Bayesian semi-parametric soccer analysis

Michael Wamberg[1] & Bjarke Arnskjær Hastrup[1], {s151602, s146791}@student.dtu.dk

Abstract
Time-to-event data analysis is a set of methods used to analyze the expected duration of time until an event of interest occurs. In this project we apply the time-to-event analysis methods to analyze the intensity with which goal scoring occurs in soccer matches. We present different semi-parametric time-to-event models from the class of Cox proportional hazards models and extend these to incorporate non-linear effects by use of the highly flexible Gaussian processes from the Bayesian non-parametrics toolbox. All models are estimated in Stan, which provides an efficient Hamiltonian Monte Carlo sampler and further supports fast and scalable variational inference.

Keywords
Soccer — Time-to-event analysis — Gaussian Processes — MCMC — Variational inference — Stan

[1] *Technical University of Denmark - DTU*

Introduction

In this project we apply time-to-event analysis to soccer data, in order to examine the dynamics of goal scoring in soccer matches. Time-to-event data analysis has mostly been applied to medical research where the set of methods for analyzing time-to-event data is known as survival analysis. The time-to-event framework can however be applied to many other probabilistic settings where data follows some sort of arrival process, but in this project we treat goal scoring in soccer matches as the main event of interest. It is appealing to analyze the goal scoring dynamics using time-to-event analysis, since these methods considers the temporal dimensions of the data.

In order to evaluate the goal scoring intensities (or *hazard rates* which we rather unintentionally and unwillingly adopt from the survival analysis vocabulary and quite possibly have used throughout this report and notebook) we will be building on the Cox proportional hazards model [1] with both fixed and time dependent covariates and expand this model by incorporating our time-to-event analysis into the Bayesian framework. After an introduction to the chosen dataset and we provide a quick recap survival analysis essentials, and show how to arrive at an analytical expression for the loglikelihood with which these semi-parametric models are easily implemented in Stan. Not all variations of the models were easily identifiable. Only a subset of all tested models are therefore presented and evaluated quantitatively in this report.

1. Methods

Our dataset consists of 2132 matches from the Portuguese Primeira Liga (612 matches), the English Premier League (760 matches) and the Italian Serie A (760 matches). That amounts to 2 full seasons in each league, the 2017/2018 and 2018/2019 seasons. The dataset contains information about the strength of the teams as well as the times when the teams scored. The goal times are scraped from the website *http://www.whoscored.com*. While in order to assess the strength of the teams, we use betting odds obtained from *http://www.football-data.co.uk* (namely closing odds from the bookmaker Pinnacle[1]). To get the most accurate strength measurement of the teams, we normalize the inverse odds, i.e. remove Pinnacle's over-round and define a **skill_gap** covariate based on these recovered probabilities. Initially we will focus on this covariate and in the following we determine its influence on the scoring dynamics. To do so, we first outline the main idea of what is called the equivalent Poisson model.

1.1 Quick time-to-event analysis recap

Let the goal time be represented as a continuous non-negative variable, T, with density $f(t)$. Then we define a no-goal probability function

$$P(T > t) = S(t) = \exp\left(-\int_0^t \lambda(u)du\right), \tag{1}$$

that is a function of the cumulative hazard, $\int_0^t \lambda(u)du$, experienced up until time t. In the proportional hazard model the hazard function is given by

$$\lambda(t) = \lambda_0(t)\exp(G(\mathbf{x},\beta)), \tag{2}$$

where $\lambda_0(t)$ is a so-called baseline hazard function identical for all matches, β is a vector of regression coefficients and $G(\mathbf{x},\beta)$ is a function. The second term is expressed using the exponential function because the hazard must be positive.

[1] The closing odds provides the most efficient representation of the match outcome probabilities [2].

With constant covariates this model implies that the ratio between the hazards of two matches remains constant over time. It is most common to have $G(\mathbf{x}^T\beta) = \mathbf{x}^T\beta$.

1.2 The equivalent Poisson model

In the semi-parametric Cox proportional hazards model which will be implemented in the next section, we divide the time interval into T intervals, $(0,s_1],(s_1,s_2],\ldots,(s_{T_1},s_T]$, and assume the baseline hazard to be constant throughout each interval, i.e. $\lambda_0(t) = \lambda_j$ if $t \in (s_{j-1},s_j]$. The hazard in the $i'th$ match at time interval j is then

$$\lambda_{ij} = \lambda_j \exp(\mathbf{x_i^T}\beta). \tag{3}$$

The likelihood contribution from $i'th$ match with home goal time, y_i, and binary censoring variable

$$v_i = \begin{cases} 1 & \text{for goal,} \\ 0 & \text{for censoring (no goals),} \end{cases} \tag{4}$$

is given by

$$L_i(\beta,\lambda) = f_i(y_i)^{v_i} S_i(y_i)^{(1-v_i)}. \tag{5}$$

Now let

$$\delta_{ij} = \begin{cases} 1 & \text{if } y_i \in I_j = [s_{j-1},s_j), \\ 0 & \text{otherwise,} \end{cases} \tag{6}$$

meaning that $\delta_{ij} = 1$ for the time interval, j, when either a home goal is scored or the match has ended. Then, by using the piecewise constant hazard function of equation (3) in equation (5),

$$\begin{aligned} L_i(\beta,\lambda) =& (\lambda_j \exp(\mathbf{x}_i^T,\beta))^{\delta_{ij}v_i} \exp\left\{-\delta_{ij}\left[\lambda_j(y_i - s_{j-1})\right.\right. \\ &\left.\left. + \sum_{g=1}^{j-1} \lambda_g (s_g - s_{g-1})\right] \exp(\mathbf{x}_i^T\beta)\right\}. \end{aligned} \tag{7}$$

This likelihood can be rewritten [3]

$$\begin{aligned} L_i(\beta,\lambda) =& \Pi_{g=1}^{j} \exp\left\{-\lambda_g t_{i,g} \exp(\mathbf{x}_i^T\beta)\right\} \\ &\times (\lambda_g t_{i,g} \exp(\mathbf{x}_i^T\beta))^{\delta_{i,g}v_i} / (\delta_{i,g}v_i)!, \end{aligned} \tag{8}$$

where $t_{i,g}$ is the amount of time match i is under exposure of its hazard, $\lambda_j \exp(\mathbf{x}_i^T\beta)$. We can recognize the likelihood in equation (8) as being equivalent to the likelihood of j independent Poisson random variables (recall the parametric form of the probability mass function of a Poisson random variable, $\frac{\mu e^{-\mu}}{k!}$). Thus, for each match, i, we consider for each interval from $g_i \in \{1,\ldots,j(i)\}$ (here $j(i)$ denotes the index of the interval in which there was scored in the match) the observation, $\delta_{i,g}v_i$, as drawn from a Poisson random variable with mean (and variance) given by $\mu_{i,g} = \lambda_g t_{i,g} \exp(\mathbf{x}_i^T\beta)$. This is obviously a rather strange model, since a Poisson variable can take any non-negative integer value and not just one and zero, as is the case for $\delta_{i,g}v_i$. The smart trick here is that if we do inference on this pseudo model, we are simultaneously doing inference on the model in equation (7). To summarize, we have $\sum_i^N j(i)$ Poisson observations with

$$\log(\mu_{i,g}) = \log(t_{i,g}) + \log(\lambda_g) + \mathbf{x}_i^T\beta, \tag{9}$$

with $g_i > j(i)$ left out (we don't attempt to estimate hazards when no exposure is present!) of the data set since these can cause numerical issues, although these observations shouldn't theoretically interfere with the parameter estimation.

2. Results

2.1 Models

Next we present our implemented models. For model selection we use the PSIS-LOO method[2].

Model 0: The first model we investigate is the Cox proportional hazard model presented in Section 1.2, here with baseline hazards, $\lambda_j, j = \{1,\ldots,T\}$, all drawn independently as $\lambda_j \sim \Gamma(0.1,0.1)$. For the regression coefficient $\beta \sim N(0,3)$ which is sufficiently vague considering the restricted scale of the potential increase in goal intensity observed between a weak and a poor team relative to two equal opponents. To get a quick feel for the magnitude of the hazards Figure 1 shows the baseline hazards estimated for each time interval along with the corresponding MCMC samples. For the **skill_gap** covariate, we obtain the estimate $\beta = 1.00$, corresponding to a 65% increase in home goal intensity for a skill gap of 50% (i.e. when the inverse odds are respectively 65 % and 15 % on the home and away win). See Figure 2 for skill gap visualization. As reported in the notebook, this model achieves a PSIS-LOO log predictive density of -1606 and 24.7 effective parameters.

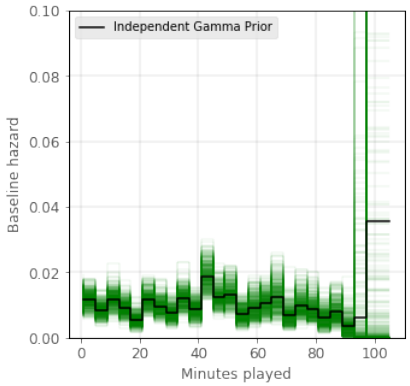

Figure 1. Estimated baseline hazard for Model 0.

Model 1: Here we attempt to improve Model 1 by using an autoregressive baseline prior λ_0, and keeping the **skill_gap** covariate. As can be seen in the notebook, this correlation can help to regularize the inference problem and hence reduce the

[2] see [4] and the Jupyter notebook for a detailed description

total number of effective parameters in the model. We observe a reduction from 24.7 to 22.0, although the log predictive density remains at the same level.

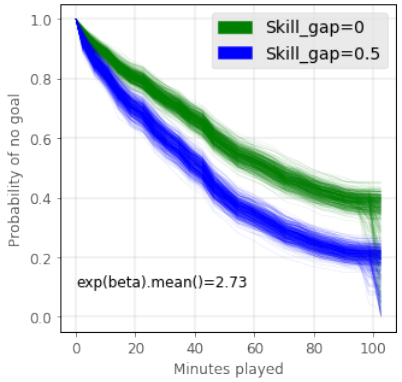

Figure 2. Estimated probability of 'no first goal' for Model 0.

Model 2: Now we add a dummy variable that indicates whether a goal has been scored or not, simply to measure if a goal accelerates the intensity of another goal. This variable will be denoted *x_goal*. The inclusion of this variable doesn't seem to improve the fit, as can be seen on the log predictive density barplot in Figure 3. The coefficient is approximately -0.4 corresponding to 33% reduction in goal intensity (due to taking the exponential).

Model 3: In this model we introduce a flexible time dependency of the regression coefficient for the covariate presented in Model 2, such that

$$\lambda_{ij} = \lambda_j \exp(\beta_1 \cdot skill\_gap + \beta_2(t) \cdot x\_goal). \tag{10}$$

The additional flexibility improves the fit but at the cost of a doubling in the effective number of parameters. (Note: By mistake the model was run without specifying a prior distribution over $\beta_2(t)$, and was therefore allowed to fluctuate wildly.)

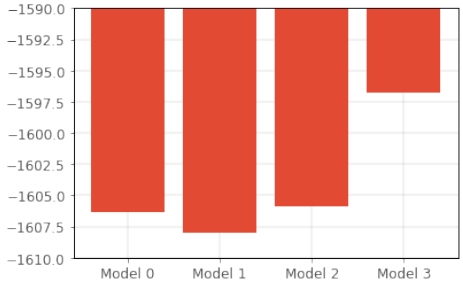

Figure 3. Log predictive densities for Models 0,1,2,3.

Model 4: In this model we perform Cox proportional hazards regression with time dependent features. The model builds on the previous models, to which we add more covariates, consisting of two new goal features: time of goal and time since goal. These are obviously only meant to model the temporal variations in the intensity of the second goal. The model was harder to fit, and was only estimated with Automatic Differentiation Variational Inference (ADVI) [5]. From the estimated coefficients in Figure 4 we obtain the result that (second) goal intensity is highest right after a goal and lower further away (see x_time_since_goal).

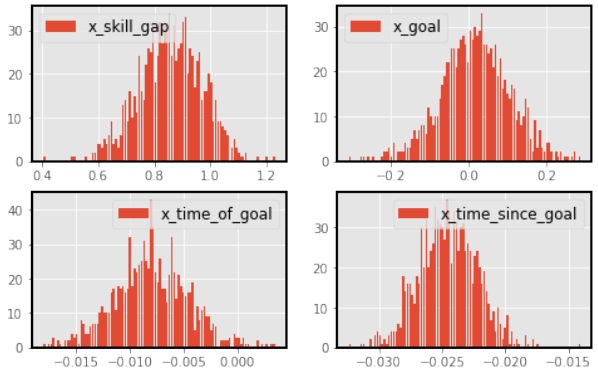

Figure 4. Estimated regression coefficients for the covariates in Model 4.

Model 5: Now the Cox model is being extended to a more general form with the (log) baseline now modelled as a continuous latent process. This is done by introducing a Gaussian process prior over the baseline such that $\log \lambda_0 \sim GP(0, K(t,t'))$. We use the squared exponential as covariance function

$$K(t,t') = \alpha^2 \exp\left(-\frac{(t-t')^2}{2\rho^2}\right) + \delta_{i,j}\sigma^2, \tag{11}$$

with α being the scale of the output values, ρ the characteristic length scale of the input (duration of game in minutes) and σ the scale of the output noise. As was also the case with Model 4, we had a hard time getting Stan to estimate the parameters in our model using MCMC. Thus, we also used ADVI in order to fit this model, therefore, the model was not evaluated with PSIS-LOO. In Figure 5 is shown the baseline generated using the Gaussian process predictive inference, from which it can be seen that it is much more smooth and continuous compared to the baseline estimates from the other models.

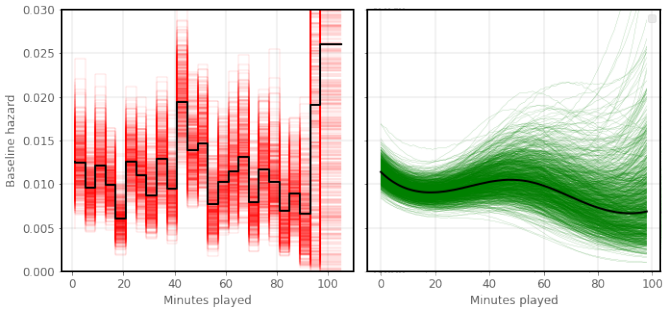

Figure 5. Right: Baseline samples generated using Gaussian process predictive inference (Model 5). For comparison we show the piecewise constant baseline with autoregressive prior (Model 2).

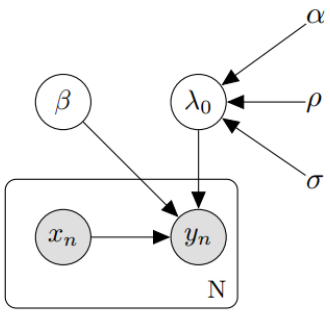

Figure 6. The probabilistic graphical model for y_n. Notice that this is merely a convenient abstraction, as explained in the text.

In order to obtain an understanding of the observed phenomena, we look into the data generating process by writing the full generative model for our final model.

Generative model

draw $\alpha \sim \text{half-normal}(0,5)$
draw $\rho \sim \text{inv-gamma}(10,1)$
draw $\sigma \sim N(0,1)$
draw $\beta \sim N(0,3)$
draw $\log \lambda_0 \sim GP(0, K(t,t'))$

for each $n \in \{1,\ldots,N\}$:
    draw $y_n \sim \tilde{S}(t)$

$y_n \sim \tilde{S}(t)$ indicates that each y is drawn from the numerically estimated no-goal probability function. The data generating process is also visualized in the probabilistic graphical model given in Figure 6. Notice, that the PGM offers a nice intuitive representation of the generative structue of the model, but isn't entirely correct, since there is no direct parametric mapping from x, β and λ_0 onto y, but more importantly, it doesn't convey how the Stan model is actually estimated. To do so we would have to write up a PGM/generative model which explicitly draws realizations for each time interval for every match and such a generative model would have to be built in a blocking mechanism preventing the first goal from being scored twice.

3. Conclusion and discussion

In this project, we introduce 5 different time-event models with increasing complexity. The models are all able to produce reliable results, with the Cox proportional hazards model containing a time dependent regression coefficient for the *x_goal* variable being the best performing model according to the PSIS-LOO evaluation. The introduction of the Gaussian process prior over the baseline also showed promising results, but its performance was not evaluated with PSIS-LOO in this project. The obvious model extension for future work is to introduce the latent predictor η_i depending on all or many of the covariates x_i, which logically could be represented with a GP-prior, thus enabling non-linear covariate effects. Having performed predictive inference with Gaussian processes using automatic relevance determination in the Stan code, this extension would be low hanging fruit and very exciting to pursue.

Future work will consist of implementing new covariates into the models, especially the time-dependent ones such as shots on goal during the match, substitutions, ball possession, etc. With the increasingly comprehensive live data collection, the possibilities for further extensions are great.

Finally, it could be interesting to measure our models against the bookmakers, since they have to be considered as the most efficient to assess the evolution of soccer matches. However, with the models presented in this project, we dare not take up the challenge yet and we also will not encourage our readers to do so, but we believe that with this project we have presented a good foundation for the further development of some strong analytical tools for assessing the course of action in soccer matches.

References

[1] Cox D R. Regression models and life-tables. Journal of the Royal Statistical Society. Series B (Methodological) 1972; 34: 187-220

[2] Using Pinnacle.com's Closing Line to Predict Profits https://football-data.co.uk/blog/pinnacle_efficiency.php

[3] Adam Branscum Timothy E. Hanson Ronald Christensen, Wesley Johnson, "Bayesian ideas and data analysis: An introduction for scientists and statisticians," International Statistical Review, vol. 79, no. 2, pp. 285–286, 2011.

[4] Vehtari, A., Gelman, A., Gabry, J. (2017). "Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC". Statistics and Computing. 27(5):1413–1432.

[5] Kucukelbir, Alp, Rajesh Ranganath, Andrew Gelman, and David M. Blei. 2015. "Automatic Variational Inference in Stan." arXiv 1506.03431