Bayesian Soccer Penalties

Benedikt Farag
Department of Statistics and Data Science
Yale University

December 16, 2025

Abstract

Penalty kicks in soccer are typically associated with a high probability of scoring a goal. However, various factors come into play such as goalkeeper and taker abilities along with the location of the shot in the goal's plane. Here we develop a Bayesian framework to determine the optimal aim point $\boldsymbol{\mu}$ which maximizes the taker's expected probability of scoring a goal, conditioned on the aiming error which quantifies the taker's shooting precision. Using a dataset of 1,361 penalties from professional soccer leagues, we estimate the posterior *goal probability surface* using a Bayesian Generalized Additive Mixed Model (GAMM) for binary outcomes. We then quantify the taker's error by using their observed spread using a Location-Scale model, which serves as an approximation for the aiming error. To determine the optimal strategy, we convolve the goal probability surface with the taker's error profile, which results in a *risk-adjusted probability surface*. Our results demonstrate that the optimal aim point is not static but varies significantly based on player consistency. High-precision takers can optimize their probability of scoring by targeting in the upper corners while low-precision takers should avoid those regions. This approach provides a useful method to maximize a taker's probability, given their aiming error distribution. The next step of this work should be to perform controlled experiments to quantify the aiming error profile accurately, providing players with their own optimized strategy. Our code is available at: `https://github.com/benedikt20/bayesian-penalties`.

1 Introduction

Penalty kicks in soccer are often considered an event that should most likely end up as a goal for the taker's team. However, in some cases, the goalkeeper saves or the taker shoots off target. For such a binary event (either a `goal` or `no goal`), are there any optimal locations over the goal frame where players should aim to maximize their probability of scoring a goal? This optimization can vary for different combinations of goalkeepers and takers, based on their strengths and weaknesses.

In this study, we develop a Bayesian framework to model penalty kick outcomes in a 2D space over the goal frame. First, we estimate the *goal probability surface* based on the observed ball location, i.e. $P(\text{goal} \mid \text{location})$. However, this does not account for the uncertainty that takers typically do not shoot exactly where they intended to shoot. In professional soccer, this *aiming error* is moderate but still exists. To account for this uncertainty, we can use a taker's precision covariance matrix $\boldsymbol{\Sigma}$ (with parameters σ_y, σ_z, ρ_{yz}) to estimate the aim locations which maximize the probability of scoring a goal.

We estimate the probability of scoring a goal given an aim location, i.e. $P(\text{goal} \mid \text{aim})$, by convolving the taker's error distribution over the goal probability surface. This approach transforms the problem from analyzing where shots landed to recommendations about where players should

aim, revealing that the optimal strategy shifts significantly depending on a player's precision parameters.

2 Methods

2.1 Dataset

For the analysis we use a scraped dataset from the StatsBomb database [3] with 1361 penalties from 3464 soccer matches. The data includes the ball's observed location (y, z) in the plane of the goal, with y horizontal and z vertical. Other relevant features are summarized in Table 1. In this study we will only be focusing on the shot location, outcome, along with taker and goalkeeper identities.

Table 1: Relevant features in the dataset

Description	**Feature**
Penalty outcome	`outcome`
Minute of the match	`minute`
Shot location on the goal	`y_end`, `y_end`
Taker and goalkeeper name	`taker`, `goalkeeper`
Goalkeeper results	`goalkeeper_action` `goalkeeper_outcome`
Game information	`match_date`, `home_team`, `away_team`

The outcome categories in the raw data included five[1] categories. Figure 1 shows the distribution of the data, colored by outcome.

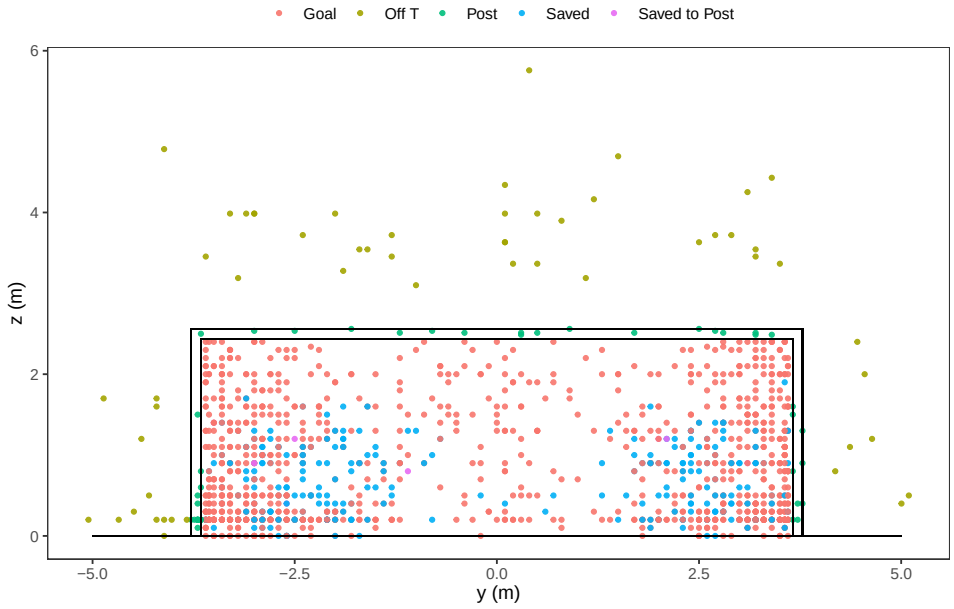


Figure 1: Shot locations of penalty kicks in the dataset

[1]The outcome class "wayward" were invalid penalty kicks, which were removed from the data

The categories were simplified to three main categories (from the raw categories) as follows:

Goal : Goal | **Saved** : Saved, Saved to Post | **Off T** : Off T, Post

2.2 Binary Outcome Model (Naïve Model)

To estimate the probability of scoring given the location of the shot,

$$P(\text{goal} \mid \text{location})$$

we condition on the goal frame, excluding shots that hit the post or were off target. It would not be meaningful to estimate the probability of scoring for locations outside the goal frame, which should intuitively be zero.

So we construct a binary variable $Y_i \in \{0, 1\}$ where $Y_i = 1$ denotes a goal and $Y_i = 0$ denotes no goal (saved). We fit a Bayesian Generalized Additive Mixed Model (GAMM) using the Bernoulli family with a logit link function. We utilize a tensor-product smooth spline function $t_2(y, z)$ which enables uneven scaling between y and z (since the goal frame is not rectangular).

We include random intercepts for taker and goalkeeper in our model, to account for individual skill levels:

$$\text{logit}(p_i) = \ln \frac{p_i}{1 - p_i} = f(y_i, z_i) + \alpha_{\text{gk}[i]} + \beta_{\text{taker}[i]} \tag{1}$$

where

$$\begin{aligned} p_i = P(Y_i | y_i, z_i) &\quad \text{is the probability of scoring at the location } (y_i, z_i), \\ \alpha_{\text{gk}[i]} \sim N(0, \sigma^2_{\text{gk}}) &\quad \text{are the random intercepts for goalkeepers} \\ \beta_{\text{taker}[i]} \sim N(0, \sigma^2_{\text{taker}}) &\quad \text{are the random intercepts for takers} \end{aligned}$$

This model estimates a probability surface over the goal frame for any combination of taker and goalkeeper. We calculate the probability surface for an average taker and goalkeeper by setting the random effects to zero (with $\alpha = \beta = 0$), to obtain a goal probability surface $S(y, z)$, which is based on the the population-level mean estimates.

The model (1) was implemented in brms package [1]. We used default weakly-informative normal priors and sampled from the posterior distribution using 4 chains, each with 2000 iterations (1000 warmup steps), with the model successfully converging ($\hat{R} < 1.01$).

2.3 Taker Location-Scale Model

The naïve binary outcome model (Eq. (1)) quantifies the success probability, but totally neglects that the resulting location is not the same as the aimed location. This aiming error is hard to quantify, without controlled experiments where the intended target is known. Here, we approximate this error using the *observed shooting spread* of players. However, this introduces a potential bias since the observed spread consists of *strategic variance* (aiming at different locations) and *shooting variance* (missing the aimed target). For takers that intentionally shoot over a wide range of location will observe a high aiming error, which is misleading. This will be considered for the interpretation of the results.

To quantify this aiming error (as shooting spread), we subset the data to only include takers with $n \geq 5$ number of penalties. We assume that the shot locations follow a Student-t distribution, which is more robust to outliers than the normal distribution (due to the heavier tails). We model the shot location (y_{abs}, z) as a multivariate Student-t distribution where both the location (aim point)

and scale (precision) varies by each taker. We use the absolute horizontal coordinate $y_{abs} = |y|$ to account for the symmetry of the goal, so left- and right-shots exhibit comparable precision patterns. The model is specified as

$$\begin{pmatrix} |y_i| \\ z_i \end{pmatrix} \sim \mathcal{T}(\nu, \boldsymbol{\mu_i}, \boldsymbol{\Sigma_i}) \tag{2}$$

where

$$\boldsymbol{\mu_i} = \begin{pmatrix} \mu_y \\ \mu_z \end{pmatrix} + \begin{pmatrix} u_{y,\text{taker}[i]} \\ u_{z,\text{taker}[i]} \end{pmatrix}$$

is the location vector with fixed global intercept and random effects

$$\mathbf{u}_{\text{taker}} \sim N(\mathbf{0}, \boldsymbol{\Sigma}_{\text{taker}}).$$

We model the standard deviations σ_y and σ_z on log-scale to ensure positivity, for every taker:

$$\boldsymbol{\Sigma_i} = \begin{pmatrix} \sigma_{y,i}^2 & \rho_{yz}\sigma_{y,i}\sigma_{z,i} \\ \rho_{yz}\sigma_{y,i}\sigma_{z,i} & \sigma_{z,i}^2 \end{pmatrix} \tag{3}$$

where

$$\log(\sigma_{y,i}) = \gamma_{y,0} + v_{y,\text{taker}[i]} \tag{4a}$$

$$\log(\sigma_{z,i}) = \gamma_{z,0} + v_{z,\text{taker}[i]} \tag{4b}$$

where ρ_{yz} is the correlation between horizontal and vertical errors, $\gamma_{y,0}$ and $\gamma_{z,0}$ are the population level precisions and

$$v_{y,\text{taker}} \sim N(0, \tau^2) \quad \text{and} \quad v_{z,\text{taker}} \sim N(0, \tau^2)$$

are the taker-specific precision deviations. This model hence estimates the location and scale of the takers, while also accounting for the correlation structure between the horizontal and vertical errors.

The model was fitted within brms with the same iteration parameters as the Naïve model. The taker-precisions were hence obtained as

$$\sigma_{y,j} = \exp(\gamma_{y,0} + v_{y,j}) \quad \text{and} \quad \sigma_{z,j} = \exp(\gamma_{z,0} + v_{z,j})$$

along with a population-level correlation ρ_{yz}. The dispersion of the players were then compared using a *consistency score* τ as the product of the two error terms:

$$\tau_j = \sigma_{y,j} \cdot \sigma_{z,j} \tag{5}$$

2.4 Risk-Adjusted Probability

The naïve binary model in Eq. (1) estimates the conditional probability of scoring $S(y, z) = P(\text{goal} \mid \text{location})$ at every coordinate over the goal frame, meaning that takers are completely accurate in shooting where they aim to shoot. This is a very strong assumption that does not capture the reality well. A more realistic set up is to calculate the probability of scoring a goal given a particular aim point $\boldsymbol{\mu} = (\mu_y, \mu_z)$, and the aiming error is characterized with the covariance matrix

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_y^2 & \rho_{yz}\sigma_y\sigma_z \\ \rho_{yz}\sigma_y\sigma_z & \sigma_z^2 \end{pmatrix}. \tag{6}$$

The probability of scoring given a particular aim point $\boldsymbol{\mu}^*$ can be obtained by convolving the shooting error distribution over the goal probability surface $S(y, z)$ as

$$P(\text{goal} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \iint_{\mathcal{A}} S(y, z) \cdot f(y, z \mid \boldsymbol{\mu}, \boldsymbol{\Sigma})\, dy\, dz \tag{7}$$

where $f(\cdot)$ is the probability density function of the shot locations, approximated as bivariate normal distribution, and $\mathcal{A}$ is the goal frame area.

However, the integral in Equation (7) over the goal frame does not effectively account for the ground. Vertical errors that result in $z < 0$ do not miss, since the ball can't go "under the goal". To account for this, we decompose the integral into air (above ground) and ground components as:

$$P(\text{goal} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \approx \underbrace{\iint_{\mathcal{A}} S(y, z) f(y, z | \boldsymbol{\mu}, \boldsymbol{\Sigma}) dA}_{\text{Air component}} + \underbrace{P(Z < 0 | \mu_z, \sigma_z) \cdot S(y, 0)}_{\text{Ground component}} \tag{8}$$

where the *Air component* integrates over shots that land within the goal frame and the *Ground component* is the probability mass of shot location $z < 0$, which we take as the naïve success probability $S(y, 0)$ at the corresponding y. This effectively rewards low aim locations, since downward errors do not imply a miss.

The integral in Equation (8) was discretized as a summation over the discretized goal frame as a grid of 150×50 ($N_y \times N_z$) rectangles:

$$P(\text{goal} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \approx \sum_{k=1}^{N_y} \sum_{l=1}^{N_z} S(y_k, z_l) \cdot f(y_k, z_l \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) \cdot \Delta A + \Phi\left(\frac{0 - \mu_z}{\sigma_z}\right) \cdot S(\mu_y, 0) \tag{9}$$

3 Results

Here we will summarize the results of the analysis described in the method section above, for the results of the two models. Then we will perform an *aiming error parameter study* using the convolutional framework outlined above.

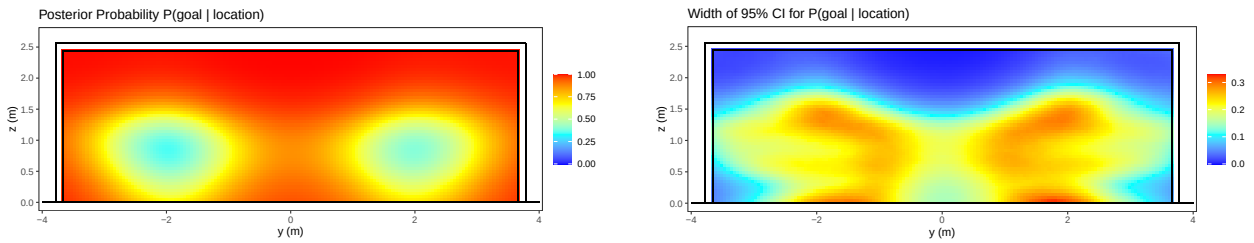
3.1 Naïve Probability Surface

The Bayesian GAMM binary model described in Section 2.2 was fitted to the data, which estimated the goal probability conditioned on the ball's location in the plane of the goal, i.e.

$$P(\text{goal} \mid \text{location}).$$

The model, see Equation (1), included random intercepts for the 709 takers and 373 goalkeepers in the dataset. The probability surface was marginalized over the random effects (with $\alpha = \beta = 0$), representing an average taker versus an average goalkeeper. The resulting probability surface $S(y, z)$ along with corresponding 95% credible intervals is shown in Figure 2.

The posterior distribution in Figure 2a is fairly symmetric about the horizontal center $y = 0$. The average probability of scoring on the left side is 0.813, compared to 0.825 on the right side. The 95% credible interval for this difference includes zero, indicating that there is no credible evidence that either of the sides have a higher probability of scoring.



(a) Posterior distribution

(b) 95% credible intervals

Figure 2: Marginalized posterior distribution $P(\text{goal} \mid \text{location})$ and 95% credible intervals of the Naïve binary model

3.2 Consistency of Takers

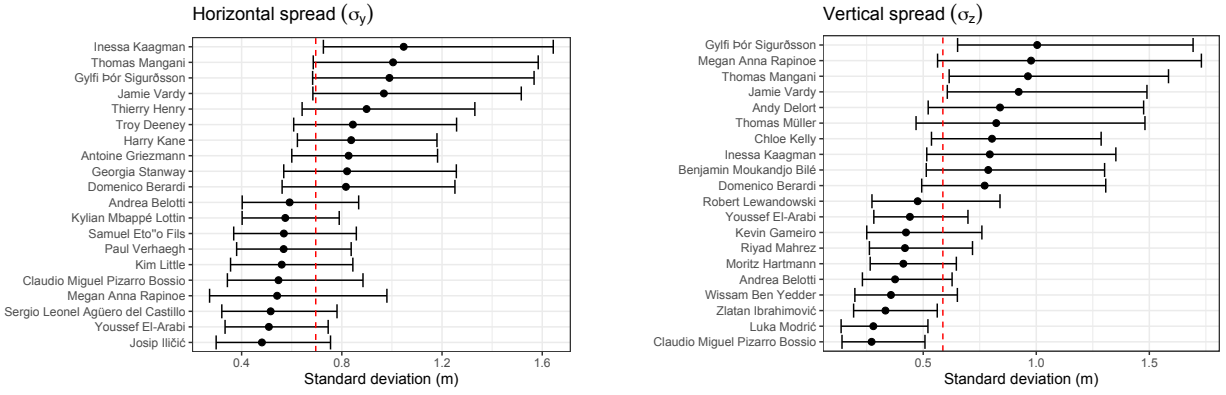
As described in Section 2.3, the data was filtered to contain only penalties from takers who had $n \geq 5$ penalties in the dataset. The resulting subset D_s contained 407 penalties from 47 unique takers.

The Location-Scale model in Equation (2) used a Student-t distribution to estimate the mean aim point $\boldsymbol{\mu}$ and the shooting precision $\boldsymbol{\Sigma}$. The model estimated the degrees of freedom as $\nu = 6.86$, with 95% credible intervals [4.07; 12.67]. This relatively low estimate clearly justifies the choice of the Student's-t distribution over normal distribution (which would imply $\nu \to \infty$), which is likely attributed to outliers in the dispersion distribution.

The standard deviation for each player on the meter scale can be obtained from Equation (4) as

$$\sigma_y = \exp(\gamma_{y,0} + v_{y,\text{taker}})$$

using the estimated population-level intercept $\gamma_{y,0}$ and the random effects for takers $v_{y,\text{taker}}$, and similarly for the vertical spread σ_z. Figure 3 shows the estimated spreads in the horizontal and vertical directions, showing the ten highest and lowest estimates.



(a) Horizontal spread σ_y

(b) Vertical spread σ_z

Figure 3: Estimated dispersion spread for takers: 10 highest and 10 lowest estimates

Figure 3 reveals that the credible intervals do not include the population average for several takers. For the horizontal spread σ_y, *Inessa Kaagman* has the highest estimate with intervals that do not include the average (the red line). More takers appear with distinguishable spreads from

the average for the vertical errors, both high and low ones. It is noticeable that some players (e.g., Gylfi Þór Sigurðsson, Thomas Mangani and Jamie Vardy) have high spreads in both horizontal and vertical directions.

However, this does not directly mean that takers with high spread are more inaccurate, as discussed before. This simply means that their shot locations vary more over the goal compared to other takers. However this gives a sense on how much individual player's shot locations vary.

Figure 4 shows the horizontal and vertical spread for the 47 takers in the subset D_s used for this analysis. The spreads do follow a linear trend, with increased horizontal spread resulting in greater vertical spread. However, Megan Rapinoe (labeled on the plot) deviates from this trend and has a high vertical spread σ_z but a low horizontal spread σ_y, suggesting that she tended to be consistent with horizontal location, but varied considerably in the vertical dimension.

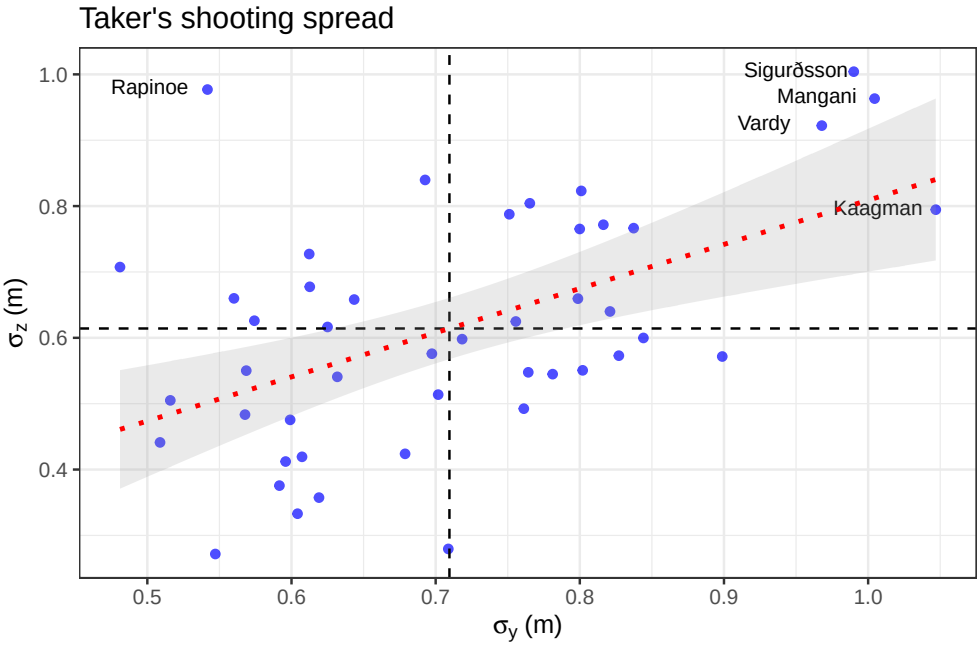


Figure 4: Taker's shooting spread in horizontal and vertical direction (dashed black lines indicate average standard deviations)

The average of taker's spreads in y and z directions, i.e. $\bar{\sigma}_y$ and $\bar{\sigma}_z$ (derived from Eq. (4)), are shown in Table 2 along with the population-level ρ_{yz} (which was estimated globally without any random effects). These parameters suggest typical values which can define covariance matrix Σ to determine the aim shooting error, as defined in the introduction.

Table 2: Population averages for covariance matrix parameters

Parameter	**Value**
$\overline{\sigma}_y$	0.7096
$\overline{\sigma}_z$	0.6142
ρ_{yz}	-0.2004

3.3 Aiming Error Parameter Study

With the marginalized probability surface $S(y, z)$ from the Naïve binary model and the information about the spreads σ_y and σ_z we can estimate the probability of scoring given an aim location

$$P(\text{goal} \mid \text{aim}).$$

By selecting parameters σ_y^*, σ_z^* and ρ_{yz}^* we can construct a covariance matrix $\boldsymbol{\Sigma}^*$ to represent various precision profiles. With a given aim point $\boldsymbol{\mu} = (\mu_y, \mu_z)$ we can use the convolution formula in Equation (8) to quantify the risk-adjusted probability surface $S^R(y, z)$. This will reveal how how shooting accuracy influences the optimal strategy to maximize the probability of scoring.

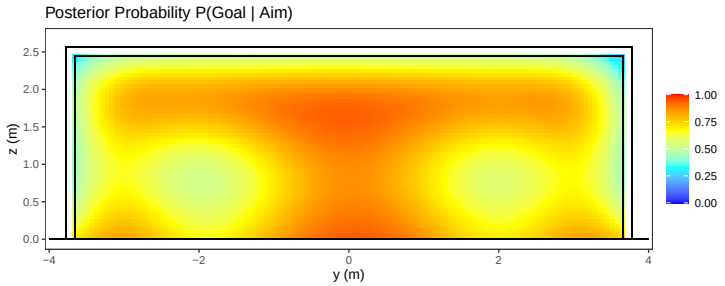
We explore six cases based on the shooting spread observed in Figure 4, which are summarized in Table 3. For the first three cases, we keep the correlation fixed at $\rho_{yz} = -0.20$ across all three cases to focus on the effect of the shooting error σ on the optimal strategy. The negative correlation implies a slight dependence between horizontal and vertical errors, where shots that have a high horizontal deviation tend to be lower in elevation, and vice versa for shots that have low horizontal deviance (the goal center) tend to be higher vertically. This is very intuitive for penalty kicks.

For the latter three cases we explore the effect of the correlation ρ_{yz} on the results. We fix the standard deviations at their means from Table 2 with $\overline{\sigma}_y$ and $\overline{\sigma}_z$, while we vary the correlation ρ_{yz} for positive and negative values as well as $\rho_{yz} = 0$ which implies independence of the two errors (since the joint distribution is a bivariate normal).

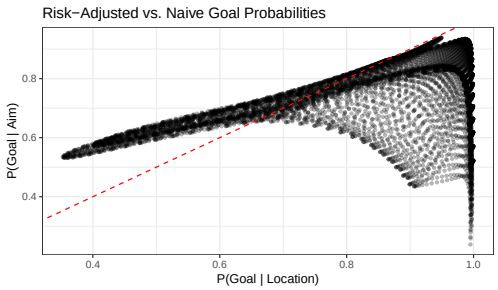
Table 3: Parametric study of aiming error

ID	**Case**	σ_y (m)	σ_z (m)	ρ_{yz}
Spread Cases				
S_{low}	Low Spread	**0.50**	**0.40**	-0.20
S_{avg}	Average Spread	**0.71**	**0.61**	-0.20
S_{high}	High Spread	**1.00**	**1.00**	-0.20
Correlation Cases				
C_{neg}	Negative Corr	0.71	0.61	**-0.9**
$C_{\text{indep.}}$	Independent	0.71	0.61	**0**
C_{pos}	Positive Corr	0.71	0.61	**+0.9**

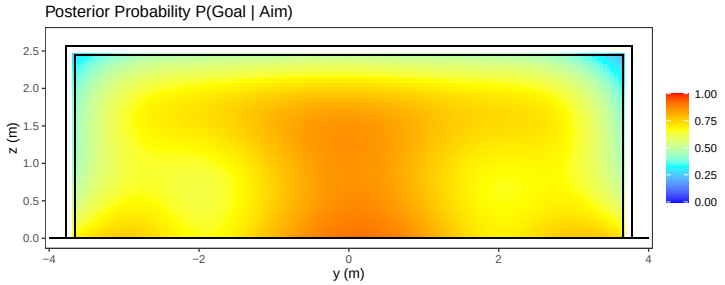
Figure 5 shows the probability surface for the risk-adjusted probability $P(\text{goal} \mid \text{aim})$ along with a scatterplot comparing the risk-adjusted probabilities versus the Naïve probabilities (from $S(y, z)$ in Figure 2a). Figure 6 shows a comparable visualization for the correlation cases in Table 3.



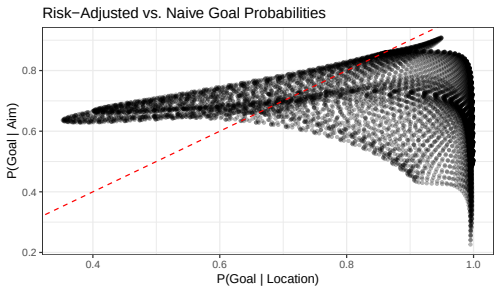
(a) S_{low}: Probability surface



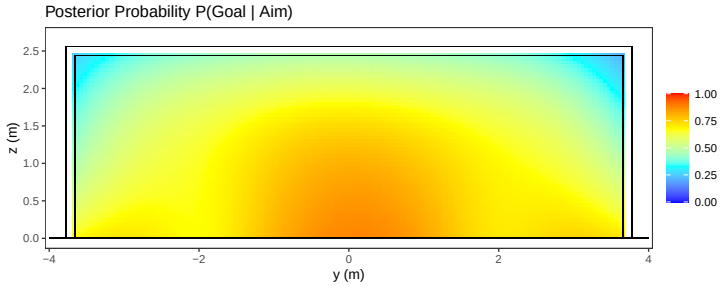
(b) S_{low}: Risk-adjusted vs. Naïve



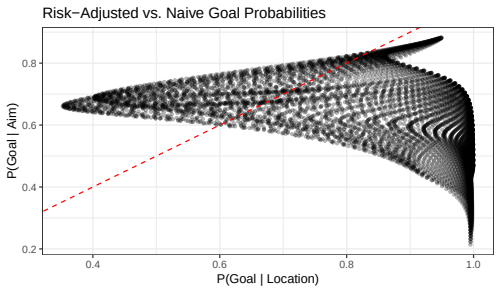
(c) S_{avg}: Probability surface



(d) S_{avg}: Risk-adjusted vs. Naïve



(e) S_{high}: Probability surface



(f) S_{high}: Risk-adjusted vs. Naïve

Figure 5: Parametric study of aiming error for spread cases: Low Spread S_{low} (a, b), Average Spread S_{avg} (c, d), High Spread S_{high} (e, f)

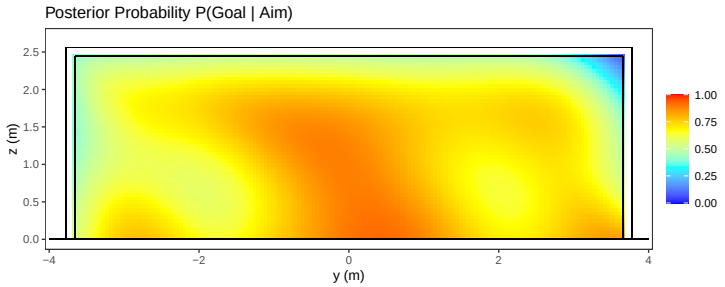
Figure 5 reveals a crucial difference between the three cases with varying σ_y and σ_z. The low-spread case (S_{low}) with $\sigma_y = 0.50m$ and $\sigma_y = 0.40m$ shows a heatmap showing the success probability of scoring at each *aim location*. The zones near the goal's boundary have lower probability, which is intuitive given the risk of shooting off target. Additionally, the ground constraint from the convolution in Equation (8) arises, favoring zones near the ground.

The comparison of the risk-adjusted probabilities to the naïve probabilities shows that some probabilities actually increase, which are the two zones around $(y, z) \approx (\pm 2, 0.5)$. This is due to the aiming error, smoothing out the probability in those regions. However, most of the risk-adjusted probabilities are lower compared to the naïve model.

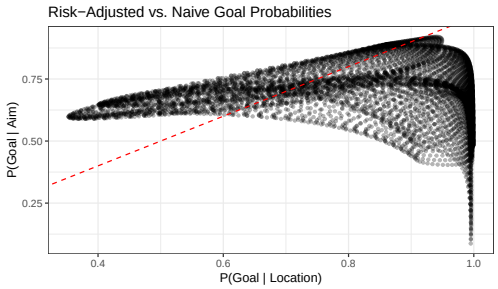
The average spread case (S_{avg}) tells a similar story, but with diminishing probabilities near the edges of the goal. The center of the goal results in the highest probability. In the high spread case (S_{high}), the aiming error is very high, resulting in a sharp probability decrease from the center of

the goal. This aiming error seems to be artificially high, as it suggests that if a taker wants to maximize their probability of scoring, they should essentially just shoot directly at the middle of the goal. This would not really be meaningful in professional soccer, given their high-level skills.

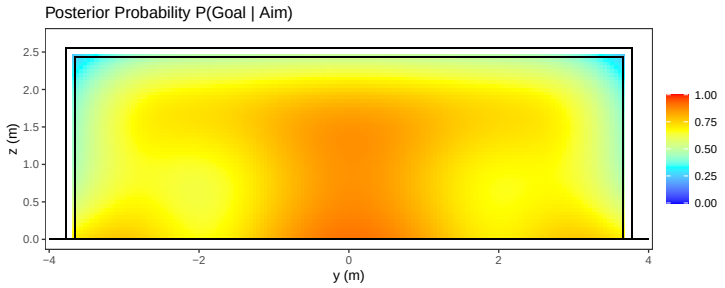
It is also interesting to see how the risk-adjusted probabilities vary with the naïve probabilities (subfigures b, d and f). Increasing aiming error tends to deviate the naïve probabilities more, which is expected.



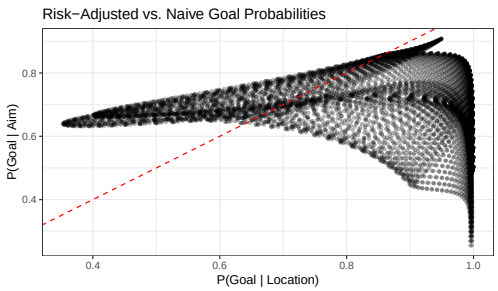
(a) C_{neg}: Probability surface



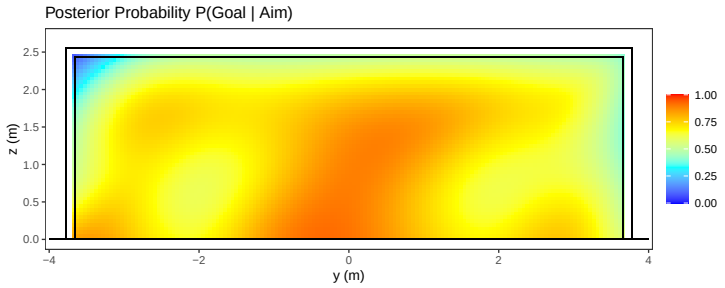
(b) C_{neg}: Risk-adjusted vs. Naïve



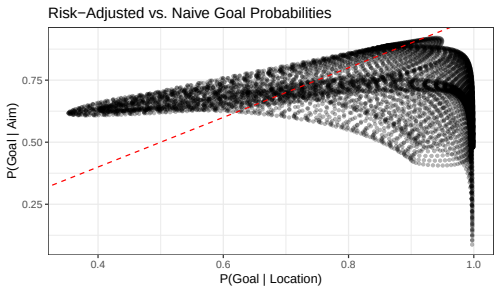
(c) $C_{\text{indep.}}$: Probability surface



(d) $C_{\text{indep.}}$: Risk-adjusted vs. Naïve



(e) C_{pos}: Probability surface



(f) C_{pos}: Risk-adjusted vs. Naïve

Figure 6: Parametric study of aiming error for correlation cases: Negative Corr C_{neg} (a, b), Independent $C_{\text{indep.}}$ (c, d), Positive Corr C_{pos} (e, f)

Figure 6 compares the effects of the correlation coefficient ρ_{yz}. The strong negative correlation in the C_{neg} case of $\rho_{yz} = -0.9$ means that a horizontal error to the right is associated with a downward vertical error, while a leftward error is associated with an upward vertical error. This correlation structure introduces a distinct asymmetry to the probability surface, as we observe from the probability surface. The top-left corner results in a moderate probability of scoring whereas

the top-right corner results in a sharp probability decrease.

In the independent case $C_{\text{indep.}}$ this asymmetry is no longer present and the probability is symmetric around $y = 0$. The probability surface is comparable to the S_{avg} case in Figure 5, suggesting that the correlation $\rho_{yz} = -0.2$ does not have a large influence on the risk-adjusted probability.

The third case C_{pos} mimics a reflection of the negative correlation case. The heatmap now exhibits a up-rightward shape, which is caused by the elliptical shape of the aiming error probability mass function. Positive horizontal errors are coupled with positive vertical errors and conversely, errors to the left exhibit lower trajectories.

The comparison plots of the risk-adjusted probabilities (in subplots b, d and f) indicate a very comparable transition of probabilities compared to the Naïve model. The negative and positive correlated cases in sub-figures b and f (in Fig. 6) appear to have a very a comparable behavior to the naïve model, which is also comparable to the independent case in sub-figure d.

4 Conclusions

Here we have effectively developed a Bayesian framework to evaluate the probability of scoring from a penalty kick. The binary naïve model estimates the probability surface over the goal, with random effects for goalkeeper and taker. In this study we marginalized over those random effects to consider the "average" matchup between goalkeeper and taker as $S(y, z)$. This does not include the aiming error for each taker, where the final location of the shot (y^{obs}, z^{obs}) deviates from the intended aim point $\boldsymbol{\mu} = (\mu_y, \mu_z)$.

To account for this aiming error, we derived a formula to convolve the aiming error $f(\cdot)$ over the posterior surface $S(y, z)$ from the naïve model, in Eq. (8). We represented the aiming error as a bivariate normal distribution

$$f(y, z \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathcal{N}\left(\begin{pmatrix} y \\ z \end{pmatrix} \middle| \, \boldsymbol{\mu}, \boldsymbol{\Sigma}\right) \tag{10}$$

where the $\boldsymbol{\mu}$ is the aim point and the covariance matrix is as in Eq. (6), characterized by the standard deviations σ_y and σ_z along with the correlation ρ_{yz}. The biggest challenge is to estimate $\boldsymbol{\Sigma}$, since the data with observed locations does not provide any information about the intended aim point behind each data point.

We fitted a Location-Scale model to estimate the spread from a subset of the data, with penalties from players which had at least 5 observations in the original dataset. However, this is a very strong assumption to use this observed spread approximate the aiming error for each player. Some players might aim to various locations of the goal, which results in a wide standard deviations. But this does not necessarily mean that this player is inaccurate, but rather that they exhibit high *strategic variance* by targeting different areas of the goal. However, we used this observed spread as an approximation here, which might not be representative of the aiming error.

In our parametric study, we were able to estimate the risk-adjusted probability surface

$$S^R(y, z|\sigma_y, \sigma_z, \rho_{yz}),$$

effectively accounting for the aiming error. As the spread increased, the high-probability regions centered at the goal. The correlation coefficient ρ_{yz} also affected the probability surface, where large-magnitude correlations introduced diagonal patterns in the surface S^R while the independent case were fairly symmetric.

For a given covariance matrix $\mathbf{\Sigma}$, the optimized strategy for a player can be computed as the location which maximizes the risk-adjusted probability over the goal, i.e. a player should aim at $\boldsymbol{\mu}^* = (\mu_y, \mu_z)$ where

$$\boldsymbol{\mu}^* = \underset{\mu_y, \mu_z}{\arg\max}\ S^R(\mu_y, \mu_z \mid \sigma_y, \sigma_z, \rho_{yz}) \tag{11}$$

to optimize their strategy of scoring a goal.

5 Future Work

While we have developed a framework to compute the risk-adjusted probability surface S^R, the challenge still remains to determine the covariance matrix $\mathbf{\Sigma}$. The methodology of using the observed spread as an approximate is a strong assumption, which may not accurately reflect the aiming error. We list some ideas for future work:

1. To accurately estimate the covariance matrix $\mathbf{\Sigma}$, the most direct way is to perform **controlled experiments**. Soccer players would take penalties, with some given aim point, and the ball's location would be tracked with computer vision (see [2]). Collecting such data could be used to estimate the covariance matrix, which is likely a function of the plane i.e. $\mathbf{\Sigma}(y, z)$.

2. Including other features in the covariance matrix estimation, such as psychological pressure or game-score, can make the covariance more accurate for in-game aiming errors. Some players might have better control in stressful moments in a game.

3. While the dataset has an adequate quality, another dataset of higher quality would be good. The current dataset spans over 25 years of data, which does not have fully accurate ball's locations in the plane of the goal. A high-quality dataset from professional soccer leagues in recent past is desirable which would have currently playing takers and goalkeepers, which could be used for the random effects of the naïve model. This would make the framework more practical for current use in professional soccer.

With incorporating these future directions, especially controlled experiments, this work can be expanded and used to serve as a practical applicable tool for optimizing taker's strategies to score from a penalty kick.

References

[1] Paul-Christian Bürkner. brms: An r package for bayesian multilevel models using stan. *Journal of statistical software*, 80:1–28, 2017.

[2] PR Kamble, AG Keskar, and KM Bhurchandi. A deep learning ball tracking system in soccer videos. *Opto-Electronics Review*, 27(1):58–69, 2019.

[3] StatsBomb. Open data, 2025. Last accessed October 23rd 2025.