight into how the sports book adjusts the true score to a profit-maximizing posted line.

In summary, through reparameterization we transform the problem of reconciling the

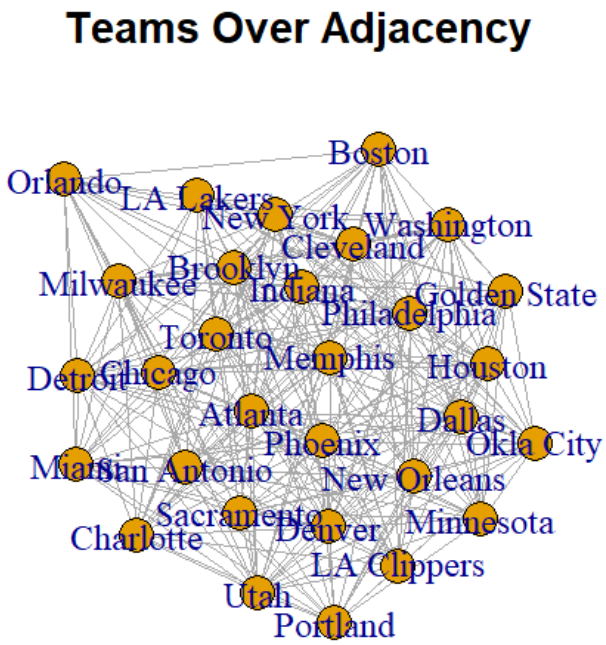

Figure 2.1: Social network representation of the adjacency matrix for NBA teams. An edge (value of 1) indicates that the combined score in a game exceeded the posted Over/Under line.

true score $T_{i,j}$ and the posted line L into one of estimating the latent interactions u_i and u_j (scaled by Λ) that drive the bias b_{ij}. This approach not only captures the dyadic relationships between teams in a social network framework but also reveals the latent structure underlying betting outcomes, enabling us to identify which side (Over or Under) the book has positioned to win.

2.4 Bayesian Hierarchical Model

We implement a Bayesian hierarchical model for the latent variables in our sampling model in Equation (2.1). This approach places prior distributions on our parameters and latent factors allowing us to incorporate uncertainty, for inference on the full posterior distribution.

Recall our sampling model from Equation (2.1),

$$y_{ij} \mid \zeta, \Lambda, \mathbf{U} \sim \text{Bernoulli}\Big(\text{expit}\big(\zeta + u_i \Lambda u_j^T\big)\Big),$$

where:

- ζ : global baseline parameter,
- $\mathbf{U}$: $n \times k$ matrix of latent factor vectors, with u_i representing the k-dimensional latent factors for node i.
- Λ :$k \times k$ diagonal matrix with eigenvalues $\lambda_1, \ldots, \lambda_k$ on the diagonal, scaling the interaction between the latent factors.

In the Bayesian hierarchical formulation, we place prior distributions on the parameters ζ, Λ, and $\mathbf{U}$, as follows:

$$\begin{aligned}
\zeta &\sim N\big(\mu_\zeta, \sigma_\zeta^2\big), \\
u_i &\sim N\big(\mathbf{0}, \sigma_u^2 I_k\big), \quad i = 1, \ldots, n, \\
\lambda_k &\sim N\big(\mu_\lambda, \sigma_\lambda^2\big), \quad l = 1, \ldots, k.
\end{aligned}$$

- μ_ζ and σ_ζ^2 are the mean and variance for the global baseline.
- σ_u^2 is the variance of the latent factors (with I_k as the k-dimensional identity matrix).
- μ_λ and σ_λ^2 govern the distribution of the eigenvalue scaling parameters.

We also place priors on the variance of our hyperparameters,

$$\sigma_\zeta^2 \sim \text{IGamma}(a_\zeta, b_\zeta),$$

$$\sigma_u^2 \sim \text{IGamma}(a_u, b_u),$$

$$\sigma_\lambda^2 \sim \text{IGamma}(a_\lambda, b_\lambda),$$

where a_ζ, b_ζ, a_u, b_u, and a_λ, b_λ are shape and rate parameters.

The full Bayesian hierarchical model for our Over/Under betting data is given by,

$$\begin{aligned}
y_{ij} \mid \zeta, \Lambda, \mathbf{U} &\sim \text{Bernoulli}\Big(\text{expit}\big(\zeta + u_i \Lambda u_j^T\big)\Big), \quad 1 \le i < j \le n, \\
\zeta &\sim N\big(\mu_\zeta, \sigma_\zeta^2\big), \\
u_i &\sim N\big(\mathbf{0}, \sigma_u^2 I_k\big), \quad i = 1, \ldots, n, \\
\lambda_l &\sim N\big(\mu_\lambda, \sigma_\lambda^2\big), \quad l = 1, \ldots, k, \\
\sigma_\zeta^2 &\sim \text{IGamma}(a_\zeta, b_\zeta), \\
\sigma_u^2 &\sim \text{IGamma}(a_u, b_u), \\
\sigma_\lambda^2 &\sim \text{IGamma}(a_\lambda, b_\lambda).
\end{aligned} \tag{2.2}$$

In the context of Over/Under betting, this framework therefore models the probability

of a game exceeding the posted line as a function of latent team factors, but also quantifies the uncertainty in these latent relationships. Imposing this hierarchical structure allows betting outcomes to provide inference on reconciling the bookmakers' true predictions with their profit-maximizing posted lines.

2.4.1 Hamiltonian Monte Carlo

Stan [5] uses Hamiltonian Monte Carlo (HMC) [6] to generate samples from the posterior distribution. HMC differs from conventional Markov Chain Monte Carlo (MCMC) methods by utilizing gradient information from the posterior distribution to guide its sampling trajectory, rather than relying on standard random-walk proposals. This gradient-based approach results in more effective exploration of the parameter space and a faster convergence to the posterior distribution, making it particularly suitable for our Over/Under data.

The Hamiltonian function is defined as follows,

$$H(\theta, r) = U(\theta) + K(r),$$

where θ is a vector of parameters to be sampled and r is an auxiliary momentum variable that facilitate exploration of parameter space. In the context of our Over/Under approach, HMC needs to explore the space of all possible θ values to determine how likely each parameter setting is, based on the Over/Under results observed in past games. This function $U(\theta)$ is described as the negative log posterior function:

$$U(\theta) = -\log p(\theta \mid y),$$

This transforms the posterior $p(\theta \mid y)$, which combines the prior belief about games Over/Under results for specific team matches with the likelihood of the observed data

The kinetic energy of the Hamiltonian $K(r)$ is typically taken to be the log-likelihood of a Gaussian distribution,

$$K(r) = \frac{1}{2} r^T M^{-1} r,$$

where M is a positive definite matrix which reflects the posterior covariance structure and employs the movement of the sampler to be guided by the gradients of $-\log p(\theta \mid y)$.

The sampling trajectories in HMC follow from Hamilton's equation of motion. This equation describes the evolution of the system in our parameter space,

$$\frac{d\theta}{dt} = \frac{\partial H}{\partial r} = M^{-1} r, \quad \frac{dr}{dt} = -\frac{\partial H}{\partial \theta} = -\nabla U(\theta).$$

These differential equations guide each proposal, so the parameters are drawn systematically to follow the shape of our posterior distribution for the Over/Under data.

Since the exact solutions of Hamilton's equations are typically intractable, Stan approximates these equations using leapfrog integration scheme,

$$r \leftarrow r - \frac{\epsilon}{2} \nabla U(\theta),$$

$$\theta \leftarrow \theta + \epsilon M^{-1} r,$$

$$r \leftarrow r - \frac{\epsilon}{2} \nabla U(\theta),$$

where ϵ represents the size of the discrete integration step, this leapfrog step is key to preserving the detailed balance condition in the MCMC scheme. After the leapfrog steps, a proposed new state (θ^*, r^*) is accepted with the Metropolis acceptance probability:

$$\alpha = \min\left[1, \exp\left(H(\theta, r) - H(\theta^*, r^*)\right)\right].$$

This acceptance criterion ensures that the parameter updates consistently reflect improved explanatory power for Over/Under outcomes, enhancing models' reliability and predictive accuracy.

3 Simulation Studies

In this section, we describe our simulation study designed to evaluate the performance and behavior of the eigenmodel under controlled conditions. This simulated dataset allows us to verify the model's reliability before applying it to real data. We also compare the simulated data with a traditional logistic regression and a simpler latent regression, to benchmark our ability to distinguish the predictive performance gain from using our eigenmodel compared to others.

3.1 Simulating Data from Network Eigenmodel

We begin by constructing an $n \times n$ adjacency matrix to represent a network of $n = 30$ NBA teams, following the hierarchical model introduced earlier. To generate the true underlying latent structure that our MCMC algorithm will attempt to recover, we first randomly sample the hyperparameters governing the variances of our priors:

$$\sigma_0^2 \sim \text{Inverse-Gamma}(a_\sigma,\, b_\sigma),$$

$$\kappa_0^2 \sim \text{Inverse-Gamma}(a_\kappa,\, b_\kappa),$$

$$w_0^2 \sim \text{Inverse-Gamma}(a_w,\, b_w).$$

Using these standard deviations (σ_0, κ_0, and w_0), we then sample the true latent parameters:

$$\mathbf{U}_0 \in \mathbb{R}^{n \times K} \sim \mathcal{N}\big(\mathbf{0},\, \sigma_0^2\, \mathbf{I}\big),$$

$$\mathbf{\Lambda}_0 \in \mathbb{R}^{K} \sim \mathcal{N}\big(\mathbf{0},\, \kappa_0^2\, \mathbf{I}\big),$$

$$\zeta_0 \sim \mathcal{N}\big(0,\, w_0^2\big).$$

These sampled priors define the true latent structure that will be used to generate the simulated network.

3.1.1 Generating the Simulated Network

Using the true latent parameters, we construct our simulated network data. For each pair of teams (i, j), we compute the linear predictor:

$$\theta_{ij} = \zeta_0 + u_{0,i}\Lambda_0 u_{0,j}^T,$$

and then draw the binary outcome y_{ij} from a Bernoulli distribution with success probability given by the expit function:

$$y_{ij} \sim \text{Bernoulli}\Big(\frac{e^{\theta_{ij}}}{1 + e^{\theta_{ij}}}\Big).$$

The resulting binary outcomes are stored in the adjacency matrix $\mathbf{Y}$, which represents our simulated network of NBA teams. Figure 3.1 shows the network structure generated from $\mathbf{Y}$.

3.1.2 HMC Diagnostics, Posterior Predictive Checks, and Reliability

We generate samples from the posterior distribution via HMC sampling. To assess convergence and the reliability of our model estimates, the generated samples, we examined several diagnostics:

- **Trace Plots:** These display the sampled values of each parameter over iterations, allowing us to assess chain mixing and convergence.

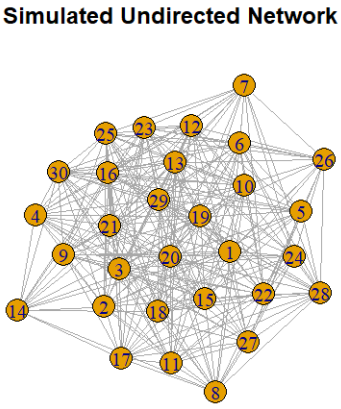

Figure 3.1: Graph of the simulated NBA network: nodes represent teams and edges indicate 'Over' outcomes.

- $\hat{R}$: Values close to 1 indicate that multiple chains have converged to the same target distribution.

- **Effective Sample Size (ESS):** Larger ESS values suggest that the chains are well-mixed, yielding a sufficient number of effectively independent samples.

Figure 3.3 and Figure 3.2 display the trace plots for key parameters, demonstrating that the chains mix well and converge.

We further evaluated the reliability of our model using posterior predictive checks. Figure 3.4 compares the density estimated from the posterior predictive distribution with the true simulation density (red). The close alignment between these densities confirms that our model reliably recovers the underlying distribution.

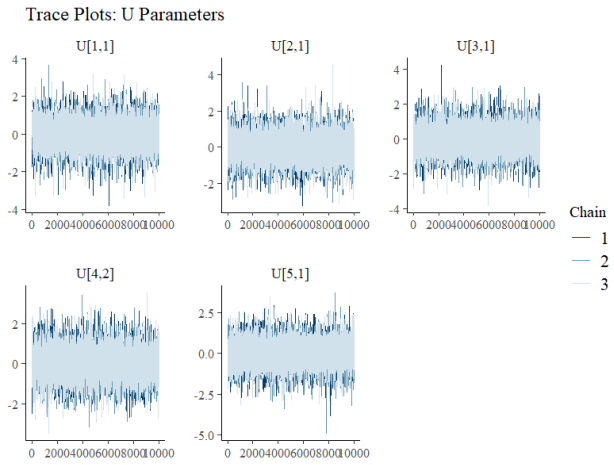

Figure 3.2: Trace plots for model parameters, indicating good mixing and convergence.

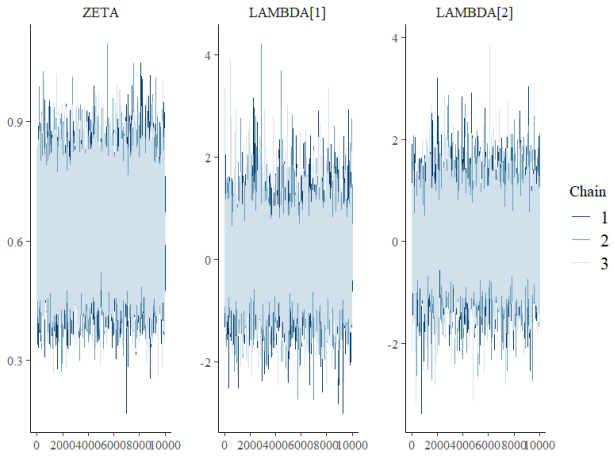

Figure 3.3: Trace plots for model parameters, indicating good mixing and convergence.

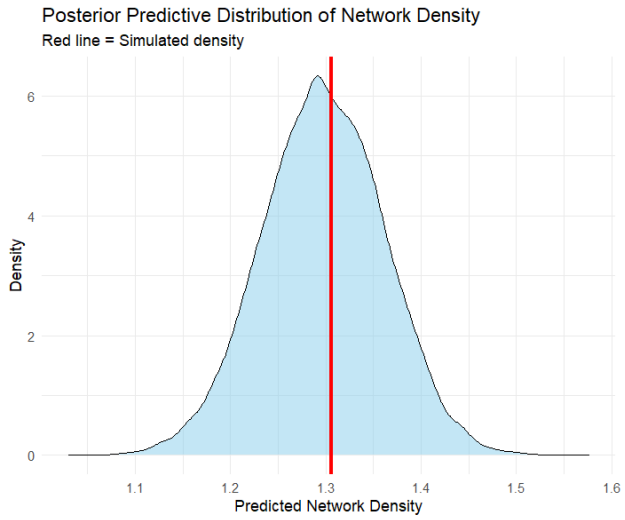

Figure 3.4: Posterior predictive density of eigenmodel simulation compared to the true simulation density (red).

3.1.3 Latent Space Visualization

Finally, the eigenmodel assigns each team a position in a low-dimensional latent space. Figure 3.5 displays the posterior mean latent positions for the simulated data. Teams that are closer together in this latent space have a higher probability of forming an edge (meaning, producing an Over outcome), while those farther apart are less likely to do so. This visualization provides an intuitive representation of the underlying latent structure captured by the model.

Overall, the combination of MCMC diagnostics, posterior predictive checks, and latent space visualization provides strong evidence that the eigenmodel provides a good fit for our data. The consistency of the diagnostic measures and the alignment between the posterior predictive and true densities indicate that our simulation framework effectively captures the underlying structure of the data, and the model can be trusted for real-data analysis.

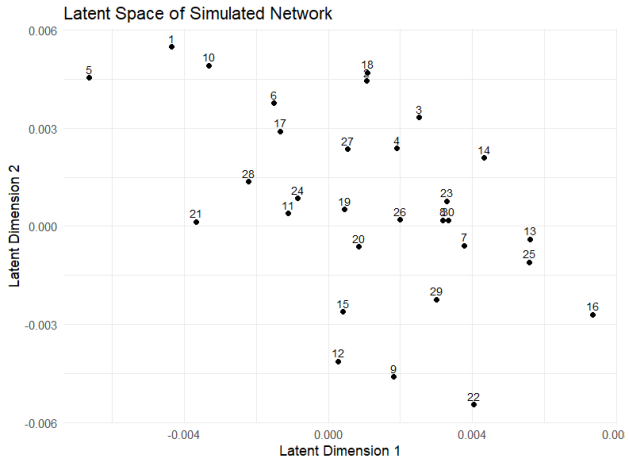

Figure 3.5: Posterior mean latent positions of teams in the simulated eigenmodel latent space.

3.2 Covariate Eigenmodel

In typical sports betting models covariate information is highly utilized, therefore we extend our eigenmodel to include covariates in the same manner that [7] introduced, since we are confident in the fit of an Over/Under adjacency matrix to the eigenmodel. To verify that this model is appropriate for the given data, we generate a simulation beginning with the construction of an $n \times n$ adjacency matrix to represent the network of $n = 30$ NBA teams, incorporating both underlying latent structures and observed team-level covariate effects.

To generate the simulation data we will use as a benchmark to check our models' reliability, we first sample the hyper-parameters that govern the variances of our priors:

$$\sigma_0^2 \sim \text{Inverse-Gamma}(a_\sigma,\, b_\sigma),$$

$$\kappa_0^2 \sim \text{Inverse-Gamma}(a_\kappa,\, b_\kappa),$$

$$w_0^2 \sim \text{Inverse-Gamma}(a_w,\, b_w).$$

Using these Variances, we then sample the true latent parameters:

$$\mathbf{U}_0 \in \mathbb{R}^{n \times K} \sim \mathcal{N}\big(\mathbf{0},\, \sigma_0^2\, \mathbf{I}\big),$$

$$\mathbf{\Lambda}_0 \in \mathbb{R}^{K} \sim \mathcal{N}\big(\mathbf{0},\, \kappa_0^2\, \mathbf{I}\big),$$

$$\zeta_0 \sim \mathcal{N}\big(0,\, w_0^2\big).$$

We then simulate the team-level covariates. For each team i where $i = 1 \ldots, n$, we generate a vector $X_i \in \mathbb{R}^P$ from a standard multivariate normal distribution $X_i \sim \mathcal{N}\big(\mathbf{0},\, \mathbf{I}\big)$ where the true covariate effects are represented by a coefficient vector $\boldsymbol{\beta}_0 \in \mathbb{R}^P$. Using the true latent and covariate parameters, we generate our simulated network. For each pair of teams i, j for $i \neq j$, the linear predictor follows:

$$\theta_{ij} = \zeta_0 + \boldsymbol{\beta}_0^\top\, (X_i + X_j) + \mathbf{U}_{0,i}\, \mathbf{\Lambda}_0\, \mathbf{U}_{0,j}^\top,$$

The probability of an edge is given by the logistic transformation:

$$p_{ij} = \frac{1}{1 + \exp(-\theta_{ij})}.$$

The binary outcome for each dyad is then drawn from a Bernoulli distribution:

$$y_{ij} \mid p_{ij} \sim \text{Bernoulli}(p_{ij}),$$

We then store the resulting outcomes in an adjacency matrix $\mathbf{Y}$, which allowed us to process the simulated data in a STAN implementation of the co-variate eigenmodel for HMC sam-

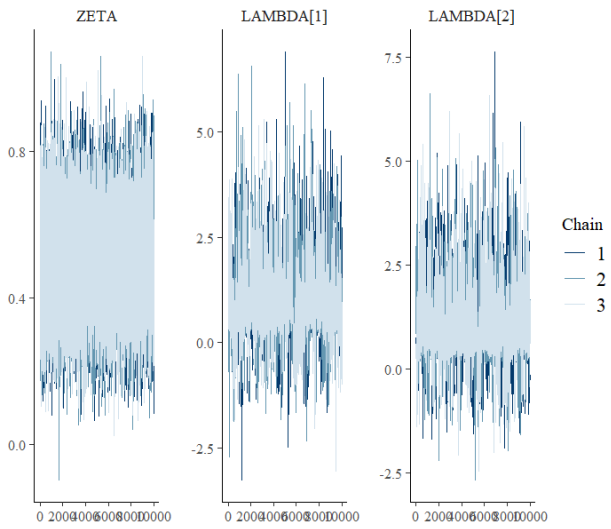

Figure 3.6: HMC trace plots for the intercept parameters ζ and λ in the covariate-augmented eigenmodel.

pling. The reliability of this model was verified through posterior checks identical to those performed in the previous simulation. Specifically, we examined trace plots for parameters (Figures 3.6, 3.7, and 3.8), evaluated $\hat{R}$ values, and computed effective sample sizes (ESS), all of which demonstrated acceptable results.

3.3 Simple Logistic Regression

In contrast to the eigenmodel, which leverages latent factors to capture dyadic relationships, a simple logistic regression model only has the ability to incorporate observed characteristics for teams i and j into the edge formation process. Without the ability to extract latent factors, the model cannot recover the nuanced, team-specific information that the eigenmodel provides.

We apply the simulated network to model this simpler framework, by assuming that the probability of an edge between any two teams i and j is given solely by a baseline log-odds

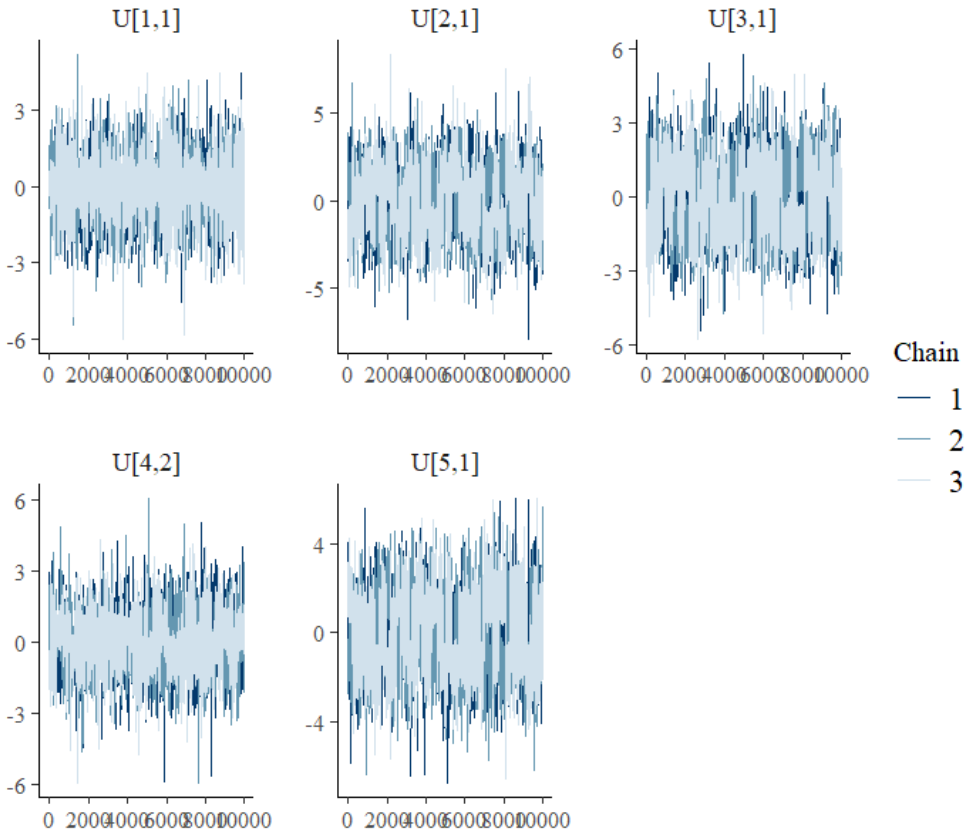

Figure 3.7: Trace plots for the first two dimensions of the latent position matrix U across HMC iterations.

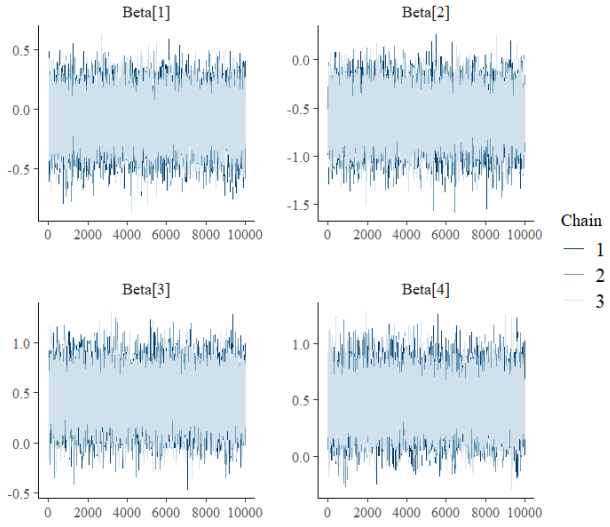

Figure 3.8: HMC trace diagnostics for the regression coefficients β in the covariate eigenmodel.

parameter ζ with co-variates $\beta^t X_{ij}$. Specifically, the sampling model is defined as,

$$y_{i,j} \sim \text{Bernoulli}\big(p_{ij}\big),$$

where $\text{logit}(p_{ij}) = \zeta + \beta^t X_{ij}$. Here ζ is the baseline log-odds parameter, *beta* quantifies the effect of the observed co-variate, and $X_{i,j}$ is the co-variate information for the dyad i, j. This formulation allows the model to adjust the edge probability based on only measurable characteristics.

The hierarchical structure of this model's parameters follow,

$$w^2 \sim \text{Inverse-Gamma}(a_w,\, b_w), \quad \zeta \sim \mathcal{N}(0,\, w^2), \quad \beta \sim \mathcal{N}(0,\, I_P).$$

We reuse the same generated data from the covariate Eigen simulation to ensure comparability between the models. We then, processed the simulated data using a Stan implementation of the simple logistic regression model. We then verified the reliability of this model by performing the same posterior checks as in the Eigen simulation, including examining trace plots, $\hat{R}$ values, and effective sample sizes (ESS).

Overall, the simple logistic regression model serves as a baseline, highlighting the advantages of the eigenmodel's ability to capture latent dyadic relationships in network data. Figure 3.9 displays the posterior predictive density of our simple logistic regression model, with the true network density indicated by a red dashed line. As seen in the plot, the predicted density is centered far from the true value. This misalignment highlights the limitations of the simple logistic regression approach,

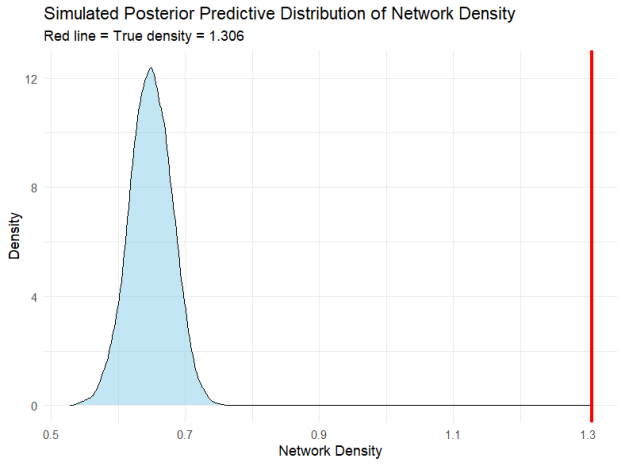

Figure 3.9: Posterior predictive density of the simple logistic regression model (blue) compared to the true network density (red dashed line).

3.4 Latent Additive Logistic Regression

We will also compare the eigenmodel with the Latent Additive Logistic Regression (LALR) model introduced in [8]. The LALR allows us to incorporate latent node-specific factors as additive terms in a logistic framework, enabling node-specific latent information to be captured, providing an extension from the simple logistic model, but still lacking the

further complexity of an eigenmodel. The sampling model and hierarchal structure follows,

$$
\begin{aligned}
y_{ij} &\sim \text{Bernoulli}(\theta_{ij}), \quad i < j, \\
\text{logit}(\theta_{ij}) &= \zeta + \beta^t X_{ij} + u_i + u_j, \\
\zeta &\sim \mathcal{N}(0,\, w^2), \\
\beta &\sim \mathcal{N}(0,\, I), \\
u_i &\sim \mathcal{N}(0,\, \sigma^2), \quad i = 1, \ldots, n, \\
w^2 &\sim \text{Inv-Gamma}(a_w,\, b_w), \\
\sigma^2 &\sim \text{Inv-Gamma}(a_\sigma,\, b_\sigma).
\end{aligned}
\tag{3.1}
$$

The sampling model utilizes the same prior distributions for parameters as established in the previous models, ensuring consistency between simulations.

We represent teams as nodes indexed by $N = 1, \ldots, n$. Edges between nodes are defined as binary indicators, specifically capturing whether a matchup between teams i and j exceeded the predicted total (*Over* result). To compare the amount of information captured by different approaches, we processed the same simulated data using a Stan implementation of the LALR model. The reliability of this model was verified through posterior checks identical to those performed in the Eigen simulation. Specifically, we examined trace plots (Figures 3.10 and 3.11), evaluated $\hat{R}$ values, and computed effective sample sizes (ESS), all of which demonstrated acceptable results.

We then plotted the density of the predicted network for this model against the true simulated density. Figure 3.12 clearly demonstrates that this model does not recover the

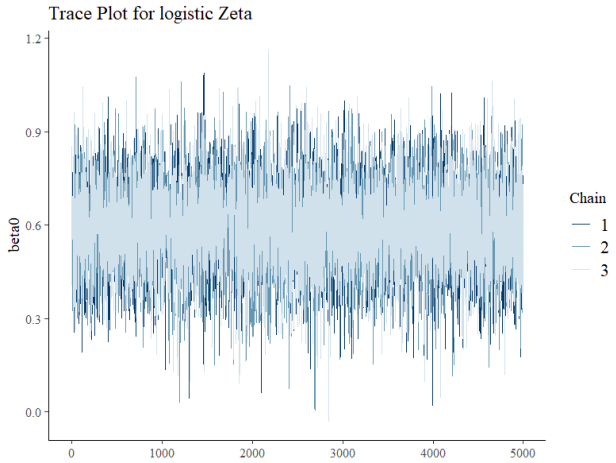

Figure 3.10: Trace plot for the global intercept parameter ζ from the LALR model in Equation (3.1).

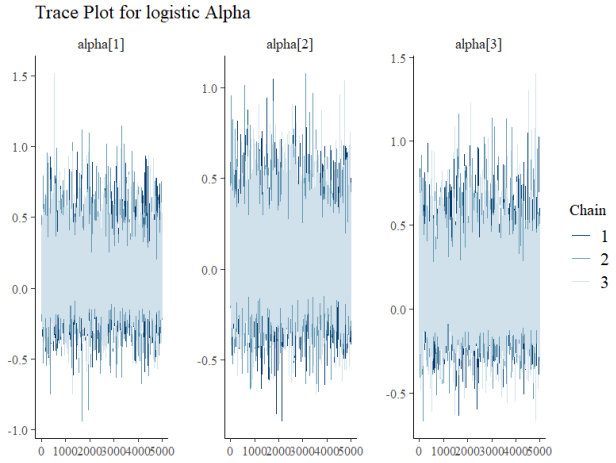

Figure 3.11: Trace plot for the node-specific latent parameters u_i from the LALR model.

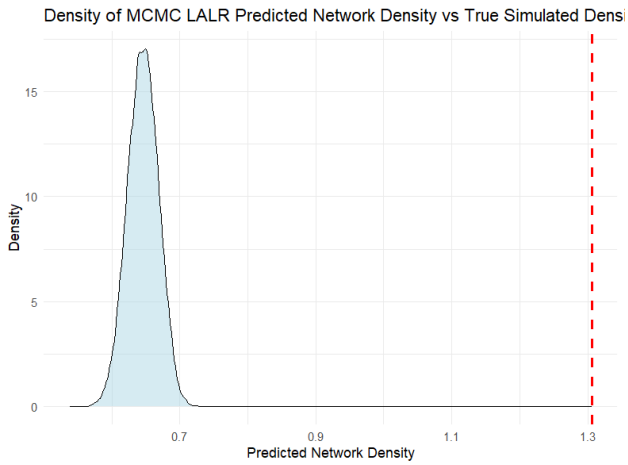

Figure 3.12: Comparison of the posterior predictive distribution of network density to the true simulated density (red line), highlighting the poor recovery performance of the LALR model.

true density effectively.

3.5 Simulation Conclusion

From the three posterior predictive plots, it is clear that introducing latent factors from the simple logistic model to the LALR model improves density prediction, although neither proves fully reliable to fit this data. However, extending from the LALR model to the eigenmodel shows significantly improved predictive reliability, confirming the motivation for the use of an eigenmodel.

4 Over/Under Betting Model for 2023 NBA Regular Season

The adaptation of the covariate eigenmodel to Over/Under betting can be applied across many sports, and we will be modeling this with the NBA. The data set utilized for our experiment comprises historical NBA betting and performance data for the 2023-2024 season. This publicly shared data set was compiled from [9]. Additionally, the covariate information for our experiment was obtained from [10] and includes offensive average statistics for each team during the 2022-2023 season, such as points per game (PPG), field goal percentages (FG%), three-point percentages (3P%), rebounds (REB), and assists (AST).

4.1 Modeling Limitations and Simplifications

One limitation of modeling the data in this manner is that teams play each other multiple times during the season, and our binary representation fails to capture the nuances of each matchup. To address this issue, we sort the data by the number of matches and construct a unique adjacency matrix for each game count, denoted as $M_1, M_2, \ldots, M_m$, where m represents the total number of match-ups between any two teams in a season. Figure 4.1 displays the network representation of the first match, M_1, adjacency matrix.

Additionally, we constructed a matrix of the offensive statistics for each team, which

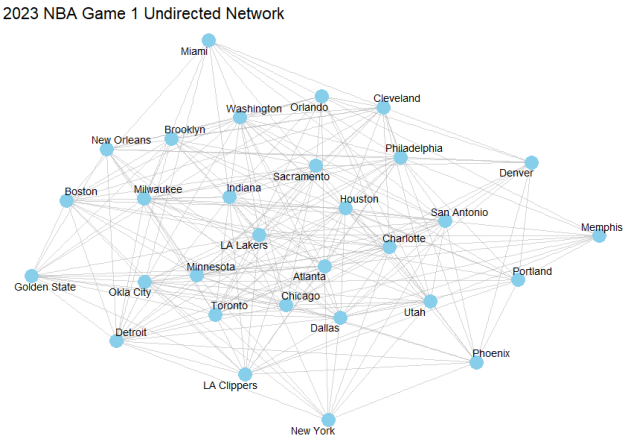

Figure 4.1: Network representation of the first encounter for each team pair (M_1), where edges indicate an 'Over' outcome in that matchup.

serves as covariate information for each dyad between team i and j:

$$\mathbf{X} = \begin{bmatrix} \text{PPG}_1 & \text{FG\%}_1 & \text{3P\%}_1 & \text{REB}_1 & \text{AST}_1 \\ \text{PPG}_2 & \text{FG\%}_2 & \text{3P\%}_2 & \text{REB}_2 & \text{AST}_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \text{PPG}_n & \text{FG\%}_n & \text{3P\%}_n & \text{REB}_n & \text{AST}_n \end{bmatrix}$$

where $n = 30$ is the total number of NBA teams.

For any matchup between team i and team j, the combined covariate information is given by the average:

$$X_{i,j} = 0.5\,(x_i + x_j)$$

We then processed the M_1 adjacency matrix using our Stan implementation of the covariate eigenmodel. Where the sampling model is defined as:

$$Y_{i,j} \sim \text{Bernoulli}\left(\text{logit}^{-1}\left[\zeta + \beta^\top X_{i,j} + u_i^\top \Lambda u_j\right]\right),$$

where,

- ζ is an intercept term which determines the baseline probability of a game being Over.
- β is a vector of regression coefficients associated with the covariates.
- $X_{i,j} = 0.5\,(x_i + x_j)$ denotes the averaged covariate vector for team i and j.
- u_i and u_j are latent factor vectors for team i and j.
- Λ is the diagonal scaling matrix which controls the stochastic equivlence of the model.

We then performed the same MCMC diagnostic tests as in the simulation section. Trace plots, $\hat{R}$ plots, and effective sample size (ESS) values all indicate that the covariate eigenmodel is appropriate for the given data.

Figure 4.2 compares the predicted density to the true density of the network, demonstrating that our model reliably recovers the true network density for the first meeting between teams.

One of the key motivations for using the covariate eigenmodel is its ability to uncover latent structures among teams. To facilitate the interpretation of network features, we applied K-means clustering with four clusters to the latent embedding plot for M_1. Figure 4.3 displays this plot, where each colored convex hull highlights a different cluster, and each point corresponds to an NBA team. Teams that fall within the same cluster exhibit similar latent factor values, suggesting that they share underlying characteristics such as playing style or overall competitiveness [3]. In addition, nodes highlighted in black represent teams with high homophily. Here, homophily is measured using a nearest neighbor metric based on the average Euclidean distance between nodes in latent space. Teams are indicated to have

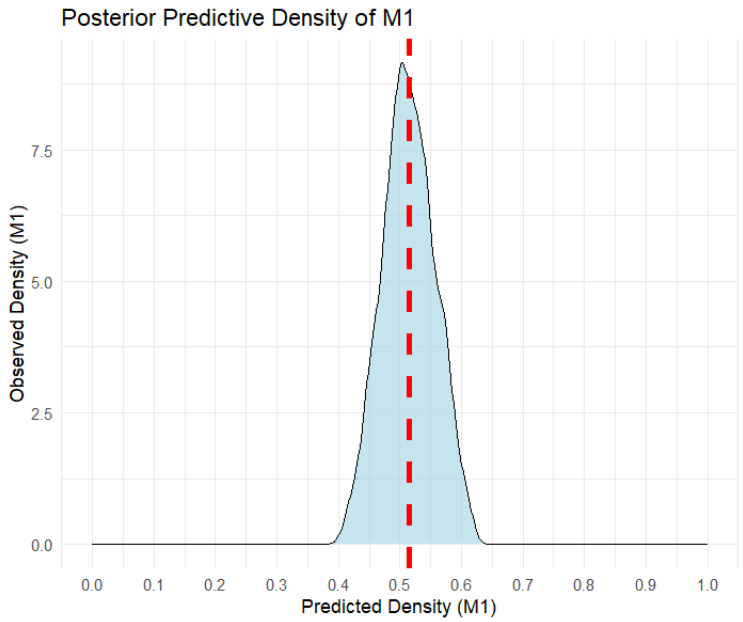

Figure 4.2: Comparison of the predicted and true (red line) network density for the first meeting between teams.

strong local connectivity based on the lower their average neighbor distance is, therefore, these are considered to have highly homophily. This displays the ability of the eigenmodel to capture homophily since the majority of our high homophily nodes are within the same cluster.

To assess homophily among Under-outcome pairs, we inverted our adjacency matrix, meaning:

$$\mathbf{M}_{\text{under}} \;=\; 1 - \mathbf{M}_1,$$

and then recomputed each team's average Euclidean distance to its Under neighbors as described previously. Figure 4.4 displays the resulting latent embedding with the same four k-means clusters, but now the black circles mark teams whose Under-neighbor distances fall below the 25th percentile, indicating strong Under homophily. As we can see from the two plots, the clustering of homophily nodes, as over or under, are within similar regions in

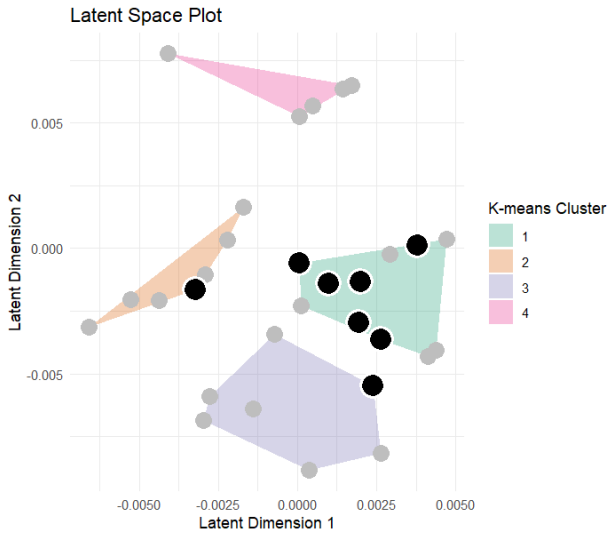

Figure 4.3: M_1 latent embedding with four k-means clusters (colored hulls) and high Over homophily pairs (bottom 25th percentile of over-neighbor distance) denoted with black circles

latent space. This consistency confirms that the covariate eigenmodel effectively captures homophily patterns across both Over and Under betting results. While the overlap of nodes in both structures indicate that these nodes are highly sensitive to their opponents strength in ability to drive over or under results.

To test the predictive power of our posterior samples, we use the M_1 samples as a training set and the M_2 observed outcomes as a test set. To visualize the predictive power of our model, Figure 4.5 displays the predicted probabilities from M_1 posterior distribution as a heat map over the adjacency matrix of M_2 where the observed "Over" outcomes are denoted as blue "O" markers. As seen in the heat map, higher probabilities in M_1 correspond to more accurate "Over" predictions in M_2, while lower probabilities are indicative of more accurate "Under" predictions.

Table 4.1 displays a confusion matrix for M_2 game outcomes using our predicted probabilities from M_1's posterior distribution. We assessed the model's discriminative power

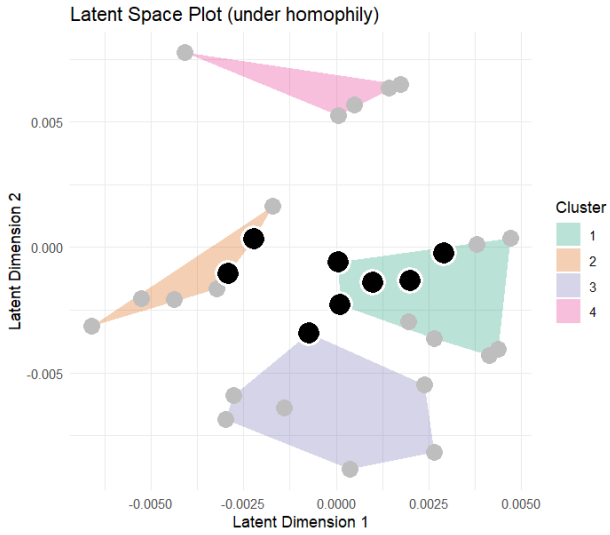

Figure 4.4: M_1 latent embedding with four k-means clusters (colored hulls) and high Under homophily pairs (bottom 25th percentile of under-neighbor distance) denoted with black circles

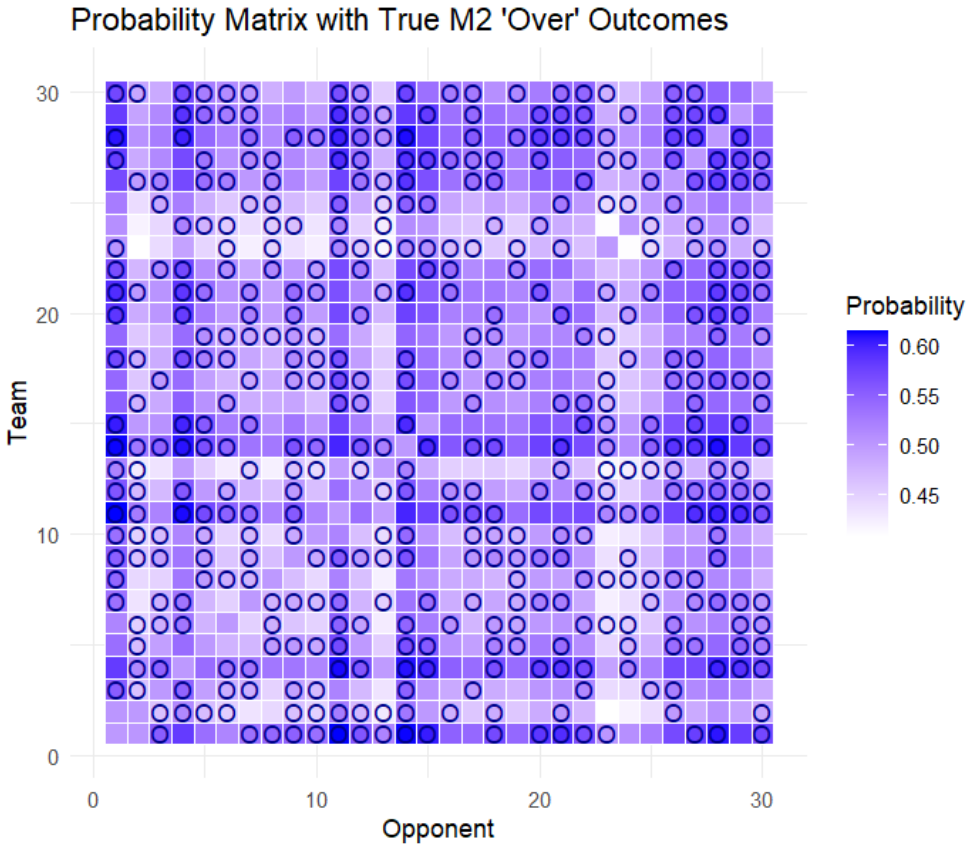

Figure 4.5: M1 posterior probabilities heat map matrix over M2 (blue "O") matrix

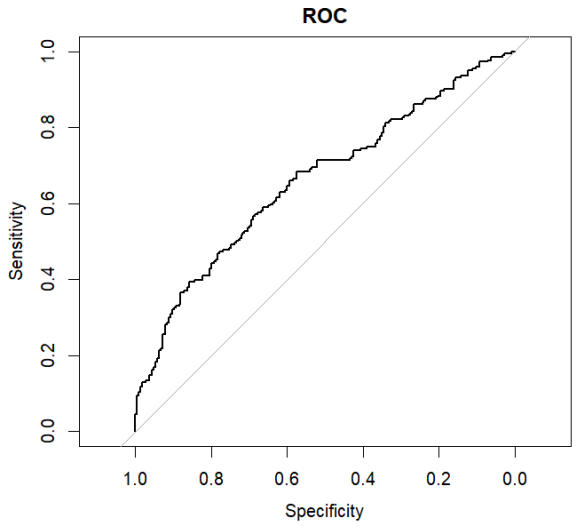

Figure 4.6: ROC curve for M_2 true outcomes from M_1 posterior predictive probabilities

using the Receiver Operating Characteristic (ROC) curve displayed with Figure 4.6, which plots the true positive rate against the false positive rate. The area under the curve (AUC) then indicates how well the model distinguished between positive and negative outcomes, we got an AUC value of 0.6607.

Prediction / Reference	**0**	**1**
0	260	142
1	192	306

Table 4.1: Confusion matrix for M_2 true outcomes based on M_1 predicted probabilities.

From a statistical standpoint, these values may appear moderate; however, in the context of sports betting, the primary objective is not necessarily achieving overall accuracy but rather reliably identifying instances where the model detects a betting advantage. Therefore, our modest accuracy can translate into systemically signaling opportunities for profit in the betting market.

Based on these observations, we set different cutoffs to assess M_1's prediction power for M_2. We evaluated the accuracy of the M_2 predictions using different thresholds, the

results are displayed in Table 4.2, which were found in R. These outcomes demonstrate that our model achieves performance that provides a betting advantage when predicting an "Over" result with specific levels of certainty or higher, and an "Under" result when the certainty of an "Over" prediction is at the stated threshold or less. We applied the same cutoff to our simple logistic regression model for the given data, to provide further insight into the additional information our eigenmodel can provide.

	$\Pr(\text{Over}) \leq \tau$			$\Pr(\text{Over}) \geq \tau$			
Threshold τ	0.43	0.44	0.45	0.57	0.58	0.59	0.60
EM Accuracy	0.70	0.81	0.72	0.78	0.87	0.94	1.00
LR Accuracy	0.51	0.51	0.51	0.49	0.49	0.49	0.49

Table 4.2: Comparison of model accuracies at selected posterior-probability thresholds for eigenmodel (EM) and Simple Logistic Regression model (LR) posterior predictions from M_1 on the true data from M_2.

5 Conclusion and Future Work

In conclusion, the use of latent space models in Over/Under sports betting provides a robust framework for gaining a competitive betting edge. Our approach, while effective, inherently contains limitations that lead to natural extensions for more sophisticated and nuanced models that enhance predictive accuracy. The separation of matrices per match gives a natural extension to sequential updating, where one could use the MCMC posterior samples from M_1 as prior information for processing M_2 in our Stan implementation of the covariate eigenmodel. This can done following the same hierarchical structure as before, with minimal changes to the prior information. Extensions rooted in the eigenmodel, such as the AMEN proposed by [8], offers the ability to incorporate additional complexities. As well as, temporal extensions, such as the sequential updating extension, or the utilization of weighted matrices would allow for further steps towards capturing more accurate predictions in sports betting Over/Under outcomes.

Bibliography

[1] E. L. Bishop, "Murphy v. ncaa," *Ohio NUL Rev.*, vol. 45, p. 239, 2019.

[2] K. Nowicki and T. A. B. Snijders, "Estimation and prediction for stochastic blockstructures," *Journal of the American statistical association*, vol. 96, no. 455, pp. 1077–1087, 2001.

[3] P. D. Hoff, A. E. Raftery, and M. S. Handcock, "Latent space approaches to social network analysis," *Journal of the american Statistical association*, vol. 97, no. 460, pp. 1090–1098, 2002.

[4] P. D. Hoff, "Bilinear mixed-effects models for dyadic data," *Journal of the american Statistical association*, vol. 100, no. 469, pp. 286–295, 2005.

[5] B. Carpenter, A. Gelman, M. D. Hoffman, D. Lee, B. Goodrich, M. Betancourt, M. Brubaker, J. Guo, P. Li, and A. Riddell, "Stan: A probabilistic programming language," *Journal of statistical software*, vol. 76, pp. 1–32, 2017.

[6] R. M. Neal *et al.*, "Mcmc using hamiltonian dynamics," *Handbook of markov chain monte carlo*, vol. 2, no. 11, p. 2, 2011.

[7] P. Hoff, "Modeling homophily and stochastic equivalence in symmetric relational data," *Advances in neural information processing systems*, vol. 20, 2007.

[8] ——, "Additive and multiplicative effects network models," *Statistical Science*, vol. 36, no. 1, pp. 34–50, 2021.

[9] OddsShark. Nba database. Accessed: 15 April 2025. [Online]. Available: https://www.oddsshark.com/nba/database

[10] (2022) Nba team offense statistics 2022-2023. Accessed: 2023-04-07. [Online]. Available: https://www.covers.com/sport/basketball/nba/statistics/team-offense/2022-2023