Risk-neutral Derivation of In-play Bet Prices by Modeling Football Matches

A Weibull Point Process Approach

Morten Andersen

Master's Thesis, Mathematics-Economics

Dept. of Mathematical Sciences

Aalborg University

Dept. of Mathematical Sciences
Skjernvej 4A
DK-9220 Aalborg Ø
https://www.math.aau.dk

AALBORG UNIVERSITY
STUDENT REPORT

Title:
Risk-neutral Derivation of In-play Bet Prices by Modeling Football Matches: A Weibull Point Process Approach

Theme:
Mathematical Finance in Sports Betting

Project Period:
Spring Semester 2020

Project Group:
Group 1.211c

Participant(s):
Morten Andersen

Supervisor(s):
Esben Høg

Page Numbers: 99

Date of Completion:
2020-06-02

Abstract:

In this thesis, we consider a risk-neutral approach to in-play betting using Weibull-based point processes to model the underlying football match which drives the bet prices. We investigate if the Fundamental Theorems of Asset Pricing are applicable in football betting markets by calibrating model prices to actual bet prices observed on a betting exchange. We do this by first analyzing the suitability of the Weibull process and the Weibull renewal process for modeling football goals, and then by formulating a risk-neutral valuation framework based on these models. We obtain promising results in both models when considering their limitations, however, we do not obtain conclusive evidence for the existence of a risk-neutral measure in these market models. However, by studying our results, we do find promising directions for further advancements in the application of the Fundamental Theorems of Assets Pricing in in-play football betting.

Preface

This master's thesis is compiled in the spring of 2020 by Morten Andersen, a student in the master's program in Mathematics-Economics at the Department of Mathematical Sciences, Aalborg University, Aalborg, Denmark. This document is typeset with LaTeX. Data used in this thesis is obtained from Premier League's season archives, Premier League (2020), and from Betfair's historical exchange data, Betfair (2020). Computations, modeling, and illustrations are carried out using the R programming language, R Core Team (2020), along with the packages:

- `dplyr` Wickham et al. (2020) and `tibble` Müller and Wickham (2019) for data processing
- `ggplot2` Wickham (2016), `ggformula` Kaplan and Pruim (2020), and `kableExtra` Zhu (2019) for visualizations and tables
- `Countr` Kharrat et al. (2019), `rpgm` Baradel (2018), and `MASS` Venables and Ripley (2002) for statistical modeling
- `foreach` Microsoft and Weston (2020) and `doParallel` Microsoft and Weston (2019) for parallel computing

Some R scripts will be available at `https://github.com/MoAnd/MastersThesis_Public`, additional scripts can be shared upon request.

The author would like to extend his greatest appreciation to his thesis supervisor, Esben Høg, and thank him for his inputs throughout the period.

Aalborg University, 2020-06-02

Morten Andersen

<moande15@student.aau.dk>

To Mathias

Nothing worth celebrating comes easy.
- Unknown.

Contents

Introduction 1

Live betting, or in-play betting, has seen massive growth in popularity over the last decade or so. Such bets are traded in real-time prior to and during an association football match (henceforth, simply referred to as football). The bet prices are mostly driven by the goals scored in the underlying game. This is comparable to financial markets where the price of an option varies in accordance with the price changes of the underlying instrument. Bet prices generally move in a way such that the prices move evenly between goals and then jump to a new level at a goal time. Since bet prices behave much like options in other financial markets, we seek to apply some general asset pricing theory to the betting markets. In particular, we wish to investigate and demonstrate the application of the Fundamental Theorems of Asset Pricing to the in-play football betting market. In doing so, we will also need to investigate some statistical patterns of football goals in order to model the underlying football match.

Betting Markets

Betting markets have historically only consisted of bookmakers setting the odds for people to bet on. Traditionally, the bookmakers only supplied fixed-odds bets, also known as pre-game bets, in which the bettor is not allowed to place bets during a football match. In-play betting changed that and now allows bettors to place bets after the game has started. Furthermore, the invention and later growth of the betting exchanges have rapidly expanded the in-play betting markets, such that is has overtaken the classic pre-game market in popularity and revenues. Betting exchanges are much like regular financial exchanges, where you can buy and sell bets. The Betfair exchange was launched back in 2000 and started seeing a huge growth in popularity towards the end of the decade. As of the time of writing, the Betfair betting exchange is the world's largest betting exchange, however, several prominent competitors have appeared in recent time.[1]

[1] Nordsted (2009), Brown and Yang (2017), and Divos et al. (2018)

Risk-neutral, Risk-management, & Hedging

The Fundamental Theorems of Asset Pricing constitute the foundation of the risk-neutral framework of mathematical finance and derivative pricing. The first fundamental theorem specifies that a market is arbitrage-free if and only if there exists a probability measure under which the underlying asset prices are martingales. The second fundamental theorem states that the market is complete if and only if the martingale measure is unique. A risk-neutral framework is beneficial when pricing derivatives as it aims at finding a probability measure such that people's varying degrees of risk, already included in the observed prices, can be quantified. Typically, the field of derivative pricing can be divided into two branches; the actual pricing of derivatives, and the risk-management branch. While the pricing of derivatives is fairly self-explanatory, the risk-management discipline is slightly more subtle. Risk-management generally aims at identifying and measuring risk, as well as reducing risk. An example of the two branches in a betting framework could be; a person wanting to act as a bookmaker and thus needs to correctly price the bets, or a bettor wanting to identify and cover his/her risk while in the market. In general, our focus will lie in the risk-management branch, since we will view the market prices as already correctly priced in order to find model parameters suitable for risk-management purposes, such as hedging strategies.

Problem Statement

In-play bet prices on football are driven by the underlying match, which can be represented by score processes. Using general counting/point process theory we will formulate such score processes, in order to model the underlying football match, using Weibull-based counting processes. Then, we will investigate if the Fundamental Theorems of Asset Pricing are applicable in the in-play football betting markets. Specifically, we will formulate a risk-neutral valuation framework for the pricing of football bets, and then calibrate the model prices to actual market prices.

Connection with Existing Work

This thesis was originally inspired by the paper of Divos et al. (2018), in which they develop a risk-neutral framework for pricing and hedging in-play football bets using a market model composed of homogeneous Poisson processes. This led to the author's former semester project at Aalborg University, Andersen and Maillard (2019), in which they implement, reconstruct, and discuss the concepts of Divos et al.'s work. The work of Divos et al. (2018) and Andersen and Maillard (2019) sparked several ideas in the author's mind, some of which are presented here. This thesis can thus be seen as a slight continuation of the work presented in Andersen and Maillard (2019); however, prior reading of which is not necessary.

In the broader spectrum of papers on football modeling and betting, several authors have dealt with the distribution of goals. The paper by Maher (1982) is by many regarded as the first of its kind. Later contributions from Dixon and Coles (1997), Dixon and Robinson (1998), Karlis

and Ntzoufras (2000), Karlis and Ntzoufras (2003), and I. McHale and Scarf (2007) must also be regarded as significant work in the field, all of which dealt with some form of extension of Maher's original model to describe the distribution of goals. Of more recent work should the papers of I. McHale and Scarf (2011), Koopman and Lit (2015), Feng, Polson, and Xu (2016), Boshnakov, Kharrat, and I. G. McHale (2017), and Divos et al. (2018) have a place on the interested reader's reading list; many of which, deal with some form of correlation between the scores in football, or some form of dynamic models.

Thesis Structure

This thesis is structured in the following way; Chapter 2 presents some general theory on point processes. Here, we also introduce the types of processes, we will be dealing with throughout the thesis, namely two types of Weibull-based point processes. Chapter 3 concerns the modeling of football goals. Here, we wish to identify and describe some statistical patterns in football goals and put these in relation to the proposed processes. In Chapter 4, we develop a general market model in which a risk-neutral framework for pricing and hedging of football bets can be carried out. We also present specific model dynamics of the Weibull-based market models. Chapter 5 introduces the historical betting exchange data, and the cleaning hereof, as well as presents a general overview and discussion of the Betfair exchange. We furthermore state the calibration procedure and outline the findings of this. In Chapter 6, we discuss the independence assumption originally imposed in the general market model and propose a natural extension, as well as the results of applying this. Finally, Chapter 7 reviews and concludes the findings of the thesis.

Theory on Point Processes 2

In this chapter, we introduce some concepts and results from the general theory on point and renewal processes, along with an introduction to a few specific processes to be used in later chapters' modeling parts. Section 2.1 covers basic definitions of stochastic processes, information, and martingales. Section 2.2 contains some introductory material on point/counting processes, and Section 2.3 introduces the intensity notion of point/counting processes, as well as the concept of a compensator. Section 2.4 states some results associated with the intensity of a counting process. Section 2.5 shows the notion of renewal processes and renewal theory in connection to the more general view of point processes. Section 2.6 presents the Poisson processes and its characteristics, and Section 2.7 introduces specific point/counting processes based on the Weibull distribution and some characteristics related to these.

2.1 Stochastic Processes

This section is based on Tankov and Cont (2004, Sec. 2.4), with some minor additions from Rinne (2008, Sec. 4.1) & Jeanblanc, Yor, and Chesney (2009, Sec. 1.1.10).

We begin this section by recalling the definition of a stochastic process and associated spaces.

Definition 2.1 (Stochastic Process)
Let $(\Omega, \mathscr{F}, \mathbb{P})$ be a probability space and $\mathbb{T} \subseteq \mathbb{R}$ be an arbitrary, but non-random and non-empty index set. A function $X : \Omega \times \mathbb{T} \to \mathbb{R}$ is called a one-dimensional, real-valued *stochastic process*. The set of realizations of X, i.e. $\{X(t, \omega) \mid \omega \in \Omega, t \in \mathbb{T}\}$ is called the *state space*, and $\mathbb{T}$ is the *parameter space*.

A stochastic process is thus a family of random variables index by $\mathbb{T}$, and we usually denote it as $X = (X_t)_{t \in \mathbb{T}}$. The parameter space may be either discrete or continuous, and we will henceforth use the term "time" when talking about the parameter space. For each realization of the randomness ω, the trajectory $t \mapsto X_t(\omega)$ defines a function of time, called the *sample path* of the stochastic process, which we denote $X_\bullet(\omega)$. In order to account for discontinuities, we need a class of functions that allows this property.

Definition 2.2 (Càdlàg)
Let $\mathbb{T} \in \{[0,T] \mid T < \infty\} \cup \{[0,\infty)\}$ be an interval. A one-dimensional function $f : \mathbb{T} \to \mathbb{R}$ is said to be *càdlàg* if it is right-continuous with left limits, i.e. for each $t \in \mathbb{T}$ the limits

$$f(t-) = \lim_{s \nearrow t} f(s),$$
$$f(t+) = \lim_{s \searrow t} f(s),$$

exist and $f(t) = f(t+)$.

We call a one-dimensional stochastic process $(X_t)_{t\in\mathbb{T}}$ on the probability space $(\Omega, \mathscr{F}, \mathbb{P})$ for càdlàg if the sample path of X is càdlàg for all $\omega \in \Omega$. Sometimes we will also be talking about *càglàd* functions, which are simply the opposite of càdlàg, namely left-continuous with right limits.

We see that any continuous function is obviously càdlàg (and càglàd), however, as stated, a càdlàg function can have discontinuities. Consider the following:

$$\Delta f(t) := f(t) - f(t-).$$

If t is a discontinuity point, we have that $\Delta f(t) \neq 0$, otherwise $\Delta f(t) = 0$. We note that the use of càdlàg functions makes sense in regard to observable events. E.g. consider a goal in a football match. If the goal is scored at time t, then the goal was not scored at $t-$. Denoting the count of goals by a càdlàg function will then make sure that when we are at the time of the goal, t, the goal is counted. On the contrary, if we use a càglàd function to describe the count of goals, then standing at t the goal will not be counted until $t+$.

2.1.1 Information, Histories, & Martingales

When t depicts time, we need to account for the notions of information, history, and predictability in the stochastic framework. Consider a dynamic context in which time flows. More information is revealed to the observer as time passes, i.e. information that is considered random at t may not be random at some future point in time. Thus, we need a time-dependent component in our probability space $(\Omega, \mathscr{F}, \mathbb{P})$ to integrate this feature.

Definition 2.3 (Filtration)
A *filtration* or *information flow* on $(\Omega, \mathscr{F}, \mathbb{P})$ is an increasing family of σ-algebras $\mathbf{F} = (\mathscr{F}_t)_{t\in\mathbb{T}}$ such that $\forall t \geq s \geq 0,\ \mathscr{F}_s \subseteq \mathscr{F}_t \subseteq \mathscr{F}$.

From Definition 2.3, we see that $\mathscr{F}_t$ is interpreted as the information known at time t, which increases as time passes. An event $A \in \mathscr{F}_t$ is such that an observer can decide, given the information $\mathscr{F}_t$, if the event has happened. Likewise, an *$\mathscr{F}_t$-measurable random variable* is a random variable whose value is unveiled at time t. A probability space $(\Omega, \mathscr{F}, \mathbb{P})$ equipped with a filtration is said to be a *filtered probability space*, and is denoted $(\Omega, \mathscr{F}, \mathbf{F}, \mathbb{P})$.

Definition 2.4 (Adapted Process)
Let $(\Omega, \mathscr{F}, \mathbf{F}, \mathbb{P})$ be a filtered probability space. The stochastic process X is called $\mathbf{F}$-*adapted*[1] if for each $t \in \mathbb{T}$, the random variable X_t is $\mathscr{F}_t$-measurable.

From Definition 2.4, we see that an adapted process is a process whose value at time t is unveiled by the filtration $\mathscr{F}_t$. If past values of a stochastic process X is the only available observations, then the filtration is represented by the specific filtration presented in the following definition.

Definition 2.5 (Natural Filtration)
The *history* or *natural filtration* of a stochastic process X on $(\Omega, \mathscr{F}, \mathbb{P})$ is the filtration $(\mathscr{F}_t^X)_{t\in\mathbb{T}}$ where $\mathscr{F}_t^X$ is the σ-algebra generated by the past values of the process, completed by the null sets:

$$\mathscr{F}_t^X = \sigma(X_s, s \in [0,t]) \bigvee \mathcal{N}, \tag{2.1}$$

where $\mathcal{N} := \{A \in \mathscr{F} \mid \mathbb{P}(A) = 0\}$.

The $\bigvee$ notation used in Definition 2.5 is the so-called *join*, and in general it has the meaning:

$$\bigvee_{i=1}^{n} \mathscr{F}_t^i = \sigma\left(\bigcup_{i=1}^{n} \mathscr{F}_t^i\right). \tag{2.2}$$

(2.2) relates to the fact that the union of a collection of σ-algebras is not necessarily a σ-algebra, however, it generates a σ-algebra which is the join as stated.

We can think of the natural filtration as all the information we can extract from the observed sample path of a stochastic process up to, and including, time t.

Definition 2.6 (Martingale)
Let $(\Omega, \mathscr{F}, \mathbf{F}, \mathbb{P})$ be a filtered probability space. A stochastic process X is said to be a $(\mathbb{P}, \mathscr{F}_t)$-*martingale* if X is $\mathscr{F}_t$-adapted, X is $\mathbb{P}$-integrable, i.e. $\mathbb{E}[|X_t|] < \infty$ for any $t \in \mathbb{T}$, and for all $t > s$:

$$\mathbb{E}[X_t \mid \mathscr{F}_s] = X_s, \quad \mathbb{P}\text{-a.s..} \tag{2.3}$$

Furthermore, X is called a $(\mathbb{P}, \mathscr{F}_t)$-*submartingale* if X is $\mathscr{F}_t$-adapted, X is $\mathbb{P}$-integrable, and for all $t > s$:

$$\mathbb{E}[X_t \mid \mathscr{F}_s] \geq X_s, \quad \mathbb{P}\text{-a.s..} \tag{2.4}$$

From the definition of a martingale, it is easy to see that the best prediction of a martingale's future value is its present value. We should point out that the notion of martingales depends

[1] Also denoted as $\mathscr{F}_t$-adapted.

on the filtration and the probability measure, hence the $(\mathbb{P}, \mathscr{F}_t)$ emphasis in the definition. However, when the filtration and measure are implicit we will sometimes drop the emphasis from the notation. Also, when dealing with several probability measures, but the filtration is implicit, we will simply speak of $\mathbb{P}$-martingales and the likes.

The discussion of càdlàg vs. càglàd presented below Definition 2.2 motivates the definition of useful σ-algebra on $\mathbb{T} \times \Omega$.

Definition 2.7 (Predictable)
The *predictable σ-algebra* is the σ-algebra $\mathcal{P}$ generated on $\mathbb{T}\times\Omega$ by all adapted left-continuous processes. A function $X : \mathbb{T} \times \Omega \to \mathbb{R}$ which is measurable with respect to $\mathcal{P}$ is called a one-dimensional *predictable process*.

We see that a predictable process is thus a process whose value at t is "announced by the preceding values", i.e. unveiled at the prior time $t-$. We finish this section by defining some useful notation on the increments of a stochastic process; first, recall that two random variables, X, Y, is equal in distribution, denoted $X \overset{\mathrm{d}}{=} Y$, if:

$$\mathbb{P}(X \leq x) = \mathbb{P}(Y \leq x)\,, \quad \forall x.$$

Definition 2.8 (Stationary & Independent Increments)
A stochastic process X has *independent increments* if for any pair $(s,t) \in \mathbb{R}^2_+$, the random variable $X_{t+s} - X_s$ is independent of $\mathscr{F}^X_s$.

A stochastic process X has *stationary increments* if for any pair $(s,t) \in \mathbb{R}^2_+$

$$X_{t+s} - X_s \overset{\mathrm{d}}{=} X_t.$$

A process is stationary if for all fixed $s > 0$,

$$(X_{t+s} - X_s,\ t \geq 0) \overset{\mathrm{d}}{=} (X_t,\ t \geq 0)\,.$$

2.2 Point Processes

This section is based on Sigman (2009, Sec. 2.1), Hautsch (2012, Sec. 4.1.2), & Björk (2011, Ch. 3).

In this section, we define point and counting processes and show how these are related to each other and to the theory of stochastic processes presented in Section 2.1. We also present the intensity notion of such processes and related results.

Definition 2.9 (Point Process)
A *simple point process* $\psi = \{t_n : n \in \mathbb{N}\}$ is a sequence of strictly increasing points

$$0 < t_1 < t_2 < \cdots t_n \cdots , \tag{2.5}$$

with $\lim_{n\to\infty} t_n = \infty$. We say that t_n is the *arrival time* of the nth arrival (event). Furthermore, we sometimes allow a point t_0 at the origin and define $t_0 := 0$. If the t_n's are random variables defined on some probability space $(\Omega, \mathscr{F}, \mathbb{P})$, then ψ is termed a simple random point process.

Remark: When we consider the case where we define $t_0 := 0$, we write $\psi = \{t_n : n \in \mathbb{N}_0\}$.

In Definition 2.9, the word "simple" refers to the fact that no more than one arrival can happen at the same time, which is stated in (2.5). Henceforth, we will only deal with simple point processes, and we will usually omit the "simple" in front.

Definition 2.10 (Interarrival Time)
Let $\psi = \{t_n : n \in \mathbb{N}_0\}$ be a simple point process. The nth *interarrival time* of ψ is then given by

$$\mathcal{T}_n = t_n - t_{n-1}, \quad n \in \mathbb{N}. \tag{2.6}$$

Remark: When a point is not defined at the origin, the first interarrival time $\mathcal{T}_1$ is not defined.

By definition we have the following:

$$t_n = \sum_{i=1}^{n} \mathcal{T}_i. \tag{2.7}$$

Definition 2.11 (Counting Process)
Let ψ be a simple point process. By defining $N_0 := 0$, we let N_t denote the number of points in the interval $(0, t]$, i.e.

$$N_t = \max \{n : t_n \leq t\} . \tag{2.8}$$

$N = (N_t)_{t \geq 0}$ is then referred to as the *counting process* for ψ.

Remark: The term *nonexplosive* is sometimes used to described a counting process corresponding to a simple point process.

We note that when ψ is a random point process, the corresponding counting process N_t is a càdlàg stochastic process with state space $\mathbb{N}_0$ and parameter space $\mathbb{T} = \mathbb{R}_{\geq 0}$, in which the sample paths are step-functions with upwards jumps of magnitude 1. Moreover, we have that the (random) variable N_t may also be expressed in terms of the indicator function:

$$N_t = \sum_{n=0}^{\infty} \mathbb{1} (t_n \leq t) .$$

The fundamental relationship between the counting process N and the point process ψ is that for each n and t the following holds by simple reasoning:

$$N_t \geq n \iff t_n \leq t. \tag{2.9}$$

Therefore, each process contains sufficient information to reconstruct the other.

2.3 Intensities and Compensators

This section is based on Hautsch (2012, Sec. 4.1.2) & Aalen (1978) with some minor additions from Karr (1991), Segall and Kailath (1975), Brémaud (1981), Jeanblanc, Yor, and Chesney (2009), and Daley and Vere-Jones (2003).

Here, we introduce essential components in the theory of point processes; namely the intensity process[2] and the notion of a compensator.

[2]The notation is not consistent in all papers or textbooks; other common names are conditional intensity function and stochastic intensity.

Definition 2.12 (Intensity Process)
Let N be an $\mathscr{F}_t$-adapted counting process for a simple point process ψ on a filtered probability space $(\Omega, \mathscr{F}, \mathbf{F}, \mathbb{P})$, and assume that $\lambda = (\lambda_t)_{t\geq 0}$ is a non-negative càglàd (stochastic) process defined by

$$\lambda_t = \lim_{\Delta \searrow 0} \frac{1}{\Delta} \mathbb{E}\left[N_{t+\Delta} - N_t \mid \mathscr{F}_t\right], \quad \forall t \geq 0. \tag{2.10}$$

Then, the process λ is called the *intensity process* of the counting process N with respect to the filtration $\mathbf{F}$ and probability measure $\mathbb{P}$.

The intensity process characterizes the evolution of the counting process N conditioned on some filtration $\mathbf{F}$. That is, Definition 2.12 reveals the intensity process as the instantaneous arrival rate of an event in t conditioned on some filtration $\mathbf{F}$.

Typically, we consider the case of the natural filtration; $\mathscr{F}_t = \mathscr{F}_t^N$, however, the intensity notion also allows for broader filtration choices, e.g. one that includes unobservable factors. We will sometimes emphasize the filtration used by writing the intensity as $\lambda_t^{\mathscr{F}_t}$.

The use of càglàd (left-continuous with right limits) in Definition 2.12 is associated with predictability, that is, if we consider a discontinuity point, then the intensity at that point should be defined by the information before that point, and not by what happens at the point itself.[3] Furthermore, according to Björk, the càglàd assumption is important in ensuring the uniqueness of the intensity process.[4]

Due to the assumption of an underlying simple point process, or correspondingly a nonexplosive counting process, an alternative expression for the intensity is, according to Hautsch, given by

$$\lambda_t = \lim_{\Delta \searrow 0} \frac{1}{\Delta} \mathbb{P}(N_{t+\Delta} - N_t > 0 \mid \mathscr{F}_t), \tag{2.11}$$

which can be related to the conditional probability per unit time to observe an event in the next instant, given the conditioning information.

Since N_t is non-decreasing we have that it is a submartingale:

$$N_s \leq \mathbb{E}\left[N_t \mid \mathscr{F}_s\right], \; s < t,$$

and according to the Doob–Meyer decomposition any (locally bounded) $(\mathbb{P}, \mathscr{F}_t)$-submartingale N_t can be decomposed into a unique zero-mean $(\mathbb{P}, \mathscr{F}_t)$-martingale M_t and a unique $(\mathbb{P}, \mathscr{F}_t)$-predictable cumulative process $\Lambda(t)$, i.e.,

$$N_t = M_t + \Lambda(t), \tag{2.12}$$

where $\Lambda(t)$ is called the *compensator* and is defined by

[3] For a further treatment of the importance of predictability of the intensity process, see Daley and Vere-Jones (2003).

[4] For a formal discussion of uniqueness of the intensity process see Björk (2011).

$$\Lambda(t) := \int_0^t \lambda_u du. \tag{2.13}$$

From (2.12) we see that the stochastic process $M = (M_t)_{t\geq 0}$, characterized by

$$M_t = N_t - \Lambda(t), \tag{2.14}$$

is a $(\mathbb{P}, \mathscr{F}_t)$-martingale. Accordingly, we obtain

$$\mathbb{E}\left[N_t \mid \mathscr{F}_s\right] = \mathbb{E}\left[\int_0^t \lambda_u du \mid \mathscr{F}_s\right], \ \mathbb{P}\text{-a.s.},$$

or

$$\mathbb{E}\left[N_t - N_s \mid \mathscr{F}_s\right] = \mathbb{E}\left[\int_s^t \lambda_u du \mid \mathscr{F}_s\right] \ \mathbb{P}\text{-a.s.} \tag{2.15}$$

Equation (2.15) yields an alternative implicit definition of the intensity process, and we see from it that the expected number of events in an interval $(s, t]$ given $\mathscr{F}_s$ is computed as the conditional expectation of the integrated intensity.

The existence of an intensity process for a counting process is a technical and non-trivial problem that is out of the scope for this thesis[5], however, Björk states that only counting processes for which the compensator $\Lambda(t)$ is absolutely continuous have an intensity process. This implies that if we restrict our focus to counting processes with an intensity process, we exclude processes with jumps at predetermined times, i.e. we only focus on random point processes.

2.3.1 Hazard Function

We briefly give a description of the *hazard function*. The hazard function is a counterpart to the intensity, which is used extensively in traditional duration and survival analysis. However, in such a framework, generally there does not exists a history of the process before the beginning of a spell and therefore, the hazard function is defined by

$$\begin{aligned} h(t) &:= \frac{f(t)}{1 - F(t)} \\ &= \lim_{\Delta \searrow 0} \frac{1}{\Delta} \mathbb{P}(t \leq \mathcal{T} < t + \Delta \mid \mathcal{T} \geq t), \end{aligned} \tag{2.16}$$

where $\mathcal{T}$ is an interarrival time, i.e. a positive random variable, whose distribution function F admits the density $f = F'$.

From (2.16) we see that the intensity and hazard function portray the same concept. However, the distinction is that the hazard function is defined in terms of the interarrival time $\mathcal{T}$.

[5] Daley and Vere-Jones (2003, Sec. 3.3) and Karr (1991, Sec. 2.4) have comprehensive discussions on the existence of a intensity process for counting processes.

2.4 Girsanov's Theorem and Other Results

This section is based on Sokol and Hansen (2015), Brémaud (1981, Sec. VI.2), & Björk (2011, Ch. 3).

We want to find the dynamics of the intensity process when we change the probability measure. For this, we use Girsanov's theorem as presented below. However, first, we introduce some useful notation.

Definition 2.13 (λ-Compatibility)
We say that the process $\widetilde{\lambda} = (\widetilde{\lambda}_t)_{t\in\mathbb{T}}$ is λ-*compatible* if it holds for all $\omega \in \Omega$ that $\widetilde{\lambda}_t(\omega) = 0$ whenever $\lambda_t(\omega) = 0$, and if the process γ defined by $\gamma_t = \frac{\widetilde{\lambda}_t}{\lambda_t}$ is locally bounded.

Remark: In Definition 2.13, we use the convention that zero divided by zero is equal to one.

Theorem 2.14 (Girsanov's Theorem for Counting Processes)**.** *Let $\mathbb{T} \in \{[0,T] \mid T < \infty\}$ and let N be an adapted counting process on the filtered probability space $(\Omega, \mathscr{F}, \mathbf{F}, \mathbb{P})$. Assume that N has the $\mathscr{F}_t$-intensity process $\lambda = (\lambda_t)_{t\in\mathbb{T}}$. Let h be a predictable process such that*

$$h_t \geq -1, \ \mathbb{P}-\text{a.s.} \tag{2.17}$$

and define the process L by

$$\begin{cases} dL_t &= L_{t-} h_t (dN_t - \lambda_t dt), \\ L_0 &= 1, \end{cases} \tag{2.18}$$

on the interval $\mathbb{T}$. Assume furthermore that

$$\mathbb{E}_{\mathbb{P}}[L_T] = 1. \tag{2.19}$$

Now define a new probability measure $\mathbb{Q}$ absolutely continuous in reference to $\mathbb{P}$ on $\mathscr{F}_T$ by

$$\frac{d\mathbb{Q}}{d\mathbb{P}} = L_T. \tag{2.20}$$

Then N has the $\mathbb{Q}$-intensity $\widetilde{\lambda} = (\widetilde{\lambda}_t)_{t\in\mathbb{T}}$, given by

$$\widetilde{\lambda}_t = \lambda_t(1 + h_t). \tag{2.21}$$

Proof. Omitted. Can be found in Björk (2011, Thm. 5.1.1, p. 41).

Remark: Girsanov's theorem can be naturally extended to the multivariate case, in which case; $L_t = \prod_{i=1}^{k} L_t^i$, where L_t^i is given by (2.18). For more on the multivariate case of Girsanov's theorem, see Brémaud (1981, Thm. T2 & T3, pp. 165-167).

Girsanov's theorem is an important result in risk-neutral pricing as it describes the dynamics of a stochastic process under a change of measure, e.g. a risk-neutral measure. An extended

discussion on the change of measure of counting processes can be found in Brémaud (1981, Ch. VI) and Sokol and Hansen (2015).

As a side note of Girsanov's theorem, we obviously have that $\mathbb{Q}$ is absolutely continuous to $\mathbb{P}$, however, a natural question to ask is whether or not they are also equivalent. This is the case when the Radon-Nikodym derivative is almost surely positive. The following lemma from Sokol and Hansen states a condition for when this is the case.

Lemma 2.15. *If the set of zeroes of $\widetilde{\lambda}$ has Lebesgue measure zero, L is almost surely positive.*

Proof. Omitted.

Furthermore, we also follow Sokol and Hansen and state sufficient criteria for when the process L, (2.18), is a martingale. Using the notation $\log_+ x := \max\{0, \log x\}$ $x \geq 0$, with the convention that the logarithm of zero is minus infinity, the result is as follows.

Theorem 2.16. *Assume that λ and $\widetilde{\lambda}$ are non-negative, predictable, and locally bounded. Assume that $\widetilde{\lambda}$ is λ-compatible. It holds that L is a martingale if there exists an $\varepsilon > 0$ such that for $0 \leq u \leq t$, with $t - u \leq \varepsilon$, one of the following two conditions are satisfied:*

$$\mathbb{E}\left[\exp\left(\int_u^t (\gamma_s \log \gamma_s - (\gamma_s - 1)\lambda_s)\, ds\right)\right] < \infty \quad \text{or} \tag{2.22}$$

$$\mathbb{E}\left[\exp\left(\int_u^t \lambda_s ds + \int_u^t \log_+ \gamma_s dN_s\right)\right] < \infty. \tag{2.23}$$

The direct use of Theorem 2.16 is an existence result for counting processes of simple point processes, i.e. nonexplosive counting processes, with particular intensities, as we shall use later. The is due to the change of measure obtained from the martingale property of L yields the existence of a nonexplosive counting process with intensity process $\widetilde{\lambda}$ on a bounded time interval $\mathbb{T} \in \{[0, T] \mid T < \infty\}$.

2.4.1 Filtrations and Uniqueness of Measure

Lastly, we show some more results pertaining to the intensity process; a result on the intensity process on different filtrations and a result of the uniqueness of measures based on intensities.

As discussed in Section 2.3, we should emphasize that the intensity notion is tied to a particular choice of filtration. If we, however, have two different filtrations $\mathbf{F}$ and $\mathbf{G}$ and a counting process N adapted to both $\mathbf{F}$ and $\mathbf{G}$ there is no reason to suspect that the $\mathbf{F}$-intensity will coincide with the $\mathbf{G}$-intensity. In general, there is no interesting connection between the two intensities, however, in the special case, when $\mathbf{G}$ is a sub-filtration of $\mathbf{F}$, we have the following result.

Proposition 2.17 (Intensities on Different Filtration)**.** *Let $\mathbf{F} = (\mathscr{F}_t)_{t\geq 0}$ and $\mathbf{G} = (\mathscr{G}_t)_{t\geq 0}$ be filtrations with $\mathscr{F}_t^N \subseteq \mathscr{G}_t \subseteq \mathscr{F}_t$ for all t, and assume that a counting process N has intensity process $(\lambda_t^{\mathscr{F}_t})_{t\geq 0}$ with respect to $\mathbf{F}$. Then the intensity with respect to $\mathbf{G}$ is given by $\lambda_t^{\mathscr{G}_t} = \mathbb{E}\left[\lambda_t^{\mathscr{F}_t} \mid \mathscr{G}_t\right]$.*

Proof. Omitted. Can be found in Segall and Kailath (1975, p. 137) or Björk (2011, p. 35).

Proposition 2.17 states that we can obtain the intensity of a counting process with respect to a smaller than the original filtration, $\mathbf{G}$, by a conditional expectation of the intensity in respect to the original intensity $\mathbf{F}$. Finally, we state a uniqueness of probability measures in the special case of the natural filtration.

Theorem 2.18 (Uniqueness of Measures)**.** *Let $\mathbb{P}$ and $\mathbb{Q}$ be two probability measures on $(\Omega, \mathscr{F})$, and let N be a counting process such that for some $\mathscr{F}_t^N$-intensity process $(\lambda_t)_{t\geq 0}$, N admits the $\left(\mathbb{P}, \mathscr{F}_t^N\right)$-intensity λ_t and the $\left(\mathbb{Q}, \mathscr{F}_t^N\right)$-intensity λ_t. Then $\mathbb{P}$ and $\mathbb{Q}$ coincide on the events of $\mathscr{F}_\infty^N$.*

Proof. Omitted. Can be found in Brémaud (1981, p. 64) (the multivariate case) or Karr (1991, p. 63).

Theorem 2.18 roughly states that to a given intensity process with respect to the natural filtration of the counting process, corresponds one probability measure at most.

2.5 Renewal Processes

This section is based on Ross (2019, Sec. 7.1-7.2) & Cha and Finkelstein (2018, Sec. 3.1), with minor additions from Sigman (2009, Sec. 1.2) & Hautsch (2012, Sec. 4.1.4).

In Section 2.2 we saw the relationship between point and counting processes and recognized that such processes may be specified in terms of its counts in certain intervals and in terms of the interarrival times of the point process. For some specific processes, one of these specifications might be more intuitive than the other. In this section, we present processes which have a simple specification in terms of the interarrival times.

Definition 2.19 (Renewal Process)
A random simple point process $\psi = \{t_n : n \in \mathbb{N}_0\}$ for which the interarrival times $\{\mathcal{T}_n : n \in \mathbb{N}\}$ form an independent and identically distributed (i.i.d.) sequence is called a *renewal process*. We then refer to t_n as the nth *renewal epoch* and $F(x) = \mathbb{P}(\mathcal{T} \leq x)\,,\ x \geq 0$, denotes the common interarrival time distribution.

To avoid trivialities we generally assume that $F(0) < 1$, hence ensuring that $t_n \to \infty$ almost surely. Furthermore, we generally say that the renewal process is an *ordinary renewal process* if we have a point at the origin; $t_0 = 0$, i.e. $\mathcal{T}_1$ also has distribution F. We will henceforth only work with ordinary renewal processes, and hence we will drop the word "ordinary" from the description.

We note that the i.i.d. assumption of the interarrival times with a common distribution is in direct contrast to the general counting process, in which, with the exception of the homogeneous

Poisson process (see Section 2.6), the interarrival times have differing distributions.

When dealing with renewal processes it is beneficial to define the left-continuous counting process for ψ:

$$\check{N}_t = \sum_{n=0}^{\infty} \mathbb{1}\left(t_n < t\right). \tag{2.24}$$

This is such that we can define the *backwards recurrence time* at time t, which is given by the process $Z = (Z_t)_{t\geq 0}$ where

$$Z_t = t - t_{\check{N}_t}. \tag{2.25}$$

The backwards recurrence time is the time elapsed since the last point, and is therefore a càglàd (stochastic) function that grows linearly in time with discrete jumps back to zero after each point t_n. We also note that

$$Z_{t_n} = t_n - t_{n-1} = \mathcal{T}_n.$$

Now, recall the intensity function (2.10). We see that when there exists no history of the counting process before the beginning of a spell, which is the case for renewal processes since the interarrival times are i.i.d. random variables, we can restrict the intensity function to the filtration generated by the backwards recurrence time, $(\mathscr{F}_t^Z)_{t\geq 0}$. Furthermore, recall the hazard function (2.16) which was defined in terms of the interarrival times. We see that the intensity of a renewal process then coincides with the hazard function evaluated at Z_t, i.e. we have that

$$\lambda_t^{\mathscr{F}_t^Z} = h(Z_t).$$

Thus, we have that a renewal process possesses the simplest history, i.e. the time elapsed since the last renewal. This has the effect that the previous renewals do not influence the times of future renewals.

2.5.1 Renewal Function

Despite the simplistic appearance of renewal processes, probabilistic description and properties of such processes are not straightforward. Consider for example the question of attaining the probability mass function of N_t, the random variable describing the number of renewals in $(0, t]$.

Using (2.9) and (2.7) the probability that there are exactly n events in $(0, t]$ is given by

$$\begin{aligned} \mathbb{P}(N_t = n) &= \mathbb{P}(N_t \geq n) - \mathbb{P}(N_t \geq n+1) \\ &= \mathbb{P}(t_n \leq t) - \mathbb{P}(t_{n+1} \leq t) \\ &= F_n(t) - F_{n+1}(t), \end{aligned} \tag{2.26}$$

where $F_n(t)$ is the n-fold convolution of $F(t)$ with itself and by definition $F_0(t) = 1,\ F_1(t) = F(t)$. This is due to the result that the distribution of a sum of i.i.d. random variables can be expressed by the corresponding convolution.

The following function plays a fundamental role in renewal theory and is a function of t that gives the expected number of renewals in $(0, t]$.

Definition 2.20 (Renewal Function)
Let N be the counting process of a renewal process ψ. The *renewal function* or *mean function* is defined by the following expectation:

$$H(t) = \mathbb{E}[N_t]. \tag{2.27}$$

According to Ross (2019), it can be shown that the renewal function $H(t)$ uniquely determines the renewal process. Particularly, we see a one-to-one correspondence between the interarrival time distribution F and the renewal function $H(t)$. Furthermore, Ross (2019) also states that $H(t) < \infty$ for all $t < \infty$.

From (2.26) and (2.27) it follows that $H(t)$ can be expressed as the infinite sum of convolutions:

$$H(t) = \mathbb{E}[N_t] = \sum_{n=1}^{\infty} n\mathbb{P}(N_t = n) = \sum_{n=1}^{\infty} n\left(F_n(t) - F_{n+1}(t)\right) = \sum_{n=1}^{\infty} F_n(t). \tag{2.28}$$

Since a renewal process has history that affect future arrivals, with the obvious exception of the homogeneous Poisson process (see Section 2.6), it cannot posses the Markov property. Consequently, its increments are not independent. It does, however, have Markovian points. Those are the renewal epochs. Therefore, we can employ a renewal-type reasoning in analytical descriptions of the main renewal indices. Specifically, by assuming that the interarrival time distribution F is continuous with density function f, the existence of the renewal epochs allows us to write the following integral equation for $H(t)$:

$$H(t) = F(t) + \int_0^t H(t - x)f(x)dx \tag{2.29}$$

(2.29) is called the *renewal equation* and can sometimes be solved to obtain the renewal function. We obtain (2.29) by noting that the following holds:

$$H(t) = \mathbb{E}[N_t] = \int_0^{\infty} \mathbb{E}[N_t \mid \mathcal{T}_1 = x]\, f(x)dx. \tag{2.30}$$

Now suppose that the first renewal happens at time $x \leq t$, then we can utilize that a renewal process probabilistically restarts at a renewal epoch. It then follows that the number of renewals by time t must have the same distribution as one plus the number of renewals in the first $t - x$ time units. Therefore,

$$\mathbb{E}[N_t \mid \mathcal{T}_1 = x] = 1 + \mathbb{E}[N_t - N_x], \quad x \leq t.$$

We also clearly have that $\mathbb{E}[N_t \mid \mathcal{T}_1 = x] = 0$ when $x > t$. Now (2.29) follows from (2.30) and that $\mathbb{E}[N_t - N_x] = H(t - x)$.

Thus, we see that attaining the renewal function, and thus specifying the renewal process, for a finite interval involves solutions of the corresponding renewal equation that in several occasions

should be done numerically.

2.6 Poisson Process

This section is based on Tankov and Cont (2004, Sec. 2.5), Jeanblanc, Yor, and Chesney (2009, Ch. 8), Klebaner (2005, Sec. 9.4), & Pedersen (2017, Ch. 1), with minor additions from Sigman (2009) and Björk (2011).

In this section, we present a specific class of point/counting processes that is essential in any work related to such; namely, we introduce the Poisson process in its homogeneous and inhomogeneous form. Henceforth, we will call the homogeneous Poisson process simply by the term "Poisson process", and use the full terminology for the inhomogeneous variant, or when emphasis is needed.

2.6.1 Homogeneous Poisson Process

There are several equivalent definitions of the Poisson process. In the following, we present the one which focuses on the counting process part.

Definition 2.21 (Poisson Process)
Let $(\Omega, \mathscr{F}, \mathbf{F}, \mathbb{P})$ be a filtered probability space and $\lambda > 0$ a constant. The process $N = (N_t)_{t\geq 0}$ is called a *homogeneous Poisson process* with intensity λ with respect to the filtration $\mathbf{F}$ if it satisfies the following conditions:

(i) $N_0 = 0$ a.s.

(ii) $t \mapsto N_t(\omega)$ is càdlàg, non-decreasing, and $\mathbb{N}_0$-valued for all $\omega \in \Omega$.

(iii) N is adapted to $\mathbf{F}$.

(iv) $N_t - N_s \sim \text{Poi}\left(\lambda(t-s)\right)$ for $0 \leq s \leq t$.

(v) $(N_t)_{t\geq 0}$ has independent increments.

Note that (i) and (ii) in Definition 2.21 ensures that the process is a counting process, (iii) is such that we can observe the outcome of the process, (iv) states that an arbitrary increment is Poisson distributed with λ times the time difference between t and s, from which also stationarity follows, and finally, (v) is rather self-explanatory.

An equally valid definition of the Poisson process would be to define it in terms of the interarrival times, in which the definition then becomes something in line of the following: Let ψ be a point process with independent interarrival times $\{\mathcal{T}_n : n \in \mathbb{N}\}$ that follows an exponential distribution with parameter λ, and furthermore:

$$\mathbb{P}(N_t = n) = \mathrm{e}^{-\lambda t}\frac{(\lambda t)^n}{n!}.$$

It then follows from the memoryless property of the exponential distribution, Proposition A.5, that a Poisson process is also a renewal process.

Due to (2.7) we see that the distribution of t_n, the nth arrival time, is the nth-fold convolution of the exponential distribution and is thus gamma distributed with parameters n, λ, and its density is given by

$$f_n(t) = \lambda \mathrm{e}^{-\lambda t}\frac{(\lambda t)^{n-1}}{(n-1)!}, \quad t \geq 0,$$

where $f_1(t) = f(t) = \lambda \mathrm{e}^{-\lambda t}$ is the exponential density function. Lastly, the following proposition states that the only counting processes with stationary and independent increments are Poisson processes.

Proposition 2.22. *Let $N = (N_t)_{t\geq 0}$ be a counting process with stationary and independent increments. Then N is a homogeneous Poisson process.*

Proof. Omitted. See Tankov and Cont (2004, pp. 54–55).

2.6.2 Inhomogeneous Poisson Process

Let us generalize the Poisson process, namely, we want to remove the rather strict stationarity property of the homogeneous Poisson process.

Definition 2.23 (Inhomogeneous Poisson Process)
Let $(\Omega, \mathscr{F}, \mathbf{F}, \mathbb{P})$ be a filtered probability space and let $(\lambda_t)_{t\geq 0}$ be a deterministic intensity process, namely an $\mathbb{R}_+$-valued Borel function satisfying $\Lambda(t) < \infty$, $\forall t$ and $\Lambda(\infty) := \int_0^\infty \lambda_u du = \infty$. A process N is then called an *inhomogeneous Poisson process* with intensity λ_t with respect to $\mathbf{F}$ if is satisfies the following conditions:

(i) $N_0 = 0$ a.s.

(ii) $t \mapsto N_t(\omega)$ is càdlàg, non-decreasing, and $\mathbb{N}_0$-valued for all $\omega \in \Omega$.

(iii) N is adapted to $\mathbf{F}$.

(iv) $N_t - N_s \sim \mathrm{Poi}\left(\int_s^t \lambda_u du\right)$ for $0 \leq s \leq t$.

(v) $(N_t)_{t\geq 0}$ has independent increments.

Remark: It is also possible to define an inhomogeneous Poisson process with stochastic intensity (see e.g. Jeanblanc, Yor, and Chesney (2009, p. 476)), however, such process are not of particular interest in this thesis.

From Definition 2.23 it is easy to see that the inhomogeneous Poisson process also encloses the homogeneous Poisson process, namely with the simple choice $\lambda_u = \lambda$, i.e. $\Lambda(t) = \lambda t$. Furthermore, due to N_t having the Poisson distribution with parameter $\int_0^t \lambda_u du$, it follows that

$$\mathbb{E}[N_t] = \int_0^t \lambda_u du, \tag{2.31}$$

$$\mathbb{V}\mathrm{ar}[N_t] = \int_0^t \lambda_u du. \tag{2.32}$$

2.6.3 Time-change Transformation

A central result in martingale-based theory of point process is the (random) time change theorem that allows the transformation of a broad class of point processes to a unit-rate homogeneous Poisson process.

Theorem 2.24 (Change of Time)**.** *Let $(\Omega, \mathscr{F}, \mathbf{F}, \mathbb{P})$ be a filtered probability space, and let $N = (N_t)_{t\geq 0}$ be a counting process with intensity process $\lambda = (\lambda_t)_{t\geq 0}$ in respect to $\mathbf{F}$ that satisfies*

$$\int_0^\infty \lambda_t dt = \infty.$$

Define for all t, the stopping-time τ_t as the solution to

$$\int_0^{\tau_t} \lambda_s ds = t. \tag{2.33}$$

Then, the counting process $\widetilde{N}_t = N_{\tau_t}$ is a homogeneous Poisson process with intensity $\lambda = 1$.

Proof. Omitted. See Brémaud (1981, pp. 41–42).

Hence, (2.33) corresponds to a change of the time scale from t to τ_t transforming N_t into a unit-rate Poisson process $\widetilde{N}_t = N_{\tau_t}$. This result can be used for simulation or model diagnostic purposes, something which is discussed further in Hautsch (2012, Sec. 4.1.5).

Using Theorem 2.24, we see that we can construct an inhomogeneous Poisson process from a deterministic time-changed homogeneous Poisson process. We have the compensator of an inhomogeneous Poisson process $\Lambda(t)$ and consider a Poisson process $\widehat{N}$ with constant intensity equal to 1. Then $N_t = \widehat{N}_{\Lambda(t)}$ is an inhomogeneous Poisson process with intensity $\Lambda(t)$.

2.6.4 Watanabe's Theorem

We finish this section with two results that give a way to determine if a counting process has independent increments.

Theorem 2.25 (Watanabe's Theorem)**.** *Let N be a counting process with a continuous deterministic compensator $\Lambda(t)$. Then it has independent Poisson distributed increments, i.e., the distribution of $N_t - N_s \sim \mathrm{Poi}\left(\int_s^t \lambda_u du\right)$, $0 \leq s < t$.*

Proof. Omitted. Can be found in Klebaner (2005, pp. 256–257).

Theorem 2.25 can be generalized to discontinuous but deterministic compensators. The proof can be found in Shiryaev and Liptser (2001, Ch. 18), where the form of the distribution of the increments is also given along with an extensive discussion.

Theorem 2.26. *Let N be a counting process with a deterministic compensator $\Lambda(t)$. Then it has independent increments.*

2.7 Weibull-based Point Processes

This section is based on Murthy, Xie, and Jiang (2004, Ch. 15), Boshnakov, Kharrat, and I. G. McHale (2017), & Rinne (2008, Sec. 4.3-4.4) with some minor additions from Casarin (2005).

We are now ready to present explicit processes that will be important for the modeling part in later chapters; namely, here we introduce some point processes based on the Weibull distribution and present some characteristics of these.

2.7.1 Weibull Process

First, we consider a point/counting process in which the intensity is a continuous deterministic function of time given by the so-called *Weibull intensity function*. A consequence of this, cf. Theorem 2.25, is that the process has independent Poisson increments, i.e. is an inhomogeneous Poisson process.

Definition 2.27 (Weibull Process)
Let N be an inhomogeneous Poisson process with the deterministic intensity process $\lambda = (\lambda_t)_{t\geq 0}$ characterized by:

$$\lambda_t = \alpha\beta t^{\beta-1}, \tag{2.34}$$

with $\alpha, \beta \in \mathbb{R}_+$. Then N is called a *Weibull process*.

Remark: The Weibull process has many names in the literature, e.g. power law process, Rasch-Weibull process, Weibull intensity function, and Weibull-Poisson process.

The Weibull intensity (2.34) explains many of the names used for this process. We should point out that what is Weibull distributed in the Weibull process is the arrival time of the first event t_1, whereas the arrival of $t_2, t_3, \ldots$ and the interarrival times $\{\mathcal{T}_n\}$ for $n \geq 2$ are not Weibull distributed.

Using (2.34) in (iv) of Definition 2.23, we can obtain

$$
\begin{aligned}
N_t - N_s &\sim \text{Poi}\left(\int_s^t \alpha\beta u^{\beta-1} du\right) \\
&= \text{Poi}\left(\alpha t^\beta - \alpha s^\beta\right) \\
&= \text{Poi}\left(\alpha\left(t^\beta - s^\beta\right)\right).
\end{aligned} \tag{2.35}
$$

Furthermore, we also have that

$$
\begin{aligned}
\mathbb{E}\left[N_t - N_s \mid \mathscr{F}_s\right] &= \mathbb{E}\left[\int_s^t \lambda_u du \mid \mathscr{F}_s\right] \\
&= \mathbb{E}\left[\alpha\left(t^\beta - s^\beta\right) \mid \mathscr{F}_s\right] \\
&= \alpha\left(t^\beta - s^\beta\right).
\end{aligned} \tag{2.36}
$$

2.7.2 Weibull Renewal Process

We now consider another Weibull-based point process; namely a renewal process in which the interarrival times are Weibull distributed.

Definition 2.28 (Weibull Renewal Process)
Let $\alpha, \beta \in \mathbb{R}_+$ and let N be a renewal process with interarrival times $\{\mathcal{T}_n : n \in \mathbb{N}\}$ distributed according to a Weibull distribution; $\mathcal{T}_n \sim \text{Weibull}(\alpha, \beta)$, $n = 1, 2, \ldots$. Then the process N is called a *Weibull renewal process*.

The intensity process of a Weibull renewal process is given by the hazard function evaluated at the backwards recurrence time:

$$
\lambda_t = h(Z_t) = \alpha\beta\left(Z_t\right)^{\beta-1}, \tag{2.37}
$$

that is, for $t \in (t_n, t_{n+1}]$ the intensity is given by $\alpha\beta\left(t - t_n\right)^{\beta-1}$.

We then have that the expected number of events in an arbitrary interval given the information available is given by:

$$
\begin{aligned}
\mathbb{E}\left[N_t - N_s \mid \mathscr{F}_s\right] &= \mathbb{E}\left[\int_s^t \lambda_u du \mid \mathscr{F}_s\right] \\
&= \mathbb{E}\left[\int_s^t \alpha\beta(Z_u)^{\beta-1} du \mid \mathscr{F}_s\right] \\
&= \alpha\beta\,\mathbb{E}\left[\int_s^t (Z_u)^{\beta-1} du \mid \mathscr{F}_s\right].
\end{aligned} \tag{2.38}
$$

Weibull Count Model

The emerging distribution of N_t, i.e. the pmf given by (2.26), when N is a Weibull renewal process is known as the *Weibull Count Model*, and was first derived in McShane et al. (2008). The Weibull Count Model is given by

$$\mathbb{P}(N_t = n) = \sum_{j=n}^{\infty} \frac{(-1)^{n+j} \left(\alpha t^{\beta}\right)^j \varsigma_j^n}{\Gamma\left(\beta j + 1\right)}, \tag{2.39}$$

where $\Gamma(\cdot)$ is the gamma function and ς_j is given by:

$$\varsigma_j^0 = \frac{\Gamma\left(\beta j + 1\right)}{\Gamma\left(j + 1\right)}, \quad j = 0, 1, 2, \ldots$$

and

$$\varsigma_j^{n+1} = \sum_{m=n}^{j-1} \varsigma_m^n \frac{\Gamma\left(\alpha j - \alpha m + 1\right)}{\Gamma\left(j - m + 1\right)},$$

for $n = 1, 2, \ldots$ and $j = n + 1, n + 2, n + 3, \ldots$.

Furthermore, the expected value, i.e. the renewal function, of the Weibull Count Model is given by

$$\mathbb{E}\left[N_t\right] = \sum_{n=1}^{\infty} \sum_{j=n}^{\infty} \frac{n(-1)^{n+j} \left(\alpha t^{\beta}\right)^j \varsigma_j^n}{\Gamma\left(\beta j + 1\right)}, \tag{2.40}$$

and the variance by

$$\mathbb{V}\mathrm{ar}\left[N_t\right] = \sum_{n=2}^{\infty} \sum_{j=n}^{\infty} \frac{n^2(-1)^{n+j} \left(\alpha t^{\beta}\right)^j \varsigma_j^n}{\Gamma\left(\beta j + 1\right)} - \left(\sum_{n=1}^{\infty} \sum_{j=n}^{\infty} \frac{n(-1)^{n+j} \left(\alpha t^{\beta}\right)^j \varsigma_j^n}{\Gamma\left(\beta j + 1\right)} \right)^2. \tag{2.41}$$

In the next chapter, we will investigate and assess the usability of the two proposed Weibull-based counting process on football matches. Specifically, we want to analyze the statistical properties of the processes in connection with the observed properties of football goals.

Football Match Characteristics 3

In this chapter, we analyze and show some empirical characteristics of football goals, and discuss how these relate to the characteristics of the proposed Weibull-based counting processes. In general, there are three main football characteristics that we are interested in; the distribution of goals, the intensity of goals throughout the match, and the distribution of waiting times of goals. Section 3.1 concerns with the distribution of goals, Section 3.2 casts light upon the intensity of goals throughout the game, and Section 3.3 deals with the distribution of the waiting times of football goals.

3.1 Distribution of Goals

In this section, we investigate the empirical distributions of goals to see if they show some interesting patterns, and if we can obtain similar features with the distribution of the proposed models, i.e. the probability mass function arising from the counting processes via $N_T - N_0$. We start by presenting a histogram of the number of goals by each side in each match for the English Premier Leagues. We consider all Premier League matches from August 2004 through May 2019, and neglect the possibility that football is evolving throughout time; it is simply more important for us to obtain a large sample. This histogram is shown in Figure 3.1.

From Figure 3.1 we see that the home teams tend to score more goals than the away team. In about two-thirds of the games, the away team only scored one or zero goals. In about a third of the games the home team scores one goal, and in about a quarter of the games the home team either scores 0 or 2 goals. We also see that the most frequent observation for the home team is 1 goal, followed by 2 goals, whereas the most frequent observation for the away team is no goals followed closely by 1 goal. In Table 3.2 we show a summary of the mean and variance of the goals by each side.

We should note that each individual match, if played over and over, may have characteristics that behave very different from the general patterns. Such scenarios can never be observed in the real world, therefore, the only data we have available is the collective of these matches played only once. That is, to make some general assumptions of a football game, we implicitly assume that all games follows the same patterns and have the same characteristics as the masses. This may

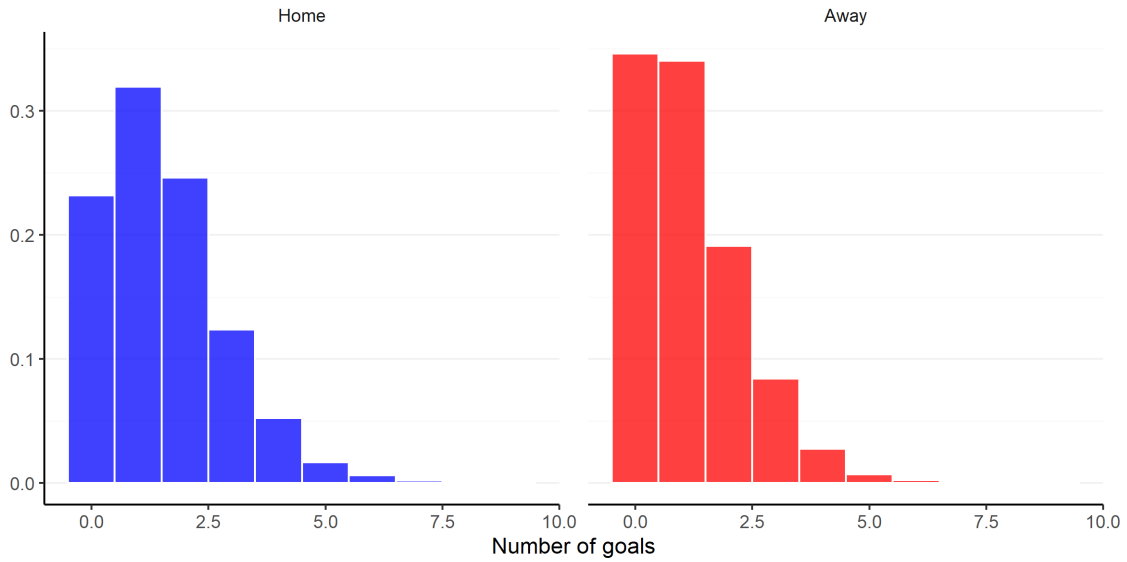


Figure 3.1: Histogram of number of goals in all English Premier League matches between August 2004 and May 2019, separated by home and away team.

	Home	Away
Mean	1.536491	1.138246
Variance	1.706863	1.287430

Table 3.2: Sample means and variances of the English Premier League football data for each side.

not be applicable for all matches, and thus we should be weary in making too strong conclusions based on the general patterns.

3.1.1 Poisson Distribution

Let us now have a look at how well the Poisson distribution describes the empirical distribution. Right off the bat, we see a potential misfit; the sample variances are larger than the sample means, this is known as *overdispersion* and since the Poisson distribution only has one free parameter, it does not allow for the variance to be set independently of the mean. Despite this lack of variability, let us try to fit a Poisson distribution to the empirical data. This fit is shown in Figure 3.3, which confirms the initial observation; the Poisson distribution cannot accurately portray the sample data. Despite the Poisson distribution fit being obviously flawed, it does yield a decent fit given that it only has one free parameter, i.e. it captures most of the features displayed in the sample.

For the sake of statistical support of the initial observation, we perform a goodness-of-fit test, namely the Pearson's chi-squared test, with results shown in Table 3.4. In this test we seek not to reject the null hypothesis stating that the sample data follows a Poisson distribution, i.e. we want large p-values. From the goodness-of-fit results, we conclude that the sample data is very unlikely to originate from Poisson distributions.

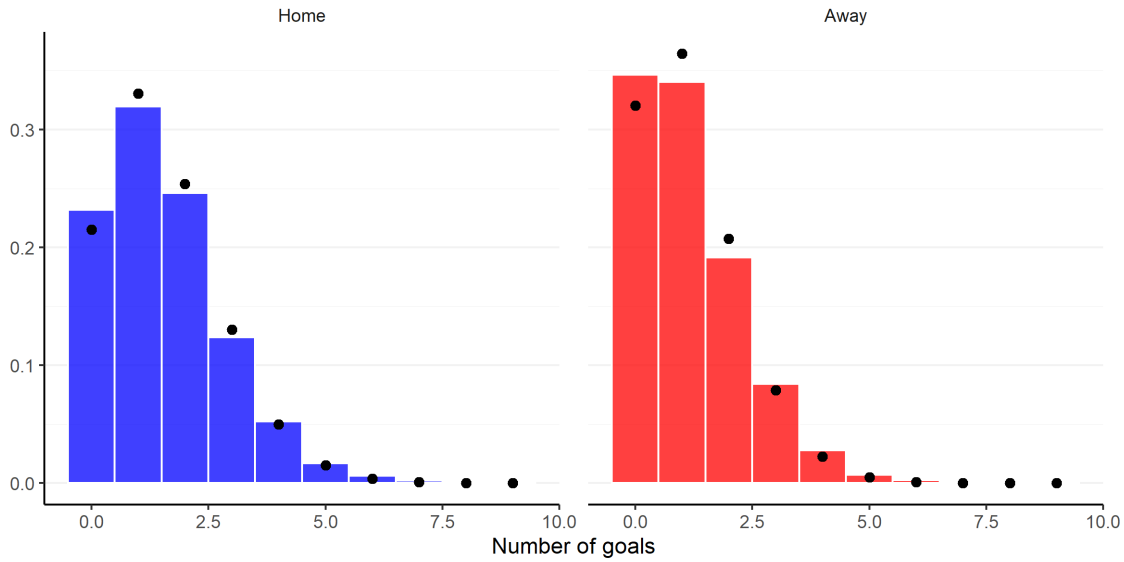


Figure 3.3: Goal histograms with fitted Poisson probability mass function for each side.

	χ^2	df	$\mathbb{P}(>\chi^2)$
Home	51.8994	9	4.722805e-08
Away	55.5388	7	1.165871e-09

Table 3.4: Pearson's chi-squared goodness-of-fit test results for the null hypothesis of Poisson distributed data with parameters equal to the respective sample means.

3.1.2 Weibull Count Model

Recall the Weibull Count Model presented in Section 2.7.2; we now show the fit of the this distribution to the sample data. The Weibull Count Model has two free parameters, the scale parameter α and the shape parameter β, and it encapsulates the Poisson distribution with $\beta = 1$. We should thus expect a better fit from the Weibull Count Model. We note that $\beta < 1$ corresponds to overdispersion and $\beta > 1$ corresponds to underdispersion. We fit the Weibull Count Model to the sample and show the results in Table 3.5 and Figure 3.6.

	Home	Away
α	1.4507183	1.0583299
β	0.9062415	0.8491849

Table 3.5: Maximum likelihood estimates for the parameters of the Weibull Count Model for each side.

From Figure 3.6 we observe a much better fit to the sample data than with the Poisson distribution. This was also expected due to the extra free parameter in the Weibull Count Model. Again, we perform a goodness-of-fit test to obtain concrete statistical evidence; the results of which is shown in Table 3.7. Here, we see that we cannot reject the null hypothesis that the data originates from Weibull Count Models. We also see that the home p-value is fairly low compared to the away p-value, meaning that the away scores seems much more reasonable

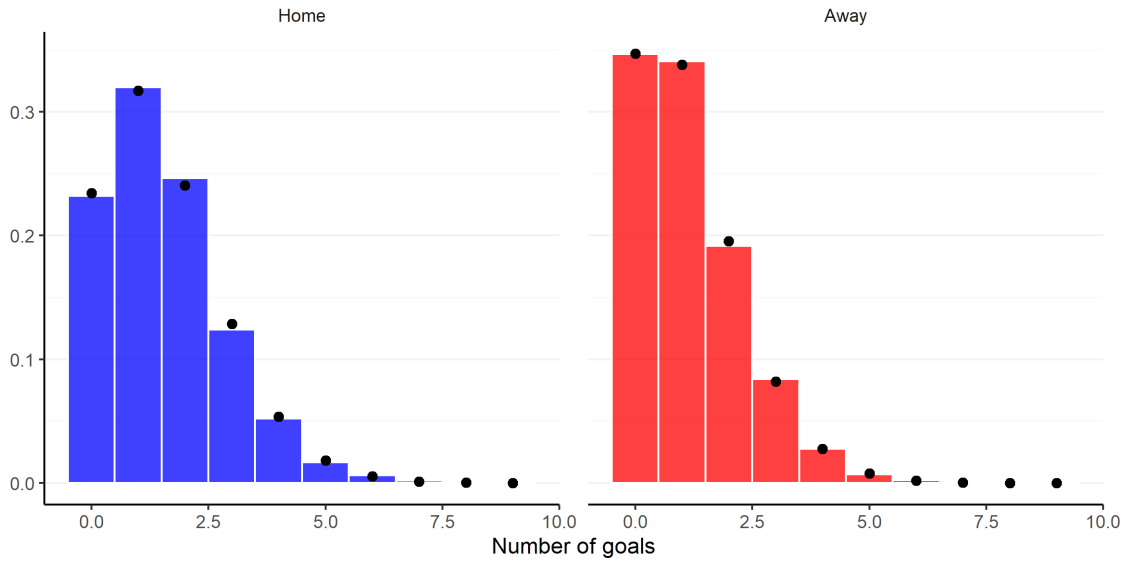


Figure 3.6: Goal histograms with fitted Weibull Count Model probability mass functions for each side.

to stem from the Weibull Count Model than the home scores.

	χ^2	df	$\mathbb{P}(>\chi^2)$
Home	12.6591	8	0.1241
Away	3.2492	6	0.7770

Table 3.7: Pearson's chi-squared goodness-of-fit test results for the null hypothesis of Weibull Count Model distributed data with parameters equal to the respective parameters provided in Table 3.5.

3.1.3 Goal Differences

In addition to presenting the two distributions individually, we also present how they compare to the goal differences. By studying the goal difference, we can effectively get an impression of possible correlation in the scores. If no correlation exists the goal differences should simply exhibit the same patterns as that of the difference of the individual distributions. We first show the empirical distribution of the goal differences by the histogram presented in Figure 3.8.

Let us now see how the empirical goal differences stack up with the proposed distributions. First, we note that the distribution of the difference between two Poisson distributed random variables follows a *Skellam distribution*.[1] The distribution of the difference between two Weibull Count Model distributed random variable have not been specified to the best of the author's knowledge, but we can use Monte Carlo simulation to determine the distribution. We show the histogram with the overlain probability mass functions of two random variables with distributions given as in Section 3.1.1 and Section 3.1.2, respectively, in Figure 3.9.

From Figure 3.9 it seems that both the Skellam distribution and the Weibull Count difference distribution do a decent job of describing the empirical goal difference distribution. The Skellam

[1] Skellam (1946)

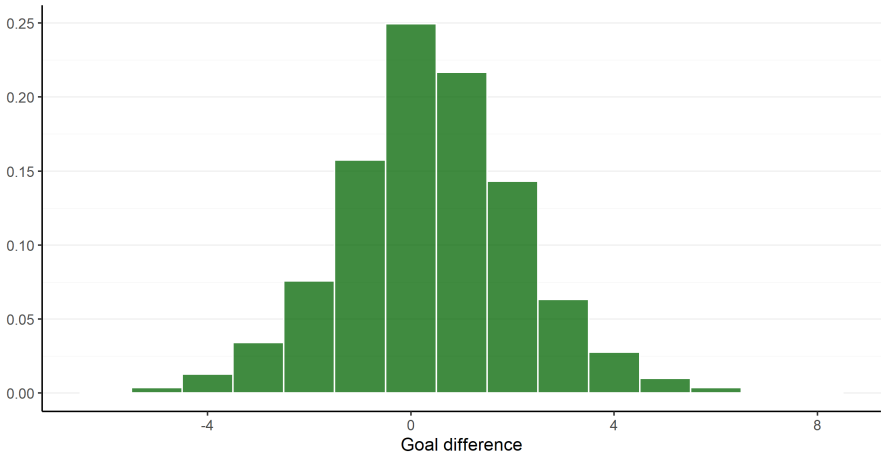


Figure 3.8: Histogram of goal differences in all English Premier League matches between August 2004 and May 2019.

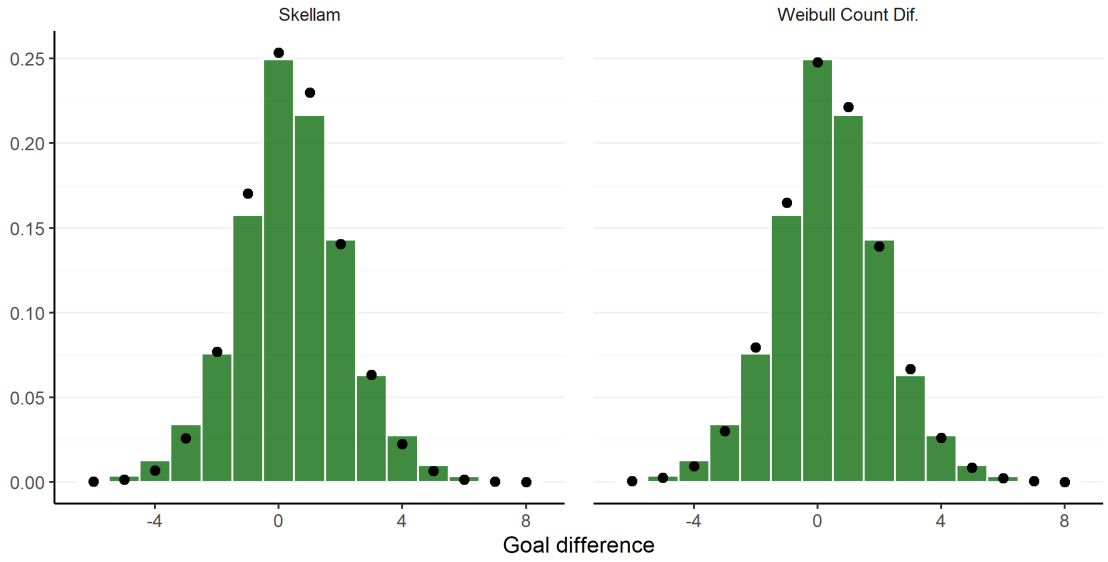


Figure 3.9: Histograms of goal differences with overlain theoretical probability mass functions of the differences of two random variables with Poisson and Weibull Count Model distribution, respectively, and with parameters as in Section 3.1.1 and Section 3.1.2, respectively.

seems a little worse than the Weibull Count difference, however, this the probably connected with the fact that the Skellam only has two free parameters (the original two of the Poisson distributions) and the Weibull Count difference has four free parameters. It does, however, look like the Skellam is slightly more accurate than what appears from Figure 3.3, and contrary, the Weibull Count difference seems slightly worse than what is portrayed in Figure 3.6. This could indicate some general miss-specification of both the Poisson and Weibull Count Model or some possible correlation in the scores. Especially, since the Skellam distribution seems to be a better fit to the goal difference than the Poisson distribution was to the individual score distributions, we can infer that some common factor is probably affecting the teams. Again, we also perform a goodness-of-fit tests to obtain sound statistical results. The results of these tests are shown in Table 3.10; both of which conclude that we should reject the null hypotheses.

	χ^2	df	$\mathbb{P}(>\chi^2)$
Skellam	148.5563	13	$<$ 2.2e-16
Weibull Count Dif.	29.2161	11	0.0021

Table 3.10: Pearson's chi-squared goodness-of-fit test results for the null hypothesis of Skellam distributed data and Weibull Count difference distributed data with parameters equal to the respective parameters provided in Section 3.1.1 and Section 3.1.2, respectively.

It is notable that Figure 3.9 shows no sign of the alleged *draw-inflation* as discussed in e.g. Karlis and Ntzoufras (2003) that states that more draws are generally observed in football than what the models suggest. We actually observe very good fits to the frequency of draws for both models, with the Weibull Count Model suggesting almost a perfect fit of draws, and the Poisson only slightly overestimate draws, i.e. opposite to the draw-inflation hypothesis.

In conclusion, we find that the two distributions are not able to perfectly explain the empirical score distributions, but, visually, they seem to both do fairly decent jobs considering their limitations, e.g. parameter freedom and possible correlation.

3.2 Goal Intensity

In this section, we investigate the empirical goal intensity by analyzing the distribution of goal times and compare this to the theoretical distribution using a specified intensity of the proposed Weibull-based counting processes. To do this analysis, we gathered the minute of each goal in all English Premier Leagues matches between August 2004 and May 2019. In Figure 3.11 we show a histogram of these observations in which each bins constitutes three minutes, i.e. a binwidth of 3. We also display an overlain *kernel density estimation*[2] of the data to get an indication of the empirical density. Since, we do not have records of exact game length for all these matches, we have decided to remove all goals scored in the first half's stoppage time and all goals scored after the 93:00 minute-mark in this analysis in order to obtain fairly consistent data. To justify this decision, we note that stoppage time in the first half tends to be fairly limited in most football games and that almost all matches tend to have at least three minutes of added time in the second half.

Figure 3.11 shows that the empirical distribution of goal times for the home and away sides are very similar with an increasing tendency throughout the game. We will in the following examine if the observed goal time densities are consistent with the maximum likelihood estimates of the parameters of their respective score distributions, i.e. we want to evaluate the counting processes leading to the distribution presented in the last section, specifically, we want to examine the intensities of these counting process in relation to the observed goal times.

[2]See e.g. Hastie, Tibshirani, and Friedman (2009, pp. 208–210).

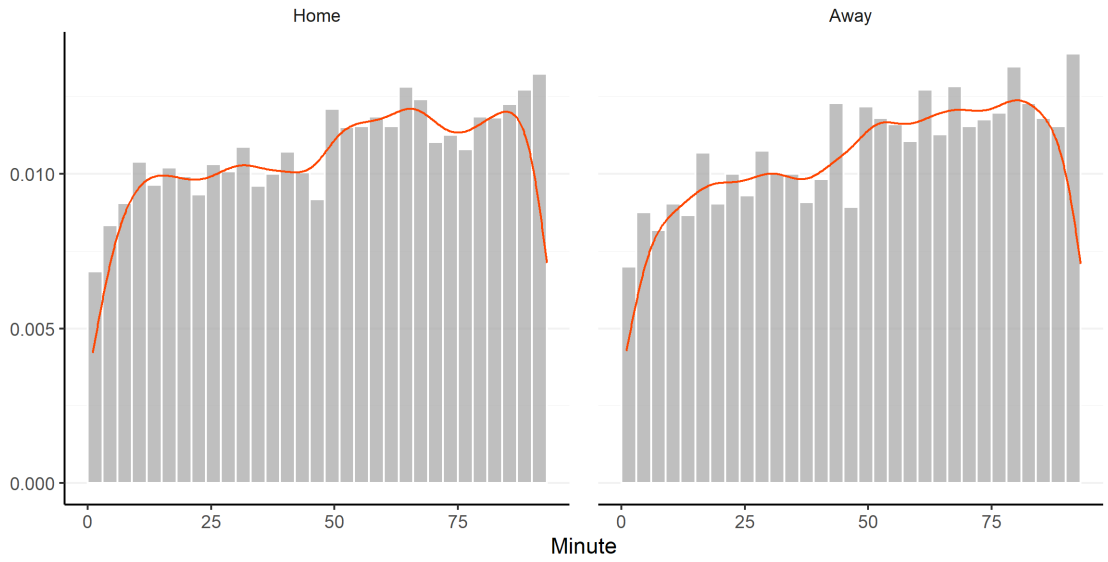


Figure 3.11: Histograms of goal times for each side in all English Premier League matches between August 2004 and May 2019 with a binwidth of 3 and overlain kernel density estimations (orange line).

3.2.1 Weibull Process

We start by looking at the Weibull process. Recall the intensity of the Weibull process given by $\lambda_t = \alpha\beta t^{\beta-1}$. We want to evaluate how well a theoretical goal time distribution from a Weibull process with this intensity can be fitted to the observed goal times, with the added restriction that it must also be consistent with the score distribution presented in Section 3.1.1. In other words, assuming that the end of the game is at $t = T = 1$, we must have that $\alpha\beta 1^{\beta-1} = 1.536491$ for the home team, and $\alpha\beta 1^{\beta-1} = 1.138246$ for the away team, meaning that we essentially only has one free parameter in each case. We can now use these intensities to fit the theoretical goal time distribution to some aggregated level of the actual frequency of goals, e.g. one-minute intervals or three-minute intervals as in the histogram. This presents an optimization problem that can easily be solved, and the results using a root-mean-squared error optimization function and an aggregation level of one minute are presented in Table 3.12 and Figure 3.13.

	Home	Away
α	1.364683	0.991171
β	1.125896	1.148385

Table 3.12: Fitted parameters of the Weibull intensity in the Weibull process for each side for the goal time distribution.

Figure 3.13 shows promising results for the Weibull process. We see that the intensity fit to both the home and away side seems to be quite good. In general, wee see that the Weibull intensity are able to capture the overall tendencies of the empirical goal time densities, and also have parameters that are consistent with the parameters obtained for the score distributions in Section 3.1.1. From the parameter estimates in Table 3.12, we also see that the shape parameter between the two sides are fairly similar, also concurring with our initial observation that the

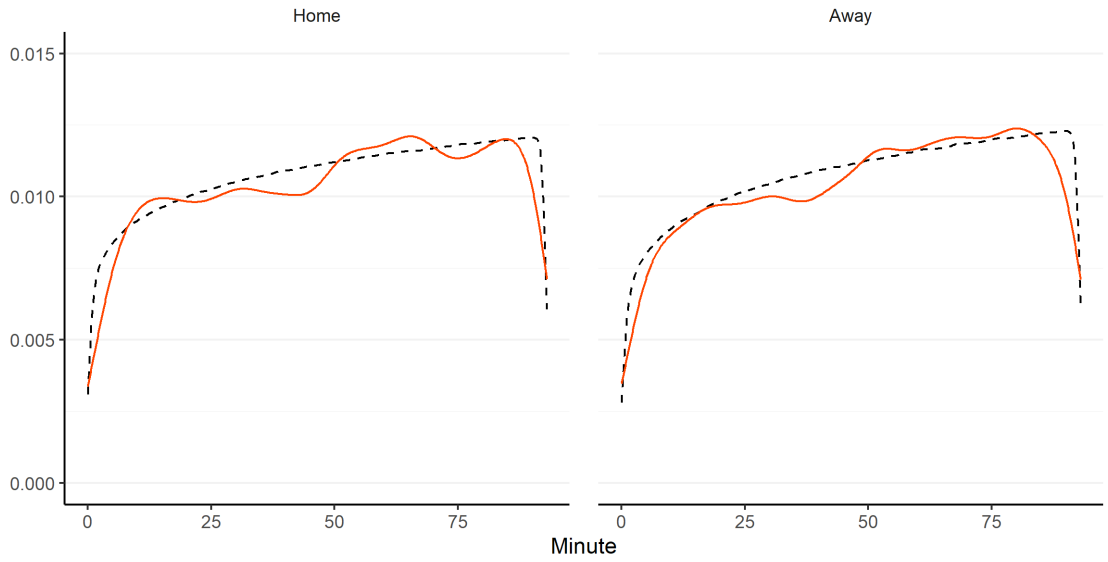


Figure 3.13: Fitted theoretical goal time distributions (dotted line) of mean-valued restricted Weibull process intensities to the empirical goal time distributions (solid orange line) using a root-mean-squared error optimization function.

overall shape of the two empirical densities seems to be similar.

3.2.2 Weibull Renewal Process

Now we examine the goal time distribution arriving from a Weibull renewal process. Again, we want to check if the parameters can be consistent with the score distributions. We note that due to the relationship between the renewal process and the score distribution, the unique parameters of the Weibull renewal process leading to the Weibull Count Model have already been determined, i.e. the only parameters of the Weibull renewal process that are consistent with the Weibull Count Model score distributions in Section 3.1.2 are the maximum likelihood estimates presented in Table 3.5. In Figure 3.14, we show the theoretical goal time densities in comparison with the empirical densities for the Weibull renewal processes of the home and away team, respectively.

From the theoretical goal time distributions in Figure 3.14, we immediately see a problem. The parameters of the Weibull renewal process leading to the score distributions presented in Section 3.1.2 are not consistent with the observed goal time distributions. In fact, we see that the shape parameters of a Weibull renewal process leading to overdispersed score distributions cannot be persistent with an increasing goal time distribution. This is, seemingly, a major drawback of the Weibull renewal process for modeling in-play football. However, the Weibull renewal process can still provide a decent fit to the empirical goal time distribution if we do not impose the score distribution restrictions. As mentioned earlier, we could imagine that a specific game, if played over and over thousands of times, would result in underdispersed score distributions and still have an increasing goal intensity, in such a case the Weibull renewal process seems to be an appropriate modeling choice. That is, we should not exclude the Weibull renewal process,

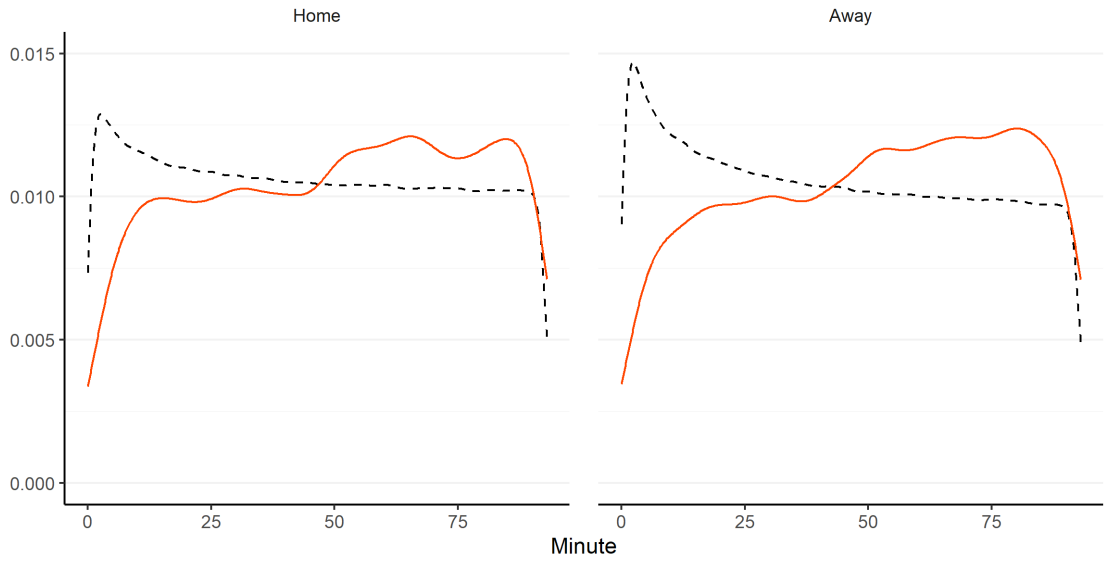


Figure 3.14: Dotted line: Theoretical goal time distributions of Weibull renewal processes consistent with the Weibull Count Model score distributions. Solid orange line: Empirical goal time distributions.

simply because the general patterns of football games are not consistent with Weibull renewal processes.

To further this discussion, consider the shape parameters of the Weibull intensity found in Table 3.12, and imposing the restriction that the mean value must be equal to the sample mean, we can find suitable scale parameters such that the theoretical goal time distributions resemble the empirical goal time distributions, as seen in Table 3.15 and Figure 3.16.

	Home	Away
α	1.659439	1.221129
β	1.125896	1.148385

Table 3.15: Parameters in the Weibull intensity of the Weibull renewal process for each side such that they are consistent with the sample means and the empirical goal time distributions.

In conclusion, our findings show that we might still be able to use the Weibull renewal process for modeling in-play football matches but we should be aware of the limitations in regard to the inconsistencies between the score distributions and the goal intensities. Also, the Weibull process, though limited by the score distribution, provide promising, and score consistent, fits to the empirical goal time distributions.

3.3 Waiting Times of Goals

In this section, we explore the distributions of waiting times for the goals in the English Premier League matches played between August 2004 and May 2019 and compare them to the theoretical distributions of the proposed models. We only focus on the distributions of the first two goals' waiting times.

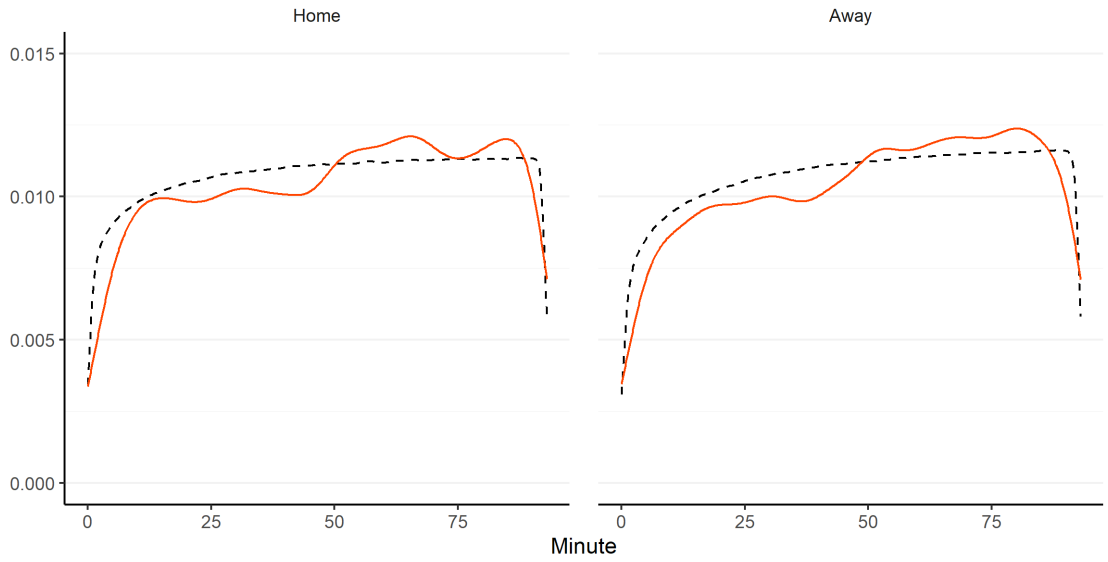


Figure 3.16: Dotted line: Theoretical goal time distributions of Weibull renewal processes with parameters stated in Table 3.15. Solid orange line: Empirical goal time distributions.

We show a histogram of the empirical waiting times of the first goal (in gray) and the waiting times for the second goal (in pink) for the home and away team in Figure 3.17. We also show the kernel density estimation in solid lines.

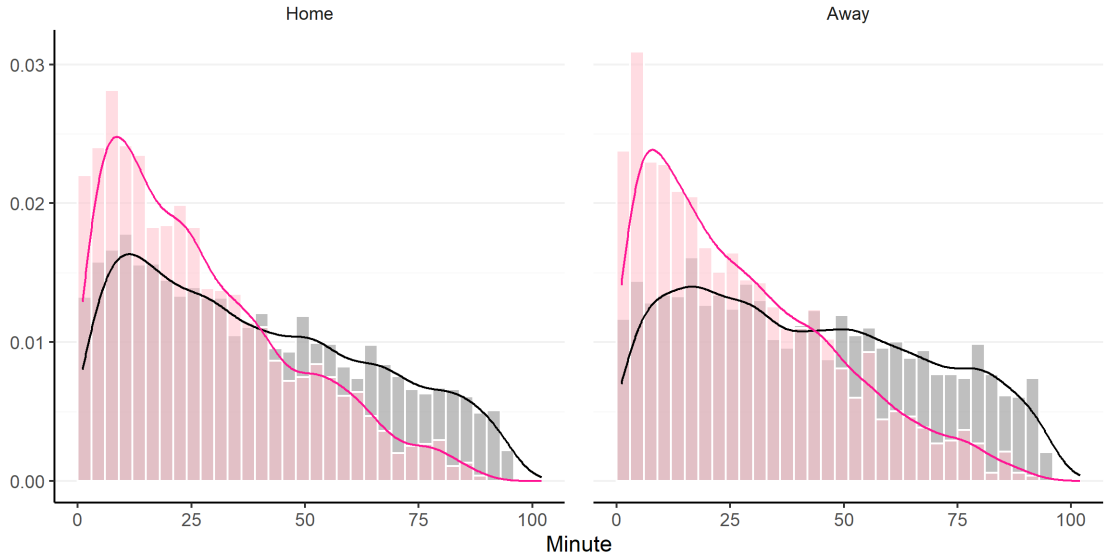


Figure 3.17: Histogram of waiting times of the first goal in gray and waiting times of the second goal in pink for each side in all English Premier League matches played between August 2004 and May 2019. The histograms have a binwidth of 3 and overlain kernel density estimations in black and purple for the waiting times of the first and second goals, respectively.

In Figure 3.17, we see that the distribution of waiting times across teams seems to be fairly similar, with a slight difference in the first goals, where the mass of the density for the home side seems to be a little further left, that is, on average, we observe a faster first goal for the home team. Both sides, however, show signs of a similar pattern for the waiting times of the second goal, with much of the mass centered at the beginning. This is not totally unexpected due to

the time truncation that the end of the football match brings; something which we shall discuss further in the following.

3.3.1 Theoretical Distributions

We are primarily interested in the distribution of goal waiting times such that we may compare the theoretical waiting time distributions of the the Weibull process and the Weibull renewal process with the observed waiting times. The fundamental relationship between counting and point processes, as explained by (2.9), means that we can specify a counting process by specifying the distribution of the waiting times. This has the implication that the theoretical distribution of waiting times of the first goal should be Weibull distributed in both processes, however, the time truncation, mentioned earlier, have an impact on the distribution. This is especially prominent, in the renewal process case, in which the waiting time distribution are i.i.d. if we let it go on forever. However, truncating it will have an effect on the distribution of waiting times. This is also the case with the Weibull process, however, here the theoretical distribution of the second (and onwards) waiting time are exponential distributed. Instead of calculating the truncated theoretical distributions, we simply simulate it. In the following, we show the simulated theoretical distributions in comparison with the observed distributions, where we use the parameters obtained in Table 3.12 and Table 3.15 for the Weibull processes and Weibull renewal processes, respectively.

Weibull Process

In Figure 3.18, we show the comparison between the theoretical density of goal waiting times under the Weibull processes, with parameters given by Table 3.12, and the estimated kernel density of the empirical goal waiting times. From the general patterns, we see many similarities between the theoretical and the empirical densities. On closer inspection, we note that the estimated kernel densities of the waiting times of the first goal (black) both seem very consistent with the general shape of the theoretical densities. There are some differences in the beginning, especially for the home team, and in the end for both teams. The endings are likely because of the simulation of the theoretical densities, where we ended every game at the 93rd-minute mark. Despite, the small inconsistencies in the beginning, the empirical densities of the waiting times to the first goals seem very consistent with the theoretical. The estimated kernel densities of the waiting times of the second goal (purple) also both seem to be very consistent with the simulated theoretical densities. Again, it is noteworthy that the slight inconsistency at the beginning is also present here. The reason for this slight right skew at the beginning is most likely due to the celebratory period right after a goal is scored, that is usually seen in a football game, and the following kickoff, which simply takes time. However, all-in-all, the estimated densities in Figure 3.18 show respectable consistency towards the simulated theoretical densities.

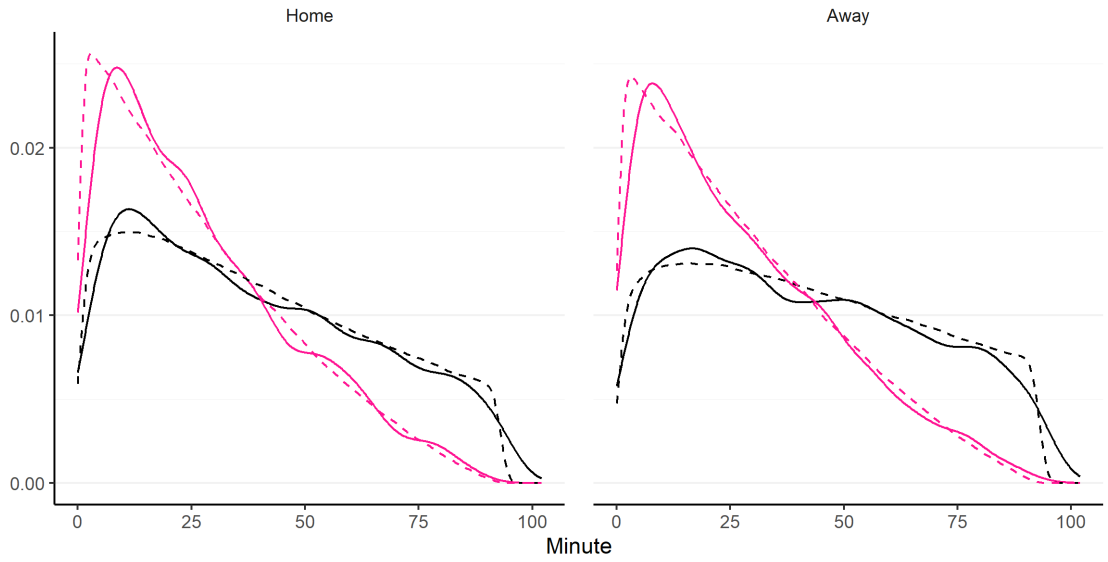


Figure 3.18: Dotted lines: Simulated theoretical densities, for each side, of the waiting times in Weibull processes with parameters given in Table 3.12. The waiting time density of the first goal is shown in black and the second goal in purple. Solid lines: Kernel density estimations of the empirical waiting times of the first (black) and second (purple) goal, respectively, for each side.

Weibull Renewal Process

In Figure 3.19, we show the comparison between the theoretical density of goal waiting times under the Weibull renewal processes, with parameters given by Table 3.15, and the estimated kernel density of the empirical goal waiting times. Again, we see many similarities between the theoretical and the empirical densities. We see that for the waiting times of the first goal, the Weibull renewal processes tend to overestimate the number of fast goals and underestimate the number of late goals. For the waiting times of the second goal, we again see that the general shapes of the simulated theoretical densities are somewhat consistent with the empirical. It is noteworthy that the general spike observed for the waiting times of the second goal, i.e. a fast second goal, is not captured well by the Weibull renewal process. This could indicate that the complete restart of the intensity, associated with a renewal process, is not an appropriate assumption. However, again all estimated densities in Figure 3.19 show reasonable consistency towards the simulated theoretical densities.

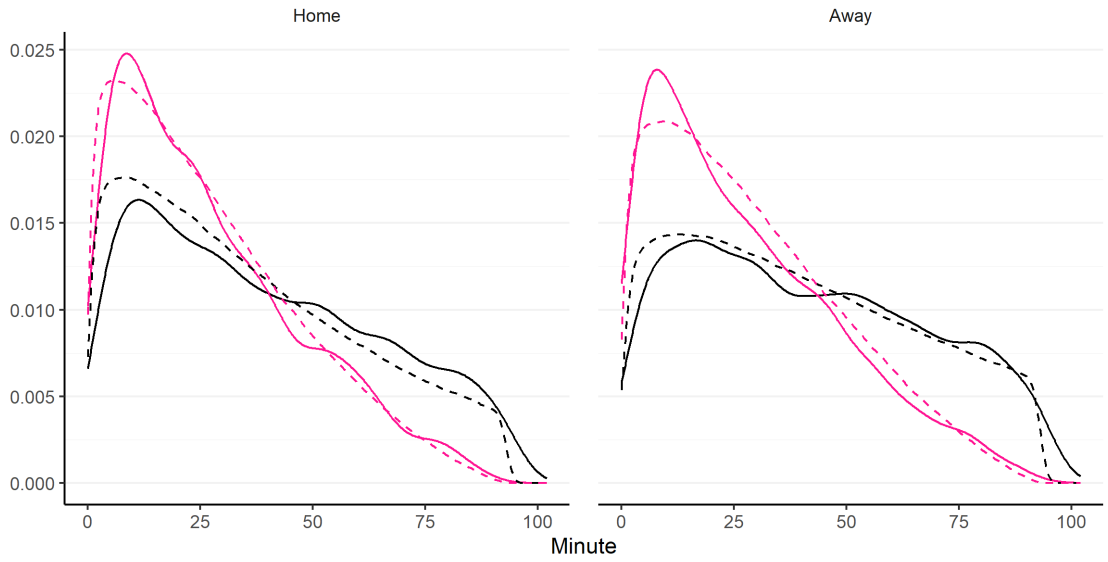


Figure 3.19: Dotted lines: Simulated theoretical densities, for each side, of the waiting times in Weibull renewal processes with parameters given in Table 3.15. The waiting time density of the first goal is shown in black and the second goal in purple. Solid lines: Kernel density estimation of the empirical waiting times of the first (black) and second (purple) goal, respectively, for each side.

From our collective findings, we may conclude that both processes seem to have some useful properties, but also some serious limitations for modeling football; most prominent of which are the distribution of goals for the Weibull process and the renewal assumption embedded in the Weibull renewal process. In the next chapter, we will present a risk-neutral valuation framework for in-play football bets, where we will also show the specific model dynamics in this framework using Weibull processes and Weibull renewal processes.

Risk-neutral Framework 4

In this chapter, we present a risk-neutral valuation framework for in-play football betting, in which we view the bets as financial derivatives on assets related to the goal processes of each team playing. We postulate a general market model for the dynamics of these assets. Section 4.1 presents the construction of the general market model and shows results in relation to this, as well as introduces several specific market models and their dynamics. Section 4.2 introduces a formal mathematical introduction to bets. Section 4.3 covers basic theory in regard to arbitrage and completeness of the market model. Section 4.4 formulates a risk-neutral pricing scheme for bets based on the model dynamics of the specific market models, and Section 4.5 introduces some hedging theory in regards to the risk-neutral framework.

4.1 General Market Model

This section is based on Divos et al. (2018, pp. 321–323), Shreve (2004, p. 228), & Andersen and Maillard (2019, pp. 12–13).

Consider a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ which carries two independent counting processes N^1 and N^2 with intensity processes μ^1 and μ^2, respectively, each equipped with its natural filtrations. Later, in Chapter 6, we will relax the independence assumption. Denote the beginning of the game as $t = 0$ and the end of the game as $t = T$, i.e. we have the parameter space $\mathbb{T} = [0, T]$, then the counting processes depict the number of goals scored by each team, where the superscript 1, corresponds to the home team and superscript 2 to the away team. The probability measure $\mathbb{P}$ is the physical probability measure.

We also assume that a liquid market exists, in which three assets, B, S^1, S^2, can be traded continuously and with no transaction costs or restrictions on short-selling or borrowing. We have that $B = (B_t)_{t\in\mathbb{T}}$ is a risk-free bond that bears no interest, a reasonable assumption due to the short time frame of a football match. $S^1 = (S^1_t)_{t\in\mathbb{T}}$ and $S^2 = (S^2_t)_{t\in\mathbb{T}}$ are assets such that their values at the end of the game are equal to the number of goals scored by the home and away teams, respectively. Let us now formally define a market model based on these descriptions.

Definition 4.1 (General Market Model)
The *general market model* is defined by the following price dynamics of the assets B, S^1, and S^2:

$$\begin{aligned} B_t &= 1 \\ S_t^1 &= N_t^1 + \mathbb{E}\left[L_T \int_t^T \lambda_u^1 du \mid \mathscr{F}_t^{N^1}\right] / L_t \\ S_t^2 &= N_t^2 + \mathbb{E}\left[L_T \int_t^T \lambda_u^2 du \mid \mathscr{F}_t^{N^2}\right] / L_t, \end{aligned} \tag{4.1}$$

where $N^i,\ i \in \{1,2\}$ are independent, nonexplosive counting process with intensity processes $\mu^i,\ i \in \{1,2\}$ and $\lambda^i = (\lambda^i)_{t\in\mathbb{T}},\ i \in \{1,2\}$, are known predictable, locally bounded, and non-negative (stochastic) processes that are μ^i-compatible and L is a known $\mathbb{P}$-martingale.

We see that the underlying process at time T is equal to the number of goals scored by each team, respectively; $S_T^i = N_T^i,\ i \in \{1,2\}$. Even though (4.1) seems rather complicated, it basically states that the assets behave as upwards-shifted compensated counting processes, with the distinction that λ^1, λ^2 are not necessarily equal to the intensity processes μ_1, μ_2, meaning that the assets are not necessarily martingales in the physical measure $\mathbb{P}$. To see this, consider the alternative form of S^i:

$$S_t^i = N_t^i - \Lambda^i(t) + \mathbb{E}\left[\frac{L_T}{L_t}\Lambda^i(T) \mid \mathscr{F}_t^{N^i}\right].$$

When $\lambda^i, i \in \{1,2\}$ are deterministic, we can further simplify the expression due to L being a $\mathbb{P}$-martingale:

$$S_t^i = N_t^i + \int_t^T \lambda_u^i du.$$

Let us now formally define a *risk-neutral probability measure*.

Definition 4.2 (Risk-Neutral Measure)
A probability measure $\mathbb{Q}$ is said to be *risk-neutral* if it satisfies the following two conditions:

(i) $\mathbb{Q}$ and $\mathbb{P}$ are equivalent i.e.

$$\forall A \in \mathscr{F},\ \mathbb{P}(A) = 0 \iff \mathbb{Q}(A) = 0. \tag{4.2}$$

(ii) Under $\mathbb{Q}$, the discounted asset prices are martingales, i.e.

$$\mathbb{E}_{\mathbb{Q}}\left[\widetilde{S}_t^i \mid \mathscr{F}_s\right] = \widetilde{S}_s^i,\ \forall s < t,\ \forall i, \tag{4.3}$$

where $\widetilde{S}_t^i = \exp\left(-\int_0^t (1 - B_u)du\right) S_t^i$.

We are now in a position to present an important result stating the existence of a risk-neutral measure in the general market model.

Proposition 4.3 (Risk-Neutral Measure)**.** *Let B, S^1, and S^2 be given as in Definition 4.1. Then there exist a probability measure $\mathbb{Q}$ such that the following holds:*

(i) Under the $\mathbb{Q}$-measure the goal processes N^1 and N^2 are counting processes with intensity processes λ^1 and λ^2, respectively.

(ii) $\mathbb{Q}$ is equivalent to $\mathbb{P}$.

(iii) The asset processes B, S^1, and S^2 are $\mathbb{Q}$-martingales.

(iv) $\mathbb{Q}$ is unique.

Proof. The proof of (i) relies on Girsanov's theorem for counting processes (Theorem 2.14), which states that N_t^1 and N_t^2 are counting processes with intensity processes λ^1 and λ^2, respectively, under the probability measure $\mathbb{Q}$ which is defined by the Radon-Nikodym derivative

$$\frac{d\mathbb{Q}}{d\mathbb{P}} = L_T, \tag{4.4}$$

with

$$L_T = \prod_{i=1}^{2} \left(\exp\left(\mathrm{M}^i(T) - \Lambda^i(T)\right) \prod_{n:t_n \leq T} \frac{\mu^i_{t_n}}{\lambda^i_{t_n}} \right), \tag{4.5}$$

where $\mathrm{M}^i(T) := \int_0^T \mu_u^i du$ and t_n is the arrival times of the the nth event.

To see this first consider N^1. We apply Girsanov's Theorem to find

$$\lambda_t^1 = \mu_t^1(1 + h_t) \Longrightarrow h_t = \left(\frac{\lambda_t^1}{\mu_t^1} - 1 \right).$$

h_t is obviously greater or equal than -1, since intensities must be non-negative and due to Definition 2.13. Plugging h_t into (2.18), we obtain the following stochastic differential equation:

$$\begin{cases} dL_t &= \left(\mu_t^1 - \lambda_t^1\right) L_{t-} dt + \left(\frac{\lambda_t^1}{\mu_t^1} - 1\right) L_{t-} dN_t^1, \\ L_0 &= 1. \end{cases} \tag{4.6}$$

By denoting $\alpha = (\mu - \lambda)$, $\beta = \left(\frac{\lambda}{\mu} - 1\right)$, and $x_0 = 1$, we see that (4.6), according to Proposition B.2, has the solution

$$\begin{aligned} L_t &= \exp\left(\int_0^t (\mu_s^1 - \lambda_s^1) ds \right) \exp\left(\int_0^t \log\left(\frac{\mu_s^1}{\lambda_s^1} \right) dN_s \right) \\ &= \exp\left(\mathrm{M}^1(t) - \Lambda^1(t)\right) \exp\left(\sum_{s \leq t} \log\left(\frac{\mu_s^1}{\lambda_s^1} \right) \Delta N_s \right) \\ &= \exp\left(\mathrm{M}^1(t) - \Lambda^1(t)\right) \prod_{s \leq t} \exp\left(\log\left(\frac{\mu_s^1}{\lambda_s^1} \right) \Delta N_s \right) \\ &= \exp\left(\mathrm{M}^1(t) - \Lambda^1(t)\right) \prod_{n:t_n \leq t} \frac{\mu^1_{t_n}}{\lambda^1_{t_n}}. \end{aligned}$$

Now (4.5) follows from the multivariate case of Girsanov's theorem as discussed in the remark below Theorem 2.14. To show that $\mathbb{E}[L_T] = 1$ it suffices to show that L is a martingale. A criteria of such is presented in Theorem 2.16 which is fulfilled by the assumptions of the general market model and due to the non-explosiveness of the counting processes. The proof of (ii) follows directly from Lemma 2.15 and the assumptions. To show (iii), i.e. the $\mathbb{Q}$-martingale properties of S^1 and S^2 consider $s \leq t \leq T$:

$$\begin{aligned}
\mathbb{E}_{\mathbb{Q}}\left[S_t^1 \mid \mathscr{F}_s^{N^1}\right] &= \mathbb{E}_{\mathbb{Q}}\left[N_t^1 + \mathbb{E}\left[L_T \int_t^T \lambda_u^1 du \mid \mathscr{F}_t^{N^1}\right] / L_t \mid \mathscr{F}_s^{N^1}\right] \\
&= \mathbb{E}_{\mathbb{Q}}\left[N_t^1 \mid \mathscr{F}_s^{N^1}\right] + \mathbb{E}_{\mathbb{Q}}\left[\mathbb{E}_{\mathbb{Q}}\left[\int_t^T \lambda_u^1 du \mid \mathscr{F}_t^{N^1}\right] \mid \mathscr{F}_s^{N^1}\right] \\
&= \mathbb{E}_{\mathbb{Q}}\left[N_t^1 \mid \mathscr{F}_s^{N^1}\right] + \mathbb{E}_{\mathbb{Q}}\left[\int_t^T \lambda_u^1 du \mid \mathscr{F}_s^{N^1}\right] \\
&= \mathbb{E}_{\mathbb{Q}}\left[\int_0^t \lambda_u^1 du \mid \mathscr{F}_s^{N^1}\right] + \mathbb{E}_{\mathbb{Q}}\left[\int_t^T \lambda_u^1 du \mid \mathscr{F}_s^{N^1}\right] \\
&= \mathbb{E}_{\mathbb{Q}}\left[\int_0^T \lambda_u^1 du \mid \mathscr{F}_s^{N^1}\right] \\
&= \mathbb{E}_{\mathbb{Q}}\left[\int_0^s \lambda_u^1 du \mid \mathscr{F}_s^{N^1}\right] + \mathbb{E}_{\mathbb{Q}}\left[\int_s^T \lambda_u^1 du \mid \mathscr{F}_s^{N^1}\right] \\
&= N_s + \mathbb{E}_{\mathbb{Q}}\left[\int_s^T \lambda_u^1 du \mid \mathscr{F}_s^{N^1}\right] \\
&= S_s^1,
\end{aligned}$$

where the second equality follows from Theorem A.3. An identical calculation can be performed for S^2, proving the $\mathbb{Q}$-martingale properties of S^i, $i \in \{1, 2\}$. Due to the constant nature of B, it is a trivial martingale in every measure. Lastly, the proof of (iv) follows directly from Theorem 2.18, which states that if two measures have the same set of intensities in regard to the natural filtration, then the two measures must coincide. ■

The implication of Proposition 4.3 is that there exists a probability measure $\mathbb{Q}$ such that the price dynamics in (4.1), in this measure, are martingales, and such that $\mathbb{P}$ and $\mathbb{Q}$ are equivalent, meaning that they both agree on sets of zero-probability events. This implies that $\mathbb{Q}$ is a risk-neutral probability measure; the implication of which being that the value of a bet at any time $t \leq T$ has a precise formulation, as we shall see shortly. Furthermore, due to the choice of the natural filtration, we also have that $\mathbb{Q}$ is uniquely determined.

4.1.1 Model Dynamics

We now briefly present some specific model dynamics for when the counting processes in Definition 4.1 has distinct forms. We restrict ourselves to the cases of Weibull-based counting processes described in Section 2.7, as well as the homogeneous Poisson process for explanatory purposes.

Poisson Process

For the sake of completeness, we start by showing the original work of Divos et al., i.e., the price dynamics of the market model when the counting processes $N^i,\ i \in \{1,2\}$ are homogeneous Poisson processes with intensities $\mu^i,\ i \in \{1,2\}$ and where $\lambda^i \in \mathbb{R}_+,\ i \in \{1,2\}$ are known. The price dynamics of this market model is then given by:

$$\begin{aligned} B_t &= 1 \\ S_t^1 &= N_t^1 + \lambda^1(T-t) \\ S_t^2 &= N_t^2 + \lambda^2(T-t). \end{aligned} \tag{4.7}$$

In Figure 4.1, we show a simulation of the model dynamics under this market model, with model parameters given by:

$$\begin{aligned} \mu^1 &= \lambda^1 = 1.4 \\ \mu^2 &= \lambda^2 = 1.1. \end{aligned}$$

In this scenario, we assume, for simplicity, that the counting processes' intensities and the known constants λ^i are equal, that is, similar to the situation under the $\mathbb{Q}$ measure. In Figure 4.1a, we show the sample paths of the assets $S^i,\ i \in \{1,2\}$ in connection with the sample paths of the score processes. From the score processes' sample paths, we see that this particular simulated game ended 1-0 in the home team's favor. In Figure 4.1b, we show the intensity processes of the score processes. As per definition, we see that the intensity processes of the Poisson processes are constants.

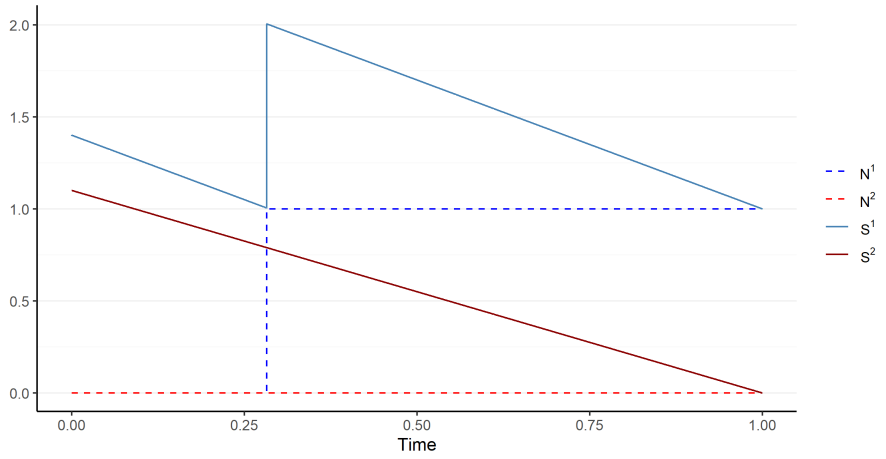
Weibull Process

Next, we show the price dynamics when the counting processes are Weibull processes with intensity processes characterized by $\mu_t^i = \widetilde{\alpha}^i \widetilde{\beta}^i t^{\widetilde{\beta}^i - 1}$ and where $\alpha^i, \beta^i \in \mathbb{R}_+,\ i \in \{1,2\}$ are known. This approach is somewhat similar to the extended market model proposed by Andersen and Maillard, in which they suggest price dynamics based on a specific form of the inhomogeneous Poisson process. The price dynamics of the market model with Weibull process model dynamics are given by:

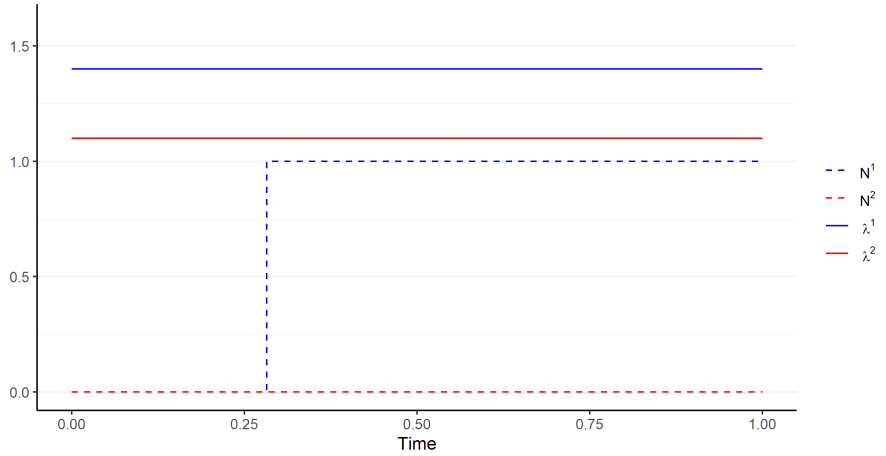
$$\begin{aligned} B_t &= 1 \\ S_t^1 &= N_t^1 + \alpha^1 \left(T^{\beta^1} - t^{\beta^1} \right) \\ S_t^2 &= N_t^2 + \alpha^2 \left(T^{\beta^2} - t^{\beta^2} \right), \end{aligned} \tag{4.8}$$

In Figure 4.2, we show a simulation of the model dynamics under the market model (4.8), with Weibull process model parameters given as follows:

$$\begin{aligned} \widetilde{\alpha}^1 &= \alpha^1 = 1.4 \\ \widetilde{\beta}^1 &= \beta^1 = 1.2 \\ \widetilde{\alpha}^2 &= \alpha^2 = 1.1 \end{aligned}$$



(a) Dotted lines are the score (counting) processes and the solid lines are the assets S^i.



(b) Dotted lines are the score (counting) processes and the solid lines are the intensity processes of the counting process.

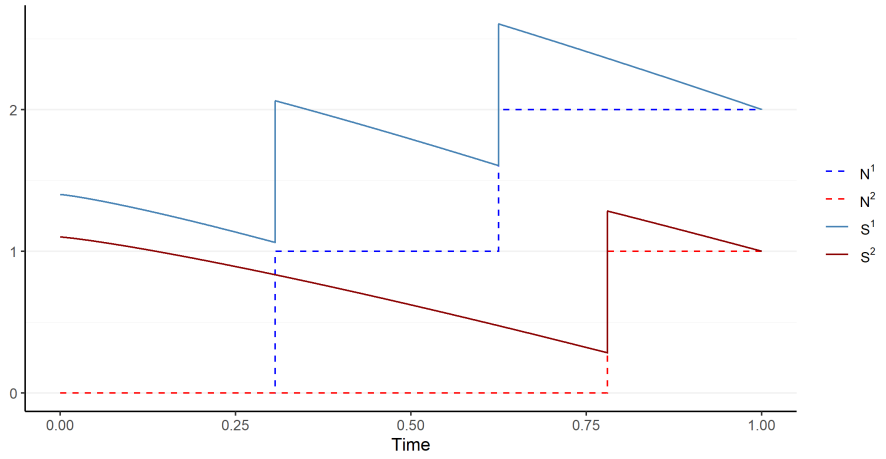
Figure 4.1: Simulated sample path of the Poisson model dynamics with parameters $\mu^1 = \lambda^1 = 1.4$ and $\mu^2 = \lambda^2 = 1.1$. Blue represents the home team and red the away team.

$$\widetilde{\beta}^2 = \beta^2 = 1.2.$$

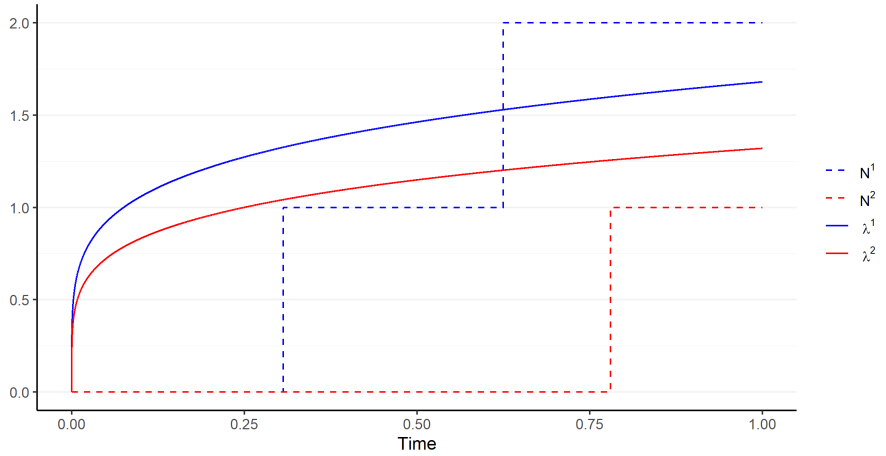
Again, for simplicity, we assume that the parameters of the intensities and the known constants are equal. This particular simulation resulted in a 2-1 win for the home team. When compared to the Poisson model dynamics, we here see a slight curvature of the assets' sample paths in Figure 4.2a, representing the increasing intensities of the score processes, i.e. the increasing likelihood of scoring a goal throughout the game, as seen in Figure 4.2b, where we also see the deterministic and increasing intensities throughout the game.

Weibull Renewal Process

Lastly, we show the price dynamics when the counting processes are Weibull renewal processes with intensity processes characterized by $\mu_t^i = \widetilde{\alpha}^i \widetilde{\beta}^i \left(Z_t^i\right)^{\widetilde{\beta}^i - 1}$ and where $\alpha^i, \beta^i \in \mathbb{R}_+,\ i \in \{1, 2\}$ are known. The price dynamics of this market model are given by:



(a) Dotted lines are the score (counting) processes and the solid lines are the assets S^i.



(b) Dotted lines are the score (counting) processes and the solid lines are the intensity processes of the counting process.

Figure 4.2: Simulated sample paths of the Weibull process model dynamics with parameters $\widetilde{\alpha}^1 = \alpha^1 = 1.4$, $\widetilde{\beta}^1 = \beta^1 = 1.2$, $\widetilde{\alpha}^2 = \alpha^2 = 1.1$, and $\widetilde{\beta}^2 = \beta^2 = 1.2$. Blue represents the home team and red the away team.

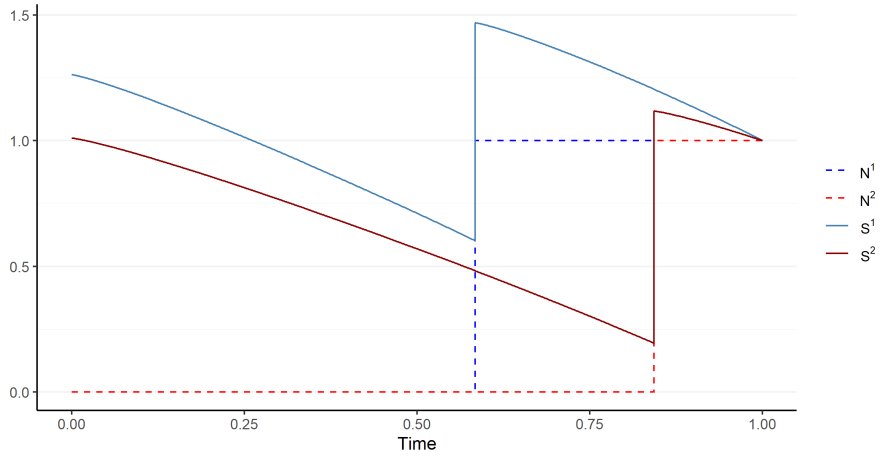
$$\begin{aligned} B_t &= 1 \\ S_t^1 &= N_t^1 + \alpha^1 \beta^1 \, \mathbb{E}\left[\int_t^T (Z_u^1)^{\beta^1 - 1} du \mid \mathscr{F}_t^{N^1} \right] \\ S_t^2 &= N_t^2 + \alpha^2 \beta^2 \, \mathbb{E}\left[\int_t^T (Z_u^2)^{\beta^2 - 1} du \mid \mathscr{F}_t^{N^2} \right], \end{aligned} \tag{4.9}$$

We recall that while the Poisson-based model dynamics of the previous two models have deterministic compensators and intensities, the market dynamics given by (4.9) are characterized by being stochastic, more specific they depend on the time since the last goal (or start of the game). In Figure 4.3, we present a simulation of a sample path of such model dynamics with parameters given by:

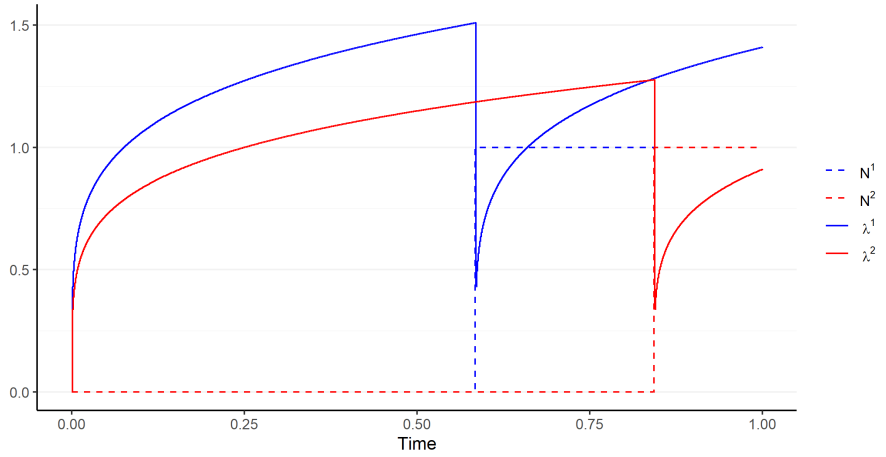
$$\begin{aligned} \widetilde{\alpha}^1 &= \alpha^1 = 1.4 \\ \widetilde{\beta}^1 &= \beta^1 = 1.2 \\ \widetilde{\alpha}^2 &= \alpha^2 = 1.1 \end{aligned}$$

$$\widetilde{\beta}^2 = \beta^2 = 1.2.$$

In Figure 4.3a we show the assets S^i, $i \in \{1, 2\}$ in relation to the sample paths of the score processes, and in Figure 4.2b, we portray the intensity processes of the simulated Weibull renewal processes. This simulated game ended in a draw with the scores 1-1.



(a) Dotted lines are the score (counting) processes and the solid lines are the assets S^i.



(b) Dotted lines are the score (counting) processes and the solid lines are the intensity processes of the counting process.

Figure 4.3: Simulated sample paths of the Weibull renewal process model dynamics with parameters $\widetilde{\alpha}^1 = \alpha^1 = 1.4$, $\widetilde{\beta}^1 = \beta^1 = 1.2$, $\widetilde{\alpha}^2 = \alpha^2 = 1.1$, and $\widetilde{\beta}^2 = \beta^2 = 1.2$. Blue represents the home team and red the away team.

4.2 Bets

This section is based on Divos et al. (2018, pp. 317–327), Björk (2009, p. 94), Tankov and Cont (2004, pp. 293–294), & Andersen and Maillard (2019, pp. 12–14).

In this section we present a mathematical representation of football bets, namely as financial derivatives. In classical finance, the distinction between an asset and a derivative is usually clear. However, in a football betting framework, such a distinction is not as clear. This is due to the fact that (almost) all bets are made on the scores, and the score process, i.e. assets S in (4.1),

are not tradable in itself, thus only the derivatives are actually tradable. One could suspect that this poses a problem in regards to the fundamental theorems of asset pricing. However, as we shall show later, the asset processes can be statically replicated by the so-called correct score bets, which are traded in practice. Furthermore, as we shall see in Proposition 4.15, any two linearly independent bets can be used as hedging instruments. Thus, according to Divos et al., making the choice of the assets S irrelevant in practice and only serves a technical purpose.

In the following, we will use the bivariate natural filtration notation of $\mathscr{G}_t = \bigvee_{i=1}^{2} \mathscr{F}_t^{N^i}$ used extensively in Brémaud (1981), as it has beneficial simplicity when stating the filtrations of the assets collectively. Here, the $\bigvee$ notation is the join as presented in (2.2). Let us now move on to the formal definition of a bet.

Definition 4.4 (Bet)
A *bet* (also known as a *contingent claim* or *derivative*) is a $\mathscr{G}_T$-measurable random variable $\mathcal{X}$. A bet is called a *simple bet* if it depends only on the final number of goals S_T^1, S_T^2, i.e. of the form

$$\mathcal{X} = \Phi\left(S_T^1, S_T^2\right), \tag{4.10}$$

where $\Phi : \mathbb{N} \times \mathbb{N} \to \mathbb{R}$ is a known scalar function, which is referred to as the *payoff function*.

Definition 4.4 states that a bet is a contract which gives the holder $\mathcal{X}$ at the time of maturity, and where the value of a bet is revealed at the maturity, thus complying with our general intuition of a bet. The issue is now to find a way to place a value of a bet for any given time $0 \leq t < T$. For this, we will use the notion of a *pricing rule* of the bet $\mathcal{X}$, denoted by $\Pi_t(\mathcal{X})$.

There are, however, some minimal conditions which $\Pi_t(\mathcal{X})$ must fulfill to qualify as a pricing rule. We must be able to calculate the value $\Pi_t(\mathcal{X})$ using the information at hand at time t if it is going to be useful, i.e. $\Pi_t(\mathcal{X})$ must be an $\mathbf{G}$-adapted process. Furthermore, we also require positiveness, i.e. a bet with non-negative payoff must also have a non-negative value:

$$\forall \omega \in \Omega,\ \mathcal{X}(\omega) \geq 0 \Longrightarrow \forall t \in [0, T],\ \ \Pi_t(\mathcal{X}) \geq 0. \tag{4.11}$$

A third condition is linearity; the value of a portfolio of bets is given by the sum of the values of its components:

$$\Pi_t\left(\sum_{j=1}^{J} \mathcal{X}_j\right) = \sum_{j=1}^{J} \Pi_t(\mathcal{X}_j). \tag{4.12}$$

All the above conditions of the pricing rule are generally upheld at a betting exchange, meaning that since all of the above conditions are fulfilled, we can use the pricing rule to place a value on a bet as we shall see in Section 4.4.

Market Value of Bets

The *market value of a bet*, also known as the *market quote*, is the price at which the bet can be bought or sold in the physical market (exchange or bookmaker) at any given time assuming that the bet pays a fixed amount of 1 unit in case it wins, and zero otherwise:

$$\mathcal{X} = \begin{cases} 1 & \text{Bet wins,} \\ 0 & \text{Otherwise.} \end{cases} \tag{4.13}$$

The market value of a bet is thus given by the reciprocal of the decimal odds, or more formally:

$$\Pi_t^{\text{Market}}(\mathcal{X}) = \frac{1}{\text{Odds}_t}. \tag{4.14}$$

4.2.1 Types of Bets

A widespread type of a simple bet $\Phi(S_T^1, S_T^2)$ is the bet that pays out 1 unit if the home team wins and zero otherwise. Analogous bets exist for an away win and a draw. Such bets are known as *match odds bet*, and these types of bets can be formally defined by:

$$\Phi_{\text{H}}(S_T^1, S_T^2) = \mathbb{1}\left(S_T^1 > S_T^2\right) \tag{4.15}$$

$$\Phi_{\text{A}}(S_T^1, S_T^2) = \mathbb{1}\left(S_T^1 < S_T^2\right) \tag{4.16}$$

$$\Phi_{\text{D}}(S_T^1, S_T^2) = \mathbb{1}\left(S_T^1 = S_T^2\right). \tag{4.17}$$

Other common types of simple bets which we will use are defined by:

$$\Phi_{\text{O}}(S_T^1, S_T^2) = \mathbb{1}\left(S_T^1 + S_T^2 > K\right), \ K \in \{0.5, 1.5, 2.5, \ldots\} \tag{4.18}$$

$$\Phi_{\text{U}}(S_T^1, S_T^2) = \mathbb{1}\left(S_T^1 + S_T^2 < K\right), \ K \in \{0.5, 1.5, 2.5, \ldots\} \tag{4.19}$$

$$\Phi_{\text{CS}}(S_T^1, S_T^2) = \mathbb{1}\left(S_T^1 = K_1, S_T^2 = K_2\right), \ K_1, K_2 \in \mathbb{N}_0. \tag{4.20}$$

Bets of the type (4.18)–(4.19) are known as *over/under bets*, and bets of the type (4.20) are known as *correct score bets*.

4.3 Arbitrage & Completeness

This section is based on Divos et al. (2018, pp. 321–323), Björk (2011, pp. 61–63), Andersen and Maillard (2019, p. 15), & Tankov and Cont (2004, pp. 296, 299–300).

In order to arrive at a general result on the risk-neutral pricing of a bet, we first introduce some advantageous definitions, notation, and results.

Definition 4.5 (Portfolio)
A *portfolio* is an $\mathscr{F}_t$-predictable vector process ϕ characterized by $\phi_t = \left(\phi_t^0, \phi_t^2, \phi_t^2\right)$ that satisfies $\int_0^t \left|\phi_s^i\right| ds < \infty$ for $i \in \{0, 1, 2\}$. The associated *value process* V^ϕ is characterized by

$$V_t^\phi = \phi_t^0 B_t + \phi_t^1 S_t^1 + \phi_t^2 S_t^2. \tag{4.21}$$

The portfolio is *self-financing* if

$$V_t^\phi = V_0^\phi + \int_0^t \phi_u^1 dS_u^1 + \int_0^t \phi_u^2 dS_u^2, \tag{4.22}$$

where $\int_0^t \phi_u^i dS_u^i,\ i \in \{1, 2\}$ is a Lebesgue-Stieltjes integral, cf. Appendix B.

In broad terms, a portfolio is self-financing if there is no outside infusion or withdrawal of money, that is, the purchase of a new asset must be paid for by the sale of an existing one. We can also state a useful result in regard to the integral of a portfolio with respect to a martingale when the portfolio is self-financing.

Proposition 4.6. *Let the asset processes $S^i,\ i \in \{1, 2\}$ be martingales, and let the portfolio ϕ be self-financing, then V^ϕ is a martingale.*

Proof. Follows directly from Proposition B.1. ■

Definition 4.7 (Arbitrage-free)
A portfolio is *arbitrage-free* if no self-financing portfolio ϕ exist such that

$$\mathbb{P}\Big(\forall t \in [0, T], V_t^\phi \geq 0\Big) = 1 \quad \text{and} \quad \mathbb{P}\Big(V_t^\phi > V_0^\phi\Big) > 0.$$

The general intuition of arbitrage is that you can make money without taking risks, or in other words, you are guaranteed not to lose money and have a positive probability of making money, which is what the two probabilities in Definition 4.7 signifies. In the business, arbitrage also goes by the name "free lunch".

Definition 4.8 (Completeness)
A market model is said to be *complete* if for every bet $\mathcal{X}$ there exists a self-financing portfolio ϕ such that $\mathcal{X} = V_T^\phi$. In this case, we say that the bet $\mathcal{X}$ is replicated by the portfolio ϕ.

Theorem 4.9. *The market model* (4.1) *is complete and arbitrage-free.*

Proof. This follows directly from Proposition 4.3 and the First and Second Fundamental Theorems of Asset Pricing, cf. Theorem B.3 & B.4. To be more specific; the First Fundamental Theorem of Asset Pricing states that the existence of a risk-neutral measure implies arbitrage-freeness, and the Second Fundamental Theorems of Asset Pricing states that a market model is complete if the risk-neutral measure is unique. ■

4.4 Risk-neutral Pricing

This section is based on Divos et al. (2018, pp. 323–325, 334), Andersen and Maillard (2019, pp. 16–17), & Tankov and Cont (2004, pp. 293–298).

As we mentioned in the paragraph below Definition 4.4, our aim is to find a "fair" pricing rule $\Pi_t(\mathcal{X})$ for the bet $\mathcal{X}$. Theorem 4.9 shows that the market generated by (4.1) is arbitrage-free, which yields the obvious result; we must have the following relation to avoid an arbitrage opportunity at the time of maturity:

$$\Pi_T(\mathcal{X}) = \mathcal{X}, \tag{4.23}$$

It should then be clear that Theorem 4.9 restricts the behavior of the pricing rule $\Pi_t(\mathcal{X})$. Furthermore, remember that we are in a setting where the filtrations used are the natural filtrations, thus $\mathscr{F}_0^N$ contains no information; now, since $\mathbb{Q}$ is a risk-neutral measure and $B_t = 1$, we must have that the only fair price at time 0 has to be given by:

$$\Pi_0(\mathcal{X}) = \mathbb{E}_{\mathbb{Q}}\left[\mathcal{X}\right]. \tag{4.24}$$

Lastly, if the pricing rule Π is not time consistent, i.e. the value at $t = 0$ of the potential payoff $\mathcal{X}$ at T is the same as the value at $t = 0$ of the payoff $\Pi_t(\mathcal{X})$ at t, then an arbitrage opportunity may arise. Thus, we have that Π should also be time consistent. Therefore, we see that, as a consequence of Theorem 4.9, the value of a bet at any time $0 \leq t \leq T$ is given in the following corollary.

Corollary 4.10 (Risk-neutral Valuation)**.** *The value of a bet at time t is equal to the risk-neutral expectation of its value at the end of the game, i.e.*

$$\Pi_t(\mathcal{X}) = \mathbb{E}_{\mathbb{Q}}\left[\mathcal{X} \mid \mathscr{G}_t\right]. \tag{4.25}$$

Proof. Follows directly from Theorem 4.9 and the above discussion. See e.g. Tankov and Cont (2004, p. 298). ■

4.4.1 Pricing Formulas

We are now in a position to present explicit pricing formulas for simple bets in some specific market models. We first state the pricing formula when in a homogeneous Poisson setting, i.e. the setting originally proposed by Divos et al.

Proposition 4.11 (Pricing Formula of a Simple Bet - Poisson Setting)**.** *Assume we are in a homogeneous Poisson process setting of the general market model. The value of a simple bet at time t with payoff function Φ is given by*

$$\Pi_t(\mathcal{X}) = \sum_{n_1=N_t^1}^{\infty} \sum_{n_2=N_t^2}^{\infty} \Phi(n_1, n_2) P\left(n_1 - N_t^1, \lambda^1(T-t)\right) P\left(n_2 - N_t^2, \lambda^2(T-t)\right), \tag{4.26}$$

where $P(N, \Lambda)$ is the Poisson probability mass function.

Proof. Follows from Corollary 4.10 and the definition of a simple bet, cf. (4.10):

$$\begin{aligned}
\Pi_t(\mathcal{X}) &= \mathbb{E}_{\mathbb{Q}}\left[\mathcal{X} \mid \mathscr{G}_t\right] \\
&= \mathbb{E}_{\mathbb{Q}}\left[\Phi\left(N_T^1, N_T^2\right) \mid \mathscr{G}_t\right] \\
&= \mathbb{E}_{\mathbb{Q}}\left[\Phi\left(N_t^1 + N_{T-t}^1, N_t^2 + N_{T-t}^2\right) \mid \mathscr{G}_t\right] \\
&= \mathbb{E}_{\mathbb{Q}}\left[\Phi\left(N_t^1 + N_{T-t}^1, N_t^2 + N_{T-t}^2\right)\right] \\
&= \sum_{n_1=N_t^1}^{\infty} \sum_{n_2=N_t^2}^{\infty} \Phi(n_1, n_2) \mathbb{Q}\left(N_{T-t}^1 = n_1 - N_t^1, N_{T-t}^2 = n_2 - N_t^2\right),
\end{aligned}$$

where the last equality follows from the law of the unconscious statistician (Proposition A.1). The result now follows due to the independence of the goal processes N^i. ■

Next, we state the pricing formula when in an inhomogeneous Poisson setting of the general market model.

Proposition 4.12 (Pricing Formula of a Simple Bet - Inhomogeneous Poisson Setting)**.** *Assume we are in an inhomogeneous Poisson process setting of the general market model. The value of a simple bet at time t with payoff function Φ is given by*

$$\Pi_t(\mathcal{X}) = \sum_{n_1=N_t^1}^{\infty} \sum_{n_2=N_t^2}^{\infty} \Phi(n_1, n_2) P\left(n_1 - N_t^1, \lambda_T^1 - \lambda_t^1\right) P\left(n_2 - N_t^2, \lambda_T^2 - \lambda_t^2\right), \tag{4.27}$$

where $P(N, \Lambda)$ is the Poisson probability mass function.

Proof. The proof is almost identical to the proof of Proposition 4.11 ■

As a consequence of Proposition 4.12, we have that the pricing formula of a simple bet, in the market model in which the goal processes are Weibull processes, i.e. the market with model dynamics given by (4.8), can be stated as follows:

$$\begin{aligned}
\Pi_t(\mathcal{X}) = \sum_{n_1=N_t^1}^{\infty} \sum_{n_2=N_t^2}^{\infty} \Phi(n_1, n_2) P\left(n_1 - N_t^1, \alpha^1\left(T^{\beta^1} - t^{\beta^1}\right)\right) \times \\
P\left(n_2 - N_t^2, \alpha^2\left(T^{\beta^2} - t^{\beta^2}\right)\right).
\end{aligned} \tag{4.28}$$

From the fact that the intensity of the Weibull renewal process is stochastic, an explicit pricing formula in such a setting is not available at the moment, to the best of our knowledge. We can however exploit the fact that we have assumed the natural filtration, in which the risk-neutral expectation at time 0 has no prior information i.e. (4.24), meaning that we can state an explicit pricing formula for simple bets at this time; namely by utilizing the explicit equation for the Weibull Count Model, as presented in (2.39):

$$\Pi_0(\mathcal{X}) = \sum_{n_1=0}^{\infty} \sum_{n_2=0}^{\infty} \Phi(n_1, n_2) P\left(n_1, \alpha^1, \beta^1\right) P\left(n_2, \alpha^2, \beta^2\right), \tag{4.29}$$

where $P(n, A, B)$ is the Weibull Count Model for a specific n and $t = 0$.

Whenever we do not have an explicit pricing formula at hand, we can use (4.25) and deploy a Monte Carlo simulation to find a decent approximation of the price.

4.5 Hedging

This section is based on Divos et al. (2018, pp. 323–327), Tankov and Cont (2004, pp. 293–298), & Andersen and Maillard (2019, pp. 17–19).

Given the replication concept of a bet arising from the completeness of the market model, we have a second approach to the pricing of a bet. Let us start by formally stating it as a direct result of Theorem 4.9.

Corollary 4.13. *The value of a bet at time t is equal to the value of the associated self-financing portfolio ϕ at time t, formally:*

$$\Pi_t(\mathcal{X}) = V_t^\phi = V_0^\phi + \int_0^t \phi_s^1 dS_s^1 + \int_0^t \phi_s^2 dS_s^2. \tag{4.30}$$

Proof. From market completeness, we have that $\mathcal{X} = V_T^\phi$, thus we have

$$\Pi_t(\mathcal{X}) = \Pi_t(V_T^\phi) = \mathbb{E}_{\mathbb{Q}}\left[V_T^\phi \mid \mathscr{G}_t\right].$$

All we have to show now is that V^ϕ is a $\mathbb{Q}$-martingale, which follows from Proposition 4.3 and Proposition 4.6. ■

Corollary 4.13 states that holding the bet and holding the replicating portfolio is equivalent from a financial point of view, i.e. that the bet's value can be perfectly replicated by the self-financing portfolio, or in other words; the bet can be perfectly hedged.

4.5.1 Replication

We are now in a position to formally state the proposition introduced in the discussion in Section 4.2; namely, the replication of a bet from any two linearly independent bets. First, we need to formally define linear independence of bets.

Definition 4.14 (Linear Independence)
Two bets $\mathcal{Z}_1$ and $\mathcal{Z}_2$ are *linearly independent* if the self-financing portfolio $\phi^1 = \left(\phi^{10}, \phi^{11}, \phi^{12}\right)$ that replicates $\mathcal{Z}_1$ is $\mathbb{P}$-almost surely linearly independent from the self-financing portfolio $\phi^2 = \left(\phi^{20}, \phi^{21}, \phi^{22}\right)$ that replicates $\mathcal{Z}_2$. Formally, at any time $t \in \mathbb{T}$ and for any constants $c_1, c_2 \in \mathbb{R}$,

$$c_1\phi_t^1 \neq c_2\phi_t^2, \quad \mathbb{P}\text{-a.s..} \tag{4.31}$$

Proposition 4.15 (Replication)**.** *Any bet $\mathcal{X}$ can be replicated by taking a dynamic position in any two linearly independent bets $\mathcal{Z}_1$ and $\mathcal{Z}_2$, formally:*

$$\Pi_t(\mathcal{X}) = \Pi_0(\mathcal{X}) + \int_0^t \left(\psi_s^1 \phi_s^{11} + \psi_s^2 \phi_s^{21}\right) dS_s^1 + \int_0^t \left(\psi_s^1 \phi_s^{12} + \psi_s^2 \phi_s^{22}\right) dS_s^2, \tag{4.32}$$

where the weights ψ_t^1, ψ_t^2 are given by the solution to the following equation:

$$\begin{bmatrix} \phi_t^{11} & \phi_t^{12} \\ \phi_t^{21} & \phi_t^{22} \end{bmatrix} \begin{bmatrix} \psi_t^1 \\ \psi_t^2 \end{bmatrix} = \begin{bmatrix} \phi_t^1 & \phi_t^2 \end{bmatrix}, \tag{4.33}$$

where $\left(\phi_t^{11}, \phi_t^{12}\right)$, $\left(\phi_t^{21}, \phi_t^{22}\right)$, and $\left(\phi_t^1, \phi_t^2\right)$ are the components of the portfolio that replicates $\mathcal{Z}_1$, $\mathcal{Z}_2$, and $\mathcal{X}$, respectively.

Proof. Substituting (4.30) in the left hand side of (4.32) verifies the proposition. ■

As briefly indicated at the beginning of Section 4.2, correct score bets can be used to statically replicate the value of other simple bets as seen in the following proposition.

Proposition 4.16 (Static Replication)**.** *The value of a simple bet at time t with payoff function Φ in terms of values of correct score bets at time t is given by*

$$\Pi_t(\mathcal{X}) = \sum_{K_1=N_t^1}^{\infty} \sum_{K_2=N_t^2}^{\infty} \Phi(K_1, K_2) \Pi_t(\Phi_{\mathrm{CS}(K_1,K_2)}), \tag{4.34}$$

where $\Pi_t(\Phi_{\mathrm{CS}(K_1,K_2)})$ denotes the value of a correct score bet at time t that pays out if the final scores are equal to (K_1, K_2).

Proof. From Corollary 4.10 and the definition of a simple bet, cf. (4.10) we have that:

$$\begin{aligned} \Pi_t(\mathcal{X}) &= \mathbb{E}_{\mathbb{Q}}\left[\mathcal{X} \mid \mathscr{G}_t\right] \\ &= \mathbb{E}_{\mathbb{Q}}\left[\Phi\left(N_T^1, N_T^2\right) \mid \mathscr{G}_t\right] \\ &= \mathbb{E}_{\mathbb{Q}}\left[\Phi\left(N_t^1 + N_{T-t}^1, N_t^2 + N_{T-t}^2\right) \mid \mathscr{G}_t\right] \end{aligned}$$

Since N_t^1, N_t^2 are discrete random variables, we have that their joint conditional density (under $\mathbb{Q}$) is given by:

$$f\left(n_1, n_2 \mid \mathscr{G}_s\right) = \mathbb{Q}\left(N_t^1 = n_1, N_t^2 = n_2 \mid \mathscr{G}_s\right).$$

We can now utilize a conditional version of the law of the unconscious statistician to arrive at:

$$\begin{aligned} &\mathbb{E}_{\mathbb{Q}}\left[\Phi\left(N_t^1 + N_{T-t}^1, N_t^2 + N_{T-t}^2\right) \mid \mathscr{G}_t\right] \\ &= \sum_{K_1=N_t^1}^{\infty} \sum_{K_2=N_t^2}^{\infty} \Phi(K_1, K_2) f\left(K_1 - N_t^1, K_2 - N_t^2 \mid \mathscr{G}_t\right) \\ &= \sum_{K_1=N_t^1}^{\infty} \sum_{K_2=N_t^2}^{\infty} \Phi(K_1, K_2) \mathbb{Q}\left(N_{T-t}^1 = K_1 - N_t^1, N_{T-t}^2 = K_2 - N_t^2 \mid \mathscr{G}_t\right). \end{aligned}$$

The results now follows from the fact that $\mathbb{Q}\left(N^1_{T-t} = K_1 - N^1_t, N^2_{T-t} = K_2 - N^2_t \mid \mathscr{G}_t\right)$, portraying the probability of a specific score given the available information, is equal to the pricing rule of a correct score bet by definition. ■

Proposition 4.16 gives us a way to replicate all simple bets from the correct score market. This has the implications that all markets consisting of simple bets, in theory, should move together. That is, in order for Proposition 4.16 to hold when the correct score market adjusts, all other markets of simple bets must also adjust correspondingly. To see an example of this consider the under 0.5 goals bet, i.e. (4.19) with $K = 0.5$. In theory, this bet corresponds to the correct score bet 0-0, since this outcome is the only outcome in which the total number of goals is below 0.5. Therefore, these two bets should be identical and should move together instantaneously when the value of either one changes. In practice, such an instantaneous move is not always the case due to liquidity concerns in the market.

To see how Proposition 4.16 can be used to replicate the assets $S^i \quad i \in \{1, 2\}$, first let $\Phi(S^1_T, S^2_T) = S^1_T$ or $\Phi(S^1_T, S^2_T) = S^2_T$ and note that $\Pi_t(S^i_T) = \mathbb{E}_{\mathbb{Q}}\left[S^i_T \mid \mathscr{G}_t\right] = S^i_t$. Now, using this in (4.34), we get:

$$S^1_t = \sum_{K_1=N^1_t}^{\infty} \sum_{K_2=N^2_t}^{\infty} K_1 \Pi_t(\Phi_{\mathrm{CS}(K_1,K_2)}), \tag{4.35}$$

$$S^2_t = \sum_{K_1=N^1_t}^{\infty} \sum_{K_2=N^2_t}^{\infty} K_2 \Pi_t(\Phi_{\mathrm{CS}(K_1,K_2)}). \tag{4.36}$$

4.5.2 Greeks

Let us now introduce the partial derivatives of a bet's value with respect to a change in time and with respect to the number of goals scored. This serves the same purpose as the Greeks in the Black-Scholes-Merton framework, that is, sensitivity analysis for changes in the underlying parameters. The Greeks are fundamental tools for hedging and risk management. We note that in the following, we sometimes want to emphasize a bet's dependence on the goals in the match, therefore we will sometimes use the explicit notation $\mathcal{X} = \mathcal{X}(N^1_t, N^2_t)$.

Definition 4.17 (Greeks)
The Greeks are the value of the following forward difference operators δ_1, δ_2 and partial derivative operator ∂_t applied to the bet value:

$$\delta_1 \Pi_t\left(\mathcal{X}(N^1_t, N^2_t)\right) = \Pi_t\left(\mathcal{X}(N^1_t + 1, N^2_t)\right) - \Pi_t\left(\mathcal{X}(N^1_t, N^2_t)\right) \tag{4.37}$$

$$\delta_2 \Pi_t\left(\mathcal{X}(N^1_t, N^2_t)\right) = \Pi_t\left(\mathcal{X}(N^1_t, N^2_t + 1)\right) - \Pi_t\left(\mathcal{X}(N^1_t, N^2_t)\right) \tag{4.38}$$

$$\partial_t \Pi_t\left(\mathcal{X}(N^1_t, N^2_t)\right) = \lim_{h \to 0} \frac{1}{h}\left[\Pi_{t+h}\left(\mathcal{X}(N^1_t, N^2_t)\right) - \Pi_t\left(\mathcal{X}(N^1_t, N^2_t)\right)\right]. \tag{4.39}$$

We note that δ_1 and δ_2 play the role of Delta in the Black-Scholes-Merton framework, i.e. they

measure the change in the price of a bet with respect to a change in the goal processes. The partial derivative operator ∂_t serve the same purpose as Theta in the Black-Scholes-Merton framework, that is, it measures time decay of a bet's value due to the passage of time.

In the next chapter, we will apply the described risk-neutral pricing theory on historic betting exchange data, in doing so, we will also present the data, the data cleaning procedure, and the origin and limitations of this data.

Model Calibration 5

In this chapter, we introduce the high-frequency historical betting data obtained from an online betting exchange and show a method of cleaning such data, as well as discussing some limitations. We also demonstrate how to calibrate our models to the betting data and show the results arising from such a procedure. Section 5.1 shows an exploratory data analysis, in which we present the betting data, the betting exchange, and related terms and mechanisms. In Section 5.2, we present the model calibration method and discuss some limitations of the method and the data. Section 5.3 shows the results of the calibration procedure with the different model dynamics on betting data from a chosen English Premier League football match.

5.1 Exploratory Data Analysis

This section is based on Brown and Yang (2017, pp. 587, 602), Bauwens, Hafner, and Laurent (2012, pp. 326–327), Barndorff-Nielsen et al. (2009, pp. C7–C8), & Nordsted (2009, p. 40).

Here, we present the historical betting data from Betfair's betting exchange as well as the cleaning of the high-frequency data. However, we first give a brief introduction and overview of betting exchanges, its lingo, and mechanisms.

5.1.1 Betting Exchange

The focus of our application is to calibrate model parameters to historical betting exchange data; more specifically, in-play English Premier League football betting data. For this, we use the Betfair betting exchange which is the largest of its kind in the world. Before the introduction of betting exchanges, there were basically only bookmakers in the betting market. On a betting exchange, bettors can bet on or bet against a given event, e.g. a team to win, and can also submit both market orders and limit orders. Market orders meet a limit order already in the book, and limit orders are placed in the book until an offsetting market order arrives.[1] Prior to the introduction of betting exchanges, bettors could not take a short position or submit limit

[1] Technically, a perfect offsetting order is not necessary due to the Betfair's cross-matching algorithm, see e.g. Berry (2017).

orders. To submit a buy order is known as *backing* a bet and to submit a sell order is known as *laying* a bet. That is, to back a position is to place a bet for something to happen, which is like a bet that you would place with a traditional bookmaker, and to lay a position is to place a bet for something not to happen, i.e., to bet against something happening. When laying a bet, you basically play the part of the traditional bookmaker.

Prices on the exchange are quoted in the form of odds. We will only deal with decimal odds, meaning that prices are quoted including the stake. For example, if a bettor places a back bet at odds 2.00, then the bettor will win 1 unit for every unit staked if the bet wins. If a bettor places a lay bet at odds 4.00, then he/she is liable to pay 3 units to the counter-party for every unit accepted, if it is a winning bet. The pricing grid in decimal odds format ranges from 1.01 to 1,000, with odds increments depending on the odds itself, see Table 5.1. The increment sizes of the odds has the impact that bet prices can, in practice, only undertake a finite number of values. We will however not be too concerned with this issue here. Another information regarding the betting exchange is that there can be no margin trading, i.e. trading using funds provided by a third party, at least not for the ordinary bettors, since all liabilities must reside with the exchange prior to any orders being submitted.

Odds from	Odds to	Increment
1.01	2.00	0.01
2.00	3.00	0.02
3.00	4.00	0.05
4.00	6.00	0.10
6.00	10.00	0.20
10.00	20.00	0.50
20.00	30.00	1.00
30.00	50.00	2.00
50.00	100.00	5.00
100.00	1,000.00	10.00

Table 5.1: Increment sizes of odds on the Betfair Exchange.

As mentioned in Section 4.2, we will use the market value of a bet as given by (4.14), i.e. the reciprocal of the decimal odds, due to its beneficial mathematical meaning. This value also goes by the term *implied probability* since it conveys the probability of the event happening as implied by the odds.[2]

To show the trading format on the Betfair Exchange, we captured a screenshot of the Betfair limit order book which is shown in Figure 5.2. The screenshot is taken from the match odds market for a Belarusian Premier League game[3] played on 2020-04-24 and only serves an illustrative purpose. The screenshot is taken a day prior to the beginning of the game and thus portrays the pre-game limit order book. A market order trade can be placed to back a team, i.e. take a

[2]Strictly speaking, the implied probability is not always given by the reciprocal of the decimal odds due to bookmakers overround, see Lindstrøm (2020), however, we ignore this issue due to the lack of intentional overround in peer-to-peer betting.

[3]Due to the outbreak of the 2019–20 coronavirus pandemic not many football leagues are playing at the time of writing, explaining the somewhat obscure choice of a Belarusian game.

long position, on the left-hand side of the book (in blue). Likewise, a market order trade to lay a team, i.e. take a short position, on the right-hand side (in red). The three best back and lay quotes are displayed with the volume available at each odds indicated below the odds.

Figure 5.2: Screenshot of the Betfair exchange's limit order book of the match odds market a day prior to a football game in the Belarusian Premier League between FC Smolevichi and FC Dinamo Minsk played on 2020-04-24.

There are transaction costs involved when trading on Betfair's exchange, more specific; you pay commission on your net winnings in a market. The commission is at 6.5%, with some discounts available based on the number of trades your have executed during the last period.

5.1.2 Market Efficiency & Information

A general assumption for the risk-neutral framework of betting markets, i.e. the general market model (4.1), is the market efficiency of bet prices. That is, we rely on bet prices to portray the true strength (or win probability) of the teams. If we cannot obtain such information from the bet prices our market model would fail to be useful. Market efficiency is thus an important issue in general, however, one that we shall not dwell too much about in this thesis. Though, we should not forget about the importance of it. Various papers[4] generally find that betting market inefficiencies are short-lived, however, much of these papers only focus on standard bookmaker markets, and not betting exchanges. Brown and Yang (2017) discuss the efficiency in horse racing betting markets on the Betfair exchange. They find that the predictive capacity of speculative trades varies throughout the trading period. They conclude that prior to the races information is stagnant and markets are mostly efficient, meaning that the average speculative trade brings limited information. However, they find that during races, speculative trades are good predictors of fundamentals.

By spending some time on the Betfair exchange, one can obtain a fairly decent knowledge of market information. In general, high liquidity on a betting market typically also means a very efficient market, however, there seems to exist illiquid markets that are not very efficient. A simple case of this phenomena is how the static replication theorem, Proposition 4.16, plays out

[4]E.g. Angelini and Angelis (2019), Williams (2005), and Deutscher, Frick, and Ötting (2018).

in reality. One can fairly easy find examples of markets not being perfect in sync with the correct score market. Whereas the Betfair cross-matching algorithm handles typical arbitrage situations within a market, the illiquidity of some markets generally creates some degree of uncertainty whether bet prices truly reflect the actual strength of the teams or not in certain markets. However, we will briefly touch on this issue later in the model calibration part in Section 5.2.

The use of the natural filtration in the market model generally agrees with Brown and Yang's findings that information is fairly stale prior to races, however, this is not always the case for football (or horse-racing). Some information may have a significant impact on the strengths (and bet prices) of the teams. Consider for example a star player being injured during practice a day before a match, and are unable to play the next day. This should have an effect on the bet prices (if the information about the injury is made public). Another situation, where the natural filtration seems inappropriate is with in-play injuries, booking, or red cards; these also have a significant impact on team strengths. Also, live bet prices tend to show an increasing chance of goal, when a dangerous set piece is occurring in the underlying game. In conclusion, the natural filtration has compelling limitations that we should be aware of.

5.1.3 Data Cleaning

The data is obtained from the Betfair Historical Exchange and is delivered as 1-second data, meaning that Betfair has sampled the odds at every second. It originally comes in `.json` format, containing several nested data frames, but after some initial pre-cleaning the data looks something as seen in Table 5.3, with a timestamp variable and best back and lay prices for each available odds.

Time	55190_BestBack	55190_BestLay	56343_BestBack	56343_BestLay
2017-12-23 12:29:58	1.6	NA	NA	7.2
2017-12-23 12:29:59	1.6	1.61	NA	NA
2017-12-23 12:30:00	NA	1.61	NA	NA
2017-12-23 12:30:01	1.6	1.61	7.0	7.4
2017-12-23 12:30:02	1.6	1.61	6.8	NA
2017-12-23 12:30:03	NA	NA	NA	NA

Table 5.3: Pre-cleaned Betfair exchange data.

The data also comes with information about the status of the betting market, e.g. suspended, open, or closed, and the in-play status, i.e. whether the game is ongoing or not. Since we only want to deal with in-play data, we define the *in-play time frame* as the time ranging from 1 minute prior to the game's beginning through the end of the match, and immediately discard any observations outside this window. Furthermore, from Table 5.3, we see that the raw data contains a lot of missing values. Partly, for this reason, we want to aggregate it to a specified level, in our case 30 seconds. Aggregation also makes the computation time of the calibration procedure much faster, and potentially more stable. We can now summarize our data cleaning procedure as follows:

Step 1 Remove all data outside the in-play time frame of the match and separate the data in pre-game and in-play data.

Step 2 Remove all data occurring during the halftime.

Step 3 Set observations to `NA` if the market is suspended.

Step 4 In the pre-game data, for the 30 observations (or desired aggregation level) leading up to the start of the match, calculate the median value, and assign this to the observation at $t = 0$.

Step 5 For each observation in the in-play data, ceiling round the time stamp to the nearest 30 second-mark (or desired aggregation level).

Step 6 For each timestamp with multiple observations, find the median of the observations, and replace the observations with this.

Step 7 Remove duplicate observations, created by the last step, such that each observation has a unique timestamp.

Step 8 Insert the current score at each time in the game.

Step 9 Inspect the data around market suspensions and set obviously flawed observations to `NA`.

This data cleaning procedure is inspired by two high-frequency data cleaning procedures proposed by Barndorff-Nielsen et al. and Bauwens, Hafner, and Laurent, along with knowledge and experience of specific problems regarding betting exchange data, such as the last step in the procedure. Step 9 is due to the market suspension mechanisms at the exchange in connection to the aggregation level, first by Betfair's sample frequency and later by our data cleaning. Betfair suspends the betting market when significant events happen in the game, such as goals, penalties, and red cards. By doing this they remove all orders in the limit order book and as such, the bettors will have to re-enter them if they still want it. This happens in order to protect the bettors as the events related to a market suspension can have a significant impact on the odds. Now, by sampling and/or aggregating the data at a certain threshold, we may obtain flawed data in which the market was actually suspended or by including data from both before and after a suspension, significantly skewing the median value. For this reason, it is good practice to inspect the data around such market suspension and place a missing value instead of an obvious erroneous value. One could also try to interpolate such observation, however, we do not consider this. Another issue regarding the market suspension is the fact that bettors have to re-enter their order again, meaning that a sample right after the suspension will likely have low liquidity, and therefore may have significantly higher variance, providing yet another reason to inspect the data in such situations.

After the cleaning procedure, the data now contains a unique timestamp, the minute of gameplay, home and away score variables, the back and lay prices, and the market mid-prices. The data generally has around 190-200 observations per match and has 115 variables, meaning that

we initially have 37 odds; 3 in the match odds market, 18 in the over/under markets (0.5 to 8.5), and 16 in the correct score markets (0-0 to 3-3). Later, we will inspect the liquidity of each bet and decide if some markets should be discarded for the given match. We show a sample of the cleaned data in Table 5.4.

Time	MinGamePlay	HomeScore	AwayScore	The Draw	The Draw_BestBack	The Draw_BestLay
2017-12-23 12:30:00	0.0	0	0	0.2409988	0.2439024	0.2380952
2017-12-23 12:30:30	0.5	0	0	0.2440476	0.2500000	0.2380952
2017-12-23 12:31:00	1.0	0	0	0.2469512	0.2500000	0.2439024
2017-12-23 12:31:30	1.5	0	0	0.2485335	0.2531646	0.2439024
2017-12-23 12:32:00	2.0	0	0	0.2485335	0.2531646	0.2439024

Table 5.4: The first five cleaned and aggregated observations of an English Premier League game played on 2017-12-23. The table only shows one kind of odds, namely the draw odds from the match odds market, in reality, it contains many more.

5.2 Calibration

This section is based on Divos et al. (2018, p. 327) & Andersen and Maillard (2019, pp. 23–25).

In this section, we discuss how to calibrate the model parameters to the historical market quotes. We follow Divos et al. fairly close and apply a least-squares approach where we consider market quotes of a set of bets and find model parameters that convey model prices for these bets that are as close as possible to the cleaned market quotes. By "model prices" we refer to the prices $\Pi_t(\mathcal{X})$ obtained via a pricing formula or Monte Carlo simulation, cf. Section 4.4. To be precise, we minimize the sum of the square of the weighted differences between the model and market mid-prices as a function of model parameters, using market back-lay spreads as weights. Due to the model parameters being of different dimensions among the models, we shall use the implicit notation $\widehat{\lambda}_i$ when emphasizing the dependence on model parameters. The reason for choosing a back-lay spread weighting is such that we take bets with a low spread into account, i.e. give them a higher weight because such mid-prices are assumed to be more certain, see e.g. Section 5.1.2. Formally, we minimize

$$R_t(\widehat{\lambda}_1, \widehat{\lambda}_2) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{\Pi_t^{\mathrm{MID}}(\mathcal{X}_i) - \Pi_t(\mathcal{X}_i)}{\Pi_t^{\mathrm{BACK}}(\mathcal{X}_i) - \Pi_t^{\mathrm{LAY}}(\mathcal{X}_i)}\right)^2}, \tag{5.1}$$

where n is the total number of bets used, $\Pi_t^{\mathrm{BACK}}(\mathcal{X}_i)$ and $\Pi_t^{\mathrm{LAY}}(\mathcal{X}_i)$ are the best market back and lay quotes of the ith type of bet at time t, $\Pi_t^{\mathrm{MID}}(\mathcal{X}_i)$ is the market mid-price of the ith bet at time t, and $\Pi_t(\mathcal{X}_i)$ is the model price of the ith bet at time t. This minimization procedure is referred to as *model calibration*.

To see how the back-lay spread weights come into play, consider a model calibration calculation with only one bet, and assume that the back price of this bet is at 0.55 and the lay price is at 0.45. This means that the market mid-price of the bet is 0.50. Now let us assume that the model price is 0.52, this means that we get a calibration error of 0.2, or in other words, the error in the model price is $\frac{1}{5}$th of the spread. On the contrary, let us now assume that we have a much

smaller spread with a back price at 0.505 and a lay price at 0.495, i.e. a much more "certain" mid-price. Now, with the same model price we get a calibration error of 2, since the difference between the mid-price and the model price is twice the size of the spread. That is, we weight the calibration such that differences between model prices and mid-prices with a low back-lay spread is punished more than with rather uncertain mid-prices.

The model calibration has been performed for the different market models, presented in Section 4.1.1, using a time step of 30 seconds during the game, cf. Section 5.1.3, and independently at each time step. We use the match odds market, over/under markets, and the correct score market with a maximum total of 37 bet types in these three categories cf. Section 5.1.3. "A maximum total of" refers to the liquidity concern, discussed above, and the fact that whenever a goal is scored in the match, some of these bets go out of play. When this happens, we discard these bets in the calibration procedure henceforth. Also, in some time steps there simply has not been an order placed, either due to market suspension or illiquidity, and thus we cannot calculate a market mid-price. Such observations have also been discarded in the model calibration, but only for the affected time step. We thus expect a slightly varying and decreasing number of bets as the match plays out (in case of goals).

In (5.1) we implicitly assume that for all bets there is always a spread, otherwise, we divide by 0. This is also the case in practice, since the back and sell price in the order book can never be the same. However, for some bets at some instances in the data, there is not a spread, due to the data aggregation. In such cases, we simply do not place a weight on this particular bet. For example, consider the bet $\mathcal{X}_n$ and assume that it has no back-lay spread for the given time step, in this case, we minimize

$$R_t(\widehat{\lambda}_1, \widehat{\lambda}_2) = \sqrt{\frac{1}{n}\left(\sum_{i=1}^{n-1}\left(\frac{\Pi_t^{\text{MID}}(\mathcal{X}_i) - \Pi_t(\mathcal{X}_i)}{\Pi_t^{\text{BACK}}(\mathcal{X}_i) - \Pi_t^{\text{LAY}}(\mathcal{X}_i)}\right)^2 + \left(\Pi_t^{\text{MID}}(\mathcal{X}_n) - \Pi_t(\mathcal{X}_n)\right)^2\right)}. \tag{5.2}$$

Likewise, we sometimes have that the data only contains missing values, e.g. during a market suspension. In such cases, we simply assign `NA` to the calibration result for this particular time step, and move on to the next time step. Lastly, we should again note that n in (5.1) and (5.2) are not fixed among time steps, cf. the discussion about "a maximum total of" above, and therefore the calibration procedure is, in general, robust to missing values.

5.2.1 Maturity

Using a fixed aggregation level, we have the discretization of time:

$$0, \Delta, 2\Delta, \cdots, T - 2\Delta, T - \Delta, T.$$

We then have that one increment has size Δ. Noting that the final time T corresponds to the fact that the game has ended, meaning that the payoff of the bet is known with certainty. We can therefore not attribute the last observation in our dataset to time T since it assumes that the game is still ongoing, albeit close to maturity. Thus, we associate the last observation in the

data with the time at $T - \Delta$, that is, the last observation in the data is at the time just prior to the end of the game. Note that we also have an observation prior to the beginning of the match at $t = 0$. From a chosen aggregation level, we can then calculate Δ as:

$$\Delta = \frac{T}{M+1},$$

where M denotes the total number of (aggregated) observations in the game. We will use the convention that the end of the game is equal to 1, i.e. $T = 1$. Consider for example an aggregation level at 30 seconds, as in our case cf. Section 5.1.3, and an total number of aggregated observations at 195, we then find the time increment to be $\Delta = \frac{1}{196} \approx 0.00510$.

As pointed out by Andersen and Maillard, the nature of stoppage time in a football match suggests a challenge in deciding the increment size. In other words, when a football game begins, one does not know precisely how many minutes/seconds the total playing time will be, meaning that, in a live setting, we cannot perfectly describe the time increment size Δ, since we do not know M. To be more specific, we cannot correctly characterize the time term, t, in the pricing of the market model, since this term relies solely on the increment size.

A way to overcome this problem is to only use 90 minutes of playing time, thus removing all stoppage time from the game, yielding a time increment size of $\Delta = 1/182$ when using a 30 seconds aggregation level. However, this poses a problem in the calculated model prices towards the end of the game. In such a situation, market participants have more information, i.e. knowledge about stoppage time, and thus actual probabilities might not be as certain as the model implies. Consider for example a calibrated model price at time 90:00 in the game, corresponding to $t = 181/182$ in our framework. Based on this assumption, the model, thinking the game will end within the next moment, will predict a rather certain outcome of the bets. Now consider a bettor watching the game and seeing that a long stoppage time of 7 minutes has been added. He/she will know that the game is far from over and thus his bets should reflect that, meaning that the market prices are not as certain as the model calculates. In practice, though, liquidity is rather scarce with maturity so close anyway.

Since we are working with historic data, we actually have knowledge about the total number of observations in the game. We will, therefore, use this in our calibration. However, as mentioned above, this approach is impossible in a live setting and may skew the model prices calculated in the beginning and middle of the game where market participants do not have perfect information on the ending of the game. Despite this, we deem this issue minuscule in comparison to the other case, since the market participants, hence market prices, should include some information about a possible, and very likely, stoppage time.

5.3 Results

We now turn our attention to the application of the theory described above. We first present the calibration results with Weibull process model dynamics and later with Weibull renewal process

model dynamics. We show the detailed results for only one match and the same match for both model dynamics. The results we show are from an English Premier League match between A.F.C. Bournemouth and Southampton F.C. played at Vitality Stadium in Bournemouth on 2017-12-03. This particular game ended 1-1, with the home goal happening in the 42nd and the away goal in the 63rd minutes of play. The game lasted for almost 96 minutes including stoppage times in both halves, meaning that our last observed prices are in the 95th-and-a-half minute. A notable observation with this particular match is that the home team, Bournemouth, was actually the underdog.

5.3.1 Weibull Process

Here, we present the results of the calibration procedure applied to the market model with Weibull process model dynamics. We start by showing the calibration error, i.e. how close the average calibrated prices are to the observed market quotes in units of the back-lay spread, cf. (5.1). To get a clearer picture of the calibration procedure, we also include a subplot of the number of bets used at each time step of the calibration such that we can inspect if this have an impact on the calibration error. This information is shown in Figure 5.5.

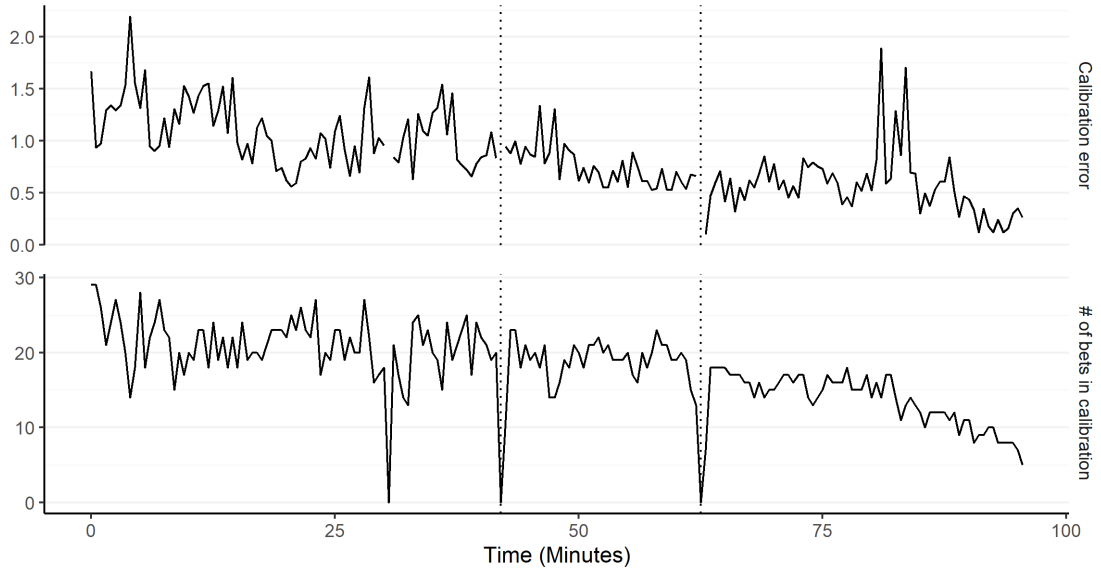


Figure 5.5: Top: Calibration error with Weibull process model dynamics in units of the back-lay spread for each time step. Bottom: Number of bets in the calibration for each time step. The vertical dotted lines indicate goal times.

In Figure 5.5, we see a fairly volatile calibration error with a decreasing trend throughout the game. We observe a spike in the beginning of the match, at around the fifth minute, where also the maximum calibration error occurs.Towards the end of the match, at around the 80th minute, we observe another large spike in the calibration error. Seemingly, there is not a clear correlation with spiky behavior and the number of bets used at that time step. Moreover, we also observe a decreasing trend in the number of bets used, this is however expected since with each goal scored the maximum bet count decreases, cf. Section 5.2. It it however notable that

the number of bets seems to decrease significantly towards the end of the game where no goals has been scored. This is likely due to the decreasing liquidity when maturity approaches, i.e. people not submitting back or lay orders on the bet. In general, the number of bets used in the given time step does not seem to impact the calibration error too much.

From the bottom part of Figure 5.5, we also see that there were three market suspensions during the game, i.e. time steps with no bets available; two of them coincide with the goal times and the other one is in the 31st minute. In the 31st minute a player took a dive in the penalty area, therefore, the market was probably suspended due to the possibility of a penalty. Instead of awarding a penalty, the referee instead booked the falling player.[5]

Lastly, we show some summary statistics of the calibration error in Table 5.6. Here, we see that the median and mean of the calibration error are 0.779 and 0.836 units of back-lay spread, respectively, and the standard deviation of the calibration error is 0.374. The fact that the median and mean are not that close also comply with the volatile behavior observed in Figure 5.5. We should note that a calibration error of 1 means that the average model price deviates from the market mid-price by 1 unit of the back-lay spread. It is also worth noting that the back and lay prices each deviate from the market mid-price with 0.5 back-lay spread by definition of the mid-price. We should thus strive for a calibration error below 0.5 in order to have the average calibrated model price within the back-lay spread of the observed prices. All-in-all, the results on calibration errors indicate a decent, yet not great, calibration performance, which is likely linked to the fact that the Poisson distribution provides a rather poor description of goals in football.

Min.	Median	Mean	Max.	Sd.
0.1031	0.7791	0.8364	2.1937	0.3739

Table 5.6: Summary statistics of the calibration errors in the market model with Weibull process model dynamics.

Next, we show the actual calibrated model prices with respect to the market back and lay quotes. In Figure 5.7 the solid lines represent the calibrated model prices of the match odds bets, i.e. home, away, and draw bets. The shaded areas show the prices within the back-lay spread for the match odds bets, that is, the edges of the shaded area corresponds to the back and lay prices at each time step.

In Figure 5.7, we observe that the calibrated model prices for the match odds bets tend to be consistently far outside the back-lay spread at the beginning of the game, indicating that the calibrated model prices are not very good for these bets. The calibrated prices seem to stabilize a bit after the first goal and even more after the second goal. This also agrees with the decreasing trend observed in the calibration error. To get a detailed view of the calibrated prices during the beginning of the game, we show a zoomed version of Figure 5.7 in Figure 5.8.

It is quite notable that the home and draw model prices are almost perfectly switched in the very

[5]For match details see Ames (2017).

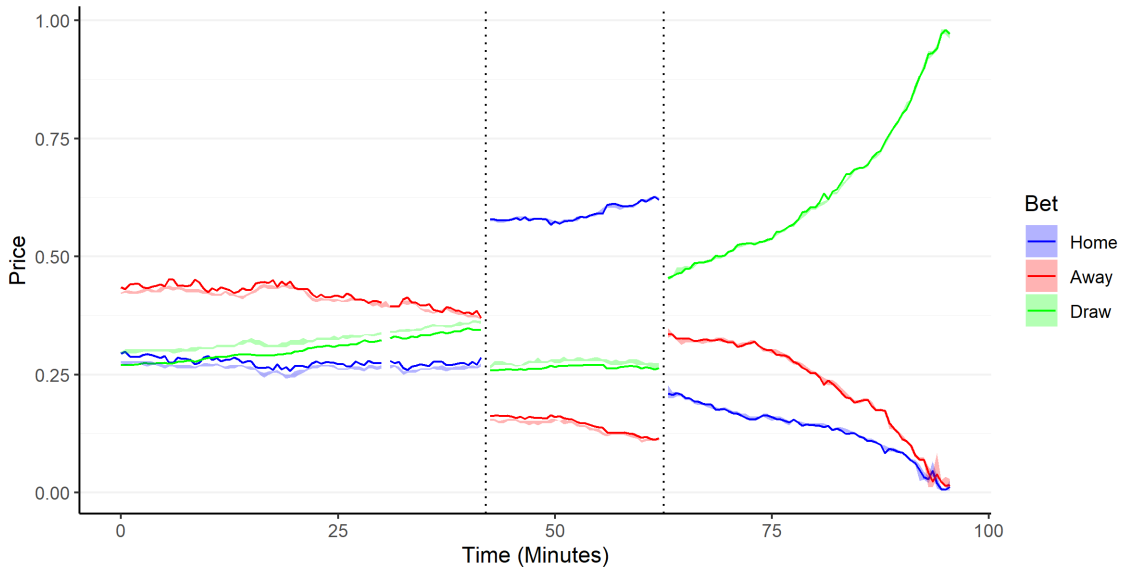


Figure 5.7: Match odds market quotes and calibrated model prices with Weibull process model dynamics. The solid lines represent the calibrated model prices of the respective bets. The edges of the shaded areas represent the back and lay prices of said bets. The vertical dotted lines indicate goal times.

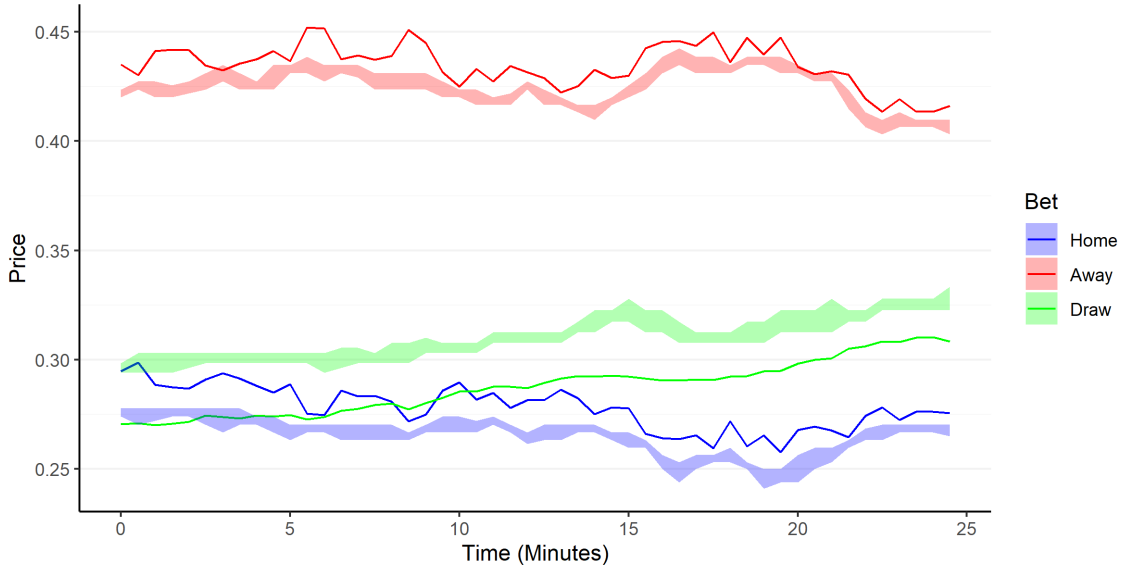


Figure 5.8: A zoomed version of Figure 5.7.

beginning as compared to the actual market quotes. Despite the seemingly poor results from the match odds market, we also take a look at some of the other markets included in the calibration. Figure 5.9 shows the calibrated model prices of the over/under 1.5 and 2.5 markets, again with respect to the back-lay spread. We again show a zoomed version in Figure 5.10, to get a more detailed view of the beginning of the match where the match odds bets had poorly calibrated prices. It looks, however, much more promising with the calibrated prices for the over/under bets. From Figure 5.9 and Figure 5.10, we see that the calibrated prices are, for the most parts, inside or very close to the back-lay spread, indicating a decent calibration performance for these types of bets.

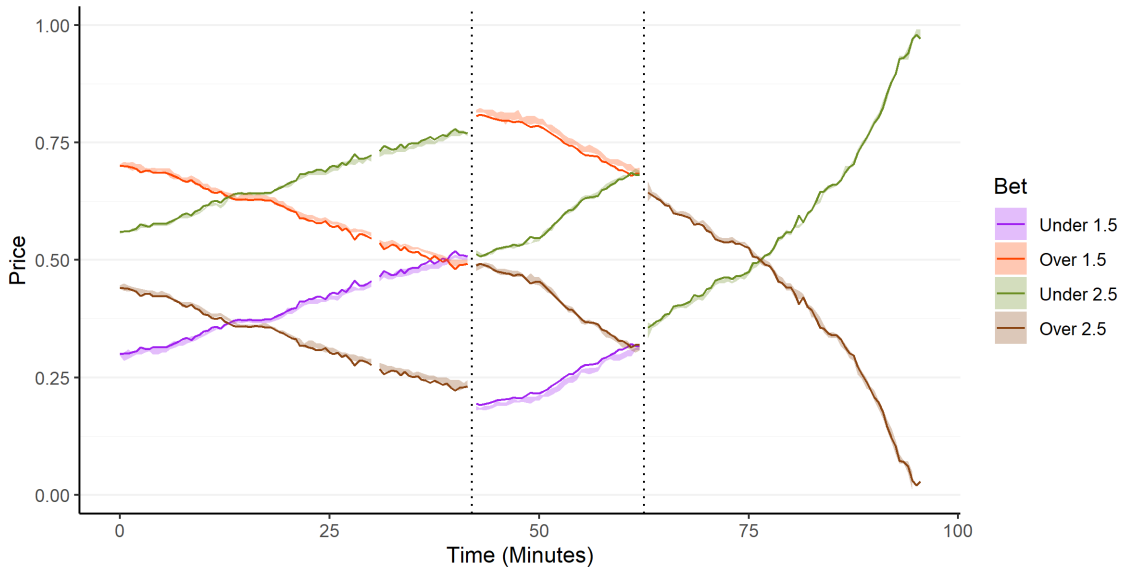


Figure 5.9: Over/under market quotes and calibrated model prices with Weibull process model dynamics. The solid lines represent the calibrated model prices of the respective bets. The edges of the shaded areas represent the back-lay prices of said bets. The vertical dotted lines indicate goal times.

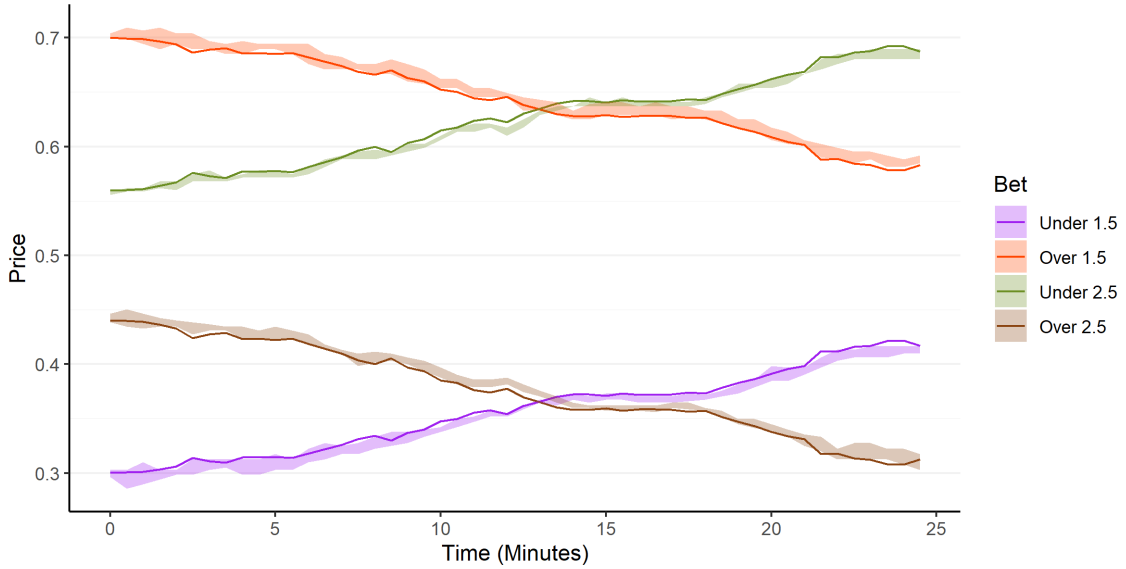


Figure 5.10: A zoomed version of Figure 5.9.

We also show a sample of the calibrated model prices for the correct score markets with respect to the back-lay spread. This is shown in Figure 5.11, with the 0-0, 0-1, 1-0, and 1-1 bets. Again, a detailed zoomed view is shown in Figure 5.12. These prices are much closer to each other, meaning that their implied probabilities are almost the same, but they generally show similar results as with the over/under markets; namely some fairly good calibrated prices throughout the match that are, for the most part, within the back-lay spread.

Another prominent feature of the correct score market is also clearly depicted; namely the illiquid nature of the correct score bets. In Figure 5.12 we see that a lot of market quotes are missing (no shading) at the beginning of the game, especially with the 1-1 and 1-0 bets. However, in

Figure 5.11, we see that these bets become more liquid when they become the current scores and time goes by. Specifically, we see that there is almost no missing data for the 1-0 bet, from the first goal until the second, and likewise, there is almost no missing data for the 1-1 bet after the second goal.

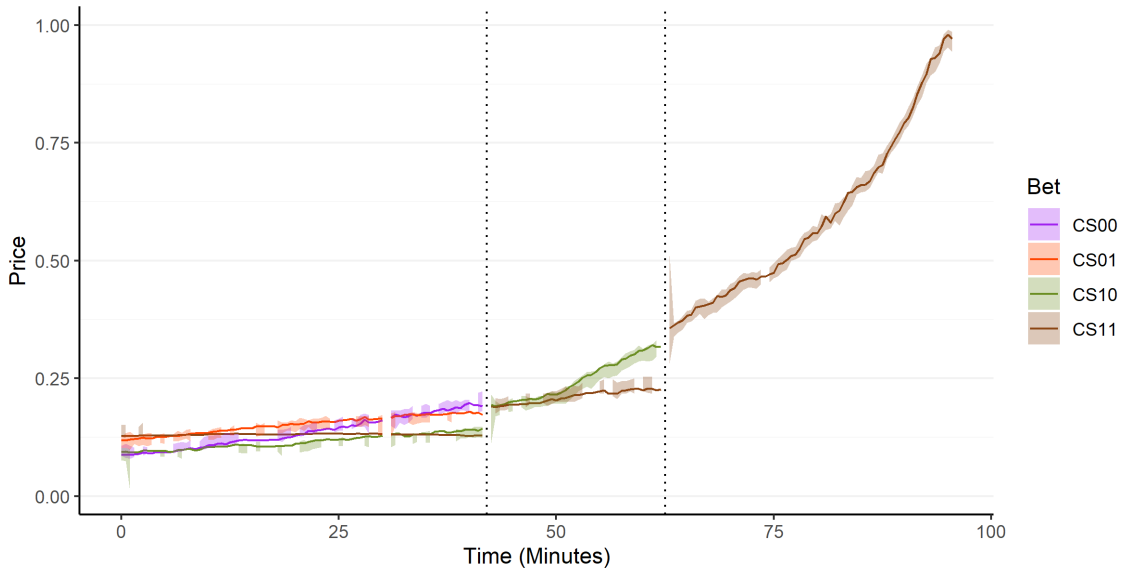


Figure 5.11: Correct score market quotes and calibrated model prices with Weibull process model dynamics. The solid lines represent the calibrated model prices of the respective bets. The edges of the shaded areas represent the back and lay prices of said bets. The vertical dotted lines indicate goal times.

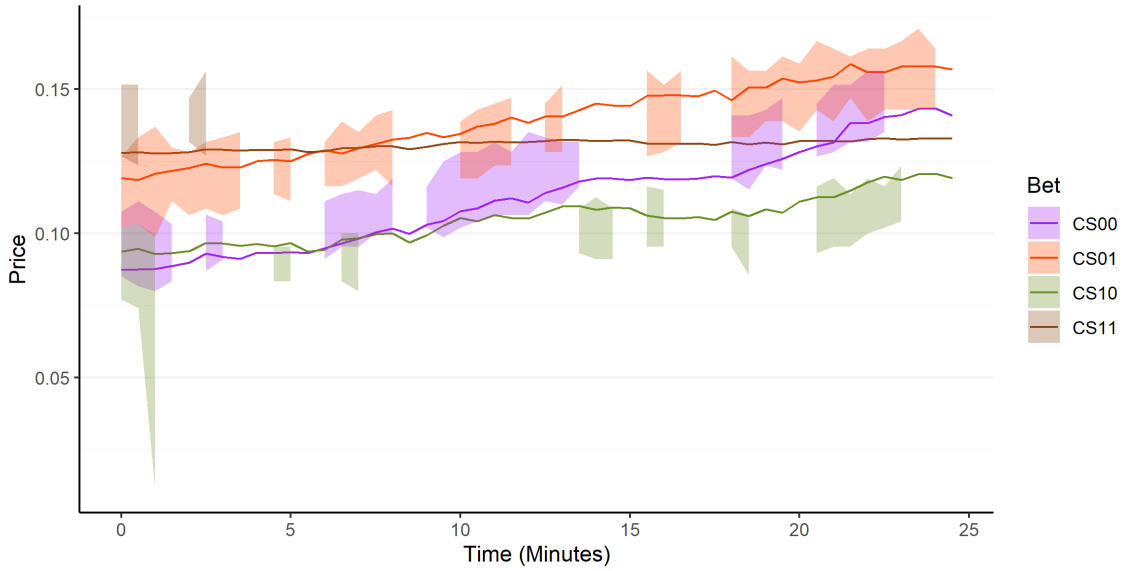


Figure 5.12: A zoomed version of Figure 5.11.

In general, with the information portrayed in Figure 5.7–5.12, we arrive at the same conclusion as with the calibration error; namely that the plots indicate fair results but with clear room for improvements. This is not totally unexpected results since we saw obvious flaws with the Poisson assumption of football scores in Chapter 3.

Parameters and Implied Intensity

Lastly, we consider the calibrated model parameters. In Figure 5.13, we see the calibrated scale and shape parameters of each team. We strive for stable parameters, however, we see that the parameters are fairly volatile, especially at the end of the game. The high volatility in the very ending of the game can likely be, somewhat, attributed to the aforementioned illiquidity of the odds in this period of the game, this is likely why we see these huge fluctuations of all parameters here. Despite, the volatile behavior, the parameters seem to have a tendency of a somewhat stable level that they fluctuate around. This is very prominent before the first goal. After this goal, there seem to be a jump in both teams' scale parameters, and the home teams shape parameters, but again, they tend to have a somewhat stable level they fluctuate around.

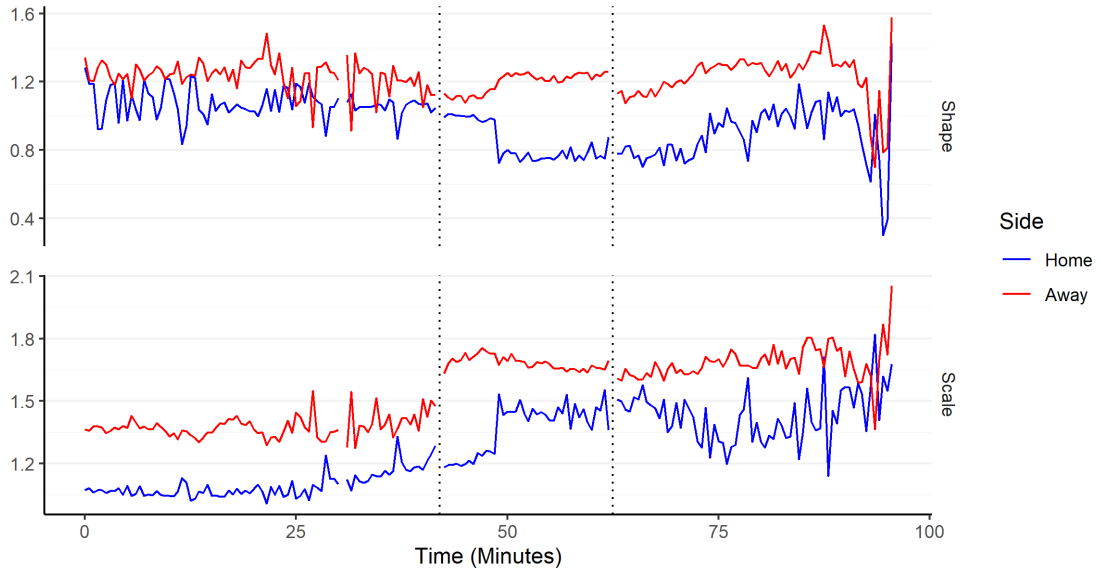


Figure 5.13: Calibrated Weibull process parameters for each side. Top: Shape parameter (β). Bottom: Scale parameter (α). The vertical dotted lines indicate goal times.

Recall, the notion of the intensity process described in Section 2.3, and the heuristic interpretation that the intensity is related to the conditional probability of scoring a goal in the next instant. We use the calibrated parameters to find the implied intensities of each team; the intensity of the Weibull process is given by (2.34), and using the calibrated model parameters in the Weibull intensity, we obtain the implied intensities as seen in Figure 5.14. We see quite volatile behaviors in the very beginning and the very end of the match, with a fairly stable period in the middle. But the general shape of the implied intensities follow an expected path, cf. Section 3.2, indicating a decent model choice, but again leaves some clear room for improvements.

5.3.2 Weibull Renewal Process

In the calibration procedure of the Weibull renewal process, we only include five types of bets; the three match odds bets and the over and under 2.5 goals bets. The reason for the choice of only five bets is due to computation time and instability in the optimization part, i.e. the

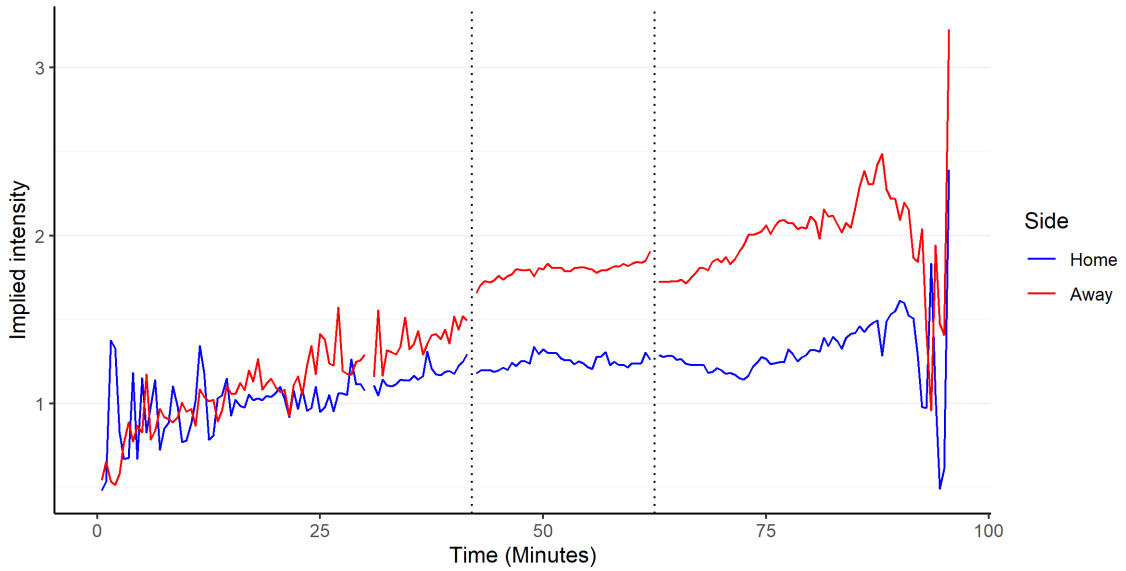


Figure 5.14: Implied Weibull process intensities for each side. The vertical dotted lines indicate goal times.

calibration procedure for the Weibull renewal process, something which shall be discussed further in Section 5.3.3. The fact that these five bets are in general the most liquid bets in the betting markets provides the reason for using, specifically, these five.

Again, we start by showing the calibration error in Figure 5.15. We do not show the number of bets used in the calibration since it is always five except in the two last time steps, where the number of bets used is four, also confirming that these markets are typically very liquid.

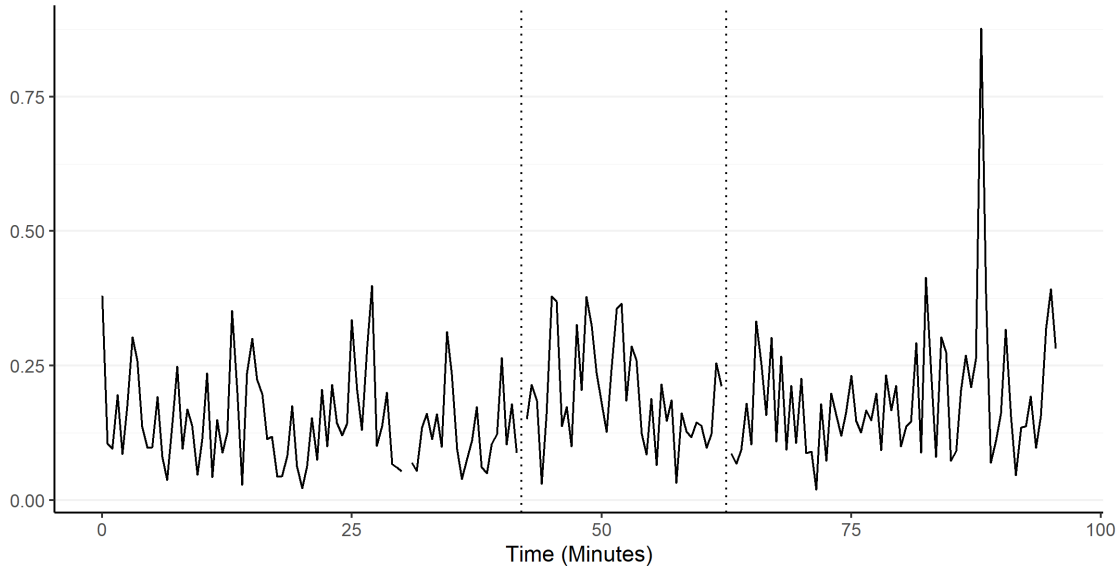


Figure 5.15: Calibration error with Weibull renewal process model dynamics in units of the back-lay spread for each time step. The vertical dotted lines indicate goal times.

In Figure 5.5, we see a very low calibration error with all errors, except one, way below the 0.5-mark. Such a low calibration error is not totally unexpected due to the added parameters

in this model, and the fact that the Weibull processes' calibration errors are embedded in the Weibull renewal processes' calibration errors. That is, we can always find parameters of the Weibull renewal processes such that their calibration error correspond to the calibration error of the Poisson process model dynamics, which has close to identical calibration errors to the Weibull process model dynamics.[6]

We do see a large spike in the calibration error in the 88th minute of play. This is seemingly not connected to a specific event in the match, and is possibly due to some illiquidity in the markets as maturity approaches. To further analyze, we show a sample of the prices around this moment in the match in Table 5.16. By inspection of the bet prices around this time, we may conclude that nothing extraordinary happened, however, some of the prices seem to "jump" slightly more than what they have been doing in the previous time steps.

MinGamePlay	HomeScore	AwayScore	Bournemouth	Southampton	The Draw	Under 2.5 Goals	Over 2.5 Goals
87.0	1	1	0.1064312	0.1747094	0.7143222	0.6968726	0.2998393
87.5	1	1	0.1010204	0.1770050	0.7233534	0.7067932	0.2921109
88.0	1	1	0.0954545	0.1739262	0.7380174	0.7326106	0.2690502
88.5	1	1	0.0910973	0.1422111	0.7547277	0.7463102	0.2550546
89.0	1	1	0.0910973	0.1361230	0.7722123	0.7576192	0.2386364

Table 5.16: Sample of the bet prices around the 88th minute of play in the Bournemouth vs. Southampton match.

Finally, we state some summary statistics of the calibration error in Table 5.17, which confirms the visual interpretation made from Figure 5.15. It is noteworthy that the minimum calibration error obtained is rather close to perfect with the average calibrated bet prices at only 0.0193 back-lay spread distance from the market mid-prices.

Min.	Median	Mean	Max.	Sd.
0.0193	0.1483	0.1704	0.8767	0.1048

Table 5.17: Summary statistics of the calibration errors in the market model with Weibull renewal process model dynamics.

In Figure 5.18, we show the calibrated model prices with respect to the market back and lay quotes; the solid lines represent the calibrated model prices of the match odds and over/under 2.5 goals bets and the shaded areas show the back-lay spread for these bets. In agreement with the calibration results, we see that the calibrated bet prices are almost all within the back-lay spreads at any given time. Furthermore, due to the high liquidity of these bets, the back-lay spread are also fairly narrow throughout the game, again indicating that the calibration of these bets show good results. Again, we also provide a zoomed version in Figure 5.19.

[6] This is due to the calibration of the time steps independently of each other, see e.g. the discussion on p. 44 in Andersen and Maillard (2019) for more on why the calibration errors of the Poisson model and the inhomogeneous Poisson model are close to identical.

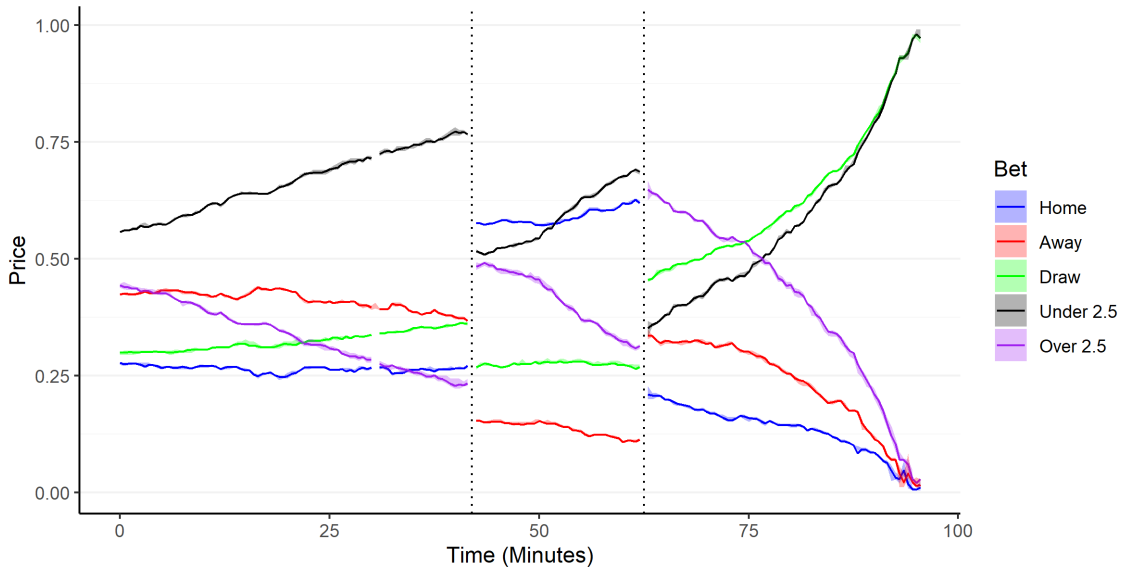


Figure 5.18: Match odds and over/under 2.5 goals market quotes and calibrated model prices with Weibull renewal process model dynamics. The solid lines represent the calibrated model prices of the respective bets. The edges of the shaded areas represent the back and lay prices of said bets. The vertical dotted lines indicate goal times.

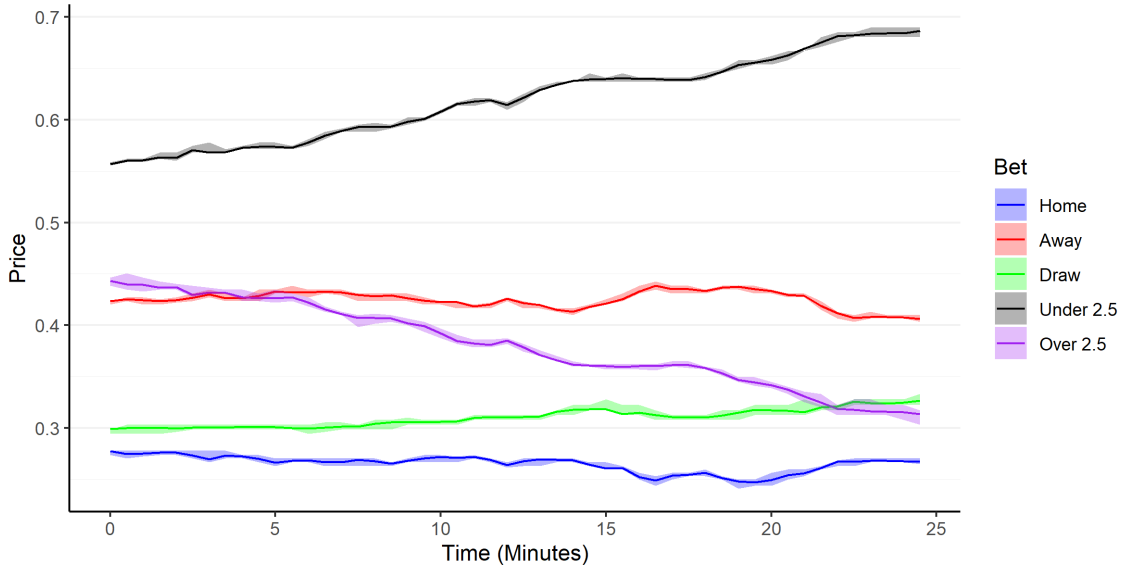


Figure 5.19: A zoomed version of Figure 5.18.

Parameters and Implied Intensity

The results of the calibration of the Weibull renewal process model seem very promising since we obtain such low calibration errors. However, we should be wary due to the added parameter freedom, which makes it possible to explain quite a lot more variation in the data, i.e. we need to consider the problem of overfitting. Therefore, it is necessary to analyze the parameters obtained from the calibration procedure in order to actually attribute the low calibration errors to the model, and not just to the increase in parameter freedom. Thus, we seek stable calibrated parameters throughout the match.

We show the calibrated Weibull renewal parameters in Figure 5.20, in which we observe very stable scale parameters of each team until each team scores a goal. After each team scores a goal, we observe quite volatile behaviors in their respective scale parameters. This is likely due to the renewal assumption, in which the intensity of the team completely starts over after a goal, which was found to be generally problematic in Section 3.3.

The shape parameters appear to be overall volatile throughout the match with a semi-stable period for each team prior to them scoring a goal. However, we see significantly more volatility of the shape parameter in the end. It is quite notable that the away side has very stable parameters just after the home goal and prior to the away goal, which tells us, in connection with the low calibration errors, that the model seems fairly appropriate in this period of the game. We also note that about halfway through the match, the shape parameters are often calibrated to close to one, indicating that the Weibull renewal model is close to a Poisson model.

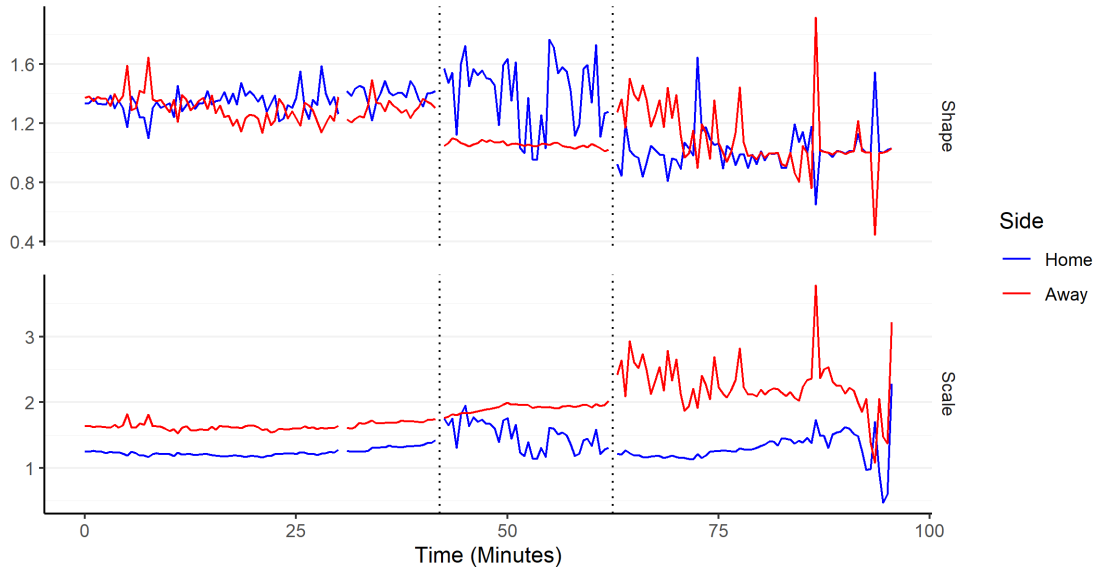


Figure 5.20: Calibrated Weibull renewal process parameters for each side. Top: Shape parameter (β). Bottom: Scale parameter (α). The vertical dotted lines indicate goal times.

To obtain a deeper understanding of the parameters and how they relate to the intensity of the two teams, we also show the implied Weibull renewal process intensities for each team. This is displayed in Figure 5.21. Here, we can clearly see that the implied intensities seem fairly stable and with expected general shapes up until the goals, after which we see quite volatile behaviors. A fact that just corroborate the observation made in the calibrated parameters in Figure 5.20, i.e. that the restart of the intensity after a goal is likely not appropriate in football modeling. This also backs the conclusion made in Section 3.3.1. However, despite the volatile behavior with the restart of the intensities, we see fairly decent results of these model dynamics, again with some obvious flaws that can most likely be attributed to the slight inaccuracy of the football modeling with the Weibull renewal process.

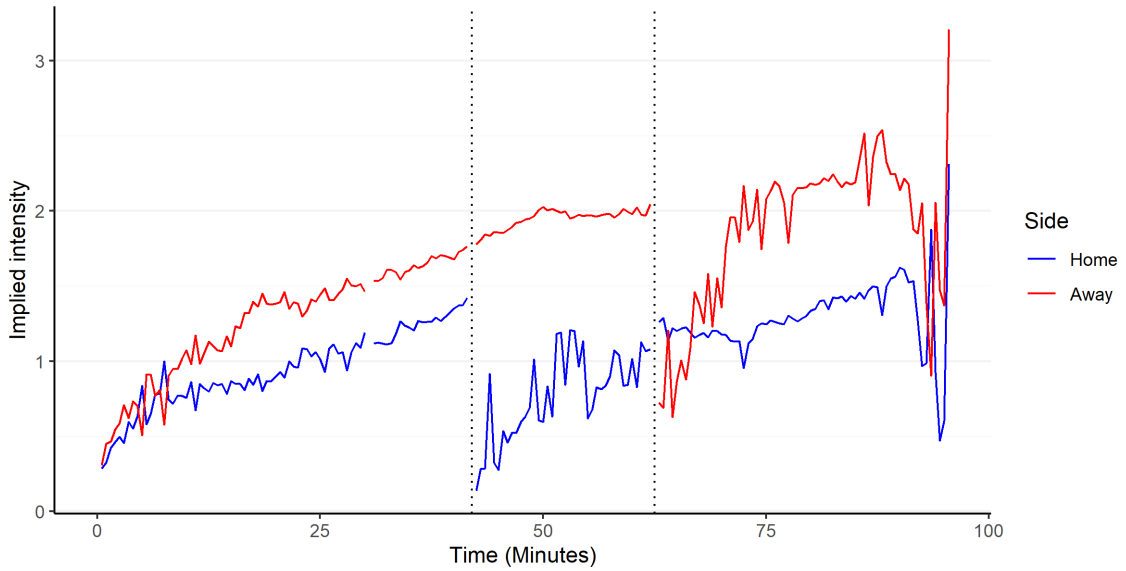


Figure 5.21: Implied Weibull renewal process intensities for each side. The vertical dotted lines indicate goal times.

Comparison with Weibull Process

As mentioned previously, the calibration error of the Weibull process model can be obtained by the Weibull renewal process with the parameters $\left(\lambda_t^1, 1, \lambda_t^2, 1\right)$, where $\lambda_t^i, \quad i \in \{1, 2\}$ are the calibrated parameters of the Poisson process model at time t, cf. (4.7). That is, we expect that the calibration error of the Weibull renewal process is always below or equal to the calibration error of the Weibull process. However, we cannot simply compare Figure 5.18 with Figure 5.7, since the two models have not been calibrated on the same data. Therefore, we need to calibrate the Weibull process model on this shrunken data set. We show the comparison of the two model's calibration errors on the same data in Figure 5.22, where the gray solid line is the calibration errors of the Weibull process model and the solid black line is the calibration errors of the Weibull renewal process model.

Figure 5.22 shows exactly what we expected; that the calibration errors of the Weibull renewal model are always equal to or below the calibration errors of the Weibull process model. For the Weibull process we observe a similar pattern as with the larger data set; namely that the calibration error starts high and then declines towards the end of the match. We note that the calibration errors of the Weibull process seem to start out very high on this data. This is likely because of the liquidity of the bets involved, i.e. that they have very low spreads. The calibration procedure then "punishes" significantly more for calibrated prices far from the observed (mid) prices; an example of this is shown in Figure 5.23, where the dashed lines are the calibrated prices under the Weibull process model and solid lines are the calibrated prices of the Weibull renewal process model.

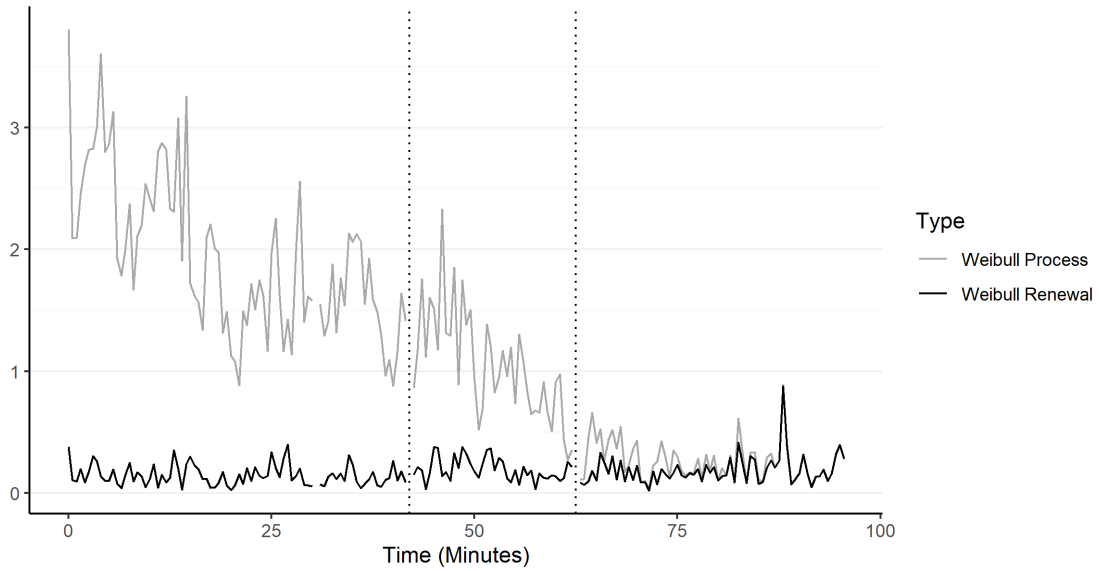


Figure 5.22: Comparison between the calibration errors of the market models with Weibull process and Weibull renewal process model dynamics, respectively, in units of the back-lay spread. The vertical dotted lines indicate goal times.

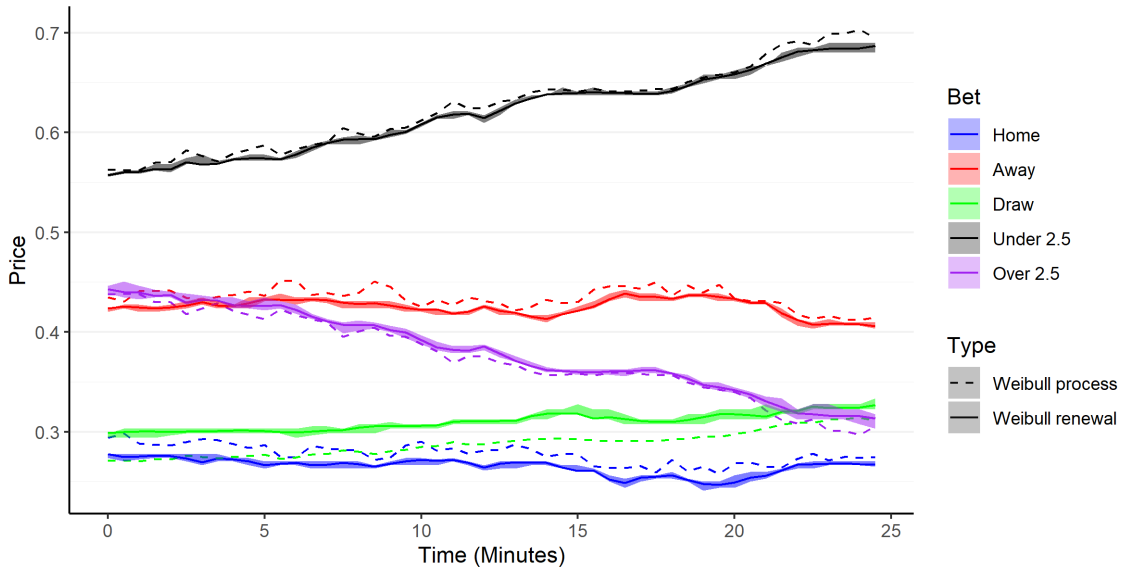


Figure 5.23: Zoomed comparison between the match odds and over/under 2.5 goals market quotes and the calibrated model prices with Weibull renewal process model dynamics (solid) and the calibrated model prices with Weibull process model dynamics (dashed). The edges of the shaded areas represent the observed back and lay prices of said bets. The vertical dotted lines indicate goal times.

5.3.3 Computation Time and Final Thoughts

Despite the significant improvement in the calibration error of the Weibull renewal process model as compared to the Weibull process, it is likely not very practical. This is due to a great increase in the computation time, and general instability of the calibration procedure even with fewer data. Let us clarify: Since we have no knowledge of any pricing formula for the Weibull renewal process while in-play, i.e. the conditional probability mass function of $N_T - N_t$, $t > 0$ given

the available information, we need to assess the probabilities using Monte Carlo simulations. This significantly increases the computation time, since a lot of simulations are needed in order to gain the desired significant digits of the probabilities, needed for a stable root-mean-squared optimization function. This stability issue is also the reason for our choice of only using five bets in the calibration procedure. Furthermore, the actual optimization of a function calculated from Monte Carlo simulations is both slow and unstable. This is because typical gradient methods are not stable in such cases, thus, we rely on derivative-free optimization methods, which tends to require even more runs of the optimization function. Also, we find that the optimizations function using Monte Carlo simulations seems to have a lot of local minimums, making the optimization even more tiresome. All-in-all, the calibration of the Weibull renewal process model shown in Section 5.3.2 took approximately 72 hours using parallelization methods on a server with 40 cores, thus, it is simply not practical for real-time betting usage.

Whereas the market model with Weibull renewal process model dynamics seems to greatly decrease the calibration errors and shows decent results of parameter stability, the Weibull process-based market model has some obvious flaws. However, what it lacks in accuracy it certainly brings in speed. Since there exists a specific pricing formula for each time step, the calculation and optimization of the model are fairly quick and are something that can be carried out in close-to real-time. This can be, somewhat, related to the use of the Black-Scholes-Merton model in regular option pricing, in which, the model assumptions are clearly flawed, but the fact that there exists a computationally fast pricing formula makes it a good guideline. This is probably also why the Black-Scholes-Merton equation is still useful today despite almost everyone disregarding the factual basis of this model. The Weibull process (or any inhomogeneous Poisson models) can most likely take a similar position in in-play betting markets.

The calibration errors of the model with Weibull renewal processes and the stable parameters before the goals provide, to some degree, evidence that a risk-neutral measure in in-play betting might not be a totally absurd idea. It is possible that by using a slightly more complicated counting process, which does not have complete renewals at the time of goals, we could obtain both stable parameters and good calibration errors. However, from our collective results on both the Weibull process- and the Weibull renewal process-based market models, we cannot conclude that a risk-neutral measure $\mathbb{Q}$ exists in the betting markets, that is, we cannot find a set of model parameters such that they are stable and consistent with all prices observed on the markets. This is a slight demotivating conclusion for risk-management and hedging purposes. We do, however, find some promising results in both cases when taking their limitations and advantages into consideration; of notable mentions are the "quick-and-dirty" evaluation that the Weibull process can bring, and the somewhat stable and consistent, but terribly slow, results occurring from the Weibull renewal process. As we have discussed previously, the assumptions of the market model are also not completely upheld in practice, again elaborating that this is not a perfect model, it does, however, seem to have great potential, especially if one is aware of the limitations. One of these limitations, namely the independence assumption of the score processes, is the subject of the next chapter.

Bivariate Model Extension 6

In this chapter, we discuss the limitations of the independence assumption imposed on the goal processes in the general market model presented in Chapter 4, and propose a bivariate model extension using a copula approach. We show calibration results for a market model with model dynamics consisting of a bivariate Weibull process and discuss these results. Section 6.1 discuss the independence assumption of football goals and presents the general bivariate market model. In Section 6.2, we present a brief introduction to copulas and how we can utilize these in a specific model dynamic. Section 6.3 shows the calibration results obtained from a bivariate copula approach to in-play football bets.

6.1 Bivariate Market Model

This section is based on Boshnakov, Kharrat, and I. G. McHale (2017, pp. 459–460).

The existence of some kind of dependence between goals scored by the two teams is widely accepted in the scientific community. The exact specification of this dependence is, however, less clear. First of all, basically all studies compiled on this subject have been on the full game, i.e. on the joint distribution of goals, and not on in-play dependence. One could argue that the dynamic nature of football teams and their in-play tactics and physical shape would make the dependence of goals vary during the game, i.e. that there are some time periods where goals tend to be more correlated than others. We also have the obvious dependence that scoring two goals within the same minute is almost impossible in football, due to the celebration and kick-off situation related to a goal in football.

Historically, the dependence has been specified, as mentioned above, in terms of the distributions of goals, for example, Dixon and Coles (1997) study the difference between the empirical joint distribution and the implied joint distribution of goals under the hypothesis that they are independent. They conclude that the distributions are in fact not independent. To deal with this they suggested an ad-hoc correction on their bivariate Poisson distribution. Dixon and Coles's main objectives with their study were to see if one could forecast football matches using a specified joint distribution and to use historical results to estimate the parameters of this distribution. Karlis and Ntzoufras (2003) used a different approach to account for the dependence; namely

by utilizing a diagonal inflated distribution to account for the alleged observation that draws seems to happen more often than the Poisson distribution suggests (which we did not find evidence of in Section 3.1). Later, I. McHale and Scarf (2007), I. McHale and Scarf (2011), and Boshnakov, Kharrat, and I. G. McHale (2017) proposed the use of copulas with some chosen marginal distributions to model the dependence between goals.

Here, we will follow the copula approach, and try to incorporate some dynamic in-play dependence between the goal processes. First, we need to set the stage for the bivariate market model. All the terminology from Chapter 2 can be readily extended to the multi-dimensional case, see e.g. Brémaud (1981) or Sokol and Hansen (2015). We will employ this fact in the extension of the risk-neutral framework developed in Chapter 4. In fact, we have already implicitly used a bivariate market model with the limiting assumption that the two processes were independent, thus making the distinction of them simple and intuitive. Our main concern is now to incorporate some form of dependence between the goal processes, and this is where the general multi-dimensional case comes in handy. Let us start by making much of the same assumptions as in Section 4.1.

Consider again a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ which carries a two-dimensional counting process $N = (N^1, N^2)^\top$ with intensity process $\mu = (\mu^1, \mu^2)^\top$, where N is equipped with the bivariate filtration $\mathscr{G}$ described earlier. Again, consider the parameter space $\mathbb{T} = [0, T]$, then the two-dimensional counting process depicts the number of goals scored by each team during the game, just as before.

Let $S = \left(S^1, S^2\right)^\top = \left((S_t^1)_{t\in\mathbb{T}}, (S_t^2)_{t\in\mathbb{T}}\right)^\top$ be such that its value at the end of the game is a two-dimensional vector equal to the number of goals scored by the home and away teams, respectively. We can now formally define a general market model based on these descriptions.

Definition 6.1 (General Market Model)
The *general market model* is defined by the following price dynamics of the assets B, S^1, and S^2:

$$\begin{aligned} B_t &= 1 \\ S_t = \begin{bmatrix} S_t^1 \\ S_t^2 \end{bmatrix} &= \begin{bmatrix} N_t^1 + \mathbb{E}\left[L_T \int_t^T \lambda_u^1 du \mid \mathscr{F}_t^{N^1}\right] / L_t \\ N_t^2 + \mathbb{E}\left[L_T \int_t^T \lambda_u^2 du \mid \mathscr{F}_t^{N^2}\right] / L_t, \end{bmatrix} \end{aligned} \tag{6.1}$$

where N is a nonexplosive bivariate counting process with intensity process μ and $\lambda = (\lambda)_{t\in\mathbb{T}}$, is a known bivariate, predictable, locally bounded, and non-negative (stochastic) process that is μ-compatible and L is a known $\mathbb{P}$-martingale.

Remark: Note that the market model (6.1) is a straightforward extension of the market model (4.1), where the only difference is the removal of the independence assumption imposed on the counting processes in (4.1) and the matrix notation.

In theory, the bivariate market model portrays a much more realistic view of the underlying

football match. It does, however, complicate things quite a bit; in order to explicitly use this framework, we have to be able to describe the dependence. Since our main focus has been on the model calibration independently at each time step, we basically only try to track and trace the observed market quotes with a mathematical model, meaning that we do not directly consider the time series aspect of it. Thus, we can ignore some of the concerns corresponding to autocorrelation and time-varying dependence in a time series, and simply use a similar approach as Boshnakov, Kharrat, and I. G. McHale (2017), in which they propose the use of a *copula* to "glue" two Weibull Count Model distributions together. Their application is focused on explaining and forecasting of the pre-game odds.

Generally, a copula approach in an autocorrelated time series framework can potentially create artificial dependence between a set of variables, but as mentioned, we can ignore these concerns due to the focus on explaining the odds at each time step independently of each other. We therefore only consider a copula approach to model the dependence of the goal processes, and furthermore, due to time constraints, we henceforth only focus on the bivariate case of the Weibull process. This is due to the simplistic nature of the independent increments of inhomogeneous Poisson processes, and the specific pricing formula it brings. There are other concerns with the use of copulas in this setting which we will discuss in Section 6.2.2.

Our choice of using copulas to model the dependence between the teams has sole basis in the work of Boshnakov, Kharrat, and I. G. McHale (2017). There are, most likely, several other interesting ways to model the dependence, unfortunately, we simply do not have time to explore them all in this thesis, which is why we only place our focus on the copula approach. In the next section, we will thus formally introduce the concept of a copula, state the main results pertaining to this, and present the specific copula we will use.

6.2 Copulas

This section is based on Boshnakov, Kharrat, and I. G. McHale (2017, p. 460),Tankov and Cont (2004, pp. 136–141), Trivedi and Zimmer (2017), & Haugh (2016, pp. 1–2).

In this section we provide a brief introduction to the copula notion. For a more rigorous approach on copulas see e.g. Nelsen (2006) and Cherubini, Luciano, and Vecchiato (2004). We start by considering a two-dimensional random vector $X = (X_1, X_2)^\top$. The law of X is typically described using its cumulative distribution function (CDF):

$$F(x_1, x_2) = \mathbb{P}(X_1 \leq x_1, X_2 \leq x_2)\,. \tag{6.2}$$

The marginals of X is the laws of X_1 and X_2 taken separately. These laws can be described using their respective distribution functions $F_1(x_1) = \mathbb{P}(X_1 \leq x_1)$ and $F_2(x_2) = \mathbb{P}(X_2 \leq x_2)$, which can also be obtained from the two-dimensional distribution function:

$$F_1(x_1) = F(x_1, \infty) \quad \text{and} \quad F_2(x_2) = F(\infty, x_2).$$

In addition to the marginals, we also need the dependence between X_1 and X_2 to fully described the distribution function $F(x_1, x_2)$. This is where copulas becomes handy.

Definition 6.2 (Copula)
A two-dimensional *copula*, $C : [0,1]^2 \to [0,1]$ is a cumulative distribution function with uniform marginals.

In other words, a copula C is the distribution of a bivariate uniform random vector. Note that one usually define a copula in a more rigorous way using terms like *grounded* and *d-increasing*, however, for the purpose of this application, we do not consider such rigorous approach. Let us now recall the definition of the *quantile function*, also known as the *generalized probability inverse*; for a CDF F the quantile function F^{-1} is defined as:

$$F^{-1}(x) := \inf \{v : F(v) \geq x\}, \quad x \in [0,1]. \tag{6.3}$$

We can now state the following well-known result.

Proposition 6.3. *If $U \sim \text{Unif}(0,1)$ and F_X is a CDF, then*

$$\mathbb{P}\big(F^{-1}(U) \leq x\big) = F_X(x).$$

In the opposite direction, if X has a continuous CDF F_X then

$$F_X(X) \sim \text{Unif}(0,1).$$

Proof. Omitted.

The quantile function thus provide a way to simulate, or "translate", random variables from a distribution F by simulating uniform random variables.

6.2.1 Sklar's Theorem

The strength of a copula approach for modeling the dependence arises from the theorem due to Sklar that states that the joint cumulative distribution function F may be expressed using a copula and its marginals. First, let F be the CDF of $X = (X_1, X_2)^\top$ with continuous and increasing marginals. Then, by Proposition 6.3 we have that the joint distribution of F_{X_1}, F_{X_2} is a copula, C_X. We can actually find an expression for C_X by noting that

$$\begin{aligned} C_X(u_1, u_2) &= \mathbb{P}(F_{X_1}(X_1) \leq u_1, F_{X_2}(X_2) \leq u_2) \\ &= \mathbb{P}\Big(X_1 \leq F_{X_1}^{-1}(u_1), X_2 \leq F_{X_2}^{-1}(u_2)\Big) \\ &= F_X(F_{X_1}^{-1}(u_1), F_{X_2}^{-1}(u_2)). \end{aligned} \tag{6.4}$$

Now, let $u_j := F_{X_j}(x_j)$, then (6.4) yields

$$F_X(x_1, x_2) = C_X\left(F_{X_1}(x_1), F_{X_2}(x_2)\right).$$

This is one side of the honored Sklar's Theorem which we now state formally.

Theorem 6.4 (Sklar's Theorem)**.** *Consider a 2-dimensional CDF F with marginals F_1, F_2. Then there exists a copula C such that*

$$F(x_1, x_2) = C\left(F_1(x_1), F_2(x_2)\right) \tag{6.5}$$

for all $x_i \in \mathbb{R}$ and $i = 1, 2$.

Proof. Omitted.

We can use the power of Sklar's Theorem to join our two marginal distributions. However, there are several copulas to choose from in order to test it in this bivariate scenario. Due to time constraints, we do not investigate or test which copulas are best suited for our needs, we simply follow Boshnakov, Kharrat, and I. G. McHale's choice by using only the *Frank copula*, which we will briefly introduce next.

Frank Copula

The Frank copula is given by

$$C(u, v) = -\frac{1}{\kappa} \log\left(1 + \frac{\left(\mathrm{e}^{-\kappa u} - 1\right)\left(\mathrm{e}^{-\kappa v} - 1\right)}{\mathrm{e}^{-\kappa} - 1}\right), \tag{6.6}$$

where $\kappa \in \mathbb{R}$ is the dependence parameter. We see that the Frank copula allows for a complete spectrum of dependence, meaning that the scope of the correlation ranges from -1 to 1. Thus, the Frank copula nests the independence copula, when $\kappa = 0$.

6.2.2 Drawbacks of Discrete Copulas

There are some issues with utilizing a copula with discrete distributions that we should be careful about. First, if the marginals F_1, F_2 are not continuous, then the corresponding copula in (6.5) is not guaranteed to be unique. This is generally not a problem in applied settings due to the fact that one's use of copulas is often based on the reason that the joint distribution is either unknown or difficult to work with. Second, and much more concerning, is that estimates of the dependence parameter are bias when either F_1 or F_2 are not continuous. Trivedi and Zimmer (2017) provides several examples of these issues with discrete copulas, and present reviews and discussions of bivariate copulas in discrete count data settings. We will not analyze these concerns deeper here; we simply state that a copula approach in our setting might not be appropriate. Despite these concerns, we progress with the copula idea, simply to obtain results on the possible correlation.

6.2.3 Pricing Formula

We are now in a position to introduce a bivariate pricing formula. Recall the pricing formula in the inhomogeneous Poisson setting, stated in Proposition 4.12. We can now extend this using the bivariate copula approach. Note that a copula, by definition, is a CDF, therefore, we need to account for this when using this method for a bivariate probability mass function calculation as in our case:

$$\begin{aligned}\mathbb{P}(X_1 = x_1, X_2 = x_2) = &\, C\left(F_1(x_1), F_2(x_2)\right) \\ &- C\left(F_1(x_1 - 1), F_2(x_2)\right) \\ &- C\left(F_1(x_1), F_2(x_2 - 1)\right) \\ &+ C\left(F_1(x_1 - 1), F_2(x_2 - 1)\right),\end{aligned} \tag{6.7}$$

for $x_i \in \mathbb{N}$ and $i = 1, 2$.

Basically, (6.7) works in the same fashion as $F(X = x) = F(X \leq x) - F(X < x)$, simply just in a discrete way. We should note that whenever, either x_1, x_2, or both are equal to zero, we cannot use (6.7) explicitly. In such cases we have

$$\mathbb{P}(X_1 = 0, X_2 = x_2) = C\left(F_1(0), F_2(x_2)\right) - C\left(F_1(0), F_2(x_2 - 1)\right), \tag{6.8}$$

$$\mathbb{P}(X_1 = x_1, X_2 = 0) = C\left(F_1(0), F_2(x_2)\right) - C\left(F_1(x_1 - 1), F_2(0)\right), \tag{6.9}$$

$$\mathbb{P}(X_1 = 0, X_2 = 0) = C\left(F_1(0), F_2(0)\right), \tag{6.10}$$

respectively.

Proposition 6.5 (Pricing Formula of a Simple Bet - Bivariate Inhomogeneous Poisson Setting)**.** *Assume we are in a bivariate inhomogeneous Poisson process setting of the general market model. The value of a simple bet at time t with payoff function Φ is given by*

$$\begin{aligned}\Pi_t(\mathcal{X}) = \sum_{n_1 = N_t^1}^{\infty} \sum_{n_2 = N_t^2}^{\infty} &\Phi(n_1, n_2) \times \\ &\Big(C\left(P\left(n_1 - N_t^1, \lambda_T^1 - \lambda_t^1\right), P\left(n_2 - N_t^2, \lambda_T^2 - \lambda_t^2\right)\right) \\ &- C\left(P\left(n_1 - N_t^1 - 1, \lambda_T^1 - \lambda_t^1\right), P\left(n_2 - N_t^2, \lambda_T^2 - \lambda_t^2\right)\right) \\ &- C\left(P\left(n_1 - N_t^1, \lambda_T^1 - \lambda_t^1\right), P\left(n_2 - N_t^2 - 1, \lambda_T^2 - \lambda_t^2\right)\right) \\ &+ C\left(P\left(n_1 - N_t^1 - 1, \lambda_T^1 - \lambda_t^1\right), P\left(n_2 - N_t^2 - 1, \lambda_T^2 - \lambda_t^2\right)\right)\Big),\end{aligned} \tag{6.11}$$

where $P(N, \Lambda)$ is the Poisson probability mass function, and where we use the conventions described by (6.8)–(6.10) in the special cases when either $n_1 - N_t^1, n_2 - N_t^2$, or both are equal to zero.

Proof. Recall that we have

$$\begin{aligned}\Pi_t(\mathcal{X}) &= \mathbb{E}_{\mathbb{Q}}\left[\mathcal{X} \mid \mathscr{G}_t\right] \\ &= \mathbb{E}_{\mathbb{Q}}\left[\Phi\left(N_t^1 + N_{T-t}^1, N_t^2 + N_{T-t}^2\right)\right]\end{aligned}$$

$$= \sum_{n_1=N_t^1}^{\infty} \sum_{n_2=N_t^2}^{\infty} \Phi(n_1, n_2) \mathbb{Q}\left(N_{T-t}^1 = n_1 - N_t^1, N_{T-t}^2 = n_2 - N_t^2\right).$$

The result then follows from (6.7). ■

Using the Frank copula and the Weibull process specification of the marginals in (6.11), we can use this in the calibration procedure, described in Section 5.2, on the in-play betting data.

6.3 Calibration Results

In this section, we present the results of applying the Frank copula Weibull process model dynamics to the in-play betting data presented Section 5.1.3. We show the results of the calibration error, the parameters, the calibrated market price vs. market quotes, and the comparison to the independent Weibull process model.

The calibration errors are shown in Figure 5.5, where we see that the calibration error generally varies within 0.5 and 1 back-lay units. We also see a slightly decreasing trend throughout the game; however, not as severe as what we observed with the independent Weibull process model. The comparison of the calibration errors between the bivariate and the independent models can be seen in Figure 6.2.

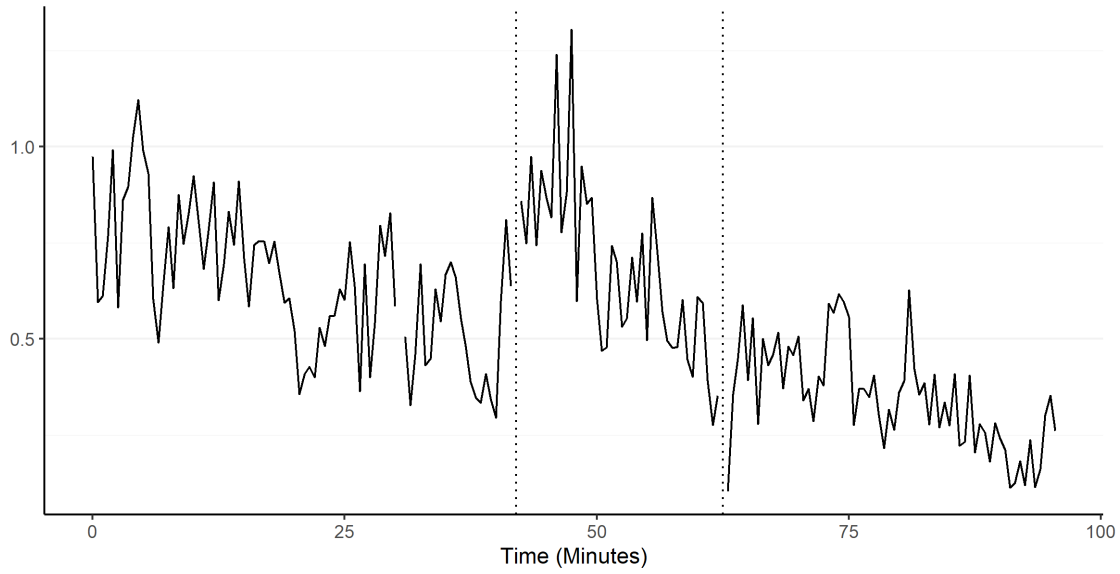


Figure 6.1: Calibration errors of the market model with Frank copula Weibull process model dynamics in units of the back-lay spread.

In Figure 6.2, we see a clear pattern that the calibration error is much lower and less spiky in the bivariate case, something that is also backed by the summary statistics, shown in Table 6.3, possibly indicating that there seems to be some sort of dependence between the two goal processes. However, such conclusions are generally tough to make from a flawed model assumption. The only period where both models seem to have about the same calibration error is the time between the first and the second goal. In this period, we also observe a small spike

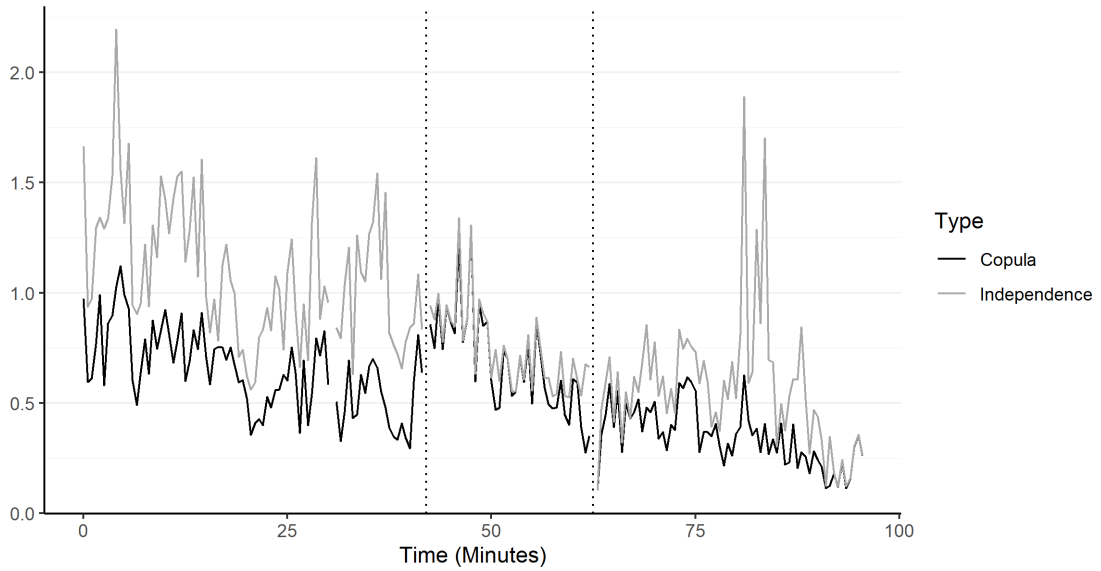


Figure 6.2: Comparison between the calibration errors of the market models with Frank copula Weibull process model dynamics and the market model with independent Weibull process model dynamics, in units of the back-lay spread. The vertical dotted lines indicate goal times.

in the calibration error. These spikes were not very prominent for the independent model as compared to the other spiky periods, however, we do not observe the other large spiky periods during the beginning and end of the game in the bivariate model.

The maximum calibration error of the bivariate model also occurs during the period between the goals; right after the halftime, in the 47th minute, to be exact. It occurs within a period of already enlarged calibration errors. A possible explanation of this enlarged calibration error observation around the halftime could be due to the fact that the underdogs were leading and the fact that betting still occurs during the halftime, despite the game being on pause; something, which is not shown in the data. The halftime betting could possibly lead some bettors to speculate and/or analyze more in-depth on the home team's chances, thus moving the market quotes significantly during the half time. However, this is only a possible explanation; it might just be an anomaly or simply that the model is not able to catch some specific feature happening during this period, after all, the Poisson assumption was found to be flawed in describing football goals.

Model Type	Min.	Median	Mean	Max.	Sd.
Bivariate	0.1030	0.5526	0.5508	1.3043	0.2329
Independent	0.1031	0.7791	0.8364	2.1937	0.3739

Table 6.3: Summary statistics of the calibration errors in a market model with Frank copula Weibull process model dynamics.

Next, we show the calibrated model prices with respect to the market back and lay quotes. In Figure 6.4 the solid lines represent the calibrated bivariate model prices of the match odds bet and the shaded areas show the back-lay spread for the match odds bets.

In Figure 6.4, we observe that the calibrated model prices for the match odds bets tend to be fairly

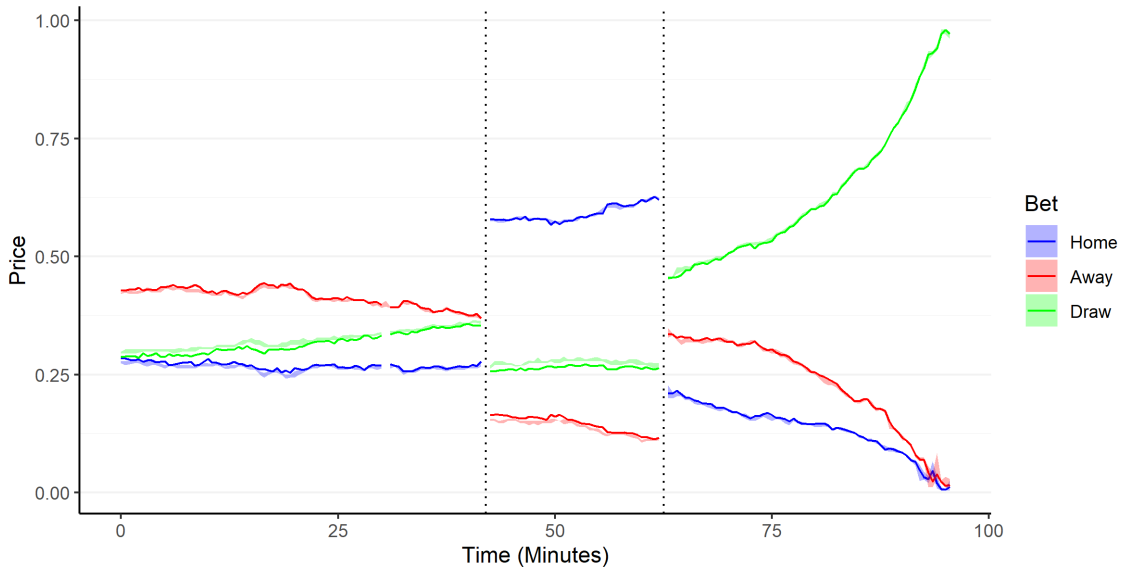


Figure 6.4: Match odds market quotes and calibrated model prices with Frank copula Weibull process model dynamics. The solid lines represent the calibrated model prices of the respective bets. The edges of the shaded areas represent the back and lay prices of said bets. The vertical dotted lines indicate goal times.

close to the back-lay spread throughout the game, confirming the calibration error conclusion that the calibrated model prices are fairly decent for these bets. The calibrated prices seem to become even better in the last third of the game, also agreeing with the slightly decreasing trend observed in the calibration errors. To get a detailed view of the calibrated prices during the beginning of the game, we also here present a zoomed version of Figure 6.4, which is shown in Figure 6.5.

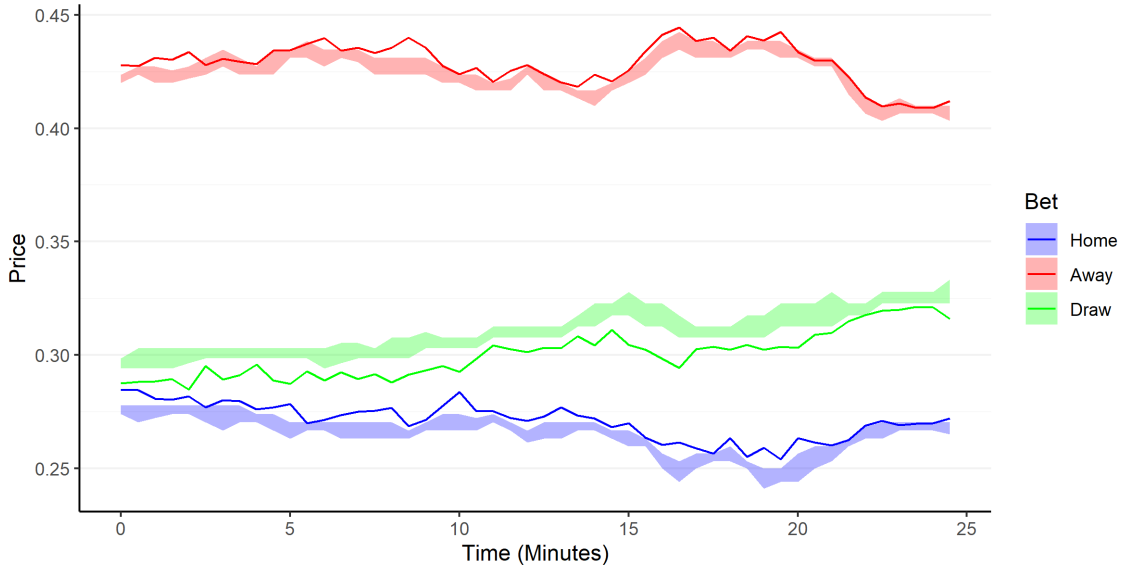


Figure 6.5: A zoomed version of Figure 6.4.

In Figure 6.5, we clearly see that the the calibrated prices are much more satisfying than the independent model's calibrated prices, however, they are still not always within the back-lay

spread, still indicating a flawed model. We also show the calibrated model prices for some over/under market quotes in Figure 6.6 and 6.7. They also seem to yield decent fits, but still indicating some flaws with the model.

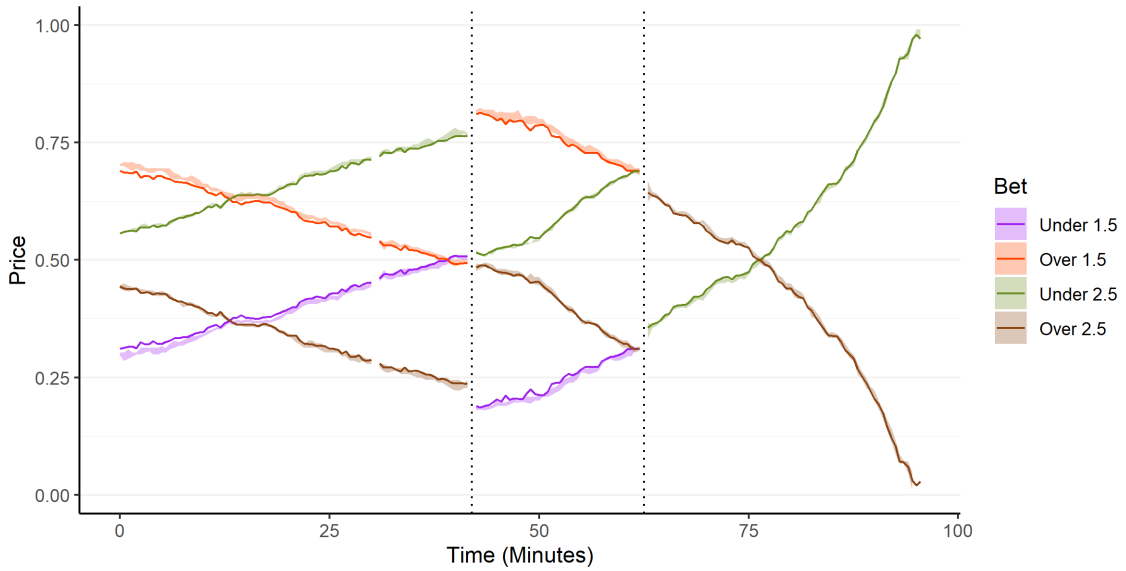


Figure 6.6: Over/under market quotes and calibrated model prices with Frank copula Weibull process model dynamics. The solid lines represent the calibrated model prices of the respective bets. The edges of the shaded areas represent the back-lay prices of said bets. The vertical dotted lines indicate goal times.

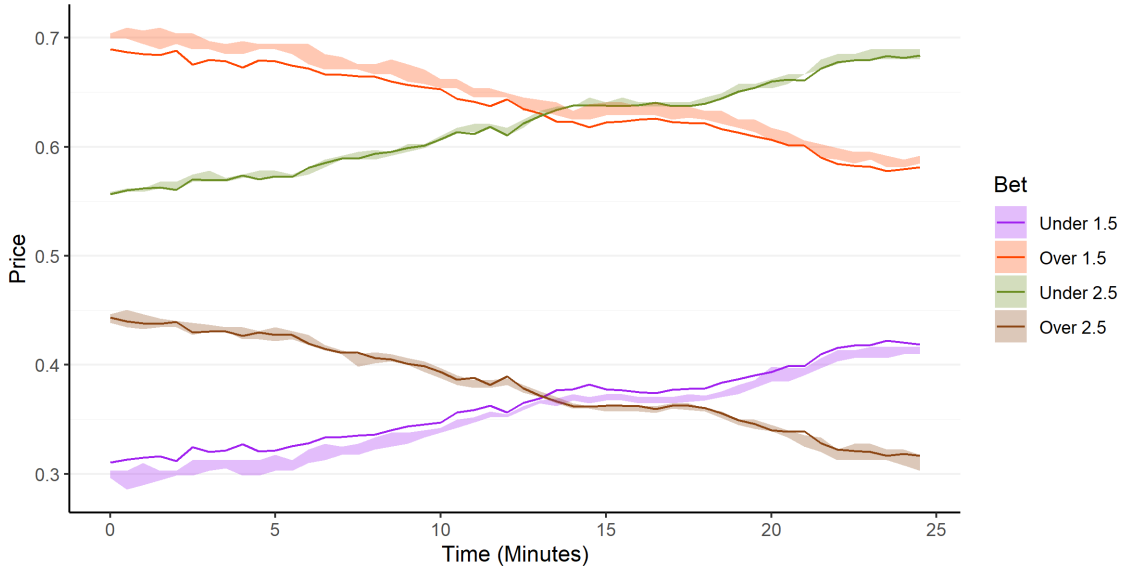


Figure 6.7: A zoomed version of Figure 6.6.

In general, the introduction of possible dependence between the goal processes seems to have increased the calibration accuracy of the model, however, as we also discussed in Section 3.1, the overall problem might lie within the Poisson assumption embedded in the Weibull process. It is also not surprising that the calibration error was significantly reduced, due to the flexibility that an extra parameter adds, as well as the fact that the original model is encapsulated in the extended model.

Parameters and Implied Intensity

Let us for the sake of clarity look at the implied parameters of the bivariate model. Figure 6.8 shows two teams' calibrated model parameters and Figure 6.9 shows the calibrated dependence parameter of the Frank copula.

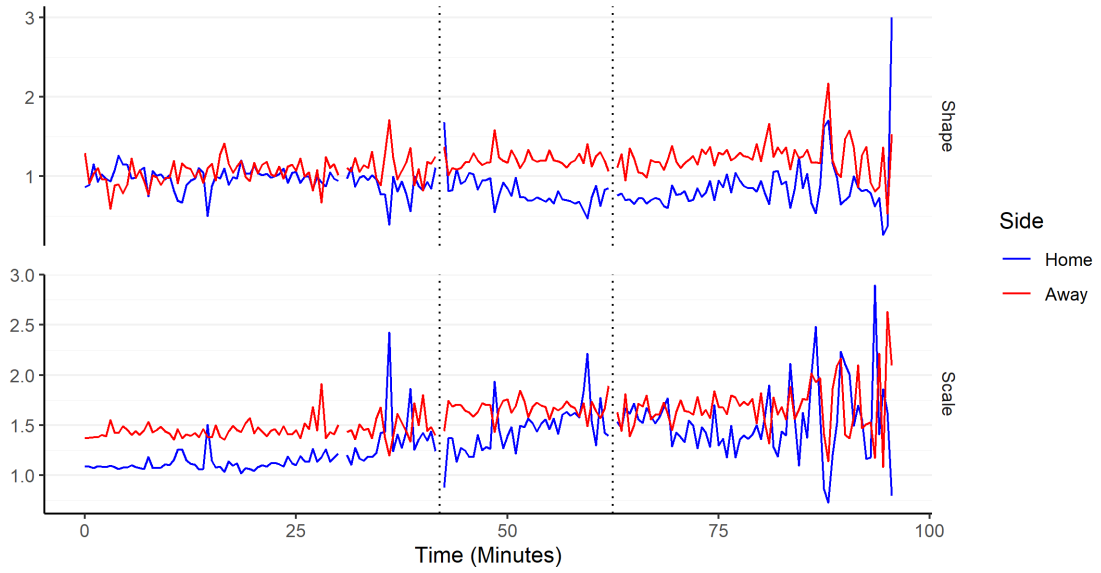


Figure 6.8: Calibrated Weibull process parameters for each side. Top: Shape parameter (β). Bottom: Scale parameter (α). The vertical dotted lines indicate goal times.

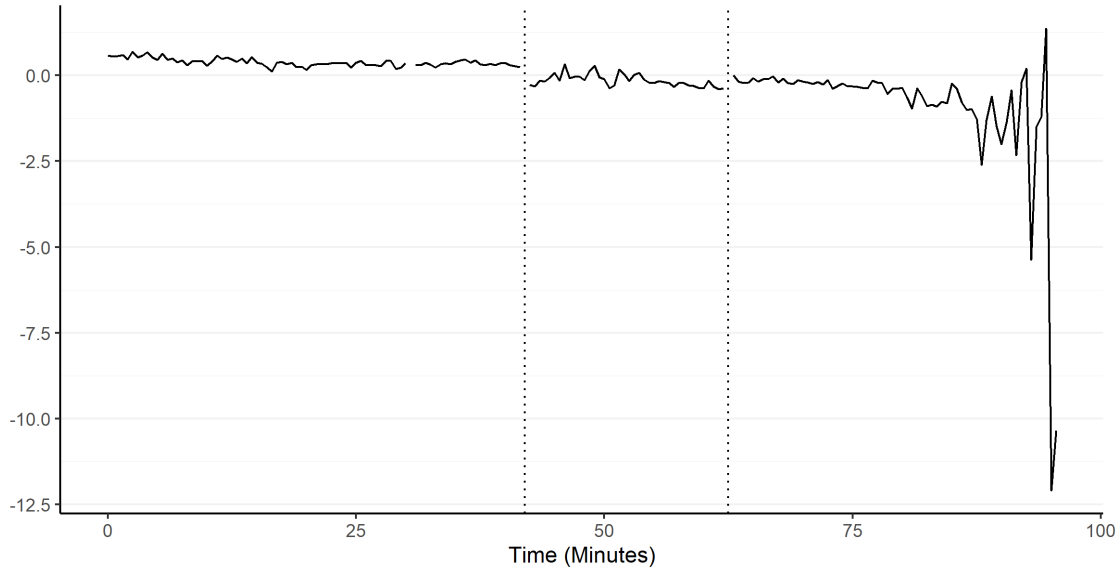


Figure 6.9: Calibrated Frank copula dependence parameter (κ) throughout the match. The vertical dotted lines indicate goal times.

It is notable that both the Weibull process parameters and the dependence parameter seem to be somewhat stable throughout the game with the obvious exception of the very end. We have already touched the reason for some unstable results and parameters during the end, but to sum it all up; it is probably due to the decreasing liquidity of the bets during this period. In Figure 6.9, we see a small positive dependence before the goal, which then turns to an almost

independence-like situation between the goals, which also explains why the calibration errors of the two models almost lined up during this period of the game. After the second goal, we actually see a slight negative dependence.

These observations could indicate that the bettors believe that if a team scores early on in this particular match, then the other team will likely also score a goal, possibly due to some match dynamics that maybe they both could not afford to lose, and thus making the trailing team likely to become more aggressive. Then after the second goal, the negative dependence could indicate a switch in the bettors' mindset, meaning that if one team would score the others would actually have less chance of also scoring, possibly due to the fast-approaching end. This is however also just a possible explanation that fits fairly well into the general football viewer's mind.

In Figure 6.10, we show the implied intensities of the two teams. Here, we observe that there are two very volatile periods at the beginning and end of the game. The volatile period, in the beginning, is fairly surprising; however, by close inspection, it is due to a dip below 1 in the shape parameter of the away team. This along with the rest of the results again suggest a fundamentally flawed model, only suitable for a fast indication.

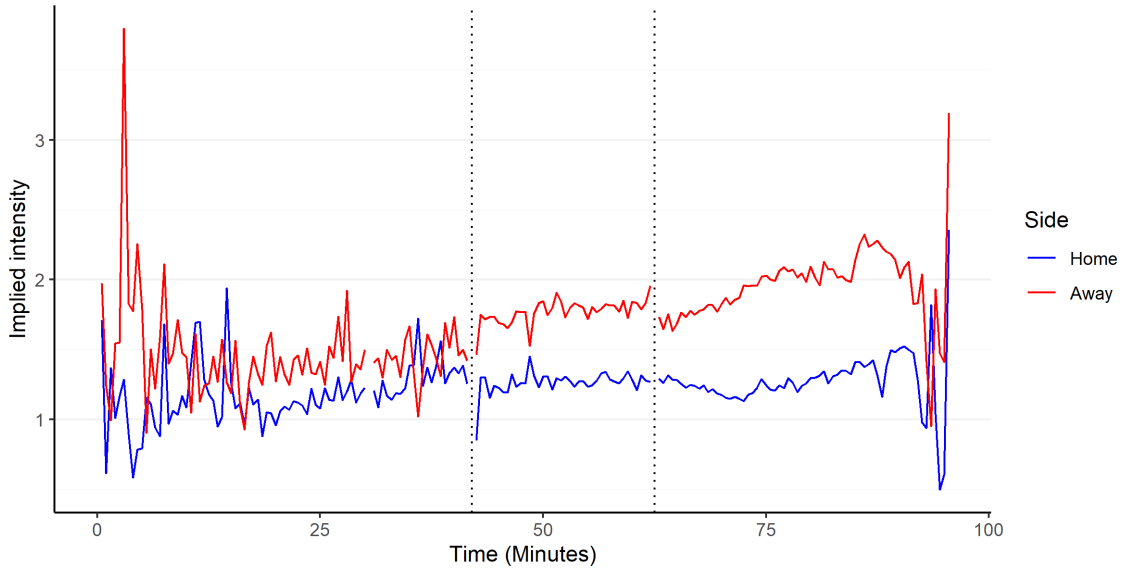


Figure 6.10: Implied Weibull process intensities for each side. The vertical dotted lines indicate goal times.

6.3.1 Final Thoughts

We see significant improvement in the overall model, with much better results than obtained from the independent Weibull process-based market model. However, we see that the bivariate model still falls short in many areas. This is likely because we still have dealt with the major flaw of the Poisson. Again, this model has a specific pricing formula, making it a good quick-and-dirty tool to obtain some information from the current state of the betting market, however, as with the independent Weibull process-based market model, the general assumptions of the football modeling are just not upheld. Therefore, again, we cannot conclude that a risk-neutral

measure $\mathbb{Q}$ exists with this market model. We do, however, note that the potential of a bivariate model can be very high, but also likely very computationally heavy. We will therefore not completely rule out the existence of a risk-neutral measure in, or the usefulness of, the market model proposed in (6.1). We have only tested one method of imposing dependence between the score processes, among the many possible methods; one of which, combined with the right score processes, could be the key to finding near-perfect results. In conclusion, imagination, creativity, and computation time appears to be the limit right now.

7 Conclusion

By studying the statistical patterns of goals scored in English Premier League matches between August 2004 and May 2019, we have demonstrated the usability and limitations of two Weibull-based counting processes for modeling football. In doing so, we first reviewed and introduced the necessary point/counting process theory. In the analysis of the football goals, we found that the distribution of goals could not be explained well with a Poisson distribution due to overdispersion in the empirical data. We did find that the Weibull Count Model was able to capture this overdispersion due to an added parameter. However, both distributions seemed to lack some consistency towards the distribution of goal difference, possibly indicating a form of correlation in the scores. We then found evidence supporting that both the Weibull process and the Weibull renewal process provide long-run goal intensities fairly consistent with the observed intensities of the goals. However, the observed goal intensities are not consistent with the theoretical intensities of the Weibull renewal process that led to the Weibull Count Model consistent with the empirical goal distribution, showing that the Weibull renewal process might be useful but that we should be aware of the possible inconsistencies between the intensity and the score distribution. We then looked at the distribution of waiting times of goals, to see how they compare to the theoretical distributions arriving from the Weibull process and the Weibull renewal process with the parameters used in the goal intensity section. Here, we found that the theoretical distributions had some consistency towards the empirical distributions. We did, however, again find flaws with both models, most notably with the renewal assumption, which does not seem to be completely appropriate for modeling football goals; there are simply too many observations of goals scored in the minutes right after a first goal is scored than the renewal assumption allows. Despite some obvious flaws in both models, we deemed the overall Weibull characteristics fairly decent towards modeling football matches.

We then proceeded to set up a risk-neutral valuation framework for in-play betting on football games, in which we first defined the general market model based on the martingale theory of the compensated counting process. We, furthermore, introduced the specific model dynamics of the market model when the underlying processes are the proposed Weibull-based counting processes. We also proved that, under the assumptions of the general market model, there exists a unique risk-neutral measure for the general market model. We then formally defined a bet and introduced the necessary pricing theory, in which we stated a result on the risk-neutral valuation of bets. We used this result to state pricing formulas for specific market model settings, e.g.

the inhomogeneous Poisson setting (Weibull process). We finished the chapter by stating some hedging and replication results in connection to bets.

Using the risk-neutral valuation result, we calibrated the model prices, as specified by the two Weibull-based dynamics of the market model, to historic betting exchange data. In doing so, we first presented an exploratory data analysis in which we discuss the betting exchange in general and the cleaning of the raw data. We originally obtained 1-second data from the exchange, however, we decided to aggregate it such that we had data with 30-second intervals instead. We then defined and performed the calibration procedure on an English Premier League match between Bournemouth and Southampton. We first showed the calibration results for the market model with Weibull process model dynamics. We obtained decent calibration results with a mean calibration error of 0.8364 and a median of 0.7791. We also noted a decreasing tendency throughout the game with some quite volatile periods. We furthermore noted that the average calibration error of 0.8364 indicated that the calibrated prices are for the most part just outside the back-lay spreads, which was also what we observed when visualizing the calibrated prices. We also looked at the calibrated model parameters and found that they were quite volatile, especially towards the end of the game. Using the calibrated model parameters, we calculated the model implied intensities which for the most part were also quite volatile. All-in-all, the Weibull process provided fair, but not conclusive results, which was also in agreement with the limitations of the model as presented in the statistical analysis of football goals. We could therefore not make any conclusions about the existence of the risk-neutral measure in this market model.

Next, we showed the calibration results for the market model with Weibull renewal process model dynamics. Here, we obtained fairly good calibration results with a mean calibration error of only 0.1704 and a median of 0.1483, meaning that almost all the calibrated prices were inside the back-lay spreads. However, due to the increase in parameters, we needed to be careful about making conclusions just yet, despite the promising results with the calibration errors. Each team's calibrated model parameters were surprisingly stable until the goals, where they became very volatile. This is likely connected to the renewal assumption where each team's intensity completely starts over at the time of their goal, which is likely not accurate in football. Again, using the calibrated model parameters, we calculated the model implied intensities which also portray the volatilities around the renewal times. All-in-all, the Weibull renewal process showed fairly good results, but again, not great results. This was also in agreement with the limitations of the model as presented in the statistical analysis of football goals. It did, however, show promising results toward the possibility of the existence of a risk-neutral measure, but we could just not make any concrete conclusions about it in this market model either. It did, however, create a promising direction for future studies on the possible application of the Fundamental Theorems of Assets Pricing to the market of in-play football betting.

Lastly, we discussed the independence assumption of the general market model and proposed an extension of the market model, in which the underlying counting processes are allowed to be correlated. Based on a recent article, we furthermore suggested the use of the Frank copula to model the dependence. We then constructed a specific pricing formula of the market model with possible correlated Weibull processes and calibrated the model prices to the historical data. We

saw clear improvements in the calibration errors, however, the proposed extension did not bring the average error below the 0.5 threshold. Also, the parameter volatility did not seem to be improved by the allowed correlation and furthermore, we observed fairly volatile behavior in the dependence parameter, especially towards the end. In conclusion, despite a clear improvement compared to the independence model, we found that the bivariate Weibull process with a Frank copula still lacked consistency in many areas, and thus we did not alter the conclusion of not enough evidence to support the existence of a risk-neutral measure with these market dynamics. The general bivariate market model does, however, also present very promising directions for further research. We do however suspect that computation time will be the Achilles' heel to further advancements.

To sum up our conclusions; we have proposed a theoretical general market model which has shown great potential in the pricing of football bets, but, as with any other pricing model, the practicality of the model is bounded by the accuracy to portray the underlying assets, which in our case is the live modeling of the underlying football game, something which the two proposed Weibull-based point processes are not perfectly able to do.

Bibliography

Aalen, Odd (1978). "Nonparametric inference for a family of counting processes". In: *The Annals of Statistics*, pp. 701–726.

Ames, Nick (2017). *Bournemouth 1-1 Southampton: Premier League - as it happened.* URL: https://www.theguardian.com/football/live/2017/dec/03/bournemouth-v-southampton-premier-league-live?page=with:block-5a23fc67ae3bcb05d202416b#block-5a23fc67ae3bcb05d202416b.

Andersen, Morten and Maillard, Marting G. R. (2019). *Mathematical Finance in Football Betting: A Risk-neutral Approach.* AAU Semester Project.

Angelini, Giovanni and Angelis, Luca De (2019). "Efficiency of online football betting markets". In: *International Journal of Forecasting* 35.2, pp. 712–721.

Baradel, Nicolas (2018). *rpgm: Fast Simulation of Normal/Exponential Random Variables and Stochastic Differential Equations / Poisson Processes.* R package version 1.1.3. URL: https://pgm-solutions.com/packages.

Barndorff-Nielsen, Ole E. et al. (2009). "Realized kernels in practice: Trades and quotes". In: *The Econometrics Journal* 12.3, pp. C1–C32.

Bauwens, Luc, Hafner, Christian M., and Laurent, Sébastien (2012). *Handbook of Volatility Models and Their Applications.* John Wiley & Sons.

Berry, Caan (2017). *How Does Cross Matching Work on Betfair? Full Explanation:* URL: https://caanberry.com/how-does-cross-matching-work-on-betfair/.

Betfair (2020). *Exchange Historical Data.* URL: https://historicdata.betfair.com.

Björk, Tomas (2009). *Arbitrage Theory in Continuous Time.* Third edition. ISBN 978-0-19-957474-2. Oxford University Press.

– (2011). *An Introduction to Point Processes from a Martingale Point of View.* KTH, Lecture Notes.

Boshnakov, Georgi, Kharrat, Tarak, and McHale, Ian G. (2017). "A Bivariate Weibull Count Model for Forecasting Association Football Scores". In: *International Journal of Forecasting* 33.2, pp. 458–466.

Brémaud, Pierre (1981). *Point Processes and Queues: Martingale Dynamics.* Springer.

Brown, Alasdair and Yang, Fuyu (2017). "The Role of Speculative Trade in Market Efficiency: Evidence from a Betting Exchange". In: *Review of Finance* 21.2, pp. 583–603.

Casarin, Roberto (2005). "Stochastic Processes in Credit Risk Modelling". In: *Available at SSRN 888633*.

Cha, Ji Hwan and Finkelstein, Maxim (2018). *Point Processes for Reliability Analysis: Shocks and Repairable Systems*. Springer.

Cherubini, Umberto, Luciano, Elisa, and Vecchiato, Walter (2004). *Copula Methods in Finance*. John Wiley & Sons.

Daley, D.J. and Vere-Jones, D. (2003). *An Introduction to the Theory of Point Processes: Volume I: Elementary Theory and Methods*. Springer.

Delbaen, Freddy and Schachermayer, Walter (1994). "A general version of the fundamental theorem of asset pricing". In: *Mathematische Annalen* 300.1, pp. 463–520.

Deutscher, Christian, Frick, Bernd, and Ötting, Marius (2018). "Betting market inefficiencies are short-lived in German professional football". In: *Applied Economics* 50.30, pp. 3240–3246.

Divos, Peter et al. (2018). "Risk-Neutral Pricing and Hedging of In-Play Football Bets". In: *Applied Mathematical Finance* 24.4, pp. 315–335.

Dixon, Mark J. and Coles, Stuart G. (1997). "Modelling association football scores and inefficiencies in the football betting market". In: *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 46.2, pp. 265–280.

Dixon, Mark J. and Robinson, Michael E. (1998). "A birth process model for association football matches". In: *Journal of the Royal Statistical Society: Series D (The Statistician)* 47.3, pp. 523–538.

Feng, Guanhao, Polson, Nicholas G., and Xu, Jianeng (2016). "The Market for English Premier League (EPL) Odds". In: *Journal of Quantitative Analysis in Sports* 12.4, pp. 167–178.

Hastie, Trevor, Tibshirani, Robert, and Friedman, Jerome (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media.

Haugh, Martin (2016). *IEOR E4602: Quantitative Risk Management - An Introduction to Copulas*. IEOR Columbia - Columbia University, Lecture Notes.

Hautsch, Nikolaus (2012). *Econometrics of Financial High-Frequency Data*. Springer Science & Business Media.

Jeanblanc, Monique, Yor, Marc, and Chesney, Marc (2009). *Mathematical Methods for Financial Markets*. Springer Science & Business Media.

Kaplan, Daniel and Pruim, Randall (2020). *ggformula: Formula Interface to the Grammar of Graphics*. R package version 0.9.4. URL: `https://CRAN.R-project.org/package=ggformula`.

Karlis, Dimitris and Ntzoufras, Ioannis (2000). "On modelling soccer data". In: *Student* 3.4, pp. 229–44.

– (2003). "Analysis of sports data by using bivariate Poisson models". In: *Journal of the Royal Statistical Society: Series D (The Statistician)* 52.3, pp. 381–393.

Karr, Alan (1991). *Point Processes and Their Statistical Inference*. Second edition. Marcel Dekker, Inc.

Kharrat, Tarak et al. (2019). "Flexible Regression Models for Count Data Based on Renewal Processes: The Countr Package". In: *Journal of Statistical Software* 90.13, pp. 1–35. DOI: 10.18637/jss.v090.i13.

Klebaner, Fima C. (2005). *Introduction to Stochastic Calculus With Applications*. Second edition. Imperial College Press.

Koopman, Siem Jan and Lit, Rutger (2015). "A dynamic bivariate Poisson model for analysing and forecasting match results in the English Premier League". In: *Journal of the Royal Statistical Society: Series A (Statistics in Society)* 178.1, pp. 167–186.

Lindstrøm, Jonas Christoffer (2020). *implied: Convert Bookmaker Odds to Probabilities*. R package version 0.3.0. URL: https://CRAN.R-project.org/package=implied.

Maher, Michael J (1982). "Modelling association football scores". In: *Statistica Neerlandica* 36.3, pp. 109–118.

McHale, Ian and Scarf, Phil (2007). "Modelling soccer matches using bivariate discrete distributions with general dependence structure". In: *Statistica Neerlandica* 61.4, pp. 432–445.

– (2011). "Modelling the dependence of goals scored by opposing teams in international soccer matches". In: *Statistical Modelling* 11.3, pp. 219–236.

McShane, Blake et al. (2008). "Count Models Based on Weibull Interarrival Times". In: *Journal of Business & Economic Statistics* 26.3, pp. 369–378.

Microsoft and Weston, Steve (2019). *doParallel: Foreach Parallel Adaptor for the 'parallel' Package*. R package version 1.0.15. URL: https://CRAN.R-project.org/package=doParallel.

– (2020). *foreach: Provides Foreach Looping Construct*. R package version 1.5.0. URL: https://CRAN.R-project.org/package=foreach.

Müller, Kirill and Wickham, Hadley (2019). *tibble: Simple Data Frames*. R package version 2.1.3. URL: https://CRAN.R-project.org/package=tibble.

Murthy, D.N. Prabhakar, Xie, Min, and Jiang, Renyan (2004). *Weibull Models*. John Wiley & Sons.

Nelsen, Roger B. (2006). *An Introduction to Copulas*. Second edition. Springer Science & Business Media.

Nordsted, Pete (2009). *Mastering Betfair: How to make serious money trading betting exchanges*. Harriman House Limited.

Pedersen, Jan (2017). *Lévy Processes*. Aarhus University.

Premier League (2020). *Premier League Football Scores, Results & Season Archives*. URL: `https://www.premierleague.com/results`.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: `https://www.R-project.org/`.

Rinne, Horst (2008). *The Weibull Distribution: A Handbook*. Chapman and Hall/CRC.

Ross, Sheldon M. (2019). *Introduction to Probability Models*. Twelfth edition. Academic Press.

Segall, Adrian and Kailath, Thomas (1975). "The Modeling of Randomly Modulated Jump Processes". In: *IEEE Transactions on Information Theory* 21.2, pp. 135–143.

Shiryaev, Albert and Liptser, Robert (2001). *Statistics of Random Processes: II. Applications*. Second edition. Springer Science & Business Media.

Shreve, Steven E. (2004). *Stochastic Calculus for Finance II: Continuous-Time Models*. Springer Science & Business Media.

Sigman, Karl (2009). *IEOR 6711: Notes on the Poisson Process*. IEOR Columbia - Columbia University, Lecture Notes.

Skellam, John G (1946). "The frequency distribution of the difference between two Poisson variates belonging to different populations." In: *Journal of the Royal Statistical Society. Series A (General)* 109.Pt 3, pp. 296–296.

Sokol, Alexander and Hansen, Niels Richard (2015). "Exponential Martingales and Changes of Measure for Counting Processes". In: *Stochastic Analysis and Applications* 33.5, pp. 823–843.

stevecheng (2013). *conditional expectation under change of measure*. URL: `https://planetmath.org/conditionalexpectationunderchangeofmeasure`.

Tankov, Peter and Cont, Rama (2004). *Financial Modelling With Jump Processes*. Chapman & HALL/CRC.

Trivedi, Pravin and Zimmer, David (2017). "A Note on Identification of Bivariate Copulas for Discrete Count Data". In: *Econometrics* 5.1, p. 10.

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S*. Fourth. ISBN 0-387-95457-0. New York: Springer. URL: `http://www.stats.ox.ac.uk/pub/MASS4`.

Wickham, Hadley (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN: 978-3-319-24277-4. URL: `https://ggplot2.tidyverse.org`.

Wickham, Hadley et al. (2020). *dplyr: A Grammar of Data Manipulation*. R package version 0.8.4. URL: https://CRAN.R-project.org/package=dplyr.

Williams, Leighton Vaughan (2005). *Information Efficiency in Financial and Betting Markets*. Cambridge University Press.

Zhu, Hao (2019). *kableExtra: Construct Complex Table with 'kable' and Pipe Syntax*. R package version 1.1.0. URL: https://CRAN.R-project.org/package=kableExtra.

Probability & Distributions A

This appendix is based on Ross (2019, pp. 42, 46), stevecheng (2013), Tankov and Cont (2004, pp. 44–47), & Murthy, Xie, and Jiang (2004, p. 10).

In this appendix, we recall some fundamental results of probability theory and present certain distributions used extensively in the thesis.

Proposition A.1 (Law of the Unconscious Statistician)**.** *Let X be a random variable and g be a real-valued function, then*

$$\mathbb{E}\left[g(X)\right] = \begin{cases} \sum_{x} g(x) f_X(x) & \text{for } X \text{ discrete} \\ \int_{-\infty}^{\infty} g(x) f_X(x) dx & \text{for } X \text{ continuous.} \end{cases}$$

Furthermore, if X and Y are random variables and g is a function of two variables, then

$$\mathbb{E}\left[g(X,Y)\right] = \begin{cases} \sum_{y} \sum_{x} g(x,y) f(x,y) & \text{for } (X,Y) \text{ discrete} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f(x,y) dx & \text{for } (X,Y) \text{ continuous.} \end{cases}$$

Conditional Expectation

Let $\mathbb{P}$ be a given probability measure on some σ-algebra $\mathscr{F}$, and $X : \Omega \to \mathbb{R}$ a real random variable with $\mathbb{E}\left[|X|\right] < \infty$.

Definition A.2 (Conditional Expectation)
Let $\mathscr{G} \subseteq \mathscr{F}$ be a sub-σ-algebra, then the *conditional expectation* of X given $\mathscr{G}$, denoted by $\mathbb{E}\left[X \mid \mathscr{G}\right]$, is any $\mathscr{G}$-measurable function $\Omega \to \mathbb{R}$ that satisfies

$$\int_G \mathbb{E}\left[X \mid \mathscr{G}\right] d\mathbb{P} = \int_G X d\mathbb{P}, \quad \text{for all } G \in \mathscr{G}.$$

Now suppose that a new probability measure $\mathbb{Q}$ is defined by

$$d\mathbb{Q} = L\, d\mathbb{P},$$

using some $\mathscr{F}$-measurable random variable L as the Radon-Nikodym derivative, i.e. we have that $L \geq 0$ almost surely, and $\mathbb{E}[L] = 1$.

Theorem A.3 (Conditional Expectation Under Change of Measure)**.** *Let $\boldsymbol{G} \subseteq \boldsymbol{F}$ be any sub-σ-algebra. For any $\boldsymbol{F}$-measurable random variable X,*

$$\mathbb{E}[L \mid \mathscr{G}]\, \mathbb{E}_{\mathbb{Q}}[X \mid \mathscr{G}] = \mathbb{E}[LX \mid \mathscr{G}].$$

Exponential Distribution

Definition A.4 (Exponential Distribution)
A positive random variable X is said to follow an *exponential distribution* with parameter $\lambda > 0$ if its probability density function has the form

$$f(x) = \lambda \mathrm{e}^{-\lambda x}\, \mathbb{1}(x \geq 0).$$

Then, we write $X \sim \mathrm{Exp}(\lambda)$.

The distribution function F of an exponential distributed random variable X is then given by

$$F(x) = 1 - \mathrm{e}^{-\lambda x}, \quad \forall x \in \mathbb{R}_+.$$

An exponential distributed random variable has a unique important property, namely, the memoryless property which we will described in the following proposition.

Proposition A.5 (Absence of Memory)**.** *Let $\mathcal{T} \geq 0$ be a random variable such that*

$$\mathbb{P}(\mathcal{T} > t + s \mid \mathcal{T} > t) = \mathbb{P}(\mathcal{T} > s), \quad \forall t, s > 0.$$

Then, $\mathcal{T}$ has an exponential distribution.

Poisson Distribution

Definition A.6 (Poisson Distribution)
A random variable X is said to follow a *Poisson distribution* with parameter λ if its probability mass function has the form

$$p(n) = \mathbb{P}(X = n) = \mathrm{e}^{-\lambda} \frac{\lambda^n}{n!}, \quad \forall n \in \mathbb{N}_0.$$

Then, we write $X \sim \mathrm{Poi}(\lambda)$.

The following proposition presents a rather special relationship between the Poisson distribution and sums of independent exponential random variables.

Proposition A.7. *If* $\{\mathcal{T}_i\}_{i\geq 1}$ *are independent exponential random variables with parameter* λ*, then for any* $t > 0$ *the random variable*

$$N_t = \inf\left\{n \geq 1, \sum_{i=1}^{n} \mathcal{T}_i > t\right\},$$

follows a Poisson distribution with parameter λt*.*

Weibull Distribution

Definition A.8 (Weibull Distribution)
A positive random variable X is said to follow a *Weibull distribution* with scale (rate) parameter $\theta > 0$ and shape parameter $\beta > 0$ if its probability density function has the form

$$f(x) = \frac{\beta}{\theta}\left(\frac{x}{\theta}\right)^{\beta-1} \exp\left(-(x/\theta)^{\beta}\right) \mathbb{1}\left(x \geq 0\right).$$

Then, we write $X \sim \text{Weibull}\,(\theta, \beta)$.

The distribution function for the Weibull distribution is

$$F(x) = 1 - \exp\left(-(x/\theta)^{\beta}\right), \quad \forall x \in \mathbb{R}_+.$$

Sometimes an alternative parameterization is employed, in which the parameters are β as above and $\alpha = \left(\frac{1}{\theta}\right)^{\beta}$, that is the density function becomes:

$$f(x) = \alpha\beta x^{\beta-1} \exp\left(-\alpha x^{\beta}\right) \mathbb{1}\left(x \geq 0\right).$$

and the distribution function becomes:

$$F(x) = 1 - \exp\left(-\alpha x^{\beta}\right), \quad \forall x \in \mathbb{R}_+.$$

Asset Pricing & SDEs B

This appendix is based on Shreve (2004, pp. 228–232), Björk (2011, pp. 20, 27), Jeanblanc, Yor, and Chesney (2009, pp. 457–459, 551–552), & Tankov and Cont (2004, pp. 298–300).

Here, we present some brief results and definitions on stochastic integrals, stochastic differential equations, and general asset pricing. Consult the cited literature for a more rigorous approach.

Stochastic Integrals

When dealing with stochastic counting processes it is beneficial to characterize the *stochastic integral* $\int_0^t C_s dN_s$, which is defined as a Stieltjes integral for every bounded measurable process (not necessarily adapted) $(C_t)_{t\geq 0}$ by:

$$\int_0^t C_s dN_s = \int_{]0,t]} C_s dN_s = \sum_{n=1}^{\infty} C_{t_n} \mathbb{1}\left(t_n \leq t\right). \tag{B.1}$$

We note that the integral (B.1) is finite due to the finite number of jumps during the interval $]0, t]$. Sometimes, we also used the notation

$$\int_0^t C_s dN_s = \sum_{s\leq t} C_s \Delta N_s.$$

Proposition B.1. *Assume that M is a martingale of bounded variation and that h is a predictable process satisfying the condition*

$$\mathbb{E}\left[\int_0^t |h_s|\, dM_s < \infty\right], \quad \forall t \geq 0.$$

Then, the process X defined by

$$X_t = \int_0^t h_s dM_s,$$

is a martingale.

Stochastic Differential Equations

A *stochastic differential equation* (SDE) is an equation of the form

$$\begin{cases} dX_t &= \mu(t, X_t)dt + \sigma(t, X_t)dW_t + \nu(t, X_t)dM_t, \\ X_0 &= x_0, \end{cases} \tag{B.2}$$

for given functions μ, σ, and ν that must satisfy some conditions[1]. The SDE (B.2) should be viewed as an informal way of expressing the corresponding integral equation

$$X_{t+s} - X_t = \int_t^{t+s} \mu(u, X_u)du + \int_t^{t+s} \sigma(u, X_u)dW_u + \int_t^{t+s} \nu(u, X_u)dM_u.$$

The continuous martingale part of the semimartingale X is $\int_0^t \sigma(u, X_u)dW_u$, and the purely discontinuous martingale part is $\int_0^t \nu(u, X_u)dM_u$.

Note that some authors prefer to write the dynamics of the SDE (B.2) using the Poisson process N instead of the compensated martingale M when dealing with jump-diffusion processes where the purely discontinuous part of the semimartingale is given by a compensated Poisson process. We shall also employ this notation in the following proposition.

Proposition B.2. *Assume that X satisfies the SDE*

$$\begin{cases} dX_t &= \alpha_t X_{t-}dt + \beta_t X_{t-}dN_t), \\ X_0 &= x_0, \end{cases} \tag{B.3}$$

where α and β are predictable processes. Then X has the solution

$$X_t = x_0 \exp\left(\int_0^t \alpha_s ds\right) \exp\left(\int_0^t \ln(1 + \beta_s)\, dN_s\right). \tag{B.4}$$

Fundamental Theorems of Asset Pricing

Theorem B.3 (First Fundamental Theorem of Asset Pricing[2])**.** *The market model defined by $(\Omega, \mathscr{F}, (\mathscr{F}_t)_{t\in[0,T]}, \mathbb{P})$ and asset prices $\left(S_t^i\right)_{t\in[0,T]}$ is arbitrage-free if and only if there exists a risk-neutral measure $\mathbb{Q}$.*

Theorem B.4 (Second Fundamental Theorem of Asset Pricing)**.** *A market defined by the assets $\left(B_t, S_t^1, S_t^2\right)_{t\in[0,T]}$, described as stochastic processes on $(\Omega, \mathscr{F}, (\mathscr{F}_t)_{t\in[0,T]}, \mathbb{P})$ is complete if and only if there is a unique risk-neutral measure $\mathbb{Q}$.*

[1] See e.g. Jeanblanc, Yor, and Chesney (2009, p. 551)

[2] A more general version of this theorem is given by Delbaen and Schachermayer (1994, Thm. 1.1, p. 467), requiring a broader definition of "no-arbitrage", namely, the "no-arbitrage with vanishing risk".