

Bayesian analysis of stochastic point processes for financial applications

Cornelius Probst

St. Cross College

University of Oxford



Thesis submitted for the degree of Doctor of Philosophy

Trinity Term 2013

Bayesian analysis of stochastic point processes for financial applications

Cornelius Probst, St. Cross College

DPhil, Trinity Term 2013

A recent application of point processes has emerged from the electronic trading of financial assets. Many securities are now traded on purely electronic exchanges where demand and supply are aggregated in limit order books. Buy and sell trades in the asset as well as quote additions and cancellations can then be interpreted as events that not only determine the shape of the order book, but also define point processes that exhibit a rich internal structure.

A large class of such point processes are those driven by a diffusive intensity process. A flexible choice with favourable analytic properties is a Cox-Ingersoll-Ross (CIR) diffusion. We adopt a Bayesian perspective on the statistical inference for these doubly stochastic processes, and focus on filtering the latent intensity process. We derive analytic results for the moment generating function of its posterior distribution. This is achieved by solving a partial differential equation for a linearised version of the filtering equation. We also establish an efficient and simple numerical evaluation of the posterior mean and variance of the intensity process. This relies on extending an equivalence result between a point process with CIR-intensity and a partially observed population process.

We apply these results to empirical datasets from foreign exchange trading. One objective is to assess whether a CIR-driven point process is a satisfactory model for the variations in trading activity. This is answered in the negative, as sudden bursts of activity impair the fit of any diffusive intensity model. Controlling for such spikes, we conclude with a discussion of the stochastic control of a market making strategy when the only information available are the times of buy and sell trades.

Contents

1	Introduction	1
2	High frequency data for limit order books	5
2.1	Point process models for financial order books	5
2.2	Two datasets from foreign exchange trading	10
2.2.1	The EBS Life Cycle data	12
2.2.2	The EBS Live dataset	16
2.2.3	Compatibility of the two datasets	17
2.2.4	Reconstruction of the entire order book	19
3	Point processes in statistical modelling	23
3.1	The intensity approach to point processes	25
3.2	Poisson processes	27
3.2.1	Deterministic intensities: homogeneous and inhomogeneous Poisson processes	27
3.2.2	Markov modulated Poisson processes	31
3.3	Point processes that are not Poisson	32
3.3.1	Interaction via the last event time	33
3.3.2	Interaction in space	35
3.3.3	Interaction in time	40
4	Continuous-time filtering of dynamical systems	43
4.1	Review of Bayesian inference for dynamical systems with Gaussian noise . .	44
4.1.1	Discrete-time inference in hidden Markov models	46

4.1.2	Continuous-time inference by conditional expectations	52
4.2	Bayesian inference for dynamical systems with point process observations .	56
4.2.1	Discretisation of time	56
4.2.2	Poisson processes driven by discrete-valued Markov processes	60
4.2.3	Inference for diffusion driven Poisson processes	65
5	Results for the filter equations for CIR-driven Poisson processes	71
5.1	Explicit solutions to the filtering equations between events	74
5.2	An equivalence result involving the death sequence in an IBD process . . .	81
6	Numerical results	95
6.1	Using the equivalence result	96
6.2	Asymptotic approximation of the first two moments	100
6.3	A particle filter for the discretised dynamical system	102
6.4	An EWMA filter	105
6.5	An application: Predicting the next event time	106
7	Parameter estimation and model fit	111
7.1	Parameter estimation	112
7.1.1	Maximum likelihood estimation	112
7.1.2	Method of Moments	115
7.2	Assessing the fit of a point process model	118
8	Event times in a limit order book	123
8.1	Parameter estimation	124
8.1.1	Using the maximum likelihood estimates	124
8.1.2	Using the method of moments	127
8.2	Assessing the model fit	128
8.3	Alternative models and adjustments	132
9	Backtesting of a market making strategy	139
9.1	Optimal market making as an intensity control problem	141

9.2	Description of the backtesting framework	144
9.3	Market making with a signal informed by the trade flow	150
9.3.1	Using the filtering results for CIR-driven point processes	154
9.3.2	Comparison of different choices for the market making signal	159
9.4	Conclusion	160

1 Introduction

A (one-dimensional) *point process* on the positive real line typically arises whenever a sequence of event times is recorded. Point processes have been studied extensively in probability theory and been used in statistical modelling for several decades. Examples include renewal processes whose events are given by the times of failure of a component in a system that is immediately renewed upon failure (Feller, 1950, Chapter XIII) or arrivals and departures from a queue waiting for service (Erlang, 1920; Kendall, 1953).

A more recent application of point processes has emerged from the electronic trading of financial assets. Due to technological progress and increased competition, many securities (such as shares, interest rate futures or currencies) are now traded on purely electronic exchanges where demand and supply are aggregated in *limit order books*. Buy and sell trades in the asset as well as quote additions and cancellations can then be interpreted as events that not only determine the shape of the order book, but also define point processes that exhibit a rich internal structure. Chapter 2 presents a brief introduction to order books and the terminology used by practitioners. We believe that comprehensive and fully realistic models for order books are too complex to make progress with, and that selected phenomena and abstract properties of the order book dynamics should be studied instead. A concrete example is the concept of *liquidity* (or trading activity) that one can identify with the arrival rates of buy and sell trades in the asset. Chapter 2 also describes two empirical datasets that represent four days of foreign exchange trading in the EUR/USD currency pair. These datasets will be revisited and analysed in subsequent chapters.

A key characteristic of many point processes, not necessarily those related to order books, is that they exhibit time or state dependent, non-stationary behaviour. Under weak regularity conditions, this fluctuation can be captured by an *intensity process* adapted to the point process. Chapter 3 reviews the mathematical theory relating to this approach, and presents several joint specifications of the intensity and the point process. The most important choice is that of a *doubly stochastic process* in which a *diffusion path* of the intensity process is chosen first; and then, conditional on that trajectory, a path of the point process evolves according to the Poisson law. Other models that allow for a bi-directional interaction, i.e. models in which the point process influences the intensity process, are also considered.

We adopt a Bayesian perspective on the analysis of such models, with particular focus on inference for the intensity process. Chapter 4 reviews classic results in filtering theory, most notably the Kalman filter. The existence of such an analytic solution to the filtering problem is exceptional and due to a fortunate description of the underlying dynamical system by a linear stochastic differential equation (SDE) with deterministic coefficients. The filtering equation for a diffusion-driven point process generally does not allow for closed-form solutions, as the intensity process associated with a point process is required to remain non-negative and must hence exhibit non-linear dynamics.

Chapter 5 demonstrates that analytic results can be established in a particular situation, namely when the intensity process is given by a *Cox-Ingersoll-Ross (CIR) diffusion*. This is a convenient choice from a modelling perspective as a CIR diffusion is positive, and has parameters for the mean and variance of its stationary distribution and for the degree of mean reversion. Our results also allow for a simple numerical implementation of a recursive filtering algorithm. This is due to a combination of two results. First, the filtering equation for the intensity process can be linearised using a technique proposed in a similar context by Boel and Beneš (1980). This results in a partial differential equation (PDE) for the posterior moment generating function that can be solved between event times. Secondly, we extend a result obtained by Clifford and Wei (1993) on the equivalence of a point process with CIR-intensity and the death process in an *immigration-birth-death (IBD)*

process. This allows for an efficient and simple numerical evaluation of posterior mean and variance of the intensity process, as a *discrete* population process can be studied instead. A comparison of this technique to other approximative filtering approaches is presented in Chapter 6.

For the remainder of the thesis, we return to the empirical datasets introduced in Chapter 2. Our objective is to assess whether a CIR-driven point process is a satisfactory model for the variations in trading activity, and how knowledge about this latent intensity process can be exploited for the purposes of *market making*.

Chapter 7 is preliminary for fitting our model to empirical data. We review the maximum likelihood estimator for CIR-driven point processes established by Wei (1995), and introduce an alternative parameter estimation based on the method of moments. Combined with an estimate of the intensity process, this yields graphical diagnostics for the model fit that are easy to interpret.

As announced, we adopt a CIR-driven point process as statistical model for the sequence of buy trades in the foreign exchange dataset introduced earlier. The estimation of the model parameters and the calculation of the mean conditional intensity is described in Chapter 8. We discover that sudden bursts of activity in the order book invalidate the parameter estimates obtained by maximum likelihood, and impair the overall fit of any diffusion driven point process model for the event times of trades. Alternative models that allow for self-excitation of the point process are considered but also found wanting.

Chapter 9 concludes by considering the question of stochastic control of a simple market making strategy. We argue that a formulation as *intensity control problem* is appropriate, as the financial institution making a market needs to adjust its quotes in such a way that the negotiated buy and sell trades occur in a pattern that is likely to yield a balanced *total inventory*. We investigate how this can be achieved in practice if the only information available are the event times of all buy and sell events.

This thesis was supervised by Dr. Peter Clifford at the Department of Statistics and financially supported by the EPSRC. Both are gratefully acknowledged.

2 High frequency data for limit order books

Technological progress has greatly improved our ability to record and process time series data of complex nature at increasingly higher frequencies. This has led to the routine accumulation of vast data sets in a wide variety of applications: In finance, hundreds of assets are traded on exchanges around the world, with prices fluctuating seemingly continuously and generating torrents of data. *High-frequency trading (HFT)* is concerned with the statistical analysis of such data to inform trading decisions *in real time*.

This chapter introduces one of the richest sources of such data, namely *(financial) order books*. Section 2.1 adopts a popular mathematical description of such order books as a system of interacting queues. A model that truly and fully mirrors the complexities of the financial markets is however too complex to make progress with. Instead we focus on models for selected phenomena and abstract properties of order book dynamics. Section 2.2 then describes two datasets representing the state and evolution of such an order book at two different levels of granularity and detail.

2.1 Point process models for financial order books

In earlier times, the finest temporal resolution of financial data was often in *tick time*: the only information available was the last traded price of the asset, and this information would only update when a trade at a different price (also called a “tick”) occurred. The

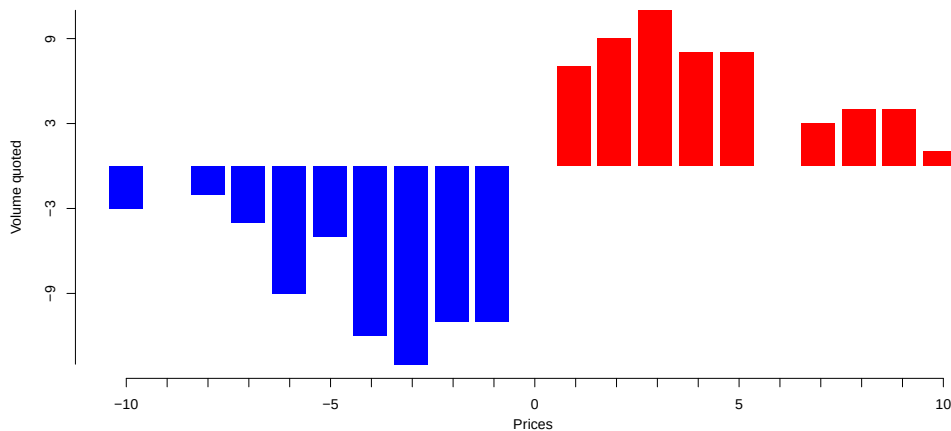


Figure 2.1 Stylised snapshot of an order book. The buy orders (“the bids”) are plotted with negative sizes at prices $-10, \dots, -1$; the sell orders (“the offers”) with positive sizes at prices $1, \dots, 10$. The best bid price is the highest price at which a non-zero bid is placed (here: -1); the best ask is the lowest price at which a non-zero offer is placed (here: $+1$). The *mid price* is the arithmetic mean of best bid and best ask; the *(bid-ask) spread* is the difference between best ask and best bid. The lack of volume quoted at prices further away from the mid price, and the hump-shape of the bid and offer side are features that appear often in real data.

popularisation of electronic exchanges, where trading is predominantly effected through *order books*, has led to much finer granularity of financial data. Such an order book aggregates the buy and sell orders for an asset that were submitted by several market participants at various prices and sizes. A stylized graphical representation of the shape of an order book is given in Figure 2.1.

Most changes in the structure and make-up of an order book can be described as arrivals to and departures from a queueing system: market participants can add to the volume already quoted at any price, or cancel any order they previously submitted. Additionally, the composition of the order book is affected by trade events: these generally occur when a trader commits to buy or sell the asset (often conditional on a certain limit price) which effects a removal of previously submitted quotes (often those that have been submitted first at the price in question). We refer to Cont et al. (2010) for a precise mathematical formulation of order books as queueing systems.

Viewed from such a microscopic perspective, changes in the traded price of an asset are hence caused by structural changes in the underlying order book, and those changes are

in turn brought about by events such as order submission, cancellation and removal. A natural probabilistic description of such events is in terms of (possibly marked) point processes; and we can now see that this re-introduces *physical time* as a key component of a statistical model for the price dynamics of an electronically traded asset.

An all-encompassing treatment of the order book dynamics is however extraordinarily difficult, if not impossible. In full generality, the order book is a multi-dimensional queueing system (namely one queue for each price level) impacted by several kinds of events (namely addition, cancellation and trade if the queue considered is the best ask or best bid; and all possibly augmented by an integer mark to indicate the number of units added, cancelled or traded). These queues do not evolve independently at all, but exhibit dynamics that are constrained and tied together by numerous economic principles and empirical facts.

Several empirical studies have been conducted to investigate these determinants and describe the resulting statistical properties of order books in detail; early contributions are due to Bouchaud et al. (2002), Potters and Bouchaud (2003) or Smith et al. (2003); a comprehensive recent survey has been compiled by Gould et al. (2012). Market participants will for instance not place a buy order at a price greater than the best offer (and analogously for sell orders, thus ensuring that bid and offer quotes remain separated); but they will also compete for high-priority queue spots close to the mid price (thus ensuring that bid and offer quotes remain close); and for some (but by no means all) assets they tend to arrange their quoted volume in a peculiar hump shape (see Figure 2.1). These few examples however just scratch the surface of the empirical study of order books and are by no means an exhaustive list.

To make any progress in this situation, a simplified model seems necessary. It is unsurprising that restrictive (and often unrealistic) model assumptions can lead to more powerful results. We give just a few examples from the current literature to illustrate this principle:

- (i) Cont et al. (2010) model quote additions and cancellations as well as trade orders at all price levels of the book using *independent Poisson processes*. This allows them to

analyse the equilibrium behaviour of this dynamical system, and to derive quantities such as the probability of “making the spread” (that is, selling and buying equal amounts of the asset while the mid-price remains constant) which is of importance to major financial institutions.

(*ii*) Cont and De Larrard (2013) describe how an interaction of the best bid queue with the remaining bids (at worse prices) can be specified in a simple fashion. This description can of course be given for the offer side as well, and the description of the order book then reduces to studying a system of two queues. Event arrivals for this system are still modelled as independent Poisson processes, permitting the derivation of results for the duration until the next price change and a diffusion limit for the mid price.

(*iii*) Other authors abandon the idea of a comprehensive order book model, and instead propose flexible and more realistic models for a single phenomenon or property of order books. Toth et al. (2011) introduce the notion of a “latent liquidity”; Hall and Hautsch (2008) describe an econometric framework to specify the intensity of “aggressive” order placements; and Large (2007) measures the “resiliency” of an order book. The last two contributions use point processes that are more general than the constant rate Poisson processes mentioned in (*i*) and (*ii*).

This last perspective on order books has strong and natural appeal: Trading rules as well as the nature of available data feeds vary greatly for different exchanges and assets; and techniques or findings for the analysis of one financial instrument may be incorrect or inapplicable for another. It seems expedient and efficient in this situation to base a statistical model of an order book (or of the corresponding asset) on quantities that have been abstracted from often idiosyncratic and contingent factors or circumstances.

We give a very simple illustration of this modelling approach using the notion of “latent liquidity”. This concept serves to quantify how much demand or supply there really is at certain prices, and over which time frame. This is related to some practitioners’ conviction that the volume visibly displayed in the order book is at best a very unreliable if

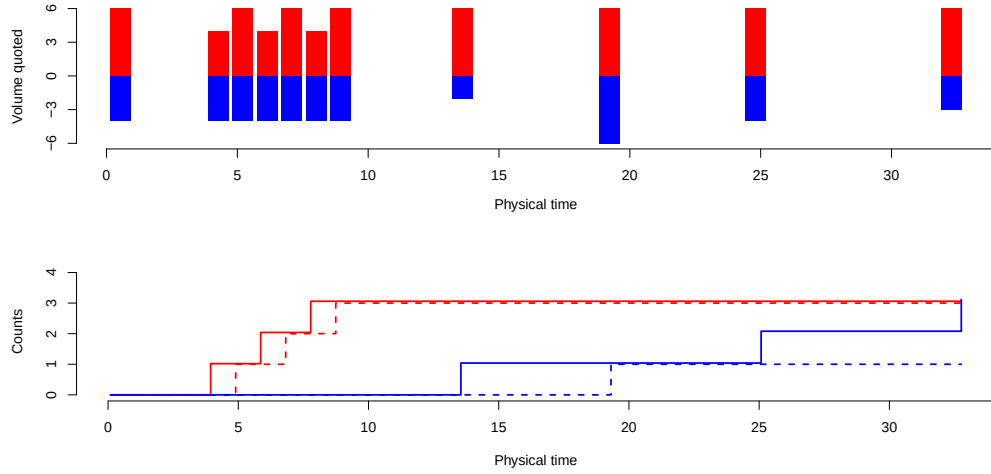


Figure 2.2 Top plot: The blue bars denote the volume quoted at the best bid, the red bars the volume quoted at the best ask. We assume no cancellations occur. Between times 4 and 9, the offered volume at the best ask is “lifted” three times but promptly replenished; the behaviour at the best bid at later time points lacks this regularity.

Bottom plot: The four-dimensional point process related to trade events and the submission of quotes thereafter (the “replenishment”). Red lines relate to the offer side, blue lines to the bid side; solid lines denote the counting processes for trades, dashed lines indicate quote submissions.

not intentionally misleading datum. Quotes in the largely unregulated foreign exchange markets are for instance often displayed only for a very brief period of time or at unreasonably large sizes; the purpose in the first case is to gather information about what other participants consider as fair price, and in the second case to pretend there was larger trading interest on one side of the book than the other. A more accurate indication of the liquidity available may become manifest in the way that best bid or best offer quotes are *replenished after trade events*.

The top plot in Figure 2.2 shows how the volume quoted at the best bid and best ask in a fictional order book evolves over time. We start with six units offered for sale and four bid for purchase at time $t = 0$. At time $t = 4$, a sell event for two units occurs (or, in trading language, the “offer is lifted”), reducing the volume offered for sale to four units. At time $t = 5$ however, that volume is promptly topped up to reach six units again. This pattern is repeated twice. At times 14, 25 and 33, three buy events also occur, but the replenishment at time $t = 19$ occurs only once and with delay. Using vague yet suggestive language, it seems the order book is rather rigid at the best ask, snapping back immediately into its previous shape once a trade has occurred. This can be seen as an indication of selling

interest at this particular price and a manifestation of “latent liquidity”. A comparable behaviour cannot be observed on the bid side where volume can be taken off without being similarly promptly replenished.

The bottom plot in Figure 2.2 suggests a first step towards a more precise, statistical description of this situation. It depicts the counting processes associated with quote additions and trades, which ignores all information relating to the *number of units* quoted or traded. The results suggest a measurement of latent liquidity at a certain price in terms of the *intensities* (made precise in Chapter 3) of the point processes for quote additions, cancellations and trades. Statistical concepts such as the mean magnitude, variability and correlation of these intensities can then be used to describe the elusive concept of latent liquidity more precisely.

As indication of the practical uses of such results, note that major financial institutions often engage in *market making*, which involves the display of prices for an asset on both sides of the book with the (sometimes contractually regulated) intent to actually buy and sell at those price levels. The role of such market makers is often described as *liquidity provision*. This hints at the importance of an accurate estimation of quantities such as latent liquidity or its close relative “latent demand” to effect this liquidity provision as efficiently and profitably as possible. We will examine this application in more detail in Chapter 9.

2.2 Two datasets from foreign exchange trading

The mathematical description of an order book as system of interacting queues (impacted by arrivals and departures) is quite realistic and accurate, but can of course not address the issues and peculiarities presented by real (financial) datasets. One of these issues is that order book information is generally distributed by data vendors and exchanges at varying levels of granularity and times of access:

- (i) *Level 1 data* just contains the two best buy and sell prices along with the total volume quoted. This information may be updated and sent to market participants whenever a change to this subset of the order book occurs, or only at specific times (say every 250 milliseconds).
- (ii) *Level 2 data* provides the total volume quoted at several price ticks (often three, five or ten, dependent on the exchange) on either side of the order book. Like Level 1 data, updates occur either instantly or with very short delays.
- (iii) *Level 3 data* is a very loosely defined term. It often refers to the full order book, giving all quotes at all price ticks. Datasets can also include information with respect to the composition of the volume quoted – that is, which market participant quoted for how many units at which time. Such data is generally not available to market participants during the trading day but (if at all) only for research purposes.

This section describes the structure and content of two datasets relating to the same asset (namely the EUR/USD currency pair), traded during the same time period (namely from the 13th to 17th November 2009) on the same exchange (namely the *Electronic Broking Services (EBS)* platform operated by ICAP, see www.ebs.com): (i) the *EBS Life Cycle data (LCD)* is Level 3 data and documents all activity in the order book. It is not available to market participants during the trading day. (ii) The *EBS Live Data (ELD)* is a Level 2 dataset. It documents aggregated snapshots of the best three price levels of the order book and is updated whenever this configuration changes. The ELD is available to market participants in real time.

We also investigate to which extent these datasets are consistent and compatible with each other. This is of importance for the joint empirical analysis of the two datasets in Chapter 9.

2.2.1 The EBS Life Cycle data

The Life Cycle dataset documents all activity in the order book, and allows the “life cycle” of any quote submitted to the exchange to be tracked. The data takes the form of a temporally ordered stream of messages sent to a central quote matching engine. Each message is of one of the following *message types*:

- (i) Quote Submission (abbreviated as QS): a bid or offer quote at a certain price and certain volume is submitted to the exchange and added to the order book.
- (ii) Quote Deletion (QD): A quote submitted earlier is cancelled (i.e. removed from the order book), possibly after some portion of it has been traded against.
- (iii) Hit Submit (HS): A market participant attempts to sell (“hit a bid”) or buy (“lift an offer”) at a specific price (usually close to the best bid or offer, and with volume displayed to buy/sell). These hit submits are not added to the order book.
- (iv) Hit Done (HAD): This message informs us whether a hit has (partially) succeeded or failed.
- (v) Deal Status Message (DSM): This message informs us of any trades that have taken place. These may be the result of successful hits or of the auto-matching of bid and offer quotes submitted at the same price.

The information provided by the data vendor (ICAP, 2008) to explain and document the data format is patchy and sometimes inaccurate. It does for instance not specify whether *partial cancellations* are allowed; the data suggests that this is indeed not permitted. It is also not entirely clear which bid or offer quotes a hit is matched against: One can for instance find a hit that is matched against a quote submitted *later* than the hit. The time span between a hit submission and its removal (see message type HAD) is usually close to 0, but can exceed 0.5 seconds in some cases.

The number of messages collected is around 1.2 million on each day. We explain what is meant by the life cycle of a quote, and how it is described by the above mentioned messages. Typically, the life of a quote begins as a market participant submits a bid or offer quote at a certain time for a specific price and a specific amount; and it may end by being cancelled by the very same market participant at a later point in time without any trade having taken place. The corresponding entries in the dataset are then of the following form:

Timestamp:	2009-09-13	Timestamp:	2009-09-13
	05:15:44.656		05:15:50.233
Message Type:	QS	Message Type:	QD
Currency Pair:	EUR/USD	Currency Pair:	EUR/USD
Gender:	OFFER	Gender:	OFFER
Quote ID:	11063732983729010426	Quote ID:	11063732983729010426
Dealer Region ID:	TY	Quote Dealt Amount:	0
Quote Price:	1.5000		
Quote Amount:	1		
Linked Amount:	0		

To illustrate the left hand table: The message type QS indicates a quote submission; the gender of a quote submission is either BID or OFFER. The quote identification number (Quote ID for brevity) is a string of numbers that is unique to the quote in question; and it reappears later in the dataset when the quote is cancelled or traded against. The location is one of TY, LN or NY for Tokyo, London, and New York respectively, and denotes the geographic origin of the market participant submitting the quote. The meaning of the linked amount is not explained in the documentation provided by EBS (ICAP, 2008). Neither the Linked Amount field nor the Dealer Region ID field are of interest to us.

The right hand table documents a quote deletion (message type QD), and shows that the offer quote submitted at 05:15:44.656 is cancelled about six seconds later, and that zero units of the initially quoted amount have been traded throughout the life of the quote.

The life cycle of a quote involving trades is far more interesting for our purposes. This is documented as follows:

Timestamp:	2009-09-13 19:15:51.586	Timestamp:	2009-09-13 19:16:17.611
Message Type:	QS	Message Type:	HS
Currency Pair:	EUR/USD	Currency Pair:	EUR/USD
Gender:	OFFER	Gender:	BUY
Quote ID:	29338937760446296423	Hit ID:	03512563710400788023
Dealer Region ID:	TY	Dealer Region ID:	TY
Quote Price:	1.4595	Hit Price:	1.4595
Quote Amount:	3	Hit Amount:	2
Linked Amount:	0		

The left hand table shows a quote submission as explained above. We denote the quote with ID 29338937760446296423 by “Offer A” for brevity. The right hand table has message type HS for “Hit Submission”, and indicates that a market participant from Tokyo later showed interest in buying two units of the asset at a price not exceeding 1.4595. The matching of these two quotes is then described as follows:

		Timestamp:	2009-09-13 19:16:17.953
		Message Type:	DSM
		Currency Pair:	EUR/USD
		Gender:	BUY
		Maker Quote ID:	29338937760446296423
		Maker Region ID:	TY
		Taker Hit ID:	03512563710400788023
Timestamp:	2009-09-13 19:16:17.817	Deal Sequence number:	1
Message Type:	HAD	Taker Region ID:	TY
Currency Pair:	EUR/USD	Deal Price:	1.4595
Gender:	BUY	Deal Amount:	2
Hit ID:	03512563710400788023	Automatch flag:	0
Missed Amount:	0		

The left hand table is a “Hit Done” message (HAD as message type) and documents that the hit submission reported earlier has been successful with all two units requested sold (and zero missed) and is no longer active. The right hand table is a “Deal Status Maker” message (DSM) and gives the definitive summary of the trade in question: two units of the asset have been bought from the maker party of Offer A by the taker party (with ID 03512563710400788023) at a price of 1.4595. Should a hit submission be executed against several maker quotes, then the priority of these matches is recorded in the field denoted “Deal Sequence Number” (the timestamps of the corresponding DSM messages are not a reliable indication for this). The purpose of the “Automatch flag” (either 0 or 1) is to indicate whether the deal negotiated originated from the matching of a hit submission or from another quote of the opposite gender: Indeed, had the exchange recorded a quote submission with gender BID instead of a hit at time 2009-09-13 19:16:17.611, the

same trade would have been negotiated, but be labelled an Automatch. Neither the Deal Sequence Number nor the Automatch flag are of interest to us.

The life cycle of Offer A continues, as the maker still offers one unit to buy. The message stream next shows that another hit at the same price as of Quote A is submitted:

Timestamp:	2009-09-13 19:16:22.957
Message Type:	HS
Currency Pair:	EUR/USD
Gender:	BUY
Hit ID:	31588160002960152648
Dealer Region ID:	TY
Hit Price:	1.4595
Hit Amount:	10

This is now matched against Quote A. As no maker is available for all units requested, the HAD message shows a cancellation of nine units. As all volume posted in Offer A has now been traded against, we also see a Quote Deletion message:

Timestamp:	2009-09-13 19:16:23.134	Timestamp:	2009-09-13 19:16:23.387
Message Type:	HAD	Message Type:	QD
Currency Pair:	EUR/USD	Currency Pair:	EUR/USD
Gender:	BUY	Gender:	OFFER
Hit ID:	31588160002960152648	Quote ID:	29338937760446296423
Missed Amount:	9	Quote Dealt Amount:	3

Finally, a DSM message for this trade is produced as well:

Timestamp:	2009-09-13 19:16:23.387
Message Type:	DSM
Currency Pair:	EUR/USD
Gender:	BUY
Maker Quote ID:	29338937760446296423
Maker Region ID:	TY
Taker Hit ID:	31588160002960152648
Deal Sequence number:	1
Taker Region ID:	TY
Deal Price:	1.4595
Deal Amount:	1
Automatch flag:	0

2.2.2 The EBS Live dataset

The EBS Live dataset provides snapshots of the volume quoted at the three best prices on either side of the book. These snapshots are irregularly spaced in time: no information is sent unless this configuration changes. (Which LCD message exactly triggers an update is however not known; see the section on compatibility of the Life Cycle dataset and EBS Live dataset for a discussion.) The data are available to market participants in real-time during the trading day, with around 200000 updates being sent per day. Each of those is of the following form:

Time EBS:	46587.607	Bid 3 Price:	1.4556
Last Sell:	0	Bid 3 Size:	23
Last Buy:	1.4559	Bid 3 Comp.:	21
Lower Bound:	1.4555	Offer 1 Price:	1.4559
Upper Bound:	1.4562	Offer 1 Size:	2
Bid 1 Price:	1.4558	Offer 1 Comp.:	2
Bid 1 Size:	8	Offer 2 Price:	1.4560
Bid 1 Comp.:	8	Offer 2 Size:	15
Bid 2 Price:	1.4557	Offer 2 Comp.:	12
Bid 2 Size:	17	Offer 3 Price:	1.4561
Bid 2 Comp.:	16	Offer 3 Size:	17
		Offer 3 Comp.:	17

The explanation of this table is as follows: Time EBS represents the times at which the exchange sends the updated snapshot to all subscribers; it is denoted in seconds from midnight of the current trading day. Last Sell and Last Buy denote the price at which the last buy or sell transaction has occurred. Lower Bound and Upper Bound serve as an indication for the liquidity contained in the order book, and are calculated as the highest (resp. lowest) price such that the total volume quoted between the Mid Price and the Lower Bound (resp. Upper Bound) exceeds a certain threshold. Bid 1 Price denotes the price of the best bid; and Bid 1 Size denotes the total volume quoted at that price. Bid 1 Comp. informs us about the “composition” of this volume, i.e. between how many market participants this volume is shared. In the above case, eight units of the asset are asked for at a price of 1.4558 by eight different market participants. The remaining fields are all treated analogously.

We call the set of prices between Bid 3 Price and Offer 3 Price (in the above case, $R = \{1.4556, 1.4557, \dots, 1.4561\}$) the *visible range* of the order book. It may of course contain

price levels not listed in the corresponding EBS Live message, namely those that carry zero volume.

2.2.3 Compatibility of the two datasets

Several (interconnected) factors complicate the joint analysis of the two datasets:

- (i) The two datasets are produced following rules and procedures that remain largely opaque to the user, and are not published by the data vendor. Some of the resulting questions and ambiguities have been mentioned above for the Life Cycle dataset.
- (ii) Technological aspects become relevant given the high-frequency nature of the data. It can for instance not be ruled out that the two datasets were processed on different machines; this raises the question whether the timestamps were produced by sufficiently well synchronised internal clocks.
- (iii) As with most datasets, the possibility of erroneous entries can not be excluded. The size and nature of the datasets however make these hard to detect with conventional techniques.

The most visible consequence of these issues is a *mismatch of time* in both datasets. As we have explained in the description of the Life Cycle dataset, the update of one of the three best prices or volumes in the order book is the result of several messages being sent back and forth between the central exchange and at least two market participants (possibly in different geographic locations); and not of a single event at a unique time. A precise definition when exactly a certain sequence of messages triggers an update in the EBS Live dataset is not available.

We give an example for the mentioned mismatch of time in the datasets for the morning of Monday, the 14th September 2009. The EBS Live dataset contains the following two consecutive entries:

Time EBS:	21600.515	Bid 3 Price:	1.4524
Last Sell:	0	Bid 3 Size:	12
Last Buy:	0	Bid 3 Comp.:	11
Lower Bound:	1.4520	Offer 1 Price:	1.4527
Upper Bound:	1.4548	Offer 1 Size:	1
Bid 1 Price:	1.4526	Offer 1 Comp.:	1
Bid 1 Size:	3	Offer 2 Price:	1.4528
Bid 1 Comp.:	3	Offer 2 Size:	8
Bid 2 Price:	1.4525	Offer 2 Comp.:	7
Bid 2 Size:	6	Offer 3 Price:	1.4529
Bid 2 Comp.:	5	Offer 3 Size:	7
		Offer 3 Comp.:	6

and

Time EBS:	21600.612	Bid 3 Price:	1.4524
Last Sell:	0	Bid 3 Size:	12
Last Buy:	0	Bid 3 Comp.:	11
Lower Bound:	1.4520	Offer 1 Price:	1.4528
Upper Bound:	1.4548	Offer 1 Size:	6
Bid 1 Price:	1.4526	Offer 1 Comp.:	6
Bid 1 Size:	2	Offer 2 Price:	1.4529
Bid 1 Comp.:	2	Offer 2 Size:	8
Bid 2 Price:	1.4525	Offer 2 Comp.:	7
Bid 2 Size:	6	Offer 3 Price:	1.4530
Bid 2 Comp.:	5	Offer 3 Size:	12
		Offer 3 Comp.:	10

The timestamps given correspond to 06:00:00.515 and 06:00:00.612. The stream of messages in the Life Cycle dataset between these timestamps is (left to right, top to bottom):

Timestamp:	2009-09-14	Timestamp:	2009-09-14
	06:00:00.517		06:00:00.518
Message Type:	QD	Message Type:	QS
Currency Pair:	EUR/USD	Currency Pair:	EUR/USD
Gender:	OFFER	Gender:	OFFER
Quote ID:	30911361340464508712	Quote ID:	01529651530161823270
Quote Dealt Amount:	0	Dealer Region ID:	TY
		Quote Price:	1.4532
		Quote Amount:	1
		Linked Amount:	0
Timestamp:	2009-09-14	Timestamp:	2009-09-14
	06:00:00.519		06:00:00.520
Message Type:	QD	Message Type:	QS
Currency Pair:	EUR/USD	Currency Pair:	EUR/USD
Gender:	BID	Gender:	BID
Quote ID:	30129512860386060499	Quote ID:	05630483923684907243
Quote Dealt Amount:	0	Dealer Region ID:	TY
		Quote Price:	1.4526
		Quote Amount:	1
		Linked Amount:	0

Timestamp:	2009-09-14	Timestamp:	2009-09-14
	06:00:00.524		06:00:00.576
Message Type:	QS	Message Type:	QD
Currency Pair:	EUR/USD	Currency Pair:	EUR/USD
Gender:	BID	Gender:	OFFER
Quote ID:	19704006753055065950	Quote ID:	17038004922912946427
Dealer Region ID:	TY	Quote Dealt Amount:	0
Quote Price:	1.4522		
Quote Amount:	1		
Linked Amount:	0		

By referring to quote messages sent earlier, it can be determined that message 1 deletes a quote at 1.4528, message 3 a quote at 1.4522, and message 6 a quote at 1.4531, all of initial size 1. It is clear that this does not explain the update in the EBS Live dataset sample above. In fact, as far as the prices in the visible range are concerned, it can be checked that all messages that contribute to the update from 06:00:00.515 to 06:00:00.612 were sent *before* the first time point (up to 100 milliseconds earlier); and that *none* of the messages sent during this timespan were yet factored into the updated message at the second later time point.

This observation suggests that the EBS Live dataset exhibits a lag relative to the Life Cycle dataset, inasmuch as it takes some time to process hit submits and quote deletion messages, and to check for matching quote submissions. How large this lag is, whether it is constant or variable, and caused by operational rules or technological reasons, is unknown.

2.2.4 Reconstruction of the entire order book

The state of the order book can in principle be reconstructed at any time by appropriate processing of the quote submission, deletion and DSM messages contained in the Life Cycle dataset. QS messages inform us about all additions to the book; those QD messages with amount dealt strictly smaller than the amount submitted inform us about removals from the book due to cancellations; and DSM messages inform us about all removals due to the trades.

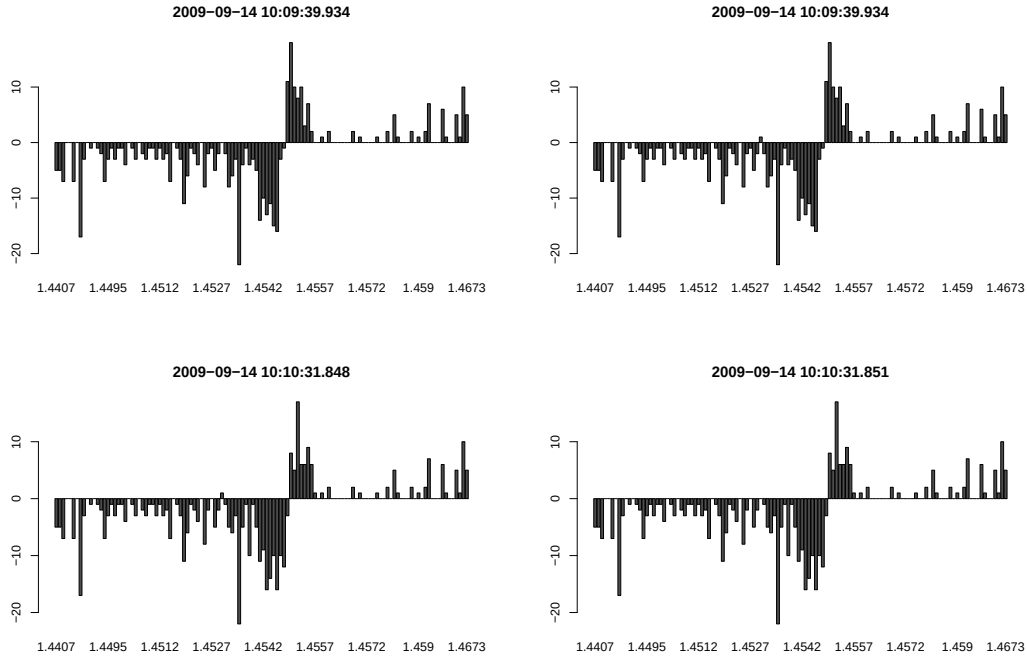


Figure 2.3 Snapshots of the order book at four different times on Monday, 14th September 2009. Note the positive offer volume displayed at a price of 1.4530 in the second and third plot, surrounded by bid quotes (left to right, top to bottom).

The order book reconstructed in this fashion however exhibits time periods of *inconsistency* – these are configurations of bid and offer quotes such that the best bid quote is at a higher price than the lowest offer quote. These states of inconsistency are at least in principle not an artefact of our reconstruction technique; in fact, the purpose of the Automatch flag in DSM messages is to indicate whenever an automatic netting of such a configuration occurs. The time necessary to arrange this netting and send the appropriate messages will then result in a period of inconsistency. This is usually of the magnitude of milliseconds, or, in times of slower market activity, up to a second.

We have however observed time periods of inconsistency in our reconstructed order book much larger than one second. On the first day for instance, a time span of about 50 seconds could be observed in which the order book configurations were inconsistent (see Figure 2.3). These time spans increase as the trading week progresses, leaving most of the reconstructed order book configurations in an inconsistent state. This suggests that a systematic error is made whose effects compound over time. This may be due to a flaw in

our reconstruction methodology or due to rules in the processing of the trading messages unknown to us. It should be pointed out that the best three prices and volumes in the reconstructed order book (following our methodology) can be compared to the snapshots provided in the EBS Live dataset. Up to occasional minor differences due to the time mismatch explained above, we have seen agreement throughout the entire first trading day and on a set of further, more sparsely sampled time points during the rest of the week. Usually only one entry in the reconstructed order book caused an inconsistent state, with price always relatively far away from the mid price.

A definite explanation of this irritating phenomenon is unavailable, as the Life Cycle dataset does not contain any snapshots of the entire book that our reconstruction could be compared to. The above mentioned successful comparison to the EBS Live dataset however makes us suspect the cause is an unknown rule in the processing or interpretation of the message stream provided in the Life Cycle dataset. This may for instance be an undocumented silent deletion of “old quotes” that have been on the book for an entire trading day.

3 Point processes in statistical modelling

A sequence $(T_n; n \in \mathbb{N})$ of random variables with values in $[0, \infty)$, defined on some common probability space $(\Omega, \mathcal{G}, \mathbb{P})$ such that $T_0 = 0$ and $T_n < T_{n+1}$, is called a (*simple*) *point process*. We define

$$N_t = \sum_{i \geq 1} \mathbb{1}(T_i \leq t)$$

with $t \geq 0$ as the *counting process* associated to (T_n) . By abuse of language, $N = (N_t; t \geq 0)$ is also called a point process.

Such a (one-dimensional) point process on the positive real line typically arises whenever (T_n) is sequence of *event times*. Examples include renewal processes whose events are given by the times of failure of a component in a system that is immediately renewed upon failure, emissions from a radioactive material, arrivals and departures from a queue waiting for service, or the times of trades in some financial asset on an electronic exchange.

It is desirable in many applications to allow for time or state dependent behaviour in such processes: the component in a system may fail at a predetermined higher rate when the system is under a heavy load rather than under normal conditions; the radioactive material will emit decreasingly fewer particles if its half-life shortens; and the frequency of arrivals at a queue might depend on its current length or the time of day. In other applications events may arrive at a rate that fluctuates randomly, so that an additional source of randomness must be incorporated in mathematical modelling.

This convenient linguistic description of such fluctuations in terms of a “rate” or “frequency” corresponds to a similarly convenient mathematical approach to point processes

in terms of an *intensity process* $\lambda = (\lambda_t; t \geq 0)$. The earliest such contributions date back to Cox (1955) (see also Cox and Lewis (1978) and Cox and Isham (1980)), and the electrical engineering literature of the early 1970s (Yashin (1970), Snyder (1972), Boel et al. (1975), Segall and Kailath (1975)). These ideas were put on a unifying rigorous mathematical basis by Brémaud (1972) and Segall (1973).

The mathematical concepts underlying this intensity approach are linked to the theory of martingales, where filtrations accurately describe how information patterns increase in time. The fundamental observation is that the *compensated process*

$$N_t - \int_0^t \lambda_s \, ds,$$

where $t > 0$, is a martingale with respect to an appropriately chosen filtration $\mathcal{F}_t$ (and subject to certain regularity conditions on N). This formulation is particularly suitable for situations where point processes are observed as dynamical time series, and estimation or control of this dynamical system has to be performed on-line. Indeed, the martingale calculus and control theory associated to such intensity-driven point processes can be formulated along the same lines as the corresponding results for dynamical systems driven by Brownian motion. Brémaud (1981, pp. xv-xviii) contains an excellent summary of this analogy; comprehensive surveys on historical contributions to the intensity approach more generally can be found in Brémaud (1981, pp. xiii-xv) or Daley and Vere-Jones (2007, Chapter 7).

Alternative approaches to model non-stationarity or other dynamics in point processes have often focussed on specifying the distribution functions for the inter-arrival times $(T_{n+1} - T_n)$, often modulated by some auxiliary process. The statistical concept relating to such a formulation is that of *Markov renewal processes*, which we briefly discuss in Section 3.3. Early examples are for instance due to Cox (1963), and general statistical techniques for the analysis of such models are summarised in Cox and Isham (1980, 3.2).

The most influential model of this kind for the times of trades in a financial asset is the “autoregressive conditional duration model” by Engle and Russell (1998), in which the

distribution of waiting times between transaction events for a financial asset is modelled conditional on previous such waiting times. Bowsher (2007) avoids this approach in favour of intensity based models; and his justification for doing so is arguably valid more generally and not only for financial data:

The specification of a full multivariate point process model may be achieved within a duration-based approach by specifying the joint conditional probability of $(T_{n+1} - T_n)$ (...) given its observed history up to (and including) the n -th event. However, progress in this direction has so far been limited – indeed, the creation of new empirical models from this starting point seems unduly difficult. I adopt a different approach in which the model is specified via the conditional intensity for each type of market event. This provides a powerful and more natural modelling framework for multivariate market event data. The common information set upon which each intensity is conditioned is updated continuously as new information arrives, thus allowing other types of event to have an immediate impact on the intensity as they occur in continuous time. When the intensities are conditioned on the natural filtration of the multivariate point process, the approach is mathematically equivalent to the duration-based one just discussed but is intuitively more appealing. The intensity-based approach also has the advantage that it is straightforward to condition additionally on the history of some other continuous time process.

This chapter is organised as follows: Section 3.1 presents the necessary mathematical concepts for an intensity based approach to point processes. Section 3.2 introduces Poisson processes; the most important example here is the diffusion driven Poisson process. These processes arise from conditioning the point process N on some realisation of an intensity path, which implies that the evolution of N does not affect the dynamics of λ . Section 3.3 relaxes this restriction and explores popular choices for a bi-directional interaction between the point process and its intensity.

3.1 The intensity approach to point processes

Point processes observed over time with varying rates of event arrivals are conveniently modelled using an *intensity process* $\lambda = (\lambda_t; 0 \leq t)$. A first intuition is that this intensity summarises the probabilistic nature of N over time intervals $[t, t + h)$ for h small.

We follow the exposition in Brémaud (1981, chapter II) and Aalen (1978) to introduce the mathematical concepts necessary to make this notion precise. Let $\mathcal{F} = (\mathcal{F}_t; t \geq 0)$ denote a filtration of $\mathcal{G}$ to which N is adapted; in many cases, we take $\mathcal{F}_t = \mathcal{H}_t^N = \sigma(N_s; 0 \leq s \leq t)$, that is, the history of N up to time t . Let N denote a point process such that $\mathbb{E}(N_t) < \infty$ for all t ; and assume that all event times are totally inaccessible, i.e. $\Pr(T_n = \tau) = 0$ for every predictable stopping time τ and each event time T_n . The Doob-Meyer decomposition then ensures that there is a unique *continuous*, non-negative and increasing process Λ_t such that

$$N_t - \Lambda_t$$

is a square-integrable martingale with respect to $\mathcal{F}_t$. We will assume that Λ_t is absolutely continuous, which implies that we can write it as

$$\Lambda_t = \int_0^t \lambda_s \, ds,$$

for some function λ_s with countably many discontinuities.

Definition 3.1. In the situation described above, we say that N_t has λ_t as its $\mathcal{F}_t$ -intensity.

It is safe for our purposes to side-step the question of uniqueness of the intensity process λ . Such a result can be obtained if we require λ to be *predictable*; details can be found in Brémaud (1981, II.4). We may also omit the qualification “ $\mathcal{F}_t$ -” for the intensity process where no confusion is possible.

The requirement that $\mathbb{E}(N_t) < \infty$ implies that the family $(N_s; 0 \leq s \leq t)$ is uniformly integrable, which allows to avoid local martingales in the Doob-Meyer decomposition. The requirement that each event time T_n be totally inaccessible ensures the continuity of Λ . We assume from now on that both requirements are always satisfied.

It can be shown (Aalen, 1978, Lemma 3.3) that we can always recover the intensity of N as

$$\lambda_{t+} = \lim_{h \downarrow 0} \frac{\mathbb{E}(N_{t+h} - N_t \mid \mathcal{F}_t)}{h}, \quad (3.1)$$

where λ_{t+} denotes the right limit. This makes precise the intuition that λ_t represents the propensity of N_t to generate another event in the time interval $(t, t + h)$ for small h .

An alternative definition of λ can be given implicitly as the appropriate object that relates the Lebesgue-Stieltjes integrator dN_s to standard Lebesgue measure ds for a class of sufficiently regular test functions. Indeed, we have

$$\mathbb{E} \left(\int_0^t C_s dN_s \right) = \mathbb{E} \left(\int_0^t C_s \lambda_s ds \right) \quad (3.2)$$

for all non-negative $\mathcal{F}_t$ -predictable processes C_t . This alternative introduction of the intensity process is more technical than the martingale characterisation above, but permits some of the restrictions imposed in Definition 3.1 to be lifted. Further details can be found in Brémaud (1981, Chapter 2).

3.2 Poisson processes

The Poisson process constitutes the most elementary class of point processes. Its distinctive feature is that increments in the counting process N over disjoint intervals are independent, and that the integrated intensity process over such intervals is sufficient to describe the probabilistic law of these increments. The most familiar situation is where λ is a deterministic or even constant function, but we will see that certain stochastic behaviour can be introduced.

We also introduce the *probability generating functional*, an important technical device for the study of point processes.

3.2.1 Deterministic intensities: homogeneous and inhomogeneous Poisson processes

In this section, let $\lambda_t = \lambda(t)$, $t > 0$, denote a deterministic, positive function. We recall the familiar definition of (inhomogeneous) Poisson processes in terms of their intensity:

Definition 3.2. Let λ denote a positive function on the positive real line such that $\int_0^t \lambda(u) du < \infty$, $t > 0$. A point process N is an (*inhomogeneous*) *Poisson process with intensity* λ if N has independent increments over disjoint intervals and

$$\Pr(N_t - N_s = k) = \frac{\left[\int_s^t \lambda(u) du \right]^k}{k!} \exp \left\{ - \int_s^t \lambda(u) du \right\}$$

for $k \in \mathbb{N}$ and all choices $s < t$.

We write $\Lambda(s, t) = \int_s^t \lambda(u) du$ for brevity. The definition implies that the increments $(N_t - N_s)$ have the Poisson distribution with mean $\Lambda(s, t)$, and hence the characteristic function

$$\mathbb{E} \left(e^{iu(N_t - N_s)} \right) = \exp \left\{ (e^{iu} - 1) \Lambda(s, t) \right\}, \quad (3.3)$$

with $u \in \mathbb{R}$, and the probability generating function

$$\mathbb{E} \left(z^{(N_t - N_s)} \right) = \exp \left\{ -\Lambda(s, t)(1 - z) \right\}$$

for $|z| \leq 1$. The classic results for homogeneous Poisson processes carry over to the heterogeneous case with appropriate modifications. We give two examples:

Proposition 3.3. Assume $N_T = n$; that is, n points occur during the interval $[0, T]$. Then these points are independent and identically distributed over $[0, T]$ with density $\lambda(t)/\Lambda(0, T)$.

Proof. (Davison, 2008, 6.5.1) Let $\{t_i; 1 \leq i \leq n\}$ denote a set of (increasing) time points in $(0, T)$; and set $t_0 = 0$ and $t_{n+1} = T$. For $h > 0$ sufficiently small, let A denote the event that there are n points in $[0, T]$, and that they occur in small intervals of width h with t_i

as starting point. We make this precise to give

$$\begin{aligned}
\Pr(A) &= \Pr(N_T = n; (N_{t_i+h} - N_{t_i}) = 1, 1 \leq i \leq n) \\
&= \Pr(N_{t_i+h} - N_{t_i} = 1, 1 \leq i \leq n) \\
&\quad \times \Pr(N_{t_i} = N_{t_{i-1}+h}, 2 \leq i \leq n+1) \Pr(N_{t_1} = N_{t_0}) \\
&= \prod_{i=1}^n \Lambda(t_i, t_i + h) \exp\{-\Lambda(t_i, t_i + h)\} \times \prod_{i=2}^{n+1} \exp\{-\Lambda(t_{i-1} + h, t_i)\} \exp\{-\Lambda(0, t_1)\}.
\end{aligned}$$

Note that the second factor converges to $\exp\{-\Lambda(0, T)\}$ for $h \rightarrow 0$. This means that

$$\begin{aligned}
\lim_{h \rightarrow 0} h^{-n} \Pr(A) &= \exp\{-\Lambda(0, T)\} \times \lim_{h \rightarrow 0} \prod_{i=1}^n h^{-1} \Lambda(t_i, t_i + h) \exp\{-\Lambda(t_i, t_i + h)\} \\
&= \exp\{-\Lambda(0, T)\} \times \lim_{h \rightarrow 0} \prod_{i=1}^n \lambda(t_i) \exp\{-\lambda(t_i)h\} + o(h) \\
&= \exp\{-\Lambda(0, T)\} \prod_{i=1}^n \lambda(t_i).
\end{aligned}$$

It now follows that

$$\lim_{h \rightarrow 0} \frac{\Pr(A)}{\Pr(N_T = n)h^n} = \frac{\lim_{h \rightarrow 0} h^{-n} \Pr(A)}{\Lambda(0, T)^n \exp\{-\Lambda(0, T)\}/n!} = n! \prod_{i=1}^n \frac{\lambda(t_i)}{\Lambda(0, T)}.$$

This is the joint probability density for the (ordered) n random variables T_i ; and we see that it coincides with the order statistics of n independent variables on $[0, T]$, each with distribution $\lambda(t)/\Lambda(0, T)$. $\square$

Proposition 3.4. The waiting time for the next event, starting at time t_0 , has density $\lambda(t_0 + t) \exp\{-\Lambda(t_0, t_0 + t)\}$, $t > 0$.

Proof. Let W denote the waiting time specified. Then, from Definition 3.2,

$$\Pr(W \leq t) = 1 - \Pr(W > t) = 1 - \Pr(N_{t_0+t} - N_{t_0} = 0) = 1 - \exp\{-\Lambda(t_0, t_0 + t)\}.$$

The result follows by differentiation with respect to t . $\square$

It is often helpful to have a generalisation of the probability generating function for a single random variable to an entire point process. The device designed for this purpose is the *probability generating functional (pgfl.)* (see Le Cam (1947)) associated with a point process N :

Definition 3.5. Let $\mathcal{V}$ denote the set of (Borel-) measurable real functions $0 \leq f \leq 1$ with compact support. This is the set of all measurable functions with values in $[0, 1]$ that vanish eventually; so if $f = (1 - \xi) \in \mathcal{V}$, then ξ equals 1 eventually. The probability generating functional G_N associated with a point process N on the real line is defined as the map

$$G_N : \{ \xi : 1 - \xi \in \mathcal{V} \} \longrightarrow \mathbb{R}$$

$$\xi \mapsto \mathbb{E} \left[\exp \left\{ \int_{\mathbb{R}} \log \xi(u) \, dN_u \right\} \right].$$

By analogy to its counterpart for single random variables, the probability generating functional characterises point processes: if N and $\tilde{N}$ are point processes with $G_N(\xi) = G_{\tilde{N}}(\xi)$ for all permissible ξ , then N and $\tilde{N}$ have the same distribution. A proof of this and further results can for instance be found in Daley and Vere-Jones (2007, sec. 5.5).

Example 3.6. The pgfl. of an inhomogeneous Poisson process N with rate function λ as in Definition 3.2 is

$$G_N(\xi) = \exp \left\{ - \int_0^\infty (1 - \xi(t)) \lambda(t) \, dt \right\}.$$

Proof. We can assume there is an interval $I = [0, T]$ such $\xi(t) = 1$ outside I . We write $\Lambda = \Lambda(0, T) = \int_0^T \lambda(t) \, dt$ for brevity. Recall the notation $N_t = \sum_{i \geq 1} \mathbb{1}(T_i \leq t)$ for increasing event times $(T_i; i \in \mathbb{N})$. We then have

$$\begin{aligned} G_N(\xi) &= \mathbb{E} \left[\exp \left\{ \int_I \log \xi(t) \, dN_t \right\} \right] \\ &= \sum_{n=0}^{\infty} \mathbb{E} \left[\exp \left\{ \int_I \log \xi(t) \, dN_t \right\} \mid N_T = n \right] \Pr(N_T = n) \\ &= \sum_{n=0}^{\infty} \Pr(N_T = n) \mathbb{E} \left[\prod_{i=1}^n \xi(T_i) \mid N_T = n \right] = \sum_{n=0}^{\infty} \Pr(N_T = n) [\mathbb{E} \xi(U)]^n, \end{aligned}$$

where U is a random variable on $[0, T]$ with density $\lambda(t)/\Lambda$, and the last equation is justified by Proposition 3.3. Now note that N_T is a $\text{Poisson}(\Lambda)$ -random variable, with pgf. $\mathbb{E}(z^{N_T}) = \exp\{-\Lambda(1 - z)\}$. We now get

$$\begin{aligned} G_N(\xi) &= \sum_{n=0}^{\infty} \Pr(N_T = n) [\mathbb{E} \xi(U)]^n = \exp\left\{-\Lambda(1 - \mathbb{E} \xi(U))\right\} \\ &= \exp\left\{-\Lambda\left(1 - \int_0^T \xi(t) \frac{\lambda(t)}{\Lambda} dt\right)\right\} = \exp\left\{-\int_0^T (1 - \xi(t)) \lambda(t) dt\right\}. \end{aligned}$$

□

3.2.2 Markov modulated Poisson processes

We now allow λ to be a stochastic process. To remain in the class of Poisson point processes, it is however necessary to restrict the interaction of λ with N : The appropriate concept is that of a *doubly stochastic process* for which a trajectory of λ is chosen first and then, conditional on that trajectory, a path of N evolves according to the Poisson law. Only λ hence drives N , and the random evolution of N has no impact upon λ . This is usually made precise by a suitable measurability condition – see Definition 3.7 below. A (continuous-time) Markov process is a popular choice for λ .

Definition 3.7. Let X denote a continuous-time Markov process such as a diffusion or a (conservative, non-explosive) Markov chain, and N a point process, both defined on $(\Omega, \mathcal{G}, \mathbb{P})$. Assume that N is adapted to a filtration $\mathcal{F}$ such that $\mathcal{F}_0 \supset \mathcal{H}_\infty^X = \sigma(X_s; s \geq 0)$. If $\lambda_t = f(t, X_t)$ for some appropriate (measurable, positive) function f , and

$$\Pr(N_{t+h} - N_t = 1 \mid \mathcal{F}_t) = \lambda_t h + o(h),$$

$$\Pr(N_{t+h} - N_t \geq 2 \mid \mathcal{F}_t) = o(h)$$

for all $t \geq 0$, then N is a doubly stochastic Poisson process *driven* (or *modulated*) by the Markov process X , with intensity process λ .

A doubly stochastic Poisson processes is often also called *Cox process*. This is due to their first appearance in Cox (1955), where “the stoppage rate of a loom [is modelled as] a smooth function of relative humidity”. There have been countless other applications since, often in the area of electrical or optical engineering: the survey by Fischer and Meier-Hellstern (1993) mentions (among many other examples and references) the overflow from trunk groups in telecommunications or the superposition of packetised voice streams, whereas Snyder and Miller (1991, Chapter 7) study photoelectronic streams.

A particular focus of this thesis is a Markov modulated Poisson process, where the intensity process driving the point process is given by the solution of the SDE

$$dX_t = -\alpha(X_t - \mu_0) dt + \beta\sqrt{X_t} dW.$$

X is a *Cox-Ingersoll-Ross (CIR) diffusion*, and remains non-negative for all t . The parameters α, μ_0, β represent the autocorrelation (or mean-reversion), the mean value, and the variance that X will exhibit. One of the many advantages of choosing a CIR-diffusion is that despite this flexible parametrisation, many quantities (such as the transition densities or the equilibrium distribution of X) still have simple analytic expressions.

3.3 Point processes that are not Poisson

The preceding characterisations of Poisson processes all emphasised that the randomness in the point process N does not “spill over” into the intensity process λ . Once such a bi-directional interaction is allowed, we leave the class of Poisson processes. This section discusses several popular ways of formulating such an interaction, and shows how certain renewal processes can be cast into the intensity framework for point processes.

3.3.1 Interaction via the last event time

An elementary and well-known example of a point process that is not Poisson is a renewal process N whose inter-arrival times

$$U_n = (T_n - T_{n-1}),$$

where $n \in \mathbb{N}^+$, are i.i.d. with a distribution function other than the exponential. We assume that the distribution of $U = U_1$ admits a density g . This ensures that N is simple and admits the following intensity process:

Proposition 3.8. For $T_n \leq t < T_{n+1}$, the renewal process N has the $\mathcal{F}_t = \mathcal{H}_t^N$ -intensity

$$\lambda(t - T_n) = \lambda_t = \frac{g(t - T_n)}{1 - \int_0^{(t-T_n)} g(x) dx}. \quad (3.4)$$

Proof. Both a formal and a heuristic proof are given in Brémaud (1981, III.2). The heuristic demonstration is as follows: For infinitesimally small dt , we have

$$\lambda_t dt = \Pr(dN_t = 1 \mid \mathcal{H}_t^N). \quad (3.5)$$

The information $\mathcal{H}_t^N$ at time $T_n \leq t < T_{n+1}$ can be decomposed as

$$\mathcal{H}_t^N = \sigma(\mathcal{H}_{T_n}^N, U_{n+1} \geq (t - T_n));$$

and we have an equality of events

$$\{dN_t = 1\} = \{U_{n+1} \in (t, t + dt] - T_n\}.$$

Plugging these two identities into Equation (3.5) and applying Bayes' rule, we obtain

$$\begin{aligned}
\lambda_t dt &= \Pr(dN_t = 1 \mid \mathcal{H}_t^N) = \Pr(U_{n+1} \in (t, t + dt] - T_n \mid \mathcal{H}_{T_n}^N, U_{n+1} \geq (t - T_n)) \\
&= \frac{\Pr(U_{n+1} \in (t, t + dt] - T_n, U_{n+1} \geq (t - T_n) \mid \mathcal{H}_{T_n}^N)}{\Pr(U_{n+1} \geq (t - T_n) \mid \mathcal{H}_{T_n}^N)} \\
&= \frac{\Pr(U_{n+1} \in (t, t + dt] - T_n \mid \mathcal{H}_{T_n}^N)}{\Pr(U_{n+1} \geq (t - T_n) \mid \mathcal{H}_{T_n}^N)} = \frac{g(t - T_n) dt}{1 - \int_0^{(t-T_n)} g(x) dx}.
\end{aligned}$$

□

If the waiting times have the $\text{Exp}(\lambda)$ -distribution, we have $g(x) = \lambda e^{-\lambda x}$ and Equation (3.4) simplifies to $\lambda_t = \lambda$. This results in a constant and hence $\mathcal{F}_0$ -measurable intensity process. This simplification is clearly unavailable for other choices of g ; and we then obtain an intensity process that is modulated by N through the time that has passed since the last event occurred.

This specification is a popular choice in lifetime models where a counting process M is associated to failure times T_n . Owing to the specific theme of those models, the intensity of such a point process is often called the *hazard rate* (the integral of which gives the *hazard function*). In applications (the earliest going back to Cox (1972)), the failure rate is often modified by a deterministic function of time, giving

$$\mu_t = m(\lambda(t - T_n), t)$$

as intensity of M for some function m . Examples include additive or multiplicative interactions between the overall time t and the time since the last event $(t - T_n)$ by means of a deterministic base rate $\mu_0(t)$; that is,

$$\mu_t = \lambda(t - T_n)\mu_0(t) \quad \text{or} \quad \mu_t = \lambda(t - T_n) + \mu_0(t).$$

Multiplicative interaction models are popular in neuroscience where a point process rep-

resents the times at which a neuron “fires” (that is, some electric discharge occurs). Immediately after such an event occurs, the probability of another discharge is zero as the electric potential is now zero; but as time passes, this potential builds up again, increasing the probability of another discharge (see Snyder and Miller (1991, 6.3) or Kass and Ventura (2001) for a discussion of such models).

3.3.2 Interaction in space

There are other ways of modelling the interaction of N with its intensity process. This section discusses two choices for which an event in N effects a change in λ : the death sequence in an *immigration-birth-death* (IBD) process, and the *Hawkes process*. Both examples admit an alternative definition as *Markov renewal process*.

Arrivals and departures in an IBD process

A very elementary (continuous time) model for the increase and decrease of the size of a population is the immigration-birth-death process (Feller, 1939). The waiting times for arrivals (“births” and “immigrations”) and departures (“deaths”) are distributed according to exponential distributions with rates dependent on the overall size of the population.

Let $Z = (Z_t; t \geq 0)$ denote such an IBD process, with parameters λ_n and μ_n for arrivals and departures dependent on the population size n . This means that conditional on $Z_t = n$, the process jumps to $(n - 1)$ with rate μ_n or to $(n + 1)$ with rate λ_n . We typically take $\lambda_n = \nu + \lambda n$, where $\lambda > 0$ is a constant birth rate and $\nu > 0$ an immigration rate, and similarly $\mu_n = \mu n$ for a constant mortality rate $\mu > 0$.

Let $\mathcal{F}_t = \mathcal{H}_t^Z$ denote the filtration generated by Z . We can then write $Z_t = A_t - D_t$ as the difference between an arrival process $A = (A_t; t \geq 0)$ with $\mathcal{F}_t$ -intensity λ_{Z_t} , and a departure process $D = (D_t; t \geq 0)$ with $\mathcal{F}_t$ -intensity μ_{Z_t} .

The departure process D will be of much interest later on, as it is related to the CIR-diffusion driven point process described in Section 3.2.2 (this relation will be made precise in Section 5.2). The above result tells us that D admits an intensity process that is directly tied to a continuous-time Markov process with values in a discrete space. This observation translates into an alternative representation of D as the counting process associated with a *Markov renewal process*:

Definition 3.9. Let $Y = (Y_n; n \in \mathbb{N})$ denote a sequence of random variables taking values in some countable space (say $\mathbb{N}$), and $U = (U_n; n \in \mathbb{N}_+)$ a sequence of random variables with values in $[0, \infty)$. We set $U_0 = 0$. The process (Y, U) is a *Markov renewal process (MRP)* if

$$\Pr(Y_{n+1} = j, U_{n+1} \leq u \mid Y_0, \dots, Y_n, U_0, \dots, U_n) = \Pr(Y_{n+1} = j, U_{n+1} \leq u \mid Y_n) \quad (3.6)$$

for all $n, i, j \in \mathbb{N}$ and $u \in [0, \infty)$.

We associate a point process to (Y, U) by setting $T_n = \sum_{i=0}^n U_m$ as the event times and defining M as the corresponding counting process. The realisations of U_n should hence be thought of as the inter-arrival times of M , modulated by the process Y .

It is often convenient to assume that (Y, U) is *time-homogeneous*. This means there is a distribution function F for the inter-arrival times such that the right hand side of Equation (3.6) satisfies

$$\Pr(Y_{n+1} = j, U_{n+1} \leq u \mid Y_n = i) = F_{ij}(u),$$

irrespective of the time index of Y . We assume again that F admits a density to ensure that the counting process N remains simple (and admits an intensity).

Setting $Y_n = Z_{V_n}$, where V_n denotes the n -th departure time in the population process Z , we can now identify the counting process M associated with (Y, U) with D . The densities corresponding to F_{ij} cannot be immediately written down as the population may increase during departures (due to the arrival process A).

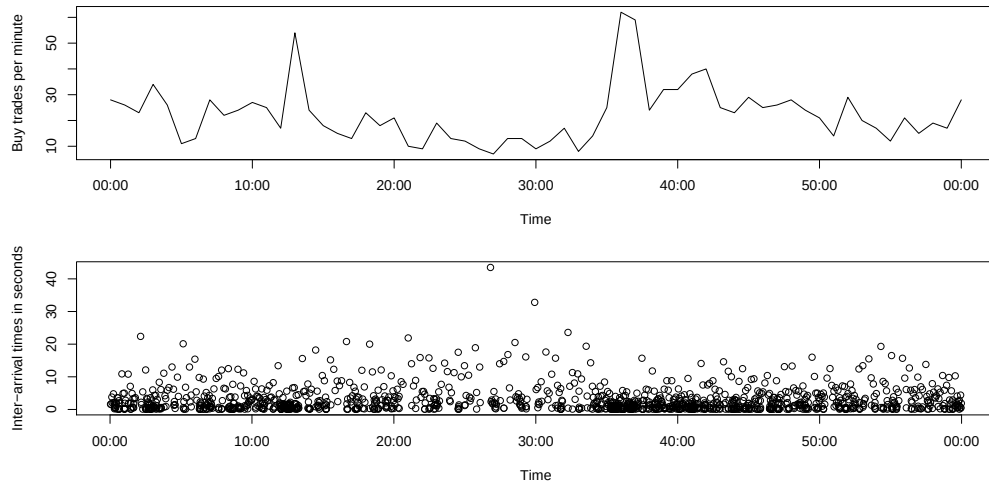


Figure 3.1 Top plot: Number of buy trades per minute in the EUR/USD currency pair, traded on the EBS exchange on Tuesday, 15 September 2009 between 1pm and 2pm. Bottom plot: Inter-arrival times for these buy events.

References to applications of MRPs have been given in the introduction, most notably Engle and Russell’s autoregressive conditional duration model for financial trade data. Figure 3.1 illustrates the difficulty in modelling these inter-arrival times directly. The bottom plot shows all inter-arrival times for buy trades in a currency pair, whereas the top plot only displays the counts aggregated over intervals of one minute. Nonetheless, as Bowsher remarked, the count data (over “small” enough intervals) suggests much more natural models as it immediately illustrates how the trading activity varies from and returns to its mean magnitude.

The Hawkes process

Hawkes processes were introduced in Hawkes (1971) to model “self-excitation” in a point process (in addition to a positive base-rate intensity). A natural application is in the modelling of contagious diseases, where a temporally clustered occurrence of newly-infected (and hence more contagious) patients can mark the beginning of an epidemic, affecting much more patients than the base-rate intensity suggests. Another recent application for Hawkes processes is in the analysis of financial data, where an action taken by one party (executing a trade, or submitting a quote) is registered by many other parties, causing

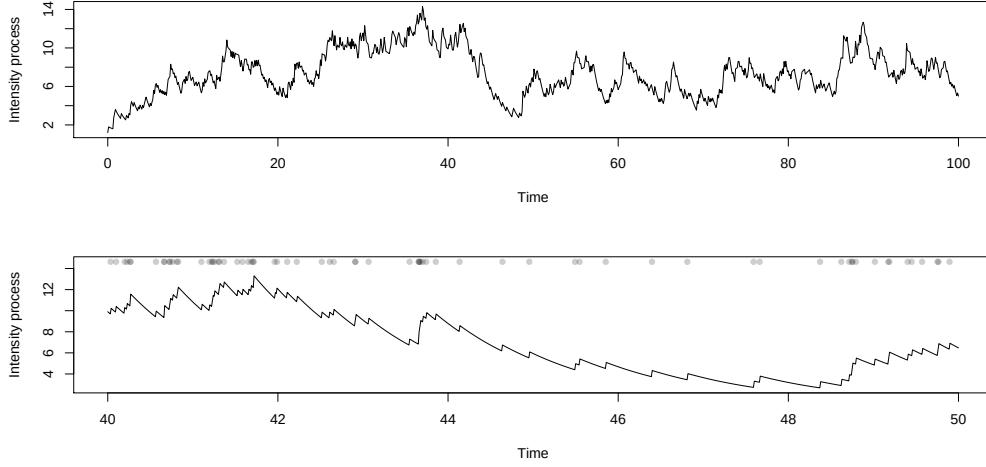


Figure 3.2 A simulated trajectory of a Hawkes process with $\nu(t) = 1.2 + 0.6 \exp(-0.8t)$ for $t \in [0, 100]$, using the procedure in Toke (2011); Ogata (1981). The upper plot displays the fluctuation of λ along the entire time window; the lower plot shows a magnified subset of times and the interaction of events (marked by semi-transparent points at the top; events temporally close correspond to darker colour shades) with the intensity process.

some of them to adapt their view on the market and to take a similar or opposing action as the initial party.

Definition 3.10. A (*univariate, linear*) *self-exciting process* (or *Hawkes process*) is a point process N with $\mathcal{F}_t = \mathcal{H}_t^N$ -intensity process

$$\lambda_t = \lambda_0(t) + \int_{-\infty}^t \nu(t-s) dN_s = \lambda_0(t) + \sum_{T_i < t} \nu(t-T_i),$$

where $\lambda_0 > 0$ is some deterministic base intensity and $\nu(t) \geq 0$ expresses the influence of previous events on the current intensity. This dependence is often exponentially decaying; that is, $\nu(t) = \alpha \exp(-\beta t)$ for $t \geq 0$ and some $\alpha, \beta > 0$, and $\nu(t) = 0$ for $t < 0$.

The exponential kernel $\nu(t)$ (along with $\lambda_0(t) = 0$ for simplicity) is a fortunate choice: If we let $Y_n = \lambda_{T_n}$ denote the intensity at event times of N , we obtain the relationship

$$Y_{n+1} = \lambda_{T_{n+1}} = 1 + e^{-(T_{n+1}-T_n)} \lambda_{T_n} = 1 + e^{-(T_{n+1}-T_n)} Y_n.$$

Much of the information contained in $\mathcal{F}_t = \mathcal{H}_t^N$ is hence irrelevant in this case: in fact only the intensity at event times determines the probabilistic law of future events in N ;

the precise pattern of past event times does not matter.

This observation again translates into an alternative characterisation of N as counting process associated with a Markov renewal process: we may take Y_n as defined above and $U_n = (T_n - T_{n-1})$. (Definition 3.9 also needs to be extended to modulating processes with values in a *continuous space*; but this is only a technical point.) The above recursion then reads as $Y_{n+1} = 1 + e^{-U_n}Y_n$. This hints at the limited use of the formulation of (exponential) Hawkes processes as Markov renewal processes. Unlike Markov modulated Poisson processes or the death sequence in an IBD process, this Markov renewal process does not depend on a hidden state. The evolution of the intensity is entirely determined by previous inter-arrival times (and hence the counting process).

This finding is reminiscent of the difference between *dynamic linear models (DLMs)* and *autoregressive moving average (ARMA) models* in classic time series analysis, where ARMA models (in fact a subclass of DLMs, see Tsay (2005, 11.3.2)) generally do not afford the statistical modeller with the same flexibility as DLMs with respect to missing values and forecasting. The first of these issues is particularly important for financial data, where data streams are subject to failures and corruptions that need to be treated in real time.

The already mentioned econometric framework established by Bowsher (2007) is based on (generalised) Hawkes processes, and indeed needs to account for this (and related) issues by introducing so-called “spillover effects” between trading periods. A model in which the intensity process is permitted to evolve autonomously in the absence of trade event data seems more appealing. This is particularly the case from a Bayesian perspective where an increase in uncertainty about the intensity can (approximately) be represented by gradually “flattening the posterior” (that is, adjusting the posterior variance) while no data is available. A more precise description of this idea can be found in Smith (1979).

3.3.3 Interaction in time

We now present a technique to allow for sharp and sudden increases in the intensity process (as afforded by the Hawkes process) in models that also feature a state whose evolution is not determined by N .

Assume we are given an intensity process with sample path $X_t = X(t) \geq 0$ with $t \geq 0$. While it is legitimate to think of this intensity as deterministic, a more useful interpretation is to adopt the setting of Definition 3.7, where X is a diffusive path measurable with respect to $\mathcal{F}_0$. Let t^* denote the last event time prior or equal to time t .

We define N as the point process with $\mathcal{F}_t = \sigma(\mathcal{F}_0, \mathcal{H}_t^N)$ -intensity

$$\lambda_t = \varphi'(t - t^*) X(\tau(t)),$$

where $t > 0$ and φ is a strictly increasing, non-negative and differentiable function that vanishes at zero; and $\tau(t)$ is a (random) time transformation based on rescaling the inter-arrival times:

$$\tau(t) = \sum_{T_n < t^*} \varphi(T_{n+1} - T_n) + \varphi(t - t^*). \quad (3.7)$$

We also define the point process

$$M_\tau = M_{\tau(t)} = N_t$$

using this time change. It is easy to see that M_τ is $\mathcal{F}_t$ -adapted and has inter-arrival times $\varphi(T_{n+1} - T_n)$.

The illustration of this construction is that the intensity process of N travels along the diffusive path X with a velocity that depends on the pattern of previous arrival times. Consider for instance the choice $\varphi(x) = \sqrt{x}$ and suppose that a large number of events occur within a short time span $[0, h]$. We then have $\tau(h) \gg 0$ (as we sum up the roots of “small” inter-arrival times) but also $M_{\tau(h)} = N_h \gg 0$. The time scale of the point process

M hence adapts to the arrival rate of events in N and progresses more quickly when events arrive shortly after one another.

In fact, M is so constructed as to exactly reverse the change of velocity indicated in Equation (3.7). This is made precise in the following result:

Proposition 3.11. The point process M_τ has intensity $X(\tau)$ (which is measurable with respect to $\mathcal{F}_0$).

Proof. Let μ_τ denote the $\mathcal{F}_\tau$ -intensity of M_τ . From Equation (3.1) and the characterisation of Markov modulated Poisson processes in Definition 3.7, we have

$$\begin{aligned}\mu_\tau &= \lim_{h \downarrow 0} h^{-1} \mathbb{E}(M_{\tau+h} - M_\tau \mid \mathcal{H}_\tau^M) = \lim_{h \downarrow 0} h^{-1} \mathbb{E}(N_{\tau^{-1}(\tau+h)} - N_\tau \mid \mathcal{H}_\tau^N) \\ &= \lim_{h \downarrow 0} h^{-1} \int_t^{\tau^{-1}(\tau+h)} \tau'(s) X(\tau(s)) \, ds = \lim_{h \downarrow 0} h^{-1} \int_\tau^{(\tau+h)} X(u) \, du = X(\tau).\end{aligned}$$

□

We can now see that N is an example of a point process that influences its intensity process *not in space but in time*. Its statistical analysis can however be reduced to the analysis of M ; inferential results for such diffusion driven Poisson processes will be established in Section 4.2.3 and Chapter 5.

4 Continuous-time filtering of dynamical systems

An objective of great interest in the analysis of point processes is the estimation of the underlying and usually unknown random intensity process λ . In many applications, the point process is observed *sequentially*, that is, the path of N is only revealed incrementally as time passes and not in its entirety. Statistical estimation in such situations typically focuses on *filtering* the intensity process; that is, deriving an estimate of the current distribution of λ , and updating that estimate as new information about the occurrence or non-occurrence of events arrives.

In the case of the (exponential) Hawkes process introduced in Section 3.3.2, we can see that this task reduces to estimating the static parameters in the model. One of the reasons for the popularity of Hawkes processes is that this can easily be effected by maximum likelihood, for the log-likelihood function is known explicitly (Ogata, 1981). The resulting estimate is however too tightly coupled to the observation process, as it does not allow for a specification of uncertainty or an autonomous evolution of λ . Given the parameters, observing N is effectively equivalent to observing λ .

It is arguably more desirable from a modelling perspective to endow λ with a dynamical structure independent of N . Such dynamics can often be described or illustrated in physical terms (such as speed, or force towards or away from certain values), which adds to the appeal of such a formulation. The resulting statistical task then becomes that of

(Bayesian) inference in *dynamical systems*, where a prior belief distribution about a hidden state driving an observation process needs to be updated into a posterior distribution once an observation becomes available.

Section 4.1 reviews elementary concepts and results for this inferential task in dynamical systems where the observations are given by random variables with values in $\mathbb{R}^d$ (for some $d \in \mathbb{N}$) and a noise term represented by Brownian motion. The Kalman filter is reviewed as the explicit solution of the filtering problem for *linear* dynamics in discrete and continuous time. This convenient situation no longer applies in Section 4.2, where dynamical systems with observations given by a point process are discussed. We will argue that a filtering approach in continuous-time is required for practical purposes; but this will lead to an updating procedure for the posterior of λ described by an infinite-dimensional differential equation with a stochastic jump term. Finite-dimensional summary statistics of the posterior distribution are neither sufficient nor can they be easily derived in the case of a diffusion driving the point process. Such exact results can only be obtained in very particular situations or by numerical or stochastic approximations.

4.1 Review of Bayesian inference for dynamical systems with Gaussian noise

The study of *dynamical systems* is pervasive throughout the applied and theoretical sciences, and encompasses subjects such as classical and statistical mechanics, population models in theoretical biology or financial modelling. A comprehensive but still partial bibliography has been compiled by Cappé (2001), containing more than 360 contributions published between 1989 and 2000. Hamilton (1989) and Kim and Nelson (2003) contain such bibliographies specifically for the area of econometrics.

A unifying characteristic of dynamical systems is the description of how a point in some (abstract) space evolves over time, and to investigate how this is affected by small changes in the system specification and random effects. If we let X_t denote the location of this point

at time t , the description of its evolution is simplified if we assume that the infinitesimal change dX_t can be described in terms of contemporaneous data – i.e. data such as the current time t or the state X_t , but no information that concerns the past such as X_s for $s < t$. The probabilistic concept corresponding to this idea is of course that of a *Markov process*; and we may for instance think of X as solution to a stochastic differential equation

$$dX_t = f(X_t, t) dt + g(X_t, t) dB_t, \quad (4.1)$$

where B_t denotes Brownian motion.

The process X is usually called the *state process* (of the dynamical system). The collection of data associated with X (the *information process* of the dynamical system) can be an intricate matter for at least two reasons:

- (i) Measurements of X are usually recorded with error or only contain an (incomplete) summary of the process data. A classical example is *target tracking*, where some object (for instance an air-plane) moves through three-dimensional space, and needs to be located using a video or thermographic camera. Besides obvious measurement errors (rooted in the design of the hardware or caused by evasive manoeuvres taken by the plane), the data are (two-dimensional) photos and take values in a lower-dimensional space than the object to be tracked. This is an example of *partial observation* that easily extends to more abstract situations.

One such example is in the area of econometrics, where the variability of the recorded prices of an asset (the data we observe) is driven by a latent, unobserved *volatility process* (the object we wish to track). See Shephard (2005) for a discussion of the unclear origins of this highly influential model.

- (ii) Measurements of X_t are necessarily only recorded at a collection of discrete time points t_n for $n \in \mathbb{N}$. This can be due to storage or recording requirements, or be related to partial information as above.

Inferring from the information process about the state process is our central objective. We

review concepts and results related to this *filtering task* both in discrete and continuous time. Examples 4.7 and 4.10 will suggest that for *linear dynamics* (ie. a linear function f in Equation (4.1)) there is often no qualitative difference between the filtering distributions obtained by “discretise (the model SDE), then filter” and “filter, then discretise (the filtering SDE)”. This fortunate situation cannot be expected in the presence of non-linearity, as will be the case in Section 4.2.

4.1.1 Discrete-time inference in hidden Markov models

Throughout this section, we assume that the data arrives at discrete (strictly increasing) times $(t_n; n \in \mathbb{N})$. We write Y_n for the data collected at time t_n , and let $Y = (Y_n; n \in \mathbb{N})$ denote the entire data process. This situation suggests to consider the evolution of X from times t_n to t_{n+1} . By abuse of notation, we write X_n for X_{t_n} . The requirement that the dynamics of X can be described in terms of contemporaneous data translates into modelling $(X_n; n \in \mathbb{N})$ as Markov chain. We use the shorthand notation $X_{n:m} = (X_n, \dots, X_m)$ for $n \leq m$, and similarly for Y .

The statistical concept designed for the analysis of such systems is that of a *hidden Markov model*. It is comprised of a state process described by a discrete-time Markov chain X , and an observation process Y that is conditionally independent of previous states and observations given the most recent state:

Definition 4.1. A *hidden Markov model (HMM)* is a bivariate, discrete time stochastic process $(X, Y) = \{(X_n, Y_n); n \in \mathbb{N}\}$ such that the *state process* X is a Markov chain with values in some measurable space $(\mathbf{X}, \mathcal{X})$, and the *observation process* Y is a sequence of random variables with values in $(\mathbf{Y}, \mathcal{Y})$ that are independent given X , and such that $(Y_n | X) \sim (Y_n | X_n)$ for $n \in \mathbb{N}$.

The terminology suggests that X is usually unobserved, and this implies that inference about X or parameters describing the model needs to be based on Y only. Nonetheless, it is of course crucial for the specification of a HMM to describe how the Markov chain

X evolves over time, and how the observations Y are related to the state process. This description is given in terms of *transition kernels*:

Definition 4.2. An *unnormalised transition kernel* T from $(\mathsf{X}, \mathcal{X})$ to $(\mathsf{Y}, \mathcal{Y})$ is a function $T : \mathsf{X} \times \mathcal{Y} \rightarrow [0, \infty]$ such that

- (i) $T(x, \cdot)$ induces a (positive) measure on $(\mathsf{Y}, \mathcal{Y})$ for all $x \in \mathsf{X}$,
- (ii) the function $x \mapsto T(x, A)$ is $\mathcal{X}$ -measurable for any measurable set $A \in \mathcal{Y}$.

If T is normalised, i.e. $T(x, \mathcal{Y}) = 1$ for all $x \in \mathsf{X}$, then T is simply called *transition kernel*.

If in addition $\mathsf{X} = \mathsf{Y}$, then T is called a *Markov transition kernel*.

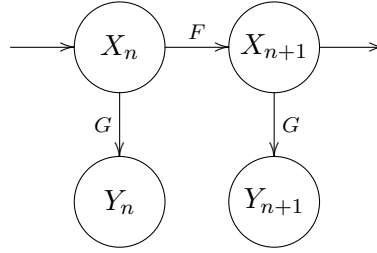


Figure 4.1 Graph of a hidden Markov model, with the arrows indicating the dependence structure, and transition kernels F and G .

Hidden Markov models have been used extensively in applied statistics and statistical modelling over the last decade. Popular applications (see for instance Cappé et al. (2005, sec. 1.3)) include speech recognition, change-point models and topics in econometrics. This breadth of applications is due to the scope and flexibility that Definition 4.1 affords the practitioner, for it allows to incorporate any prior knowledge about X_0 , and is not limited to diffusive dynamics in X or observation processes with white noise.

We give examples for hidden Markov models:

Example 4.3. In a *finite hidden Markov model*, the state space $\mathcal{X} = \{1, \dots, r\}$ and the observation space $\mathcal{Y} = \{1, \dots, s\}$ are finite. The transition kernel F is given by the transition probabilities $F(i, j) = \Pr(X_{n+1} = j \mid X_n = i)$ for $i, j \in \mathsf{X}$, and G by the conditional probabilities $G(i, j) = \Pr(Y_n = j \mid X_n = i)$ for $i \in \mathsf{X}$, $j \in \mathsf{Y}$. Finite HMMs

are particularly popular in telecommunication applications, where a speech signal is first digitised and then transmitted with error over a channel.

Definition 4.4. A *Gaussian linear state-space model* is a hidden Markov model in which $\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \mathbb{R}^{d'}$, and the state and observation processes are linked by linear functions and normal disturbances. That is,

$$\begin{aligned} X_{n+1} &= AX_n + RU_{n+1}, \\ Y_n &= BX_n + SV_n, \end{aligned}$$

where X_0 denotes a (multivariate) normal random variable with mean 0, and U_n and V_n denote white noise processes with zero mean and unit covariance matrix. A, B, R, S are (possibly time-dependent but in any case deterministic) matrices of appropriate dimensions.

Example 4.5. Gaussian linear state-space models can be obtained by discretising a linear stochastic differential equation. Indeed,

$$dX_t = -\alpha(X_t - \mu) dt + \sigma dB_t \tag{4.2}$$

can be solved to give

$$X_t = X_0 e^{-\alpha t} + \mu(1 - e^{-\alpha t}) + \int_0^t \sigma e^{\alpha(s-t)} dB_s.$$

If we choose time points $t_n \in [0, t]$ and write $\Delta_{n+1} = (t_{n+1} - t_n)$ for the time increment, it follows that $(X_{t_{n+1}} | X_{t_n})$ has a normal distribution with mean

$$X_{t_n} e^{-\alpha \Delta_{n+1}} + \mu(1 - e^{-\alpha \Delta_{n+1}}),$$

and variance $\sigma^2 / (2\alpha)(1 - e^{-2\alpha \Delta_{n+1}})$. If measurements Y_n are taken at times t_n with normal error, we are in the situation of Definition 4.4.

Several questions of interest arise in the analysis of hidden Markov models. Most notably,

given observations $Y_{0:n}$, what can be said about the distribution of $(X_{0:n} | Y_{0:n})$ or some marginal of it? This *state inference* is quite naturally connected to Bayesian ideas and techniques.

This can be easily seen for the problem of determining the *filtering distributions* $p(X_n | Y_{0:n})$. Suppose an initial distribution $p(X_0)$ is given, and assume that the dynamics for the state and observation processes are described by the conditional distribution of $p(X_{n+1} | X_n)$ and the likelihood $p(Y_n | X_n)$. Bayes' theorem immediately gives

$$p(X_{n+1} | Y_{0:n+1}) \propto p(Y_{n+1} | X_{n+1}) p(X_{n+1} | Y_{0:n}); \quad (4.3)$$

so the filtering distribution $p(X_{n+1} | Y_{0:n+1})$ is proportional to the prior $p(X_{n+1} | Y_{0:n})$ weighted by the likelihood $p(Y_{n+1} | X_{n+1})$. A recursion for the filtering distributions is available via

$$p(X_{n+1} | Y_{0:n}) = \int p(X_{n+1} | X_n) p(X_n | Y_{0:n}) dX_n. \quad (4.4)$$

These equations are truly “deceptively simple”, as there is an abundance of situations in which the above mentioned likelihood or the prior is intractable or unknown, the integration in (4.4) cannot be performed, or the resulting posterior distribution (4.3) cannot be efficiently described by a finite set of parameters (such as moments, shape or scale). Alternative approaches, usually based upon simulation and computationally intensive (most notably Markov chain Monte Carlo and sequential Monte Carlo), then have to be used to circumvent these difficulties.

There are however two important exceptions where efficient and exact algorithms are available. In the case of finite HMMs as in Example 4.3, the *Baum-Welch algorithm* (see Cappé et al. (2005, sec. 5.1.1.2)) calculates the distribution of $(X_{0:n} | Y_{0:n})$ with computational cost linear in n and quadratic in the number of states r . In the case of Gaussian linear state-space models, the celebrated *Kalman filter* provides an efficient algorithm for the recursive calculation of the filtering distributions $p(X_{n+1} | Y_{0:n+1})$.

An informal derivation of the latter algorithm is simple: If we start with a normal prior

$p(X_n | Y_{0:n})$, Equation (4.4) implies that $p(X_{n+1} | Y_{0:n})$ is a mixture of a normal distribution (namely $p(X_{n+1} | X_n)$) by another normal distribution – hence $p(X_{n+1} | Y_{0:n})$ is again normal. By conjugate priors, (4.4) then gives that the posterior at time $(n + 1)$, $p(X_{n+1} | Y_{0:n+1})$, is again normal. The many different formulations of the Kalman filter found in the literature simply calculate the first two moments of these distributions in a recursive fashion:

Algorithm 4.6 (Cappé et al. (2005), 5.2.13). With the notation of Definition 4.4, let $\mu_{n|m}$ denote the mean of $(X_n | Y_{0:m})$, and $\Sigma_{n|m}$ its variance. The first two moments of $(X_n | Y_{0:n})$ can be calculated by the following recursion:

- (1) *Prediction*: If $n = 0$, set $\mu_{n|n-1} = \mathbb{E}(X_0) = 0$ and $\Sigma_{n|n-1} = \text{Cov}(X_0)$; otherwise, set

$$\begin{aligned}\mu_{n|n-1} &= A\mu_{n-1|n-1}, \\ \Sigma_{n|n-1} &= A\Sigma_{n-1|n-1}A^t + RR^t.\end{aligned}$$

- (2) *Update*: Compute

$$\begin{aligned}\varepsilon_n &= Y_n - B\mu_{n|n-1}, \\ \Gamma_n &= B\Sigma_{n|n-1}B^t + SS^t, \\ K_n &= \Sigma_{n|n-1}B^t\Gamma_n^{-1}, \\ \mu_{n|n} &= \mu_{n|n-1} + K_n\varepsilon_n, \\ \Sigma_{n|n} &= \Sigma_{n|n-1} - K_nB\Sigma_{n|n-1}.\end{aligned}$$

The quantity ε_n is often called the *innovation* and K_n is the *Kalman gain*. This terminology is motivated by two ideas. Firstly, let

$$\text{span}(Y_0, \dots, Y_n) = \left\{ Y \in \mathbb{R}^{d'}; Y = a + \sum_{i=0}^n a_i Y_i \right\}$$

denote the linear subspace in $\mathbb{R}^{d'}$ generated by the observations $Y_{0:n}$. The innovations

span the same subspace as the observations, that is,

$$\text{span}(\varepsilon_{0:n}) = \text{span}(Y_{0:n}),$$

but unlike $Y_{0:n}$, the innovations $\varepsilon_{0:n}$ are *uncorrelated*. ε_n hence represents the part of the random variable Y_n that cannot be linearly predicted from the history $Y_{0:n-1}$ – hence the terminology “innovation” (Cappé et al., 2005, 5.2.2.2). Secondly, the posterior mean adapts to this innovation with a rate that depends on the degree of uncertainty in the state estimate and the accuracy of the observation. The quantity K_n hence describes what is “gained” from making another observation by specifying that rate of adaptation.

It can be shown (see (Cappé et al., 2005, 5.2.12), but note there is an incorrect plus-sign, or (Petrus et al., 2009, 2.10)) that for some Gaussian linear state-space models with time-invariant parameters A, B, R, S , the error covariance matrix $\Sigma_n |_{n-1}$ converges to the solution of the algebraic Riccati equation

$$\Sigma = A\Sigma A^t - A\Sigma B^t(B\Sigma B^t + SS^t)^{-1}B\Sigma A^t + RR^t.$$

This is the case in Example 4.5 which we can formulate as autoregressive process of order 1, observed in noise:

Example 4.7 (The AR(1)-model). Let $X = Y = \mathbb{R}$, and define

$$X_{n+1} = aX_n + U_{n+1},$$

$$Y_n = X_n + sV_n,$$

where $0 < a < 1$. In this case, the unique positive solution of the algebraic Riccati equation is

$$\Sigma = \frac{1}{2} \left(1 + (a^2 - 1)s^2 + \sqrt{4s^2 + (1 + s^2(a^2 - 1))^2} \right),$$

which gives an *invariant* Kalman gain of $K = \Sigma/(\Sigma + s^2)$. The recursions of the Kalman

filter in Algorithm 4.6 now give

$$\mu_{n|n} \approx a\mu_{n-1|n-1} + K\varepsilon_n$$

for n large, so by recursive application

$$\mu_{n|n} \approx K \sum_{k=0}^n a^k \varepsilon_{n-k}.$$

It follows that the posterior mean has qualitatively the same behaviour as an exponentially weighted moving average.

One of the main reasons for the popularity of the Kalman filter is its compact updating procedure: all that is required to store are the model matrices, the posterior mean vector, and the posterior covariance matrix. The class of models that can be described as Gaussian linear state-space models is however limited – the often-cited Bearings-Only problem (see for instance (Fearnhead, 1998, sec. 6.2)) or the stochastic volatility model in econometrics (see Cappé et al. (2005, Exm. 1.3.13)) are examples in which the observations are non-linear functions of the state. A standard application of the Kalman filter fails for these two models despite their otherwise simple nature; and approximative solutions of some kind (see (Fearnhead, 1998, chps. 3, 4) or (Chen, 2003) for a summary) need to be used.

4.1.2 Continuous-time inference by conditional expectations

Example 4.5 was motivated by discretising a linear SDE to make sense of data collected at discrete times. If the data could be collected continuously, this would beg the question which inferential procedure took best advantage of this additional information. To make this question precise, let

$$dX_t = AX_t dt + R dU_t, \quad X_0 = x, \tag{4.5}$$

denote a linear dynamical system in $\mathbb{R}^d$, where A and R are matrices of appropriate dimension, and U denotes d -dimensional Brownian motion. All matrices can be deterministic functions of time. The observations collected are linear transformations of the state process corrupted by “white noise”. This gives the SDE

$$dY_t = BX_t dt + S dV_t \quad (4.6)$$

for the observation process Y with values in $\mathbb{R}^{d'}$, where B, S are matrices of appropriate dimension, and V is a d' -dimensional Brownian motion independent of U .

The question of what constitutes the “best” inference with respect to X given data Y is ill-posed. It is usually understood to find a process $\hat{X} = (\hat{X}_t; t \geq 0)$ adapted to the history $\mathcal{H}_t^Y = \sigma(Y_s; 0 \leq s \leq t)$ generated by Y such that the mean square error $\mathbb{E}(X_t - \hat{X}_t)^2$ is minimised amongst all $\mathcal{H}_t^Y$ -measurable random variables. It is known that for each t , this is satisfied by the conditional expectation

$$\hat{X}_t = \mathbb{E}(X_t | \mathcal{H}_t^Y).$$

It is reasonable to assume that the change in $\hat{X}_{t+h}$ for small h is well-behaved and related to the parameters describing the linear system in (4.5) and (4.6). It can indeed be shown (for instance Øksendal (2003, Chapter 6) or (Davis, 1977, Chapter 4)) that $\hat{X}_t$ satisfies a *linear* stochastic differential equation:

Proposition 4.8. The conditional expectation $\hat{X} = (\mathbb{E}(X_t | \mathcal{H}_t^Y); t \geq 0)$ satisfies the linear SDE

$$d\hat{X}_t = (A - PB^t(SS^t)^{-1}B^t)\hat{X}_t dt + PB'(SS^t)^{-1}dY_t, \quad (4.7)$$

where $P = P(t) = \mathbb{E}((X_t - \hat{X}_t)(X_t - \hat{X}_t)^t)$ denotes the *error covariance matrix*. P satisfies

the matrix Riccati equation

$$\begin{aligned}\frac{dP}{dt} &= RR^t - PB^t(SS^t)^{-1}BP + AP + PA^t, \\ P(0) &= \text{Cov}(X_0) = P_0,\end{aligned}\tag{4.8}$$

which is *non-random and can be computed in advance*.

As in the discrete case, the (recursive) calculation of the filter $\hat{X}$ can be expressed more succinctly by introducing an innovations process. We define $Z_t = Y_t - B\hat{X}_t$, which translates into the SDE

$$dZ_t = B(X_t - \hat{X}_t)dt + SdV_t.$$

Z has orthogonal increments (that is, $\mathbb{E}((Z_t - Z_s)(Z_{t'} - Z_{s'})) = 0$ for non-overlapping intervals $[s, t]$, $[s', t']$), and spans the same subspace of functions as Y (Øksendal, 2003, 6.2). This means that $\hat{X}$ can be represented in terms of Z , and indeed we have

$$d\hat{X}_t = A\hat{X}_t dt + PB^t(SS^t)^{-1}dZ_t.$$

We demonstrate the use of the Kalman filter on two examples:

Example 4.9 (Constant Signal in white noise). Let $X_t = X_0$ for $t \geq 0$, where X_0 is a random variable with $\mathbb{E}(X_0) = 0$ and $\mathbb{E}(X_0^2) = \sigma^2$, and assume that the observation process Y satisfies the SDE

$$dY_t = X_t dt + s dV_t, \quad Y_0 = 0.$$

Equation (4.7) now gives

$$d\hat{X}_t = -P(t)s^{-2}\hat{X}_t dt + P(t)s^{-2}dY_t, \quad \hat{X}_0 = 0,$$

where the ODE of Equation (4.8) becomes

$$\frac{dP}{dt} = -P(t)s^{-2}, \quad P(0) = \sigma^2,\tag{4.9}$$

which can be solved to give $P(t) = (s^2\sigma^2)/(s^2 + \sigma^2 t)$. Setting $g(t) = \exp \left\{ \int_0^t P(u)s^{-2} du \right\}$, we get

$$d(\hat{X}_t g(t)) = P(t)s^{-2} g(t) dY_t.$$

Integration by parts with $\hat{X}_0 = Y_0 = 0$ now gives

$$\begin{aligned} \hat{X}_t g(t) &= P(t)s^{-2} g(t)Y_t - \int_0^t Y_u \frac{dP(u)s^{-2} g(u)}{du} \\ &= P(t)s^{-2} g(t)Y_t - \int_0^t Y_u s^{-2} g(u) \underbrace{\left(P(u)^2 s^{-2} + \frac{dP}{du} \right)}_{=0, \text{ Eq. 4.9}} du = P(t)s^{-2} g(t)Y_t. \end{aligned}$$

It now follows that

$$\hat{X}_t = P(t)s^{-2}Y_t = \frac{\sigma^2}{s^2 + \sigma^2 t} Y_t.$$

This means that for large t , we can put $\hat{X}_t \approx t^{-1}Y_t$, whereas for small t , the signal-to-noise ratio σ/s becomes important. For instance, if $s \gg \sigma$, the observations are mostly ignored, and $\hat{X}_t \approx \mathbb{E}(X_0) = 0$.

Example 4.10 (Ornstein-Uhlenbeck process in noise). We study the system

$$dX_t = -\alpha X_t dt + dU_t,$$

$$dY_t = X_t dt + s dV_t,$$

where $\alpha, s > 0$ are positive real numbers. We assume $\mathbb{E}(X_0) = 0$ and $\mathbb{E}(X_0^2) = \sigma^2$. The SDE for the Kalman filter corresponding to Equation (4.7) now becomes

$$d\hat{X}_t = -(\alpha + P(t)s^{-2})\hat{X}_t dt + P(t)s^{-2} dY_t, \quad \hat{X}_0 = 0,$$

where $P(t)$ is the solution of the Riccati equation

$$\frac{dP}{dt} = 1 - \frac{P^2}{s^2} - 2\alpha P, \quad P(0) = \sigma^2.$$

This has solution

$$P(t) = p_1 + \frac{(p_1 - p_2)(p_0 - p_1)}{(p_0 - p_2)e^{2\beta t} - (p_0 - p_1)},$$

where $\beta = \sqrt{\alpha^2 + 1/s^2}$, $p_1 = s^2(\beta - \alpha)$, $p_2 = -s^2(\beta + \alpha)$. For t large, it can be shown (Øksendal, 2003, sec. 6.2) that

$$\hat{X}_t \approx e^{-\beta t} \hat{X}_0 + (\beta - \alpha) \int_0^t e^{-\beta(t-s)} dY_s, \quad (4.10)$$

which shows the decaying influence of $\hat{X}_0$ and the increasing influence of more recent observations on $\hat{X}_t$. If we discretise Equation (4.10), we recover in analogy to Example 4.7 that the filtered conditional expectation has the same qualitative behaviour as an exponentially weighted moving average.

4.2 Bayesian inference for dynamical systems with point process observations

We now consider dynamical systems with observations given by a point process – specifically Markov modulated Poisson processes as introduced in Section 3.2.2 – and study how to recursively calculate the filtering distributions $p(\lambda_t | \mathcal{H}_t^N)$ of the intensity process.

Section 4.2.1 discusses a time-discretisation approach and the application of hidden Markov model techniques to observations given by the increments of N over disjoint intervals. We will see that this does not yield satisfactory results for the purposes of monitoring and controlling a dynamic system in real time. This can only be provided by a continuous-time filter, which is derived in Section 4.2.2 for Markov processes with values in a discrete space. Section 4.2.3 collects results for the situation where the intensity is given by a diffusion, which is the case arguably most important in applications.

4.2.1 Discretisation of time

An elementary approach to estimating the intensity process is based on time-discretisation and using the incremental counts of N over disjoint intervals as observations to update

the estimate for λ .

Conjugate priors for constant but unknown intensity. We start with the simple case of a point process N with unknown but constant intensity $\lambda_t = \lambda_0$. We assume given some prior belief $p_0(\lambda)$ about the value of the intensity. Discretising time in this situation leads to a rather simple hidden Markov model: choose a constant $\Delta > 0$, and consider the set of equidistant times $\{t_n = n\Delta; n \in \mathbb{N}\}$. Let $X_n = \lambda_0$ denote the constant intensity with trivial dynamics for all $n \in \mathbb{N}$, and write $Y_n = (N_{t_n} - N_{t_{n-1}})$ for the number of events between times t_n and t_{n-1} for $n \geq 1$, and $Y_0 = 0$. We obtain from Equation (4.3) that for $n \geq 1$

$$p(X_n = x \mid Y_{0:n}) \propto p(Y_n \mid X_n = x)p(X_n = x \mid Y_{0:n-1}) \propto x^{Y_n} e^{-x\Delta} p(X_{n-1} = x \mid Y_{0:n-1}).$$

We write $\text{Gamma}(k, \theta)$ for a Gamma distribution with shape k and scale θ , that is, with density proportional to $x^{k-1} e^{-x/\theta}$. If $p_0(X_0)$ is a $\text{Gamma}(\alpha, 1/\beta)$, we see that

$$p(X_1 = x \mid Y_{0:1}) \propto x^{Y_1} e^{-x\Delta} x^{\alpha-1} e^{-\beta x} = x^{\alpha+Y_1-1} e^{-(\beta+\Delta)x}.$$

By induction, it then follows that all posterior distributions $p(X_n \mid Y_{0:n})$ are Gamma distributions with shape $(\alpha + \sum_{l=1}^n Y_l) = (\alpha + N_{n\Delta})$ and scale $1/(\beta + n\Delta)$. This is a result similar to Example 4.9.

HMMs for variable intensity. The assumption of a constant intensity is of course not satisfactory. This could be addressed by modelling the intensity process as piecewise constant, with values in a finite set and possible change-points at the observation times. While inference in this model can be treated routinely with the machinery of hidden Markov models, it seems unnatural to tie the specification of change-points to the (arbitrary) discretisation constant Δ .

A more challenging and interesting example are Markov modulated Poisson processes introduced in Section 3.2.2. Let N now denote a point process driven by a continuous-

time Markov chain X with values in a finite subset $\mathcal{X} = \{1, \dots, r\}$ of the integers, and intensity at time t given by some function $\lambda(X_t)$. Again, let $Y_n = (N_{t_n} - N_{t_{n-1}})$ denote the increments during $[t_{n-1}, t_n]$, with $(t_n - t_{n-1}) = \Delta$ for all n . But clearly we have

$$(Y_n \mid X_0, X_\Delta, \dots, X_{n\Delta}) = (Y_n \mid X_{(n-1)\Delta}, X_{n\Delta}),$$

so that a formulation as hidden Markov model requires an extended state $\tilde{X}_n = (X_{(n-1)\Delta}, X_{n\Delta})$. To calculate the filtering distribution $\Pr(\tilde{X}_n \mid Y_{0:n})$, Equations (4.3) and (4.4) then require the likelihood

$$\Pr(Y_n = y \mid X_{\Delta(n-1)} = i, X_{\Delta n} = j). \quad (4.11)$$

The computation of this likelihood is feasible, but involves the computation of the exponential of a matrix with size of the order $(yr)^2$ (Fearnhead and Sherlock, 2006). This is impractical for point processes with high intensity where y can be very large.

The choice of discretisation constant. In general, the specification of Δ is not obvious. The filtering distributions of the intensity process are typically of interest in monitoring and surveillance applications; and if Δ is chosen too large, a sudden change in the intensity process might not be detected swiftly enough. If Δ is chosen small, the resulting filtering algorithm has a computational cost similar to a (yet to be devised) continuous-time algorithm, but does not reward this with a more realistic treatment of the intensity process (for instance continuous paths). For an illustration of the choice of Δ , see Figure 4.2.

This indicates a general inadequacy of time-discrete models: although N can be summarized by discrete event times, the non-occurrence of events continuously adds to the information generated by N . If we do not have access to this information, but for instance only to aggregated counts, then how severely does this impact the precision of the filtering distribution? Examples 4.7 and 4.10 suggested that for linear Gaussian systems, the posterior distributions resulting from “discretise, then filter” and “filter, then discretise”

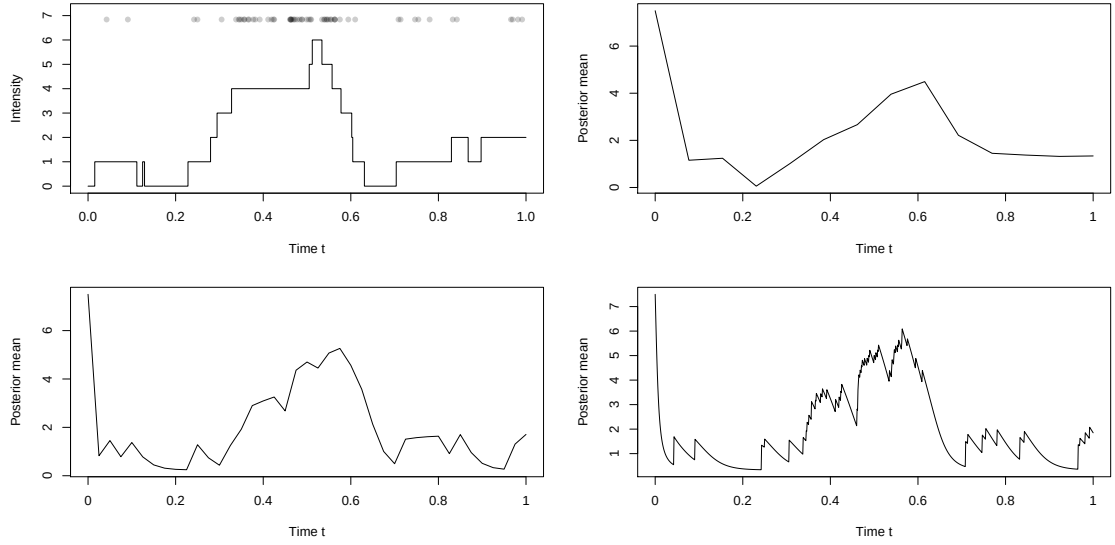


Figure 4.2 The first plot shows a simulation of a point process N modulated by a continuous-time, symmetric random walk X on $\{0, \dots, 15\}$ reflected at the boundaries (solid line). Event times are marked by semi-transparent points at the top; events temporally close correspond to darker colour shades. The subsequent plots show the mean posterior intensity $\hat{\lambda}_t = c \sum_{i=0}^{15} i \Pr(X_t = i | \mathcal{H}_t^N)$ using a non-informative prior distribution, uniform on $\{0, \dots, 15\}$. $c = 35$ is a known scaling constant.

The posterior distribution is calculated under the simplifying assumption that X was constant on intervals $[(n-1)\Delta, n\Delta)$, where $1 \leq n \leq 1/\Delta$. This ignores the problem mentioned in Equation (4.11). The discretisation constant Δ was chosen as 0.075 (top right), 0.025 (bottom left), and 0.001 (bottom right). One may argue that $\Delta = 0.075$ results in an overly smooth, and $\Delta = 0.001$ in an overly rough, jagged estimate; while $\Delta = 0.025$ appears to strike the right balance. It is however hard to justify any value of Δ beforehand; the result in Section 4.2.2 will in fact justify the “saw-tooth” appearance for $\Delta = 0.001$ as a characteristic feature of the continuous-time filter.

exhibit roughly the same behaviour even for coarser discretisation steps; and, more importantly, because we can solve a linear SDE such as (4.2), the degree of error introduced by the discretisation is known. This can usually not be said for point processes driven by some (fairly general) Markov process X . In the case of a diffusion for instance, because the intensity process $\lambda = (\lambda(X_t); t \geq 0)$ is positive, we end up with a *non-linear* SDE describing its dynamics, which renders quantities such as the likelihood in Equation (4.11) intractable.

4.2.2 Poisson processes driven by discrete-valued Markov processes

We now describe a filter in *continuous time* for Markov modulated Poisson processes, where X is a Markov process with state space $\mathcal{X} \subset \mathbb{N}$ a subset of the integers. This will not only improve upon the hidden Markov model approach explained above, but also serves as motivation for the corresponding result for a *diffusive* Markov process: in both cases, the update of the posterior distribution is described by a stochastic differential equation involving the generator of X .

We assume that the process X is time-homogeneous, and let

$$P_t(i, j) = \Pr(X_{t+s} = j \mid X_s = i)$$

denote the transition probability from state i to state j during a time span of length t . We write P_t for the matrix given by these transition probabilities. We assume P_t is continuous in t , and that P_0 is the unit matrix. It is known (see Karlin and Taylor (1999, chapter 14) for instance) that P_t is differentiable (with respect to t) for all $t \geq 0$, with the limit

$$q_i \stackrel{\text{def}}{=} -\lim_{t \rightarrow 0} \frac{dP_t(i, i)}{dt} = \lim_{t \rightarrow 0} \frac{1 - P_t(i, i)}{t} \leq \infty$$

as the rate of leaving state i , and the limits

$$q_{ij} \stackrel{\text{def}}{=} \lim_{t \rightarrow 0} \frac{dP_t(i, j)}{dt} = \lim_{t \rightarrow 0} \frac{P_t(i, j)}{t} < \infty$$

the rates of jumping from state i to state j . We assume the chain is *conservative*, that is, $\sum_{j \neq i} q_{ij} = q_i < \infty$ for all $i \in \mathcal{X}$, and define the *generator* (or *Q-matrix*) of X as the matrix Q with entries in the i -th row and j -th column

$$Q(i, j) = \begin{cases} q_{ij} & \text{for } i \neq j, \\ -q_i & \text{for } i = j. \end{cases}$$

We then know that the transition probabilities satisfy the Kolmogorov backward and forward equations

$$\frac{dP_t}{dt} = QP_t, \quad \frac{dP_t}{dt} = P_tQ. \quad (4.12)$$

Furthermore, if $\mathcal{X}$ is finite, the differential equations in (4.12) have the common unique solution $P_t = e^{tQ}$.

We now let

$$p_t(n) = \sum_{j \in \mathcal{X}} P_t(j, n) = \Pr(X_t = n)$$

denote the probability of being in state $n \in \mathcal{X}$ at time t , and write $p_t = (p_t(n); n \in \mathcal{X})$ for the full vector of probabilities. We assume the initial distribution p_0 is given. Summing over the rows in the backward equation gives a system of differential equations for $p_t(n)$, namely

$$\frac{dp_t(n)}{dt} = -q_n p_t(n) + \sum_{i \neq n} q_{in} p_t(i) = Q^t(n) p_t, \quad (4.13)$$

where $Q^t(n)$ denotes the n -th row of the transpose of Q .

We now assume the intensity of N at time t is given by some function $\lambda(X_t)$, and we wish to study the posterior probabilities

$$\hat{p}_t(n) = \Pr(X_t = n \mid \mathcal{H}_t^N).$$

The resulting description is given by a system of partial differential equations that include a jump term. The derivation below is similar to Rudemo (1972) (where $\mathcal{X}$ is assumed

finite).

Proposition 4.11. The posterior probabilities satisfy the differential equations

$$d\hat{p}_t(n) = Q^t(n) \hat{p} \, dt + \hat{p}_t(n) \left(\lambda(n) - \hat{\lambda}_t \right) \hat{\lambda}_t^{-1} \left(dN_t - \hat{\lambda}_t \, dt \right), \quad (4.14)$$

where $\hat{\lambda}_t = \sum_{k \in \mathcal{X}} \lambda(k) \hat{p}_t(k)$ denotes the posterior mean (also *causal estimate* in the literature) of λ at time t .

Proof. We consider the time interval $[t, t+h)$, and write $\Delta_h = N_{t+h} - N_t$ for the increment of N within $[t, t+h)$. Conditional on $X_t = k$, we have

$$\begin{aligned} \Pr(X_{t+h} = n, \Delta_h = 0 \mid X_t = k) &= q_{kn}h + o(h) \quad \text{for } n \neq k, \\ \Pr(X_{t+h} = k, \Delta_h = 0 \mid X_t = k) &= (1 - q_k h)(1 - \lambda(k)h) + o(h) \\ &= 1 - q_k h - \lambda(k)h + o(h), \\ \Pr(X_{t+h} = n, \Delta_h = 1 \mid X_t = k) &= q_{kn}h\lambda(k)h + o(h) \\ &= o(h) \quad \text{for } n \neq k, \\ \Pr(X_{t+h} = k, \Delta_h = 1 \mid X_t = k) &= \lambda(k)h + o(h). \end{aligned}$$

Letting h tend towards 0 hence allows us to only consider the events $(\Delta_h = \delta)$ for $\delta = 0, 1$; and this means that we have $\mathcal{H}_{t+h}^N = \sigma(\mathcal{H}_t^N, \Delta_h = \delta)$. It follows that

$$\begin{aligned} \hat{p}_{t+h}(n) &= \Pr(X_{t+h} = n \mid \mathcal{H}_{t+h}^N) = \Pr(X_{t+h} = n \mid \mathcal{H}_t^N, \Delta_h = \delta) \\ &= \frac{\Pr(X_{t+h} = n, \Delta_h = \delta \mid \mathcal{H}_t^N)}{\Pr(\Delta_h = \delta \mid \mathcal{H}_t^N)}. \end{aligned} \quad (4.15)$$

We study Equation (4.15) in the two cases mentioned:

(i) The case $(\Delta_h = 0)$: The denominator in (4.15) becomes

$$\begin{aligned}\Pr(\Delta_h = 0 \mid \mathcal{H}_t^N) &= \sum_{k \in \mathcal{X}} \Pr(\Delta_h = 0 \mid \mathcal{H}_t^N, X_t = k) \Pr(X_t = k \mid \mathcal{H}_t^N) \\ &= \sum_{k \in \mathcal{X}} (1 - \lambda(k)h + o(h)) \hat{p}_t(k) = 1 - \hat{\lambda}_t h + o(h).\end{aligned}$$

For the numerator, we get

$$\begin{aligned}\Pr(X_{t+h} = n, \Delta_h = 0 \mid \mathcal{H}_t^N) &= \sum_{k \in \mathcal{X}} \Pr(X_{t+h} = n, \Delta_h = 0 \mid \mathcal{H}_t^N, X_t = k) \Pr(X_t = k \mid \mathcal{H}_t^N) \\ &= \begin{cases} q_{kn} h \hat{p}_t(k) + o(h) & \text{if } n \neq k \\ (1 - q_n h - \lambda(n)h) \hat{p}_t(n) + o(h) & \text{if } n = k \end{cases} \\ &= h \sum_{k \neq n} q_{kn} \hat{p}_t(k) - h q_n \hat{p}_t(n) + \hat{p}_t(n) - \hat{p}_t(n) \lambda(n) h + o(h) \\ &= h Q^t(n) \hat{p}_t + \hat{p}_t(n) - h \hat{p}_t(n) \lambda(n) + o(h).\end{aligned}$$

Writing $1/(1 - \hat{\lambda}_t h) = 1 + \hat{\lambda}_t h + o(h)$, we see that Equation (4.15) becomes

$$\hat{p}_{t+h}(n) = h Q^t(n) \hat{p}_t + \hat{p}_t(n) - \lambda(n) \hat{p}_t(n) h + h \hat{\lambda}_t \hat{p}_t(n) + o(h).$$

Letting $h \rightarrow 0$, this gives

$$d\hat{p}_t(n) = Q^t(n) \hat{p}_t \, dt - (\lambda(n) - \hat{\lambda}_t) \hat{p}_t(n) \, dt, \quad (4.16)$$

which agrees with the claim made in Equation (4.14) for $dN_t = 0$.

(ii) The case $(\Delta_h = 1)$: The denominator in (4.15) now becomes

$$\begin{aligned}\Pr(\Delta_h = 1 \mid \mathcal{H}_t^N) &= \sum_{k \in \mathcal{X}} \Pr(\Delta_h = 1 \mid X_t = k, \mathcal{H}_t^N) \Pr(X_t = k \mid \mathcal{H}_t^N) \\ &= \sum_{k \in \mathcal{X}} \lambda(k) \hat{p}_t(k) h + o(h) = \hat{\lambda}_t h + o(h).\end{aligned}$$

The numerator becomes

$$\begin{aligned}\Pr(X_{t+h} = n, \Delta_h = 1) | \mathcal{H}_t^N &= \sum_{k \in \mathcal{X}} \Pr(X_{t+h} = n, \Delta_h = 1 | \mathcal{H}_t^N, X_t = k) \Pr(X_t = k | \mathcal{H}_t^N) \\ &= \lambda(n) \hat{p}_t(n) + o(h).\end{aligned}$$

Putting this together, we see that Equation (4.15) now becomes

$$\hat{p}_{t+h}(n) = \frac{\lambda(n) \hat{p}_t(n)}{\hat{\lambda}_t} + o(h); \quad \text{that is,} \quad d\hat{p}_t(n) = \hat{p}_t(n) \frac{\lambda(n) - \hat{\lambda}_t}{\hat{\lambda}_t}. \quad (4.17)$$

This agrees with Equation (4.14) at jump times, for which $dN_t = 1$ and $dt = 0$.

□

Example 4.12. We consider the case of a finite state space $\mathcal{X} = \{0, \dots, r-1\}$, and assume the Markov process is a symmetric random walk reflected at the boundaries 0 and $(r-1)$; that is,

$$Q = \begin{pmatrix} -\alpha & \alpha & 0 & \dots & 0 \\ \alpha/2 & -\alpha & \alpha/2 & 0 & \dots & 0 \\ & & & \dots & & \\ 0 & \dots & 0 & \alpha/2 & -\alpha & \alpha/2 \\ 0 & \dots & & 0 & \alpha & -\alpha \end{pmatrix}$$

for some $\alpha > 0$. Equation (4.17) now gives

$$\hat{p}_t(n) = \hat{p}_{t-}(n) + \hat{p}_{t-}(n) \frac{\lambda(n) - \hat{\lambda}_t}{\hat{\lambda}_t} \propto \hat{p}_{t-}(n) \lambda(n)$$

for a jump at time t . Let Λ denote the $r \times r$ -matrix with $\Lambda(n, n) = \lambda(n)$ for $1 \leq n \leq r$ and 0 otherwise; and I_r the $r \times r$ -identity matrix. For any interval $(t_0, t]$ in which no jumps occurs, Equation (4.16) applies to give

$$\frac{d\hat{p}_t}{dt} = (Q^t + \hat{\lambda}_t I_r - \Lambda) \hat{p}_t,$$

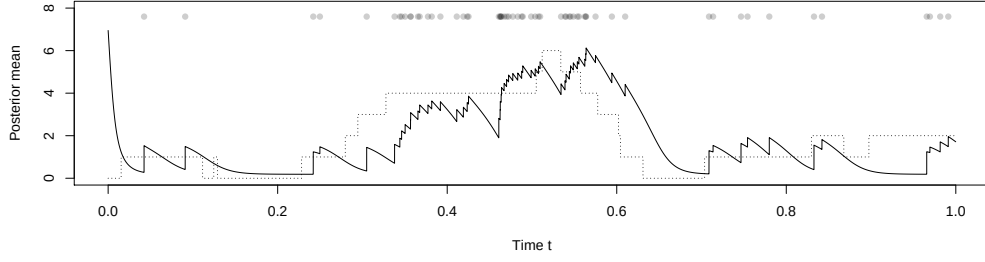


Figure 4.3 The dotted step function shows a simulation of X_t ; event times are marked by semi-transparent points at the top with events temporally close corresponding to darker colour shades. The mean of the posterior distribution $\hat{\lambda}_t$ is displayed as the solid line.

with solution

$$\hat{p}_t = \exp\{t(Q^t + \hat{\lambda}_t I_r - \Lambda)\} \hat{p}_{t_0} \propto \exp\{t(Q^t - \Lambda)\} \hat{p}_{t_0}.$$

Figure 4.3 illustrates these calculations in a particular example: The dotted step function shows a simulation of X_t with $r - 1 = 15$, $\alpha = 25$; circles represent the event times of N with intensity $\lambda(X_t) = 35X_t$ at time $t \in [0, 1]$; and the mean of the posterior distribution $\hat{\lambda}_t = \sum_{n=1}^{r-1} \lambda(n) \hat{p}_t(n)$ is displayed as solid line. This should be compared to Figure 4.2, for which the same model and simulation data was used.

4.2.3 Inference for diffusion driven Poisson processes

We now assume the Markov process driving the point process N is a (continuous) diffusion.

We assume X satisfies the stochastic differential equation

$$dX_t = f(X_t) dt + g(X_t) dW_t, \quad t \geq 0,$$

where W_t is a standard Wiener process; and that the random intensity process λ is related to X through some known (non-negative) function $\lambda_t = \lambda(X_t)$.

This implies that the intensity process is described by a *non-linear* SDE. This is a fundamental difference to Equation (4.6) for the Kalman filter. Non-linear filtering is generally

much more challenging than its linear counterpart (both for point and continuous observation processes). To describe the update of the posterior distribution, we let $\mathcal{L}$ denote the forward Kolmogorov operator associated with X , defined by

$$\mathcal{L}h = -\frac{\partial fh}{\partial x} + \frac{1}{2} \frac{\partial^2 g^2 h}{\partial x^2},$$

and denote its adjoint, the backward Kolmogorov operator by

$$\mathcal{L}^*h = f \frac{\partial h}{\partial x} + \frac{g^2}{2} \frac{\partial^2 h}{\partial x^2}.$$

We denote the probability density of X_t by $p(X_t)$. The historically first result in this situation is the point process analogue to the *Wong-Zakai equation* for dynamical systems driven by Brownian motion (Zakai, 1969), namely an SDE for the posterior density $\hat{p}_t(X_t) = p(X_t | \mathcal{H}_t^N)$:

Proposition 4.13 (Snyder (1972), III, (9)). Let $\Delta X_t = X_{t+h} - X_t$ denote the increment of X over $(t, t+h)$. Assume that

- (i) for Δt sufficiently small, $(\Delta t)^{-1} | \mathbb{E}(\exp\{i u \Delta X_t\} - 1 | X_t) | \leq g(u; t, X_t)$ a.s. where g is a function such that $\mathbb{E}(|g|) < \infty$; and that
- (ii) $(\Delta t)^{-1} \mathbb{E}(\exp\{i u \Delta X_t\} - 1 | X_t)$ converges in probability for $\Delta t \rightarrow 0$.

Then $\hat{p}_t(x)$ satisfies the SDE

$$d\hat{p}_t = \mathcal{L}\hat{p}_t dt + \hat{p}_t (\lambda_t - \hat{\lambda}_t) \hat{\lambda}_t^{-1} (dN_t - \hat{\lambda}_t dt), \quad (4.18)$$

where $\hat{\lambda}_t = \mathbb{E}(\lambda_t | \mathcal{H}_t^N) = \int \lambda(x) \hat{p}_t(x) dx$ is the causal estimate of the intensity.

The proof for this result in Snyder (1972) is based on a differential equation for the conditional characteristic function of X , obtained by directly studying its infinitesimal behaviour. A more systematic derivation of the latter result is available through the martingale approach to point processes (Brémaud, 1972) and the corresponding innovations

technique for filtering (Van Schuppen, 1973): if we wish to filter for $Z_t = h(X_t)$, and that process Z admits a semi-martingale representation

$$dZ_t = a_t dt + dM_t,$$

where a_t is progressively measurable and M_t a martingale, both with respect to some filtration $\mathcal{F}_t \supset \mathcal{H}_t^N$, then the projection $\hat{Z}_t = \mathbb{E}(Z_t \mid \mathcal{H}_t^N)$ onto $\mathcal{H}_t^N$ satisfies

$$d\hat{Z}_t = \hat{a}_t dt + \left(\frac{\widehat{Z\lambda}_{t-}}{\hat{\lambda}_{t-}} - \hat{Z}_{t-} \right) (dN_t - \hat{\lambda}_t dt),$$

where $\hat{a}_t = \mathbb{E}(a_t \mid \mathcal{H}_t^N)$. One can now for instance set $h(X_t) = e^{sX_t}$, obtain the semi-martingale representation by Itô's Lemma, and then derive the filtering SDE by the above projection rule. Further details can be found in a comprehensive but very technical presentation in Brémaud (1981, Chapter 4), or in a more accessible fashion in Leijdekker and Spreij (2011).

Equation (4.18) is an infinite-dimensional filtering equation over the support of $\hat{p}_t$, but the numerical results in Example 4.12 suggest that often a one-dimensional summary statistic S such as the posterior mean is sufficient for practical purposes. Proposition 4.13 can be used to derive a filtering equation for this statistic:

Corollary 4.14. Let $S_t = \int s(x)\hat{p}_t(x) dx$ denote a summary statistic of the posterior density of λ at time t . Then we have

$$dS_t = \mathbb{E}(\mathcal{L}^* s(X_t) \mid \mathcal{H}_t^N) dt + \mathbb{E}(s(X_t)(\lambda_t - \hat{\lambda}_t) \mid \mathcal{H}_t^N) \hat{\lambda}_t^{-1} (dN_t - \hat{\lambda}_t dt), \quad (4.19)$$

and for the particular choice of $s(x) = x$, the posterior expectation $\hat{X}_t = \mathbb{E}(X_t \mid \mathcal{H}_t^N)$ satisfies

$$d\hat{X}_t = \mathbb{E}(f(X_t) \mid \mathcal{H}_t^N) dt + \mathbb{E}((X_t - \hat{X}_t)\lambda_t \mid \mathcal{H}_t^N) \hat{\lambda}_t^{-1} (dN_t - \hat{\lambda}_t dt). \quad (4.20)$$

Proof. Assuming we can interchange differentiation and integration, we get

$$\begin{aligned}
dS &= d \int s(x) \hat{p}(x) dx = \int s(x) d\hat{p}(x) dx \\
&= \int s(x) \mathcal{L} \hat{p} dt dx + \int s(x) \hat{p}(\lambda_t - \hat{\lambda}_t) \hat{\lambda}_t^{-1} (dN_t - \hat{\lambda}_t dt) dx \\
&= \left[\int \hat{p} \mathcal{L}^* s(x) dx \right] dt + \left[\int s(x) \hat{p}(\lambda_t - \hat{\lambda}_t) \hat{\lambda}_t^{-1} dx \right] (dN_t - \hat{\lambda}_t dt) \\
&= \mathbb{E}(\mathcal{L}^* s(X_t) | \mathcal{H}_t^N) dt + \mathbb{E}(s(X_t)(\lambda_t - \hat{\lambda}_t) | \mathcal{H}_t^N) \hat{\lambda}_t^{-1} (dN_t - \hat{\lambda}_t dt).
\end{aligned}$$

For the particular choice of $s(x) = x$, this becomes

$$\begin{aligned}
d\hat{X}_t &= \mathbb{E}(\mathcal{L}^* X_t | \mathcal{H}_t^N) dt + \mathbb{E}(X_t(\lambda_t - \hat{\lambda}_t) | \mathcal{H}_t^N) \hat{\lambda}_t^{-1} (dN_t - \hat{\lambda}_t dt) \\
&= \mathbb{E}(f(X_t) | \mathcal{H}_t^N) dt + \mathbb{E}((X_t - \hat{X}_t)\lambda_t | \mathcal{H}_t^N) \hat{\lambda}_t^{-1} (dN_t - \hat{\lambda}_t dt)
\end{aligned}$$

□

This last equation can be compared to the SDE (4.7) for the Kalman filter – it can now be seen that a non-linear function f will impede further simplification; and that the compensated process $(N_t - \hat{\lambda}_t)$ now assumes the role of Brownian motion. Most importantly, unlike the Riccati equation for the Kalman filter, the posterior variance

$$\text{Var}(X | \mathcal{H}_t^N) = \mathbb{E}[(X_t - \hat{X}_t)^2 | \mathcal{H}_t^N] = \mathbb{E}[(X_t - \hat{X}_t)X_t | \mathcal{H}_t^N]$$

does not admit an explicit solution that can be calculated offline. If we have $\lambda = X$, we see how Equation (4.20) relates the first (posterior) moment to the second, the second to the third and so forth, without terminating at a quantity that can be computed.

Direct progress with Proposition 4.13 can hence only be made in very particular situations (see Section 5.1 or Boel and Beneš (1980, section IV.B and IV.C)). An approximative approach is suggested in Snyder (1972, VI.): using a Taylor approximation

$$f(x_t) = f(\hat{x}_t) + (x - \hat{x}) \frac{\partial f}{\partial x} \Big|_{x=\hat{x}} + \frac{(x - \hat{x})^2}{2} \frac{\partial^2 f}{\partial x^2} \Big|_{x=\hat{x}} + o((x - \hat{x})^2)$$

for $f(x_t)$ and analogously for $\lambda(x_t)$, and making the approximation

$$\hat{\lambda}_t = \mathbb{E}[\lambda(X_t) \mid \mathcal{H}_t^N] \approx \lambda(\mathbb{E}[X_t \mid \mathcal{H}_t^N]) = \lambda(\hat{X}_t),$$

one can derive a stochastic differential equation for an approximate conditional expectation $\tilde{X}_t$ and an approximate conditional variance $\tilde{\Sigma}_t \approx \mathbb{E}[(X_t - \hat{X}_t)^2 \mid \mathcal{H}_t^N]$. The SDE for $\tilde{X}_t$ for instance takes the form

$$d\tilde{X}_t = \left(f(\tilde{X}_t) + \frac{\tilde{\Sigma}_t \partial^2 f(x)}{2 \partial x^2} \Big|_{x=\tilde{X}_t} \right) dt + \tilde{\Sigma}_t \frac{\partial \lambda}{\partial x} \Big|_{x=\tilde{X}_t} \lambda(\tilde{X}_t)^{-1} (dN_t - \lambda(\tilde{X}_t) dt).$$

We will apply this form of approximation in Section 6.2. Alternative sequential Monte Carlo approximations are possible; we discuss these in Section 6.3.

5 Results for the filter equations for CIR-driven Poisson processes

From the perspective of statistical modelling, choosing a diffusion as Markov process driving a point process N is often the most appealing and natural choice. This is because the underlying intensity process λ typically has a direct physical interpretation which requires smooth variations instead of discontinuous jumps. There are however technical obstacles in the way of such an appealing modelling framework: as λ needs to remain positive, the SDE governing its dynamics becomes non-linear; and the transition densities and equilibrium distribution are typically intractable.

The *Cox-Ingersoll-Ross* (CIR) diffusion introduced in Section 5.1 is an exception to this rule. This diffusive process X is non-negative with SDE

$$dX_t = -\alpha(X_t - \mu_0) dt + \beta\sqrt{X_t} dW_t. \quad (5.1)$$

The parametrisation of X can directly be interpreted in terms of its rate of mean-reversion, stationary mean and variance, and there are explicit forms of the (conditional) transition densities and a Gamma distribution as its equilibrium.

Setting $\lambda = X$, we hence obtain a flexible model for the intensity process of N . This chapter adds another item to the list of favourable properties by formulating a central contribution of this thesis: for the purposes of Bayesian inference in a point process N driven by a CIR diffusion X , we can identify N with the death sequence D of an IBD

process Z . This duality between two non-identifiable latent variable models has numerous beneficial implications, and is made precise in Proposition 5.12.

Solving the filtering SDE between events. For the first step towards this result, recall that by Equation (4.18) and setting $\lambda = X$, the posterior density function $\hat{p}$ satisfies

$$d\hat{p}_t = \mathcal{L}\hat{p}_t dt + \hat{p}_t (X_t - \hat{X}_t) \hat{X}_t^{-1} (dN_t - \hat{X}_t dt),$$

where $\mathcal{L}$ is the Kolmogorov forward operator associated with the CIR-diffusion (5.1) above. In Section 5.1, a solution to this filtering SDE in terms of the Laplace transform $M(s, t)$ of $\hat{p}_t$ is derived *between events*, that is, for $dN_t = 0$ (this is Proposition 5.2). The results for this particular approach to the filtering problem are original and were unknown in the literature.

An alternative approach to the material in Section 5.1 has been developed by Leijdekker and Spreij (2011) and was unknown to us at the time of writing. The authors give an SPDE (Proposition 3.2 in their paper) that corresponds to our Proposition 5.1, but is derived using Brémaud’s filtering formula (see Section 4.2.3) and the *method of probability of reference* (Brémaud (1981, VI, 3) or Boel and Beneš (1980)). This SPDE is then solved between and at event times, which gives an elaborate recursive solution for an auxiliary function the authors use (Theorem 5.2 in their paper). A recursive solution to the conditional moment generating function

$$\hat{f}(s, t) = \frac{M(-s, t)}{M(0, t)} \tag{5.2}$$

is obtained for *all* t (with arbitrary number of events up to t), if X_0 has a $\text{Gamma}(\nu/\lambda, \phi)$ distribution, where $\phi > 0$ is an arbitrary scale parameter (Corollary 5.4 in their paper).

An evaluation of this function is however difficult numerically. The recursion at step n (that is, for $T_n < t < T_{n+1}$) involves a polynomial $P_n = P_n(s, t)$ of degree n in s that is

related to P_{n-1} by

$$P_n = B \left(P_{n-1} \mathcal{C} + \frac{\partial P_{n-1}}{\partial s} \left(\mathcal{A} - \frac{C \mathcal{C}}{B} \right) \right),$$

where $B = B(s, t)$ and $C = C(x, t, y)$ are known functions, and $\mathcal{C} = \mathcal{C}(t, T_n)$, and $\mathcal{A} = \mathcal{A}(t, T_n)$ are functions recursively defined in terms of the n -th event time. (The above presentation is a simplification; see Equation (5.12) in Leijdekker and Spreij (2011) for the precise recurrence relation.)

If at all possible, refined numerical techniques are needed to control the accumulation of error due to repeatedly approximating these functions at a growing number of event times; Leijdekker and Spreij only give numerical examples for a trajectory of N with two and three events.

Equivalence relation to the death sequence of an IBD process. An alternative approach to filter the posterior expectation $\hat{X}_t$ is hence pursued in Section 5.2. We study the death sequence D of an IBD process and give an explicit form for the probability generating function of the filtering distribution (Proposition 5.7). Using results established by Wei et al. (2002), we can then formulate and prove the announced main result.

There are several practical advantages of this equivalence result. (i) Most importantly, using the elementary filtering result established in Section 4.2.2 and the coupling of posterior expectation and posterior variance in Corollary 4.14, we can obtain sharp approximations to the posterior expectation and posterior variance of the intensity process X . This is a consequence of replacing the *real-valued* diffusion X by the *integer-valued* population process Z as latent process driving N . The resulting numerical implementation will compare favourably to an implementation of the above mentioned result by Leijdekker and Spreij: the computation of the posterior moments of X is both elementary and genuinely sequential in that only the previous filtering distribution (of the population process) is needed. The computational effort to evaluate $\hat{f}(s, t)$ in Equation (5.2) however grows with the number of events. Further details will be given in Chapter 6. (ii) Furthermore, the simulation from the predictive distribution of N , where desirable, is greatly simplified,

and (iii) the estimation of the model parameters can be effected by maximum-likelihood techniques as demonstrated by Wei (1995). This will be the subject of Chapter 7.

5.1 Explicit solutions to the filtering equations between events

We consider a diffusion driven Poisson process N (see Definition 3.7). We assume that the information collected over time is due to N only, and that the intensity process λ is already $\mathcal{F}_0$ -measurable. We write $\mathcal{H}_t^N = \sigma(N_s; 0 \leq s \leq t)$, and take $\mathcal{F}_t = \sigma(\mathcal{H}_\infty^X, \mathcal{H}_t^N)$ as filtration.

We assume the diffusion X is given as solution to the SDE

$$dX_t = -\alpha(X_t - \mu_0) dt + \beta\sqrt{X_t} dW_t,$$

where α, μ_0, β are all positive, and $t \geq 0$. X is a *Cox-Ingersoll-Ross (CIR) diffusion* and can be realised as the square of a radial Ornstein-Uhlenbeck process (see Clifford and Wei (1993, sec. 3) and Karlin and Taylor (1999, chapter 15, 13.B)). In particular, $X_t \geq 0$ for all $t \geq 0$, and we set $\lambda_t = \lambda(X_t) = X_t$ so that the diffusion drives the point process N directly. Under the assumption that $2\alpha\mu_0 > \beta^2$, it is known (Cox et al., 1985, eq. (20)) that $X_t > 0$ for all t , and that X possesses an equilibrium distribution, which is a Gamma distribution with shape parameter $2\alpha\mu_0/\beta^2$ and scale parameter $\beta^2/(2\alpha)$.

We write $(X_t | X_0 = x_0)$ for the random variable X_t , conditional on $X_0 = x_0$. Its density (a conditional transition density) can be given in explicit form (see Feller (1951); our notation follows Iacus (2008, sec. 1.13.3)). We have

$$p(x, t | x_0) = ce^{-(u-v)} \left(\frac{u}{v}\right)^{q/2} I_q(2\sqrt{uv}), \quad (5.3)$$

where $u = cx_0e^{-\alpha t}$, $v = cx$, and I_q is the modified Bessel function of the first kind of order q , that is,

$$I_q(x) = \sum_{k=0}^{\infty} \left(\frac{x}{2}\right)^{2k+q} \frac{1}{k! \Gamma(k+q+1)}.$$

This means that

$$(X_t | X_0 = x_0) \sim Z \frac{\beta^2(1 - e^{-\alpha t})}{2\alpha}, \quad (5.4)$$

where Z is a non-central Chi-squared random variable with $\frac{4\alpha\mu_0}{\beta^2}$ degrees of freedom, and non-centrality parameter

$$\frac{4\alpha x_0 e^{-\alpha t}}{\beta^2(1 - e^{-\alpha t})}.$$

For reasons that will become clear in Section 5.2, we will find it more convenient to write the SDE for X in the alternative parametrisation

$$dX_t = -(\mu - \lambda) \left(X_t - \frac{\mu\nu}{\mu - \lambda} \right) dt + \sqrt{2\lambda\mu} \sqrt{X_t} dW_t, \quad (5.5)$$

with obvious mapping of parameters. In this new parametrisation, the equilibrium distribution becomes $\text{Gamma}(\nu/\lambda, \lambda\mu/(\mu - \lambda))$, with moment generating function

$$\left(1 - \frac{\lambda\mu s}{\mu - \lambda}\right)^{-\nu/\lambda} = \left(\frac{\mu - \lambda}{\mu - \lambda - \lambda\mu s}\right)^{\nu/\lambda} \quad \text{for } s < \frac{1}{\lambda} - \frac{1}{\mu}. \quad (5.6)$$

We focus on the time between events in the point process N , and hence study the conditional probability density function $\hat{p}_t(x)$ of X_t given $N_t = 0$, with some initial distribution $\hat{p}_0$ given. By Proposition 4.13, $\hat{p}_t$ satisfies

$$\frac{d\hat{p}_t}{dt} = \mathcal{L}\hat{p}_t - (x - \hat{\lambda}_t)\hat{p}_t. \quad (5.7)$$

We define $k_t(x) \stackrel{\text{def}}{=} \exp\{-\int_0^t \hat{\lambda}_s ds\} \hat{p}_t(x)$. This allows to linearise Equation (5.7) (a technique due to Boel and Beneš (1980)) and to derive a PDE for the logarithm of the Laplace transform of k_t :

Proposition 5.1. Let $M(s, t) \stackrel{\text{def}}{=} \int_0^\infty e^{-sx} k_t(x) dx$ denote the Laplace transform of k_t for

$s, t \geq 0$, and $L(s, t) = \log M(s, t)$ its logarithm. Then we have

$$\frac{\partial L}{\partial t} = -\mu\nu s + \frac{\partial L}{\partial s} \left(1 - s(\mu - \lambda) - s^2\lambda\mu \right), \quad s, t > 0. \quad (5.8)$$

Proof. It follows directly from (5.7) that

$$\frac{dk_t(x)}{dt} = \mathcal{L}k_t(x) - xk_t(x).$$

We expand the first factor as

$$\mathcal{L}k_t(x) = (\mu - \lambda)k_t(x) + [(\mu - \lambda)x + 2\lambda\mu - \mu\nu] \frac{\partial k_t(x)}{\partial x} + \lambda\mu x \frac{\partial^2 k_t(x)}{\partial x^2}.$$

If we let $F(s)$ denote the Laplace transform corresponding to a function $f(x)$, recall that we have:

function	Laplace transform
$x f(x)$	$-F'(s)$
$f'(x)$	$sF(s) - f(0)$
$x f'(x)$	$-F(s) - sF'(s)$
$x f''(x)$	$-2sF(s) - s^2F'(s) + f(0)$

We can now calculate

$$\begin{aligned} \frac{\partial M}{\partial t} &= \int e^{-sx} \frac{dk_t(x)}{dt} dx = \int e^{-sx} [\mathcal{L}k_t(x) - xk_t(x)] dx \\ &= \int e^{-sx} [(\mu - \lambda) - x] k_t(x) dx + \int e^{-sx} [(\mu - \lambda)x + 2\lambda\mu - \mu\nu] \frac{\partial k_t(x)}{\partial x} dx \\ &\quad + \int e^{-sx} \lambda\mu x \frac{\partial^2 k_t(x)}{\partial x^2} dx \\ &= M(s, t) [(\mu - \lambda) + s(2\lambda\mu - \mu\nu) - (\mu - \lambda) - 2s\lambda\mu] \\ &\quad + \frac{\partial M(s, t)}{\partial s} [1 - s(\mu - \lambda) - s^2\lambda\mu] + k_t(0) [\mu(\nu - \lambda)]. \end{aligned}$$

We can assume that $k_t(0) = 0$ (because changing $\hat{p}_t$ at a single point does not change the Laplace transform), so we conclude

$$\begin{aligned}\frac{\partial L}{\partial t} &= \frac{1}{M} \frac{\partial M}{\partial t} = -\mu\nu s + \frac{\partial M}{M \partial s} (1 - s(\mu - \lambda) - s^2 \lambda \mu) \\ &= -\mu\nu s + \frac{\partial L}{\partial s} (1 - s(\mu - \lambda) - s^2 \lambda \mu).\end{aligned}$$

□

Let $m(s) \stackrel{\text{def}}{=} M(s, 0)$ denote the value of the Laplace transform of k_t at time $t = 0$; that is, $L(s, 0) = \log m(s)$ is the boundary condition. We can then derive a solution for $M(s, t)$:

Proposition 5.2. The Laplace transform $M(s, t) = \exp L(s, t)$ has solution

$$\begin{aligned}M(s, t) = \exp L(s, t) &= \left[\frac{\lambda + \mu}{\mu + s\lambda\mu + \lambda(1 - s\mu)e^{-t(\lambda+\mu)}} \right]^{\nu/\lambda} \cdot \exp\{-\nu t\} \\ &\quad \cdot m \left(\frac{1 + s\lambda + e^{-t(\lambda+\mu)}(s\mu - 1)}{\mu + s\lambda\mu + \lambda(1 - s\mu)e^{-t(\lambda+\mu)}} \right).\end{aligned}\tag{5.9}$$

Proof. The first auxiliary equation from (5.8) is

$$\frac{dt}{1} = \frac{ds}{s^2\lambda\mu - s(\lambda - \mu) - 1},$$

with solution

$$t = \frac{1}{\mu + \lambda} \log \left| \frac{s\mu - 1}{s\lambda + 1} \right| + c_1 = w(s) + c_1.\tag{5.10}$$

The second auxiliary equation is

$$\frac{ds}{s^2\lambda\mu - s(\lambda - \mu) - 1} = \frac{dL}{-s\mu\nu},$$

with solution

$$\begin{aligned}L &= \frac{1}{\lambda(\lambda + \mu)} \left(-\mu\nu \log|1 + s\lambda| - \nu\lambda \log|s\mu - 1| \right) + c_2 \\ &= v(s) + c_2.\end{aligned}\tag{5.11}$$

This means the general solution for $L(s, t)$ has the form

$$L(s, t) = \phi(t - w(s)) + v(s)$$

for some function ϕ . To determine ϕ , we need to find a function $s(x)$ such that the first equality in

$$\phi(x) = L(s(x), 0) - v(s(x)) = \log m(s(x)) - v(s(x))$$

is satisfied. Setting $x = x(s) = -w(s)$, it follows that $\phi(x) = \phi(-w(s)) = L(s, 0) - v(s)$, and $s(x)$ follows by inversion:

$$s(x) = \frac{1 + e^{-x(\lambda+\mu)}}{\mu - \lambda e^{-x(\lambda+\mu)}}.$$

Combining these results gives

$$\begin{aligned} M(s, t) &= \exp L(s, t) = \exp\{\phi(t - w(s)) + v(s)\} \\ &= \exp\left\{\log m[s(t - w(s))] - v[s(t - w(s))] + v(s)\right\} \\ &= \exp\left\{-v[s(t - w(s))] + v(s)\right\} m[s(t - w(s))] \\ &= \left[\frac{\lambda + \mu}{\mu + s\lambda\mu + \lambda(1 - s\mu)e^{-t(\lambda+\mu)}}\right]^{\nu/\lambda} \cdot \exp\{-\nu t\} \\ &\quad \cdot m\left(\frac{1 + s\lambda + e^{-t(\lambda+\mu)}(s\mu - 1)}{\mu + s\lambda\mu + \lambda(1 - s\mu)e^{-t(\lambda+\mu)}}\right). \end{aligned}$$

□

We can now derive an explicit description of the distribution of $(X_t | N_t = 0)$ for specific choices of the initial distribution:

Proposition 5.3. If we start with a Gamma distribution $X_0 \sim \text{Gamma}(\nu/\lambda, \lambda\mu/(\mu - \lambda))$,

then $(X_t | N_t = 0)$ has a Gamma distribution as well, namely

$$(X_t | N_t = 0) \sim \text{Gamma} \left(\nu/\lambda, \frac{\lambda\mu (\lambda + \mu e^{t(\lambda+\mu)})}{\mu^2 e^{t(\lambda+\mu)} - \lambda^2} \right).$$

Proof. We write Equation (5.9) in the form

$$M(s, t) = e^{-\nu t} A(s, t)^{\nu/\lambda} m(B(s, t)) = e^{-\nu t} A(s, t)^{\nu/\lambda} \left(1 + \frac{\lambda\mu}{\mu - \lambda} B(s, t) \right)^{-\nu/\lambda},$$

with obvious definition of $A(s, t)$ and $B(s, t)$. The Laplace transform of $(X_t | N_t = 0)$ then writes as

$$\begin{aligned} \mathbb{E}(e^{-sX_t} | N_t = 0) &= \frac{M(s, t)}{M(0, t)} = \left[\frac{A(0, t)}{A(s, t)} \frac{\mu - \lambda + \lambda\mu B(s, t)}{\mu - \lambda + \lambda\mu B(0, t)} \right]^{-\nu/\lambda} \\ &= \left[1 + \frac{\lambda\mu (\lambda + \mu e^{t(\lambda+\mu)})}{\mu^2 e^{t(\lambda+\mu)} - \lambda^2} s \right]^{-\nu/\lambda} \end{aligned}$$

for $s \geq 0$, which is the Laplace transform of a Gamma distribution with shape and scale as required. $\square$

Proposition 5.4. If the prior concentrates at the origin, that is, $X_0 = 0$, then $(X_t | N_t = 0)$ has a Gamma distribution, namely

$$(X_t | N_t = 0) \sim \text{Gamma}(\nu/\lambda, \theta(t)),$$

where

$$\theta(t) = \frac{\lambda\mu(1 - e^{-t(\lambda+\mu)})}{\mu + \lambda e^{-t(\lambda+\mu)}}.$$

Proof. Using $m(s) = \mathbb{E}(e^{-sX_0}) = 1$ as boundary condition, Equation (5.9) now gives $M(s, t) = e^{-\nu t} A(s, t)^{\nu/\lambda}$, and the Laplace transform of $(X_t | N_t = 0)$ then writes as

$$\mathbb{E}(e^{-sX_t} | N_t = 0) = \left[\frac{A(0, t)}{A(s, t)} \right]^{-\nu/\lambda} = \left[1 + \frac{\lambda\mu (1 - e^{-t(\lambda+\mu)})}{\mu + \lambda e^{-t(\lambda+\mu)}} s \right]^{-\nu/\lambda}$$

for $s \geq 0$. $\square$

Proposition 5.5. If the prior concentrates at some non-negative position $X_0 = x_0$, then $(X_t | N_t = 0)$ is a Poisson mixture of Gamma distributions. More precisely, let $S \sim \text{Pois}(x_0 \rho(t))$ denote a Poisson random variable with rate

$$x_0 \rho(t) = x_0 \frac{(\lambda + \mu)^2}{(1 - e^{-t(\lambda + \mu)}) \lambda \mu (\lambda + \mu e^{t(\lambda + \mu)})}.$$

Then, conditional on S , we have

$$(X_t | N_t = 0, S) \sim \text{Gamma}(S + \nu/\lambda, \theta(t)).$$

Proof. The boundary condition now becomes $m(s) = \mathbb{E}(e^{-sX_0}) = e^{-sx_0}$, so that the Laplace transform of $(X_t | N_t = 0)$ becomes

$$\mathbb{E}(e^{-sX_t} | N_t = 0) = \left[\frac{A(0, t)}{A(s, t)} \right]^{-\nu/\lambda} \exp\left\{-x_0(B(s, t) - B(0, t))\right\}. \quad (5.12)$$

It is straightforward to check that

$$B(s, t) - B(0, t) = \rho(t) \left(1 - \frac{1}{1 + \theta(t)s}\right),$$

so that

$$\exp\left\{-x_0(B(s, t) - B(0, t))\right\} = e^{-x_0 \rho(t)} \sum_{k=0}^{\infty} \frac{(x_0 \rho(t))^k}{k!} (1 + \theta(t)s)^{-k}.$$

This represents a mixture of Gamma distributions with shape determined by the Poisson random variable S and identical scale $\theta(t)$. From the proof of Proposition 5.4 above, we know that

$$\left[\frac{A(0, t)}{A(s, t)} \right]^{-\nu/\lambda} = (1 + \theta(t)s)^{-\nu/\lambda}; \quad (5.13)$$

that is, the first component of the Laplace transform (5.12) corresponds again to a Gamma distribution with shape ν/λ and scale $\theta(t)$. This shows the result. $\square$

5.2 An equivalence result involving the death sequence in an IBD process

We place ourselves in the situation of Section 3.3.2 and consider an immigration-birth-death process Z with rates of immigration $\lambda(Z_t)$ and death $\mu(Z_t)$ at time t , and filtration $\mathcal{F}_t = \sigma(Z_s; 0 \leq s \leq t)$. We can write (see Brémaud (1981, II.3, E7))

$$Z_t = A_t - D_t \tag{5.14}$$

for arrival and departure processes A and D with $\mathcal{F}_t$ -intensities $\lambda(Z_t)$ and $\mu(Z_t)$, respectively. We assume from now on that $\lambda(Z_t) = \lambda Z_t + \nu$ and $\mu(Z_t) = \mu Z_t$ for some $\mu > \lambda > 0$ and $\nu > 0$.

It is known (Cox and Miller, 1965, Chapter 4) that Z has an equilibrium distribution provided that $\mu > \lambda$, and that this distribution is a $\text{NegBin}(\nu/\lambda, \lambda/\mu)$, that is,

$$\Pr(Z_0 = k) = \frac{\Gamma(r + k)}{\Gamma(r)k!} (1 - p)^r p^k, \quad k \in \mathbb{N},$$

if Z_0 is in equilibrium, and with probability generating function

$$G_{Z_0}(z) = \left(\frac{1 - p}{1 - pz} \right)^{\nu/\lambda} = \left(\frac{1 - \lambda/\mu}{1 - \lambda z/\mu} \right)^{\nu/\lambda} = \left(\frac{\mu - \lambda}{\mu - \lambda z} \right)^{\nu/\lambda}, \quad |z| < \mu/\lambda.$$

Population processes have been studied for decades. A distinctive feature of our situation is that we assume Z is only *partially observed*: we are only notified of the *sequence of deaths* D in the population; and it will be key to study the conditional death rate

$$\hat{\mu}_t = \lim_{h \downarrow 0} \frac{\mathbb{E}[D_{t+h} - D_t \mid \mathcal{H}_t^D]}{h}.$$

The reason for the reparameterisation of the SDE (5.5) for the CIR-diffusion X is related to a result established by Clifford and Wei (1993) and Wei et al. (2002): the point process

D above is *statistically equivalent* to the point process N driven by X . This means that the *finite-dimensional distributions* of D and N agree, that is,

$$\Pr(N_{t_1} = n_1, \dots, N_{t_k} = n_k) = \Pr(D_{t_1} = n_1, \dots, D_{t_k} = n_k)$$

for all finite sets of times $t_1 < \dots < t_k$ and integers $n_1 \leq \dots \leq n_k$.

Key to this result is the probability generating functional $G_M(\xi)$ introduced in Section 3.2 for a point process M and a test function ξ . In Wei et al. (2002, 2.1, 2.2) it is shown that the (strong) Markov property of Z allows to express $G_D(\xi)$ in terms of a function f associated with ξ via the Riccati equation

$$\frac{df}{dt} - (\lambda + \mu)f(t) + \lambda f^2(t) = -\mu\xi(t), \quad t \geq 0, \quad (5.15)$$

with boundary condition $f(t) = 1$ for $t \geq \limsup\{u; 1 - \xi(u) > 0\}$; and that a very similar result can be obtained for G_N by a martingale and stochastic calculus argument. Agreement of G_N and G_D is obtained in three cases: (i) for Z and X in their respective equilibrium distributions, (ii) for $Z_0 = X_0 = 0$, and (iii) $X_0 = x_0 > 0$ and $Z_0 \sim \text{Pois}(x_0/\mu)$.

The aim of this section (proved in Propositions 5.12 and 5.13) is to show that these results extend to recursive filters for the population size and the value of the diffusion, respectively.

The first step towards this is a result similar to Proposition 4.13: Let

$$\hat{p}_t(n) = \Pr(\hat{Z}_t = n) = \Pr(Z_t = n \mid \mathcal{H}_t^D)$$

denote the posterior probability of the population having size n , conditional on observing the death sequence D up to time t , and let $\hat{p}_t = (\hat{p}_t(n))_{n \in \mathbb{N}_0}$ denote the full vector of probabilities. Then we have the following:

Lemma 5.6. The evolution of $\hat{p}_t$ over some interval $[t, t + \tau)$ is described by

$$\begin{aligned} d\hat{p}_t(n) = & \left(\mathbb{1}_{\{n>0\}} \hat{p}_t(n-1) \lambda(n-1) - \hat{p}_t(n) (\lambda(n) + \mu(n)) + \hat{p}_t(n+1) \mu(n+1) \right) dt \\ & + \left(\hat{p}_t(n+1) \mu(n+1) - \hat{p}_t(n) \hat{\mu}_t \right) \hat{\mu}_t^{-1} (dD_t - \hat{\mu}_t dt). \end{aligned} \quad (5.16)$$

Proof. Using the same arguments as for Proposition 4.11, we obtain the update equation

$$\begin{aligned} \hat{p}_{t+\tau}(n) &= \hat{p}_t(n+1) [\mu(n+1)\tau + o(\tau)] \Big/ \sum_{k=0}^{\infty} [\mu(k)\tau + o(\tau)] \hat{p}_t(k) \\ &= \frac{\hat{p}_t(n+1) \mu(n+1)}{\hat{\mu}_t} + o(\tau) \end{aligned}$$

if a departure occurs within $[t, t + \tau)$, and

$$\begin{aligned} \hat{p}_{t+\tau}(n) &= \frac{\hat{p}_t(n) [1 - \lambda(n)\tau - \mu(n)\tau] + \mathbb{1}_{\{n>0\}} \hat{p}_t(n-1) \lambda(n-1)\tau}{\sum_{k=0}^{\infty} (1 - \mu(k)\tau) \hat{p}_t(k)} \\ &= \hat{p}_t(n) + \hat{p}_t(n) (\hat{\mu}_t - \lambda(n) - \mu(n))\tau + \hat{p}_t(n-1) \lambda(n-1)\tau \end{aligned} \quad (5.17)$$

if no departure occurs. The result follows on subtracting $\hat{p}_t(n)$ on both sides of both equations. $\square$

We can now describe the distribution of $(Z_t | D_t = 0)$; that is, the distribution of the population size conditional on not having observed any departures:

Proposition 5.7. Let $q_t(n) = \hat{p}_t(n) \exp\{-\int_0^t \hat{\mu}_s ds\}$, and define

$$H(t, z) \stackrel{\text{def}}{=} \sum_{n=0}^{\infty} z^n q_t(n).$$

Then we have

$$H(t, z) = h(i)(i/z)^{\nu/\lambda} \exp\{\nu \mu t / \lambda\}, \quad (5.18)$$

where $h(z) = H(0, z)$ is the probability generating function of $\hat{p}_0$, and $i = i(t, z)$ a known function given in (5.21).

Proof. Taking $\tau \rightarrow 0$, Equation (5.17) yields

$$\frac{d\hat{p}_t(n)}{dt} = \hat{p}_t(n)(\hat{\mu}_t - \lambda(n) - \mu(n)) + \hat{p}(n-1)\lambda(n-1).$$

The definition of $q_t(n) = \hat{p}_t(n) \exp\{-\int_0^t \hat{\mu}_s ds\}$ allows to eliminate the posterior death rate $\hat{\mu}_t$, and the above equation becomes

$$\frac{dq_t(n)}{dt} = -q_t(n)(\lambda(n) + \mu(n)) + q_t(n-1)\lambda(n-1).$$

Multiplying by z^n and summing up gives the following PDE for $H(t, z)$:

$$\frac{\partial H}{\partial t} = z(\lambda(z-1) - \mu) \frac{\partial H}{\partial z} + \nu(z-1)H.$$

We solve this PDE by the method of characteristics. The auxiliary equations are

$$\frac{dt}{1} = -\frac{dz}{\lambda z^2 + (-\lambda - \mu)z} = \frac{dH}{\nu(z-1)H}.$$

The first auxiliary equation gives

$$\begin{aligned} t &= \frac{1}{\lambda + \mu} \int \frac{1}{z} - \frac{\lambda}{\lambda z - \lambda - \mu} dz + c_1, \\ (\lambda + \mu)t &= \log(z) - \log(\lambda z - \lambda - \mu) + c_1, \\ c_1 &= \exp\{(\lambda + \mu)t\} \frac{\lambda z - \lambda - \mu}{z}, \end{aligned}$$

where c_1 denotes a (changing) constant of integration. The second auxiliary equation writes as

$$\frac{dH}{H} = \frac{\nu(1-z)}{\lambda z^2 + (-\lambda - \mu)z} dz = -\frac{\nu}{(\lambda + \mu)z} dz - \frac{\mu\nu}{(\lambda + \mu)(\lambda z - \lambda - \mu)} dz,$$

and this gives

$$\begin{aligned}
(\lambda + \mu) \log H &= -\nu \log(z) - \frac{\mu\nu}{\lambda} \log(\lambda z - \lambda - \mu) + c_2, \\
H^{\lambda+\mu} &= \left[z(\lambda z - \lambda - \mu)^{\mu/\lambda} \right]^{-\nu} \cdot c_2, \\
H &= \left[z(\lambda z - \lambda - \mu)^{\mu/\lambda} \right]^{-\nu/(\lambda+\mu)} \cdot c_2.
\end{aligned}$$

This means the general solution is of the form

$$H(t, z) = f \left(\underbrace{\exp \left\{ (\lambda + \mu)t \right\} \frac{\lambda z - \lambda - \mu}{z}}_{=g(t, z)} \right) \underbrace{\left[z(\lambda z - \lambda - \mu)^{\mu/\lambda} \right]^{-\nu/(\lambda+\mu)}}_{A(z)} \quad (5.19)$$

for some function $f(g) = f(g(t, z))$. Using $h(z) = H(0, z)$ as boundary condition, we get

$$h(z) = f(g(0, z)) A(z) \quad (5.20)$$

We can find a function $i(t, z)$ such that $g(0, i) = g(t, z)$; in fact

$$i(t, z) = \frac{z(\lambda + \mu)}{\exp((\lambda + \mu)t)(\lambda + \mu - \lambda z) + \lambda z}. \quad (5.21)$$

We now combine Equations (5.19) and (5.20) to derive Equation (5.18):

$$H(t, z) = f(g(t, z)) A(z) = f(g(0, i)) A(z) = \frac{h(i)A(z)}{A(i)} = h(i)(i/z)^{\nu/\lambda} \exp\{\nu\mu t/\lambda\}.$$

□

We can now for instance derive the distribution of $(Z_t | D_t = 0)$ for certain initial conditions:

Corollary 5.8. If Z_0 is in equilibrium (a Negative Binomial with parameters ν/λ and λ/μ), recall that we have $h(z) = \left(\frac{1-\lambda/\mu}{1-\lambda z/\mu} \right)^{\nu/\lambda} = \left(\frac{\mu-\lambda}{\mu-\lambda z} \right)^{\nu/\lambda}$. The probability generating

function of $(Z_t \mid D_t = 0)$ is then

$$\sum_{n=0}^{\infty} \hat{p}_t(n) z^n = \frac{H(t, z)}{H(t, 1)} = \left(\frac{\mu^* - \lambda^*}{\mu^* - \lambda^* z} \right)^{\nu/\lambda}, \quad (5.22)$$

which is a $\text{NegBin}(r = \nu/\lambda, p = \lambda^*/\mu^*)$, where $\mu^* = \mu^2 + \mu\lambda$, $\lambda^* = \lambda^*(t) = \lambda\mu + \lambda^2 e^{-(\lambda+\mu)t}$.

Proof. If Z_0 is in equilibrium, then Equation (5.18) with $h(z)$ and $i = i(t, z)$ as in Proposition 5.7 gives

$$\begin{aligned} H(t, z) &= h(i(z)) \left(\frac{i}{z} \right)^{\nu/\lambda} \exp\{\nu\mu t/\lambda\} = \left[\frac{(\mu - \lambda)i}{(\mu - \lambda i)z} \right]^{\nu/\lambda} \exp\{\nu\mu t/\lambda\} \\ &= \frac{(\mu^2 - \lambda^2)^{\nu/\lambda} \exp(-\nu t) \exp\{(\lambda + \mu)\nu t/\lambda\}}{\left[(\mu - \lambda i) \left((\lambda + \mu - \lambda z) + \exp\{-(\lambda + \mu)t\} \lambda z \right) \right]^{\nu/\lambda} \left[\exp\{(\lambda + \mu)t\} \right]^{\nu/\lambda}} \\ &= \frac{(\mu^2 - \lambda^2)^{\nu/\lambda} \exp(-\nu t)}{\left[\lambda\mu + \mu^2 - \lambda\mu z - \lambda^2 z \exp\{-(\lambda + \mu)t\} \right]^{\nu/\lambda}}. \end{aligned}$$

It then follows that

$$\begin{aligned} H(t, z)/H(t, 1) &= \left[\mu^2 - \lambda^2 \exp\{-(\lambda + \mu)t\} \right]^{\nu/\lambda} \\ &\quad \bigg/ \left[\lambda\mu + \mu^2 - \lambda\mu z - \lambda^2 \exp\{-(\lambda + \mu)t\} z \right]^{\nu/\lambda} \\ &= \left[\frac{(\mu^2 + \lambda\mu) - (\lambda\mu + \lambda^2 \exp\{-(\lambda + \mu)t\})}{(\mu^2 + \lambda\mu) - (\lambda\mu + \lambda^2 \exp\{-(\lambda + \mu)t\})z} \right]^{\nu/\lambda}, \end{aligned}$$

which agrees with Equation (5.22). □

Recall that the main result of this chapter is concerned with two latent variable models that are non-identifiable from the perspective of Bayesian analysis. To describe the prior distributions for which this is the case, we introduce the concept of *compatible random variables*:

Definition 5.9. We say that an integer-valued random variable Z and a positive, con-

tinuous random variable X are *compatible* if

$$H_Z(z+1) \stackrel{\text{def}}{=} \mathbb{E} \left((z+1)^Z \right) = \mathbb{E} \left(e^{zX} \right) \stackrel{\text{def}}{=} M_X(z).$$

Example 5.10. We give examples for compatible random variables:

- (i) Let $Z \sim \text{NegBin}(r, p)$ and $X \sim \text{Gamma}(k, \theta)$. Z and X/μ are compatible if there exist parameters ν, λ, μ and t_0 such that

$$r = k = \nu/\lambda, \quad p = p(t_0) = \frac{\lambda\mu + \lambda^2 e^{-t_0(\lambda+\mu)}}{\mu^2 + \mu\lambda}, \quad \theta = \theta(t_0) = \frac{\lambda\mu (\lambda + \mu e^{t_0(\lambda+\mu)})}{\mu^2 e^{t_0(\lambda+\mu)} - \lambda^2}.$$

Note that for $t_0 = 0$, this becomes the respective equilibrium distribution.

- (ii) Let $Z \sim \text{NegBin}(r, p)$ and $X \sim \text{Gamma}(k, \theta)$. Z and X/μ are compatible if there exist parameters ν, λ, μ and t_0 such that

$$r = k = \nu/\lambda, \quad p = p(t_0) = \frac{\lambda (1 - e^{-t_0(\lambda+\mu)})}{\lambda + \mu}, \quad \theta = \theta(t_0) = \frac{\lambda\mu (1 - e^{-t_0(\lambda+\mu)})}{\mu + \lambda e^{-t_0(\lambda+\mu)}}.$$

For $t_0 = 0$, this corresponds to the “degenerate” distribution $Z = X = 0$.

- (iii) Let $Z \sim \text{Pois}(x_0/\mu)$ and $X = x_0 > 0$. Then Z and X/μ are compatible.

All cases are easily verified. Note that in the first two situations, we have $p = \theta/(\theta + \mu)$.

The announced justification for the terminology in Definition 5.9 is this:

Lemma 5.11. If we equip the population size Z_0 in the IBD-process and the initial state X_0 in the CIR-diffusion with prior distributions such that Z_0 and X_0/μ are compatible, then the death sequence D generated by Z is statistically equivalent to the point process N driven by the diffusion X .

Proof. From Wei et al. (2002, 2.1, Lemma 1), we know that the pgfl. G_D conditional on $Z_0 = k$ is

$$G_D(\xi \mid Z_0 = k) = \mathbb{E} \left[\exp \left\{ \int_{\mathbb{R}} \log \xi(u) \, dD_u \right\} \mid Z_0 = k \right] = f^k(0) \exp \left\{ -\nu \int 1 - f(t) \, dt \right\},$$

where f satisfies the Riccati Equation (5.15). The unconditional pgfl. becomes

$$\begin{aligned} G_D(\xi) &= \sum_{k \geq 0} G_D(\xi \mid Z_0 = k) \Pr(Z_0 = k) \\ &= \exp \left\{ -\nu \int 1 - f(t) \, dt \right\} \sum_{k \geq 0} \Pr(Z_0 = k) f^k(0) \\ &= \exp \left\{ -\nu \int 1 - f(t) \, dt \right\} H_{Z_0}(f(0)). \end{aligned}$$

Now, from Wei et al. (2002, 2.2, Lemma 2), it is also known that the pgfl. G_N conditional on $X_0 = x$ is

$$\begin{aligned} G_N(\xi \mid X_0 = x) &= \mathbb{E} \left[\exp \left\{ \int_{\mathbb{R}} \log \xi(u) \, dN_u \right\} \mid X_0 = x \right] \\ &= \exp \left\{ \frac{x(f(0) - 1)}{\mu} - \nu \int 1 - f(t) \, dt \right\}, \end{aligned}$$

where f again satisfies the Riccati Equation (5.15). Let P_{X_0} denote the distribution function of X_0 . The unconditional pgfl. becomes

$$\begin{aligned} G_N(\xi) &= \int_{\mathbb{R}} \exp \left\{ \frac{x(f(0) - 1)}{\mu} - \nu \int 1 - f(t) \, dt \right\} dP_{X_0}(x) \\ &= \exp \left\{ -\nu \int 1 - f(t) \, dt \right\} \mathbb{E} \left(\exp \left\{ X_0 \frac{f(0) - 1}{\mu} \right\} \right) \\ &= \exp \left\{ -\nu \int 1 - f(t) \, dt \right\} M_{X_0} \left(\frac{f(0) - 1}{\mu} \right) \end{aligned}$$

where M_{X_0} denotes the moment generating function of X_0 . We can now see that, because of compatibility as in Definition 5.9, the distributions of D and N agree. $\square$

Proposition 5.12. Suppose a point process M is observed up to some time s , giving a σ -field $\mathcal{F}_s$. Without loss of generality, we set $s = 0$. We can interpret this point process

as the death sequence D in an IBD-process Z , or as Poisson process N modulated by the CIR-diffusion X . If the conditional distributions are as in Example 5.10 (iii), that is, $(X_0 | \mathcal{F}_0) = x_0 > 0$ and $(Z_0 | \mathcal{F}_0) \sim \text{Pois}(x_0/\mu)$, then any inferential or predictive results with respect to M_t for $t \geq 0$ agree for these two interpretations.

Proof. We have compatible priors at time 0 by Example 5.10 (iii), and by Lemma 5.11, we know that the point processes D and N have the same distribution *given* $\mathcal{F}_0$. We show that under the Bayesian updating procedure, the posterior distributions of $\hat{Z}_t = (Z_t | \mathcal{F}_t)$ and $\hat{X}_t = (X_t | \mathcal{F}_t)$ remain compatible for all $t > 0$ and $\mathcal{F}_t = \sigma(\mathcal{F}_0, \mathcal{H}_t^M)$. This will imply that D and N have the same distribution *given* $\mathcal{F}_t$, which gives the result.

(i) Suppose an event occurs at time $t > 0$, so that $M_t = \lim_{s \uparrow t} M_s + 1 \stackrel{\text{def}}{=} M_{t-} + 1$, and that the distributions of $\hat{X}_s$ and $\hat{Z}_s$ are compatible at all times $s < t$. We use a subscript “ $t-$ ” to denote left limits and left derivatives. To check compatibility at time t , we need to verify that

$$H_{\hat{Z}_t}(z+1) = \frac{H(t, z+1)}{H(t, 1)} = M_{\hat{X}_t}(z/\mu) = \frac{M(-z/\mu, t)}{M(0, t)} \quad (5.23)$$

where $H_{\hat{Z}_t}$ denotes the probability generating function of $\hat{Z}_t$, H is as in Proposition 5.7, $M_{\hat{X}_t}(z/\mu)$ is the moment generating function of $\hat{X}_t$, and M the function defined in Proposition 5.2.

We study the distribution of $\hat{X}_t$ first: Equation (4.18) implies that

$$\hat{p}_t(x) \propto \lim_{s \uparrow t} x \hat{p}_s(x) = x \hat{p}_{t-}(x);$$

and we let $C_1 = \int x \hat{p}_{t-}(x) dx$ denote the normalising constant. It follows that

$$\begin{aligned}
\frac{M(-z/\mu, t)}{M(0, t)} &= \frac{C_1^{-1} \int \exp\{z/\mu x\} x \hat{p}_{t-}(x) \exp\{-\int_0^t \hat{X}_s ds\} dx}{C_1^{-1} \int x \hat{p}_{t-}(x) \exp\{-\int_0^t \hat{X}_s ds\} dx} \\
&= C_1^{-1} \int \exp\{z/\mu x\} x \hat{p}_{t-}(x) dx \\
&= C_1^{-1} \mu \int \frac{d \exp\{z/\mu x\}}{dz} \hat{p}_{t-}(x) dx \\
&= C_1^{-1} \mu \frac{d}{dz} \left[\int \exp\{z/\mu x\} \hat{p}_{t-}(x) dx \right] \\
&= \frac{\mu}{C_1} \frac{1}{M(0, t-)} \frac{\partial M(-z/\mu, t-)}{\partial z}.
\end{aligned} \tag{5.24}$$

We now consider the distribution of $\hat{Z}_t$: Equation (5.16) implies that

$$\hat{p}_t(n) \propto (n+1) \hat{p}_{t-}(n+1),$$

and we let $C_2 = \sum_{n=0}^{\infty} (n+1) \hat{p}_{t-}(n+1)$ denote the normalising constant. It follows that

$$\begin{aligned}
\frac{H(t, z+1)}{H(t, 1)} &= \frac{C_2^{-1} \sum_{n=0}^{\infty} (z+1)^n (n+1) \hat{p}_{t-}(n+1) \exp\{-\int_0^t \hat{\mu}_s ds\} dx}{C_2^{-1} \sum_{n=0}^{\infty} (n+1) \hat{p}_{t-}(n+1) \exp\{-\int_0^t \hat{\mu}_s ds\} dx} \\
&= C_2^{-1} \sum_{n=0}^{\infty} (z+1)^n (n+1) \hat{p}_{t-}(n+1) \\
&= C_2^{-1} \sum_{n=0}^{\infty} \frac{d(z+1)^{n+1}}{dz} \hat{p}_{t-}(n+1) \\
&= C_2^{-1} \frac{d}{dz} \left[\sum_{n=0}^{\infty} (z+1)^{n+1} \hat{p}_{t-}(n+1) \right] \\
&= \frac{1}{C_2} \frac{1}{H(t-, 1)} \frac{\partial H(t-, z+1)}{\partial z}.
\end{aligned} \tag{5.25}$$

We have

$$\frac{1}{M(0, t-)} \frac{\partial M(-z/\mu, t-)}{\partial z} = \frac{1}{H(t-, 1)} \frac{\partial H(t-, z+1)}{\partial z}$$

because of compatibility at all times $s < t$. For the same reason, we have

$$C_1 = \lim_{s \uparrow t} \mathbb{E}(\hat{X}_s) = \mu \lim_{s \uparrow t} \mathbb{E}(\hat{Z}_s) = \mu C_2.$$

This shows the equality of Equations (5.24) and (5.25).

(ii) We now show that the posterior distributions remain compatible at all times t where no events occur. To make this precise, let $(T_n; n \in \mathbb{N} \setminus \{0\})$ denote the event times, and set $T_0 = 0$. Consider a time t such that $T_n < t < T_{n+1}$ for some $n \in \mathbb{N}$. By Propositions 5.2 and 5.7, the criterion for compatibility at time t in Equation (5.23) translates into checking whether

$$\begin{aligned} \frac{H(t - T_n, z + 1)}{H(t - T_n, 1)} &= \frac{h(i(t - T_n, z + 1))}{h(i(t - T_n, 1))} \left[\frac{i(t - T_n, z + 1)}{(z + 1)i(t - T_n, 1)} \right]^{\nu/\lambda} \\ &\stackrel{?}{=} \frac{M(-z/\mu, t - T_n)}{M(0, t - T_n)} = \left[\frac{A(-z/\mu, t - T_n)}{A(0, t - T_n)} \right]^{\nu/\lambda} \frac{m(B(-z/\mu, t - T_n))}{m(B(0, t - T_n))}, \end{aligned} \quad (5.26)$$

where h and m now denote *compatible* boundary conditions *at time* T_n , and $A(s, t)$ is as in Proposition 5.3.

It is a matter of simple algebra to check that

$$\frac{A(-z/\mu, t)}{A(0, t)} = \frac{i(t, z + 1)}{(z + 1)i(t, 1)}$$

for all z and t ; so the criterion in Equation (5.26) really comes down to checking whether

$$\frac{h(i(t - T_n, z + 1))}{h(i(t - T_n, 1))} \stackrel{?}{=} \frac{m(B(-z/\mu, t - T_n))}{m(B(0, t - T_n))} \quad (5.27)$$

for boundary conditions h and m that are compatible at T_n .

We show this by induction on the times T_n . For $n = 0$, i.e. at time $T_0 = 0$, we have

$$h(z) = \exp\{x_0/\mu(z - 1)\} \quad \text{and} \quad m(s) = \exp\{-sx_0\}$$

as boundary conditions. Plugging this into Equation (5.27) reduces to checking whether

$$\frac{x_0}{\mu}(i(t, z+1) - i(t, 1)) \stackrel{?}{=} x_0(B(0, t) - B(-z/\mu, t)),$$

which is easily verified. By Example 5.10 (iii), this choice of h and m is compatible at time 0, so we have compatibility at all times $0 \leq t < T_1$. Part (i) of the proof now applies to extend this to all times $0 \leq t \leq T_1$.

We can now proceed to the induction step and assume compatibility at all times $t \leq T_n$ for some n ; and we wish to extend it to all $t < T_{n+1}$. This can be achieved by establishing Equation (5.27) when choosing

$$h(z) = H(T_n, z)/H(T_n, 1) \quad \text{and} \quad m(s) = M(s, T_n)/M(0, T_n)$$

as the conditions at the “new boundary” T_n . Plugging this into Equation (5.27) gives

$$\frac{H(T_n, i(t - T_n, z + 1))/H(T_n, 1)}{H(T_n, i(t - T_n, 1))/H(T_n, 1)} \stackrel{?}{=} \frac{M(B(-z/\mu, t - T_n), T_n)/M(0, T_n)}{M(B(0, t - T_n), T_n)/M(0, T_n)} \quad (5.28)$$

for $T_n < t < T_{n+1}$. By compatibility at time T_n , we know that

$$H(T_n, x + 1)/H(T_n, 1) = M(-x/\mu, T_n)/M(0, T_n)$$

for all x . Equation (5.28) is hence verified if

$$-\mu B(-z/\mu, t - T_n) + 1 = i(t - T_n, z + 1).$$

This can easily be checked.

□

We obtain the same result for two other choices of the distribution of $(X_0 | \mathcal{F}_0)$ and $(Z_0 | \mathcal{F}_0)$:

Proposition 5.13. The results of Proposition 5.12 also hold for the initial conditions in Example 5.10 (i) and (ii).

Proof. The previous proof for compatibility at event times applies without change. The proof for compatibility between event times only needs to be changed with respect to the choice of h and m at time $t = 0$. These functions now are the pgf. of a Negative Binomial and the Laplace transform of a Gamma distribution, with parameters such that they are compatible. □

6 Numerical results

For the purposes of Bayesian statistical inference in a Poisson process N driven by a CIR diffusion X , approximations to the posterior mean and variance of the underlying diffusion may provide an adequate summary. In this chapter, we investigate several numerical methods to achieve this:

- (i) The equivalence result obtained in Proposition 5.12 yields an elegant and accurate technique: The interpretation of the point process N as death sequence of a population process means we are now concerned with a latent process in a discrete space, for which the filtering distributions can be described by the technique presented in Section 4.2.2. We present the detailed calculations in Section 6.1.
- (ii) An approach that relies on the asymptotic approximation of the filtering distribution by a normal distribution is explained in Section 6.2. This technique is applicable also to specifications of the diffusion X (driving the point process N) other than a CIR diffusion, and hence quite flexible; but there obviously remain errors due to the normal approximation that cannot be reduced with increased computational effort.
- (iii) A discretisation of the dynamical system, and an approximation of the filtering distributions in the resulting hidden Markov model by a particle filter is explained in Section 6.3. Again, this is a technique that is principally applicable to other specifications of X as well; but the choice of a CIR diffusion allows for a particularly convenient and easily implemented proposal step of these filtering algorithms. For moderate number of particles used, the accuracy of the resulting approximations

seems average for the first moment, and somewhat poor for the second moment. Convergence to the results obtained by the IBD interpretation in Section 6.1 can be achieved, but only with impractical computational effort.

We conclude the chapter by comparing these results to an *exponentially weighted moving average* (EWMA) filter, and show how to (approximately) predict the next event time in a CIR-driven point process.

6.1 Using the equivalence result

In this section we will show how the results obtained in Section 5.2 and the equivalence result in Proposition 5.12 can be put to practical use, namely to obtain sharp approximations to the posterior mean and variance of the filtering distribution $p(X_t | \mathcal{H}_t^N)$, for any $t \in [0, T]$.

The key idea is to interpret the point process N as the death sequence D in an immigration-birth-death process Z with parameters ν, λ, μ (which map to α, μ_0, β as in Section 5.1). Z is a process with values in $\mathbb{N}$. We will replace it with a process $\tilde{Z}$ that takes values in the finite subset $\{0, 1, \dots, n_0\} \subset \mathbb{N}$, and consider the associated departure process $\tilde{D}$. The filtering distribution $\Pr(\tilde{Z} = n | \mathcal{H}_t^{\tilde{D}})$ for this truncation of Z can be calculated exactly and will be a close approximation to $\Pr(Z = n | \mathcal{H}_t^D)$ for n_0 large enough. Up to this finite state-space approximation, we can hence calculate the posterior population size $\mathbb{E}(Z_t | \mathcal{H}_t^D)$, which is equal to $\mu^{-1}\mathbb{E}(X_t | \mathcal{H}_t^N)$ by Proposition 5.12.

Furthermore, the relation between the first and second posterior moment for the posterior mean intensity (explained in Corollary 4.14) can then be exploited to approximate the standard error of the filtered mean intensity between event times by numerical differentiation of $\hat{Z}_t$.

Approximation of the mean posterior intensity. Lemma 5.6 has shown that the posterior distribution $\hat{p}_t(n) = \Pr(Z_t = n \mid \mathcal{H}_t^D)$ of the population size satisfies the differential equation

$$\begin{aligned} d\hat{p}_t(n) = & \left(\mathbb{1}_{\{n>0\}} \hat{p}_t(n-1) \lambda(n-1) - \hat{p}_t(n) (\lambda(n) + \mu(n)) + \hat{p}_t(n+1) \mu(n+1) \right) dt \\ & + \left(\hat{p}_t(n+1) \mu(n+1) - \hat{p}_t(n) \hat{\mu}_t \right) \hat{\mu}_t^{-1} (dD_t - \hat{\mu}_t dt), \quad (6.1) \end{aligned}$$

where $\hat{p}_0$ is the prior distribution. We introduce a finite state space approximation to the population process Z by introducing a maximum population size $n_0 > 0$: The process $\tilde{Z}$ takes values in $\{0, \dots, n_0\}$ and has infinitesimal generator

$$\tilde{Q} = \begin{pmatrix} -\lambda(0) & \lambda(0) & 0 & \dots & 0 \\ \mu(1) & -(\mu(1) + \lambda(1)) & \lambda(1) & 0 & 0 \\ & & \dots & & \\ 0 & \dots & 0 & \mu(n_0 - 1) & -(\lambda(n_0 - 1) + \mu(n_0 - 1)) & \lambda(n_0 - 1) \\ 0 & \dots & 0 & & \mu(n_0) & -\mu(n_0) \end{pmatrix},$$

where $\lambda(i) = \nu + i\lambda$ and $\mu(i) = i\mu$ for $0 \leq i \leq n_0$. We also define the $(n_0 + 1) \times (n_0 + 1)$ -diagonal matrices

$$\Lambda = (\lambda_{ij}) = \begin{cases} \lambda(i) = \nu + i\lambda & \text{if } 1 \leq i = j \leq n_0, \\ 0 & \text{otherwise,} \end{cases}$$

and

$$M = (\mu_{ij}) = \begin{cases} \mu(i) = i\mu & \text{if } 2 \leq i = j \leq n_0 + 1, \\ 0 & \text{otherwise.} \end{cases}$$

We let $\tilde{D}$ denote the departure process associated with $\tilde{Z}$, and write $\tilde{p}_t(n) = \Pr(\tilde{Z} = n \mid \mathcal{H}_t^{\tilde{D}})$ for the distribution of the population size of $\tilde{Z}$ given the departure process up to time t . Between departure times given by $\tilde{D}$, we know that this probability vector satisfies the differential equation

$$d\tilde{p}_t(n) = \left(\mathbb{1}_{\{n>0\}} \tilde{p}_t(n-1) \lambda(n-1) - \tilde{p}_t(n) (\lambda(n) + \mu(n) - \tilde{\mu}_t) \right) dt,$$

where

$$\tilde{\mu}_t = \sum_{i=0}^{n_0} \mu(i) \tilde{p}_t(i).$$

We can write this as

$$\frac{d\tilde{p}}{dt} = (\Lambda + M + \tilde{\mu}_t I_{n_0+1} + \Lambda_0), \quad (6.2)$$

where I_{n_0+1} denotes the $(n_0 + 1) \times (n_0 + 1)$ -identity matrix, and

$$\Lambda_0 = \begin{pmatrix} 0 & 0 & 0 & \dots & 0 \\ \lambda(0) & 0 & 0 & 0 & \dots & 0 \\ & & \dots & & & \\ 0 & \dots & 0 & 0 & \lambda(n_0 - 2) & 0 \\ 0 & \dots & 0 & 0 & 0 & \lambda(n_0 - 1) \end{pmatrix}.$$

The solution of Equation (6.2) is then

$$\begin{aligned} \tilde{p}_t &= \exp\{t(\Lambda + M + \tilde{\mu}_t I_{n_0+1} + \Lambda_0)\} \tilde{p}_0 \\ &\propto \exp\{t(\Lambda + M + \Lambda_0)\} \tilde{p}_0. \end{aligned} \quad (6.3)$$

At departure times of $\tilde{D}$, we have

$$d\tilde{p}_t(n) = \left(\mathbb{1}_{\{n < n_0\}} \tilde{p}_t(n+1) \mu(n+1) \tilde{\mu}_t^{-1} - \tilde{p}_t(n) \right) d\tilde{D}_t,$$

which gives

$$\begin{aligned} \tilde{p}_t(n) &= \tilde{p}_{t-}(n) + \left(\mathbb{1}_{\{n < n_0\}} \tilde{p}_{t-}(n+1) \mu(n+1) \tilde{\mu}_{t-}^{-1} - \tilde{p}_{t-}(n) \right) \\ &\propto \mathbb{1}_{\{n < n_0\}} \tilde{p}_{t-}(n+1) \mu(n+1). \end{aligned} \quad (6.4)$$

For $n_0 \gg 0$, we can replace $\tilde{D}$ by D and expect that $\tilde{p}_t$ is very close to $\hat{p}_t$. In particular, we can calculate the constant of proportionality $\tilde{\mu}_t$ in Equation (6.3) and expect that

$$\tilde{\mu}_t \approx \hat{\mu}_t = \mu \hat{Z}_t = \mathbb{E}(X_t | \mathcal{H}_t^N)$$

is a very accurate approximation to the posterior mean intensity, as the (posterior) probability mass for a population size exceeding n_0 will be small and hence have a negligible impact on the mean population size.

Approximation of the posterior standard error. Recall that in Corollary 4.14 (with $\lambda(x) = x$, and hence $\hat{\lambda}_t = \hat{X}_t$), it was established that

$$\begin{aligned} d\hat{X}_t &= \mathbb{E}\left(-\alpha(X_t - \mu_0) \mid \mathcal{H}_t^N\right) dt + \mathbb{E}\left((X_t - \hat{X}_t)X_t \mid \mathcal{H}_t^N\right) \hat{X}_t^{-1} (dN_t - \hat{X}_t dt) \\ &= -\alpha(\hat{X}_t - \mu_0) dt + \mathbb{E}\left((X_t - \hat{X}_t)^2 \mid \mathcal{H}_t^N\right) \hat{X}_t^{-1} (dN_t - \hat{X}_t dt). \end{aligned}$$

Using $d\hat{X}_t = \mu d\hat{Z}_t$, this allows to obtain an approximation to the (posterior) mean squared error *between event times*, namely

$$\begin{aligned} \text{Var}\left(X_t \mid \mathcal{H}_t^N\right) &= \mathbb{E}\left((X_t - \hat{X}_t)^2 \mid \mathcal{H}_t^N\right) \\ &= \mu \hat{Z}_t \frac{\mu d\hat{Z}_t - \alpha(\mu_0 - \mu \hat{Z}_t) dt}{dN_t - \mu \hat{Z}_t dt} = -\mu \frac{d\hat{Z}_t}{dt} + \alpha(\mu_0 - \mu \hat{Z}_t) \end{aligned}$$

by numerical differentiation of $\hat{Z}_t$. Figure 6.1 illustrates these results, using as prior for $\widetilde{Z}_0$ a $\text{NegBin}(\nu/\lambda, \lambda/\mu)$, cut off at $n_0 = 50$. This Negative Binomial is the equilibrium distribution of Z_0 , and corresponds to the equilibrium distribution of X_0 , which is a $\text{Gamma}(\nu/\lambda, \lambda\mu/(\mu - \lambda))$.

The accuracy of the results displayed in Figure 6.1 can hence be compared to the exact mean and variance up to the first event time by means of Proposition 5.3; the difference (of absolute size $< 1 \cdot 10^{-10}$) between the results seems largely due to the numerical differentiation, and not the finite state space approximation made for $\widetilde{Z}$.

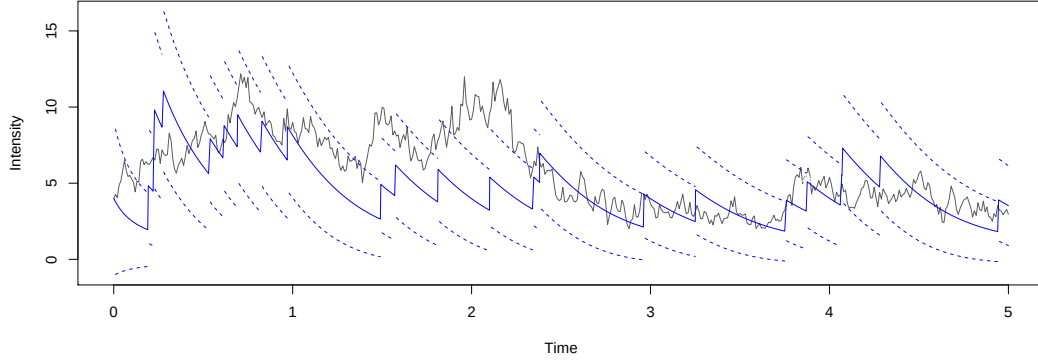


Figure 6.1 The grey solid line shows the simulated trajectory of a CIR diffusion with $\mu = 2$, $\lambda = 1.5$, $\nu = 1$. The blue solid line shows $\tilde{\mu}_t$, calculated at the union of all event times and intermediate sampling times equidistant from each other (0.01 time units as width). The blue dashed lines show the credible bounds between event times; that is, $\tilde{\mu}_t \pm \text{Var}(X_t | \mathcal{H}_t^N)^{1/2}$.

6.2 Asymptotic approximation of the first two moments

We compare these results with approximations suggested in Snyder (1972, sec. VI.). These results require that we consider the Lamperti transform

$$Y_t = L(X_t) = \frac{2\sqrt{X_t}}{\beta}$$

to obtain an SDE for Y with *constant* diffusion coefficient, namely

$$\begin{aligned} dY_t &= \left(-\frac{\alpha}{2} Y_t + \frac{4\alpha\mu_0 - \beta^2}{2\beta^2 Y_t} \right) dt + dW_t \\ &= f(Y_t) dt + dW_t. \end{aligned} \tag{6.5}$$

The results in Snyder (1972, sec. VI.) now give approximations Y_t^* and Σ_t^* for the posterior mean and variance of Y_t . These results rely on an approximation of the posterior density by a normal distribution (with error terms of higher than second order neglected); and

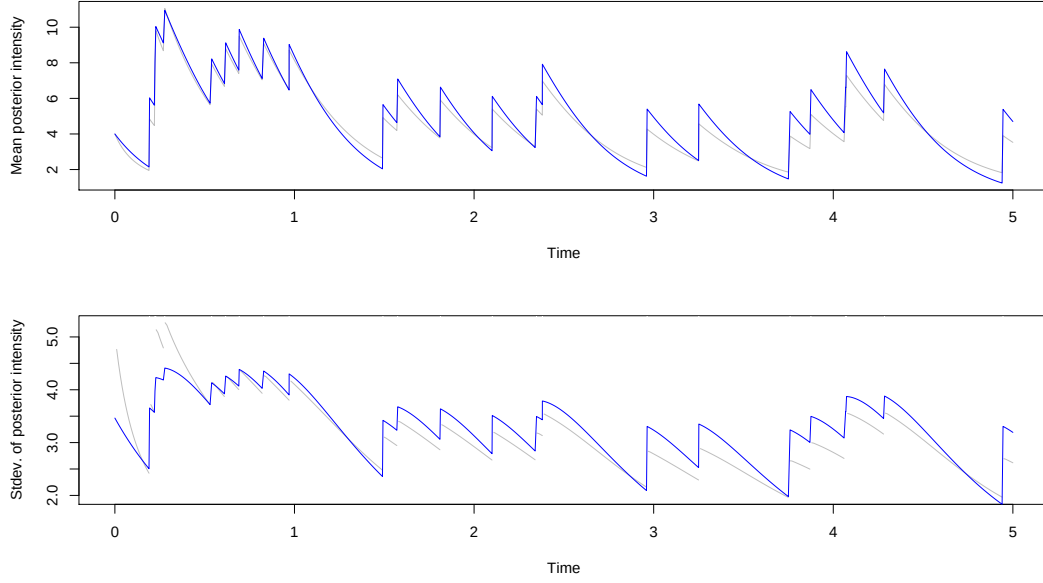


Figure 6.2 We consider the same simulated data as in Figure 6.1. The top plot shows the posterior mean intensities X_t^* in blue, and $\tilde{\mu}_t$ as reference in grey. The bottom plot shows $(\Psi_t^*)^{1/2}$ in blue, and the approximation to the posterior standard deviation discussed in Section 6.1 as reference in grey.

yield the following system of differential equations:

$$dY_t^* = \left(f(Y_t^*) + \frac{1}{2} \Sigma_t^* f''(Y_t^*) - \frac{\Sigma_t^{*2} \beta^2}{2} Y_t^* \right) dt + \frac{2\Sigma_t^*}{Y_t^*} dN_t, \quad (6.6)$$

$$d\Sigma_t^* = \left(1 + 2f'(Y_t^*)\Sigma_t^* - \frac{\Sigma_t^{*2}\beta}{2} \right) dt - \frac{2\Sigma_t^{*2}}{Y_t^{*2}} dN_t, \quad (6.7)$$

where f is the function introduced in Equation (6.5). Estimates for the first two moments of X_t are then obtained via the Delta-method (for details, see for instance Davison (2008, sec. 2.2.2.)) as

$$X_t^* = L^{-1}(Y_t^*); \quad \Psi_t^* = \left(\frac{\partial L^{-1}}{\partial y} \Big|_{y=Y_t^*} \right)^2 \Sigma_t^*. \quad (6.8)$$

An illustration of this approximation is given in Figure 6.2. The performance seems satisfactory; but it should be noted that the numerical solution of the coupled Equations (6.6) and (6.7) can be a delicate task. This is because Σ_t^* in (6.7) decreases at event times (the coefficient for dN_t is negative), which could result in negative values whenever $dN_t = 1$. Updating Y_t^* *before* Σ_t^* is advisable, because the increase in the former estimate at event times implies an increase in the (negative) coefficient for dN_t .

6.3 A particle filter for the discretised dynamical system

The filtering problem outlined at the beginning of this chapter can be also be approached by first discretising the dynamics for both X and N , and then approximating the sequence of filtering distributions in the resulting hidden Markov Model by a *particle filter*.

For a fixed observation period $[0, T]$, we let $(s_n; n \geq 0, s_0 = 0)$ denote a set of (increasing) discretisation times at which we sample the point process N . We let $\Delta_n = (s_{n+1} - s_n)$ denote the time increment between two sampling times, and write $\Delta = \Delta_n$ for brevity (although the time increments are not equidistant and do depend on n). We assume the event times are a subset of the sampling times. The hidden state of the system at step $n \geq 1$ is given by the pair $\tilde{X}_n = (X_{s_{n-1}}, X_{s_n})$, and the observation process is the increment $Y_n = (N_{s_n} - N_{s_{n-1}})$ in the point process. As we sample at least at all event times, we have $Y_n \in \{0, 1\}$; and $Y_n = 1$ implies that $N_{s_n} = \lim_{t \uparrow s_n} N_t + 1$.

The dependency graph of this hidden Markov model is then in standard form (see Figure 6.3), and we let F denote the transition kernel for the hidden variable, and write G for the likelihood.

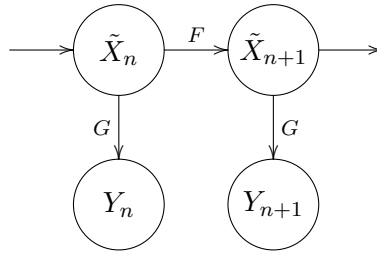


Figure 6.3 Dependency graph for the hidden Markov model resulting from discretisation at times $(s_n; n \geq 0)$.

We make the role of F and G more precise: Let $A, B \subset \mathbb{R}$ denote measurable sets, and let $x \in \mathbb{R}$. The transition kernel

$$F\left((X_{s_{n-1}}, X_{s_n} = x_0), (A, B)\right) = \mathbb{1}_{\{x_0 \in A\}} \Pr(X_{s_{n+1}} \in B \mid X_{s_n} = x_0) \quad (6.9)$$

is defined by the transition probabilities for the diffusion X ; and for the likelihood of $\tilde{X}_{n+1} = (x_0, x)$ we have

$$G((x_0, x), Y_{n+1}) = x^{\mathbb{1}_{\{Y_{n+1}=1\}}} \mathbb{E} \left(e^{-\int_{s_n}^{s_{n+1}} X_t dt} \middle| X_{s_n} = x_0, X_{s_{n+1}} = x \right). \quad (6.10)$$

We use a Sequential Importance Sampling scheme with (multinomial) Resampling (SISR, see Cappé et al. (2005, Alg. 7.3.4)) to approximate the filtering distributions

$$p(X_{s_n} \mid \mathcal{H}_{s_n}^N) = \int p(\tilde{X}_n = (X_{s_{n-1}}, X_{s_n}) \mid Y_{0:n}) dX_{s_{n-1}}$$

by the discrete measure

$$\hat{p}_n(x) = \sum_{i=1}^M w_n^i \mathbb{1}_{\{x=\xi_n^i\}}.$$

In this equation, $(\xi_n^i, w_n^i)_{1 \leq i \leq M}$ denotes the family of particles at the n -th time step, where the i -th particle has location ξ_n^i and weight w_n^i . Resampling of the particles is effected at deterministic and random times; namely at integer multiples of a fixed time interval, and also whenever an event in the point process occurs. An important design choice for a particle filter is the *proposal kernel* R from which we sample proposed particle locations $\tilde{\xi}_{n+1}^i \sim R(\xi_n^i, \cdot)$ to effect an update of $\hat{p}_n$ to $\hat{p}_{n+1}$.

Choosing the prior kernel. The simplest choice (resulting in what is commonly known as the *Bootstrap filter*, first put forward by Gordon et al. (1993)) is to use the dynamics described by the kernel F . In simplified notation (omitting the repeated component in $\tilde{X}_n$), this means that

$$\tilde{\xi}_{n+1}^i \sim R(\xi_n^i, \cdot) \sim (X_{s_{n+1}} \mid X_{s_n} = \xi_n^i),$$

i.e. the family of particles follows the (prior) dynamics of the hidden variable without taking into account the observation Y_{n+1} . This *prior kernel* can usually be exactly sampled from and is hence convenient to implement; but this choice can fail to concentrate particles in regions of high posterior probability (see Cappé et al. (2005, 7.2.2.2) for details).

By the result stated in Equation (5.4) in Section 5.1, we know that this choice for R implies that

$$\tilde{\xi}_{n+1}^i = \left(\tilde{\xi}_{n+1}^i \mid \xi_n^i \right) \sim Z \frac{\beta^2(1 - e^{-\alpha\Delta})}{2\alpha},$$

where Z is a non-central Chi-squared random variable with $\frac{4\alpha\mu_0}{\beta^2}$ degrees of freedom, and non-centrality parameter

$$\frac{4\alpha\xi_n^i e^{-\alpha\Delta}}{\beta^2(1 - e^{-\alpha\Delta})}.$$

The weights corresponding to this kernel are given by the likelihood $w_{n+1}^i \propto G\left((\xi_n^i, \tilde{\xi}_{n+1}^i), Y_{n+1}\right)$, which we approximate by

$$w_{n+1}^i \propto \left(\tilde{\xi}_{n+1}^i \right)^{\mathbb{1}_{\{Y_{n+1}=1\}}} \exp \left\{ -\frac{\tilde{\xi}_{n+1}^i + \xi_n^i}{2} \Delta \right\}.$$

Choosing the optimal kernel. An alternative to the prior kernel is the so-called *optimal kernel*, defined by

$$\tilde{\xi}_{n+1}^i \sim R\left(\xi_n^i, \cdot\right) \sim \left(X_{s_{n+1}} \mid X_{s_n} = \xi_n^i, Y_{n+1}\right).$$

We see that the proposal distribution now also depends on Y_{n+1} ; the purpose of this design is to steer particles into regions of high posterior probability and to reduce the variance of the weights. For a precise discussion of these ideas, see Cappé et al. (2005, sec. 7.2.2.2).

In the situation at hand, sampling from this optimal kernel can indeed be effected. By Proposition 5.5, we know that $\left(\tilde{\xi}_{n+1}^i \mid \xi_n^i, Y_{n+1}\right)$ is to be drawn from a Poisson mixture of Gamma distributions, where the Poisson random variable S has rate $\xi_n^i \rho(\Delta)$, and the Gamma distributions have shape $(S + \nu/\lambda + \mathbb{1}_{\{Y_{n+1}=1\}})$ and scale $\theta(\Delta)$, as defined in Proposition 5.4.

It is known that the weight w_{n+1}^i corresponding to the i -th proposed particle is independent of the location $\tilde{\xi}_{n+1}^i$, and satisfies (Cappé et al., 2005, sec. 7.2.2.2)

$$w_{n+1}^i \propto \int G\left((\xi_n^i, x), Y_{n+1}\right) F\left((\xi_{n-1}^i, \xi_n^i), \left(\xi_n^i, dx\right)\right). \quad (6.11)$$

We first assume that s_{n+1} is no event time, i.e. $Y_{n+1} = 0$. We approximate the likelihood G again by

$$G((\xi_n^i, x), Y_{n+1} = 0) \approx \exp \left\{ -\frac{\xi_n^i + x}{2} \Delta \right\}. \quad (6.12)$$

Letting $g(x; k, \psi)$ denote the density of a $\text{Gamma}(k, \psi)$ -random variable, and plugging this approximation into Equation (6.11), we get

$$\begin{aligned} (w_{n+1}^i | Y_{n+1} = 0) &\propto e^{-\xi_n^i \rho(\Delta)} \\ &\times \sum_{s=0}^{\infty} \left[\int_{\mathbb{R}_+} g(x; s + \nu/\lambda, \theta(\Delta)) \exp \left\{ -\frac{\xi_n^i + x}{2} \Delta \right\} dx \frac{(\xi_n^i \rho(\Delta))^s}{s!} \right] \\ &= e^{-\xi_n^i \rho(\Delta)} \sum_{s=0}^{\infty} \left[e^{-\xi_n^i \Delta/2} \left(\frac{2}{\Delta \theta(\Delta) + 2} \right)^{s+\nu/\lambda} \frac{(\xi_n^i \rho(\Delta))^s}{s!} \right] \\ &\propto \exp \left\{ -\xi_n^i \left(\rho(\Delta) + \Delta/2 - \frac{2\rho(\Delta)}{\Delta \theta(\Delta) + 2} \right) \right\}. \end{aligned}$$

If s_{n+1} is an event time, this weight needs to be multiplied by the expected value of the proposal $\tilde{\xi}_{n+1}^i$, which equals $\theta(\Delta) (\xi_n^i \rho(\Delta) + \nu/\lambda + 1)$. We hence get

$$w_{n+1}^i \propto (\xi_n^i \rho(\Delta) + \nu/\lambda + 1)^{\mathbb{1}_{\{Y_{n+1}=1\}}} \exp \left\{ -\xi_n^i \left(\rho(\Delta) + \Delta/2 - \frac{2\rho(\Delta)}{\Delta \theta(\Delta) + 2} \right) \right\}. \quad (6.13)$$

A comparison of the two particle filter implementations is given in Figure 6.4. The results suggest it is advantageous to choose the optimal kernel; convergence to the reference estimates from the IBD interpretation can be achieved for $M \geq 200000$ particles.

6.4 An EWMA filter

An *exponentially weighted moving average* (EWMA) filter, defined by

$$S(t, (N_s; 0 \leq s \leq t)) = S(t) = \rho \int_0^t e^{-\rho(t-u)} dN_u = \rho \sum_{T_i \leq t} e^{-\rho(t-T_i)}, \quad (6.14)$$

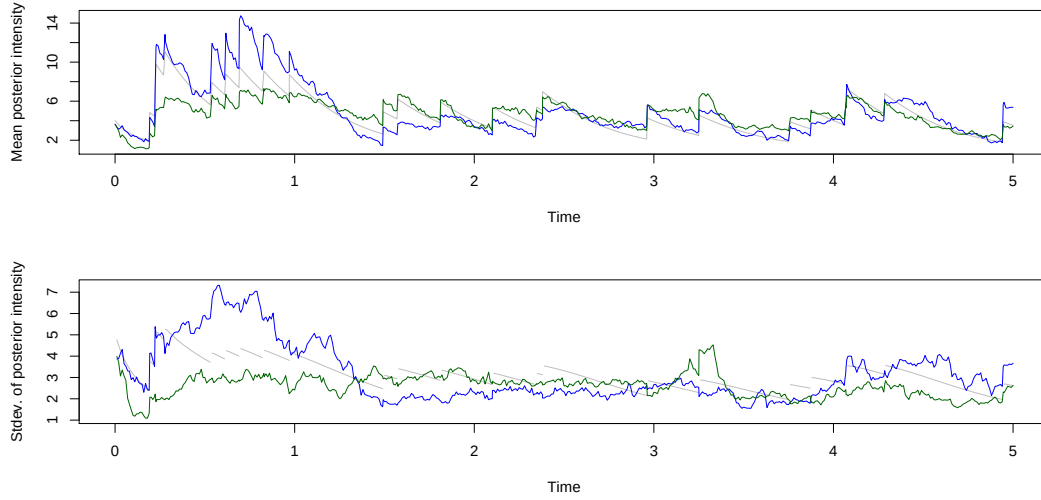


Figure 6.4 We consider the same simulated data as in Figure 6.1. The top plot shows the posterior mean $M^{-1} \sum_{i=1}^M \xi_n^i$ (we resample at every time step and so have equal weights) for the same sampling times as for the filter in Figure 6.1. Blue represents the prior kernel, green the optimal kernel, and grey the reference mean $\tilde{\mu}_t$. The lower plot shows the approximation to the posterior standard deviation (using the same colour code).

We have used $M = 250$ particles. The prior distributions at time 0 are the respective equilibrium distributions; that is, a $\text{Gamma}(2\alpha\mu_0/\beta^2 = \nu/\lambda, \beta^2/(2\alpha) = \lambda\mu/(\mu - \lambda))$ for the particle filters, and a $\text{NegBin}(\nu/\lambda, \lambda/\mu)$ for the IBD process filter.

where T_i denote the event times of N , estimates the underlying intensity X_t without taking into account the diffusion dynamics of this process. The choice of ρ is crucial for the performance of this filter, as it controls how quickly the impact of past observations decays. This clearly needs to be related to the amount of autocorrelation present in the driving process X , which is $e^{-\alpha t}$ on a time scale of t (see Wei (1995, Sec. 6.3.1) or (Iacus, 2008, Sec. 1.13.3)). This suggests to choose $\rho \approx \alpha^{-1}$, as more recent observations should receive higher weights in the presence of greater autocorrelation. The filter S , using $\rho = \alpha^{-1}$, works remarkably well, as demonstrated in Figure 6.5.

6.5 An application: Predicting the next event time

Let τ denote an inter-arrival time between two events in N , that is, $\tau = (T_{n+1} - T_n)$ for some $n \in \mathbb{N}$. We can set $T_n = 0$ for simplicity. For $0 = T_n < t < T_{n+1}$ and $h > 0$, we want

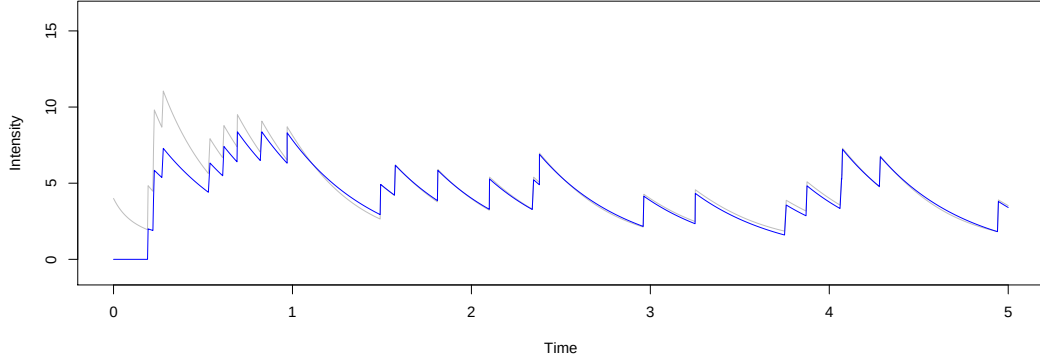


Figure 6.5 We consider the same simulated data as in Figure 6.1. The blue line represents the EWMA filter with $\rho = \alpha^{-1}$; the grey line represents the reference mean $\hat{\mu}_t$.

to (approximately) calculate the posterior survival probability

$$\Pr(\tau > t + h \mid \mathcal{H}_t^N)$$

of not seeing an event in the next h units of time, given that we have observed N up to time t . We can assume that $t < \tau$; otherwise the above expression equals 0.

Example 3.1 in Leijdekker and Spreij (2011) shows that

$$\begin{aligned} \Pr(\tau > t + h \mid \mathcal{H}_t^N) &= \exp\{-\alpha\mu_0\phi(h)\} \mathbb{E}\left(\exp\{-\psi(h)X_t\} \mid \mathcal{H}_t^N\right) \\ &= \exp\{-\alpha\mu_0\phi(h)\} M(\psi(h), t)/M(0, t), \end{aligned} \quad (6.15)$$

where M denotes the Laplace transform first introduced in Proposition 5.1, and

$$\begin{aligned} \phi(h) &= -\frac{2}{\beta^2} \log\left(\frac{2\gamma \exp\{h(\gamma + \alpha)/2\}}{\gamma - \alpha + (\gamma + \alpha) \exp\{h\gamma\}}\right), \\ \psi(h) &= \frac{2(\exp\{h\gamma\} - 1)}{\gamma - \alpha + (\gamma + \alpha) \exp\{h\gamma\}}, \end{aligned}$$

with $\gamma = \sqrt{\alpha^2 + 2\beta^2}$.

The evaluation of Equation (6.15) depends on the information given by $\mathcal{H}_0^N$ (as no events take place in $[0, t]$), which is represented by the boundary condition $m(s) = M(s, 0)$ at time 0. No analytic expression for this boundary condition may be available, in which

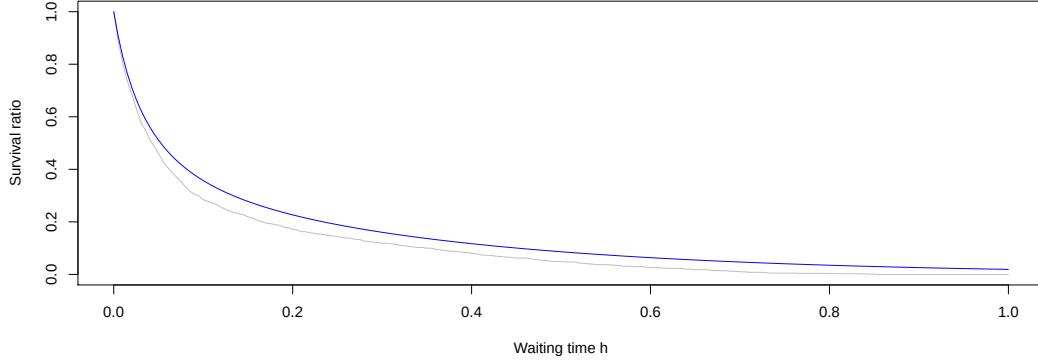


Figure 6.6 The blue solid line shows the approximate posterior probabilities for not seeing another event within the next h time units, using a Gamma approximation to the boundary condition. The grey line shows the empirical survivor function $S_{\text{emp}}(h)$.

case appropriate approximations need to be used instead. A first idea is to use the results in Section 6.1 on the filtered first two moments; and to approximate $m(s)$ by the Laplace transform of a Gamma distribution with mean and variance equal to conditional expectation $\hat{X}_0$ and conditional variance $\text{Var}(X_0 | \mathcal{H}_0^N)$.

We illustrate the accuracy of this approach by considering a simulated path $N = (N_t; 0 \leq t \leq T = 20)$ driven by a CIR diffusion with $\mu = 6$, $\lambda = 5$, $\nu = 3$. We fix some $h > 0$, and calculate approximations to Equation (6.15) at times $(t_m; 1 \leq m \leq M = 1000)$ within $[0, 20]$, using $\hat{X}_{t_m}$ and $\text{Var}(X_{t_m} | \mathcal{H}_{t_m}^N)$. We record the mean of these approximate survival probabilities. At each of the t_m , we also record the waiting time τ_m until the next event occurs in the given path of N ; and we calculate the empirical survival function $S_{\text{emp}}(h) = M^{-1} \#\{\tau_m; \tau_m > h\}$ as the average of those empirical waiting times that exceed h .

The resulting curves are plotted as functions of $h \in [0, 1]$ in Figure 6.6. The agreement is not fully convincing. This is due to the choice of a single Gamma variate to represent the boundary condition. The results in Leijdekker and Spreij (2011) show that (ignoring the initial distribution, which will have less and less impact the more events occur resp. time passes) the posterior distribution is a *mixture* of Gamma variates. A *single* Gamma variate, with moments chosen to match the posterior mean and variance, would hence be

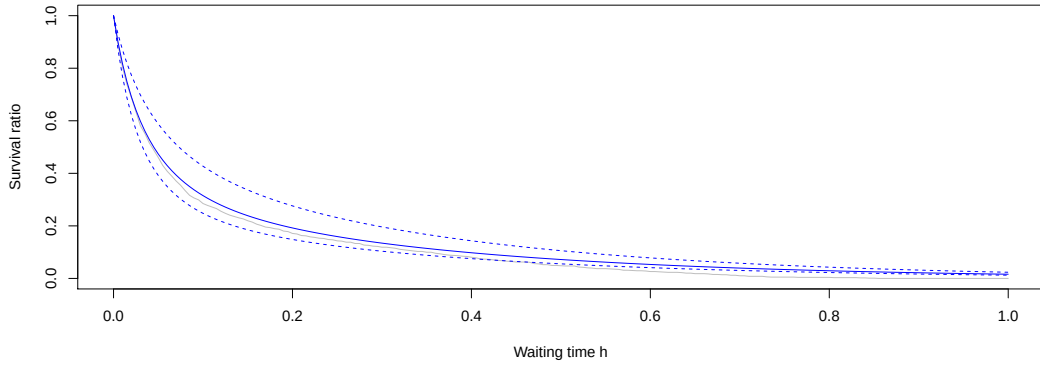


Figure 6.7 The blue solid line shows the approximate posterior probabilities for not seeing another event within the next h time units, using as boundary condition a distribution that concentrates at the posterior mean. The two dashed blue lines show these approximate posterior probabilities with boundary conditions that concentrate at the posterior mean plus/minus the posterior standard deviation.

over-dispersed relative to the target.

An alternative is to assume a boundary condition that concentrates at the posterior mean, and to assess the accuracy by choosing two other boundary conditions that concentrate at that posterior mean plus and minus the posterior standard deviation. The resulting functions are plotted in Figure 6.7. We see that this results in a closer fit, and an informative indication of those values of h for which our approximation is most sensitive to uncertainty about the intensity process X .

7 Parameter estimation and model fit

Let $N = (N_t; 0 \leq t \leq T)$ denote a finite path of an (observed) point process; and assume it has an intensity process $X = (X_t; 0 \leq t \leq T)$ given by a CIR diffusion in equilibrium, i.e. X_0 has a Gamma distribution with mean μ_0 and variance $\beta^2\mu_0/(2\alpha)$.

Throughout this chapter, we assume that α, μ_0, β are unknown. The first part of this chapter explains two techniques to estimate these parameters from the data given by N : the first approach is to use the likelihood function of the event times that can be obtained from interpreting N as death sequence of an immigration-birth-death process; and the second approach is a method of moments technique that relates the observed mean, variance and autocorrelation structure in N to the corresponding properties of X .

The second part of this chapter briefly surveys several techniques to assess the fit of a stochastic intensity model (not necessarily a CIR diffusion) for a given finite path of a point process. These techniques are based on the fact that, if the point process admits an intensity, the event times can be transformed in such a way that they become the event times of a standard Poisson process with unit rate. This *random change of time* for point processes is analogous to the inversion of the cumulative distribution function for continuous univariate random variables, and it yields graphical diagnostics that are easy to interpret.

7.1 Parameter estimation

7.1.1 Maximum likelihood estimation

The statistical equivalence between a CIR driven point process N and the death sequence D associated with an IBD population process $Y = (Y_t; 0 \leq t \leq T)$ (established in Clifford and Wei (1993)) allows the likelihood function for N to be expressed as

$$L(\mu, \lambda, \nu; T_1, \dots, T_n), \quad (7.1)$$

i.e. in terms of μ , λ and ν , and the death times $(T_k; 1 \leq k \leq n)$. In Wei (1995, Chapter 6), it is explained that

$$L(\nu, \lambda, \mu; T_1, \dots, T_n) = Q \sum_{j=0}^n a'_j \frac{\Gamma(\delta + j)}{\Gamma(\delta)} M^j,$$

where the coefficients a'_j are (complicated, but recursively defined) functions of *all* event times $(T_k; 1 \leq k \leq n)$ for each j . M only depends on T and $(T - T_n)$; and Q is a product of n terms, each involving only one of the T_k . Furthermore, a numerically stable implementation of the log-likelihood function $l = \log L$ is described. Note however that equation (6.38) in Wei (1995) relating to this implementation is incorrect; the product ends at n , not $(n + 1)$.

We wish to assess the accuracy of the maximum likelihood estimates derived from $l = \log L$ using a simulated CIR driven point process with parameters $\alpha = 0.5$, $\beta = 2.45$ and $\mu_0 = 8$; these values correspond to $\mu = 2$, $\lambda = 1.5$ and $\nu = 2$ for the associated IBD process.

Judging whether a maximum likelihood estimate for the parameters μ , λ , ν is close or far from the true values is problematic. In many applications, the interpretation of a point process N as a doubly stochastic process driven by a diffusion X is quite natural; and the death sequence D associated to the population process Y is only introduced to facilitate the analysis of relevant and interesting properties of N and X , but bears no physical meaning itself. The maximum likelihood estimates of μ , λ , ν should hence be

transformed to estimates of the parameters in the SDE for X ; or, even better, to a set of quantities that characterise the equilibrium distribution of X and the speed of mean reversion or amount of autocorrelation of X . This re-parametrisation usually allows for a much easier assessment of whether a given MLE is a good and accurate estimate of the model parameters.

We adopt the suggestion in Wei (1995) and use the equilibrium mean μ_0 and variance V of X , and the “autocorrelation coefficient” α . The latter choice is justified as (see Iacus (2008, sec. 1.13.3))

$$\text{Cov}(X_s, X_t) = \frac{\beta^2 \mu_0}{2\alpha} e^{-\alpha(t-s)},$$

and hence

$$\text{Cor}(X_s, X_t) = e^{-\alpha(t-s)}.$$

It can be checked that these parameters map to μ , λ , ν and vice-versa as

$$\begin{pmatrix} \alpha & = & \mu - \lambda \\ \mu_0 & = & \frac{\mu\nu}{\alpha} \\ V & = & \frac{\nu\lambda\mu^2}{\alpha^2} \end{pmatrix} \longleftrightarrow \begin{pmatrix} \mu & = & \frac{\alpha + \sqrt{4\alpha V/\mu_0 + \alpha^2}}{2} \\ \lambda & = & \frac{-\alpha + \sqrt{4\alpha V/\mu_0 + \alpha^2}}{2} \\ \nu & = & \lambda\mu_0^2 V \end{pmatrix};$$

but note that this correspondence is incorrectly stated in Wei (1995, Lemma 12). We will write $\theta = (\alpha, \mu_0, V)$ for brevity.

We use different lengths of time T to understand how many event times are required for a close fit of the MLE. Table 7.1 shows the values of the maximum likelihood estimates for several values of T . At least for the true parameter θ in question, these results suggest to choose T such that $N_T \geq 1000$. The inverse of the observed information matrix $j(\theta)$,

	$N_{100} = 659$	$N_{200} = 1610$	$N_{500} = 4342$	True
$\hat{\alpha}$	0.51	0.44	0.49	0.5
$\hat{\mu}_0$	6.60	8.35	8.69	8
$\hat{V}$	39.15	63.66	53.84	48

Table 7.1 Values of the MLE $\hat{\theta}$ for different values of T .

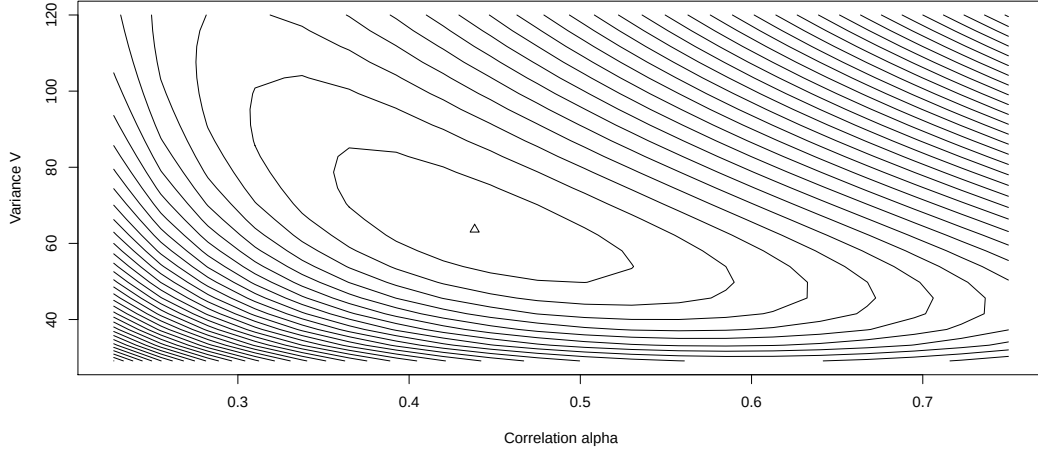


Figure 7.1 Contour lines of the profile likelihood l_p based on $(N_t; 0 \leq t \leq 200)$ as described above. The location of the MLE is marked by a triangle.

evaluated at the maximum likelihood estimate $\hat{\theta} = (\hat{\alpha}, \hat{\mu}_0, \hat{V})$, can be used to substantiate and refine this ad-hoc recommendation. For $T = 200$ for instance, we obtain the following estimate of the asymptotic covariance structure of the maximum likelihood estimate:

$$j(\hat{\theta})^{-1} = \begin{pmatrix} 0.01 & -0.08 & -1.66 \\ -0.08 & 1.53 & 23.69 \\ -1.66 & 23.69 & 446.33 \end{pmatrix}.$$

The strong positive correlation between μ_0 and V is unsurprising; but the relation between α and V deserves further attention. We follow the suggestion of Wei and study the profile log-likelihood

$$l_p(\alpha, V) = \max_{\mu_0} \log L(\alpha, \mu_0, V)$$

to remove the (usually easily and accurately estimated) mean parameter μ_0 . A contour plot of l_p for the data relating to $T = 200$ is shown in Figure 7.1 and confirms the correlation structure between α and V .

7.1.2 Method of Moments

We now describe a technique to fit the parameter θ directly by relating α , μ_0 , V (the autocorrelation, mean, and variance of X) to the *observed* autocorrelation, mean, and variance of the counts given by N .

The mean parameter μ_0 can be estimated immediately by $\widehat{\mu}_0 = N_T/T$. This is of course justified by

$$\mathbb{E}(N_T \mid X_t, 0 \leq t \leq T) = \int_0^T X_t \, dt;$$

and hence, unconditionally,

$$\mathbb{E}(N_T) = \mathbb{E} \left(\int_0^T X_t \, dt \right) = \int_0^T \mathbb{E}(X_t) \, dt = T\mu_0.$$

The estimation of the correlation structure of X via the sample correlation of N is more involved. We use a result in Grandell (1976, Chapter 7) that applies more generally to any point process with stationary intensity. We choose some fixed discretisation constant $\Delta > 0$, and assume for simplicity that $T/\Delta = n$ for some integer $n > 0$. (Otherwise we could discard some of the observations and replace T by $\lfloor T/\Delta \rfloor$.) We let $N_t^\Delta = (N_{t+\Delta} - N_t)$ denote the increment of N over the time interval $(t, t + \Delta]$.

Conditional on X_t , the increment N_t^Δ then has the Poisson distribution with intensity

$$X_t^\Delta \stackrel{\text{def}}{=} \int_t^{t+\Delta} X_t \, dt.$$

We introduce the quantity

$$\widehat{r}_k = \frac{1}{n-k} C_k^N - \mathbb{1}_{\{k=0\}} \mu_0 \Delta,$$

where

$$C_k^N = \sum_{j=1}^{n-k} \left(N_{j\Delta}^\Delta - \mu_0 \Delta \right) \left(N_{(j+k)\Delta}^\Delta - \mu_0 \Delta \right),$$

and $k \in \{0, \dots, n-1\}$. The $\widehat{r}_k$ can then be used as estimator for the covariance of X , because

$$\mathbb{E}(\widehat{r}_k) = \mathbb{E}(X_t^\Delta - \mu_0 \Delta)(X_{t+k\Delta}^\Delta - \mu_0 \Delta) \approx \Delta^2 \text{Cov}(X_t, X_{t+k\Delta}). \quad (7.2)$$

For details regarding this result and also a discussion of the variance of $\widehat{r}_k$, see Grandell (1976, sec. 7.1). Upon replacing μ_0 by $\widehat{\mu}_0$ in Equation (7.2), we obtain an estimator for the variance $V = \beta^2 \mu_0 / (2\alpha)$ of X in equilibrium, namely

$$\widehat{V} = \Delta^{-2} \widehat{r}_0.$$

We also recall that

$$\text{Cov}(X_s, X_t) = \frac{\beta^2 \mu_0}{2\alpha} e^{-\alpha(t-s)} = V \text{Cor}(X_s, X_t),$$

so the correlation estimates satisfy

$$\frac{\widehat{r}_k}{\widehat{r}_0} \approx \text{Cor}(X_t, X_{t+k\Delta}) = e^{-\alpha k \Delta}.$$

We can then derive an estimate $\widehat{\alpha}$ by fitting a correlation curve $e^{-\widehat{\alpha} k \Delta}$ to the sample correlations $\widehat{r}_k / \widehat{r}_0$ for several values of k .

This can be done in numerous way; we describe a simple proposal using the event times of a simulated CIR driven point process with parameters $\alpha = 0.5$, $\beta = 2.45$ and $\mu_0 = 8$. This corresponds to parameter values $\mu = 2$, $\lambda = 1.5$, and $\nu = 2$ for the death sequence in an IBD process. We set $T = 200$, and use the same dataset that gave $N_T = 1610$ event times in Section 7.1.1. Figure 7.2 shows the empirical autocorrelation function $\widehat{\rho}(k\Delta) = \widehat{r}_k / \widehat{r}_0$ at times $k\Delta$ for $k \in \{0, \dots, 15\}$ and $\Delta = 1$, and also the fitted curve $e^{-\widehat{\alpha} k \Delta}$. This fit has been achieved by minimising the (rather arbitrarily chosen) penalty function

$$\sum_{k=1}^{15} w_k \left(\widehat{\rho}(k\Delta) - e^{-\widehat{\alpha} k \Delta} \right)^2, \quad (7.3)$$

where $w_k = 1$ for $1 \leq k \leq 5$, $w_k = 0.5$ for $6 \leq k \leq 10$, and $w_k = 0.25$ for $11 \leq k \leq 15$

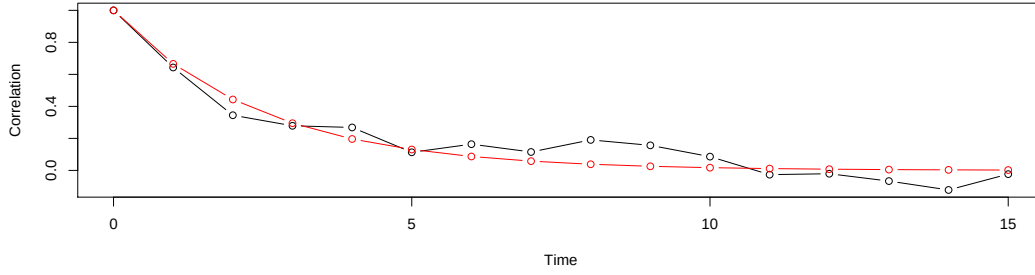


Figure 7.2 The sample autocorrelation curve $\hat{\rho}$ for $\Delta = 1$ at $k \in \{0, \dots, 15\}$ (corresponding to times $\{0, \dots, 15\}$) is in black; the fitted autocorrelation curve in red.

are decreasing weights. (This accounts for the observation that, to establish the degree of decay in autocorrelation, the first lags are much more informative than later ones.)

The parameter estimates obtained by choosing different values of Δ are summarised in Table 7.2. The performance of this technique (at least for the parameter considered) seems very satisfactory for $0.1 \leq \Delta \leq 1$. Values of $\Delta \ll 0.1$ result in a greater fluctuation of $\hat{\rho}(k\Delta)$ over the same range of k ; to compensate for this, the minimisation task in Equation (7.3) needs to be performed over more than just 15 components. Values of $\Delta \gg 1$ result in a shorter, possibly overly smooth increment process N_t^Δ ; and this may result in a systematic underestimation of the variability V of X .

It is difficult to give sound general recommendations on the choice of Δ ; but it seems advisable to have at least $n = T/\Delta = 200$ observations of the increment process N_t^Δ . In practice, the size of α will be the dominating factor for the quality of the estimate: a high mean reversion (ie. $\alpha > 1$) will imply a fast decay of the correlation in the increment counts N_t^Δ for most values of Δ ; and this loss of information can usually only be compensated for

	$\Delta = 0.1$	$\Delta = 1$	$\Delta = 2.5$	MLE	True
$\hat{\alpha}$	0.47	0.41	0.27	0.44	0.5
$\hat{\mu}_0$	8.05	8.05	8.05	8.35	8
$\hat{V}$	55.30	52.64	39.09	63.66	48

Table 7.2 Method of moments estimate for a simulated point process with $T = 200$ and $N_T = 1610$ events, using different choices for Δ .

by collecting more information, that is, by increasing T rather than attempting to change Δ .

More general solutions for the estimation of the autocorrelation function of the arrival rate in a point process than our suggestion in Equation (7.3) have been proposed in the literature. Zhang and Kou (2010) can be mentioned as a recent example; their kernel estimate of $\hat{\rho}(t)$ (Equation (3.1) in their paper) is however very similar to the quantities C_k^N defined in Grandell (1976). It is hence questionable whether a substantially improved fit from adopting their method can be achieved. The results obtained by Zhang and Kou on the convergence of their estimator to the true correlation function however provide a theoretical justification of their technique that was lacking in earlier contributions.

Finally, an alternative method of moments technique could be established by relating the theoretical autocorrelation, mean and variance of the departure process D to the observed autocorrelation, mean and variance of N . This technique would follow the spirit of the maximum likelihood estimation explained earlier inasmuch as it does not involve the diffusive process X . However, no closed form expression for the covariance of the incremental counts D_t^Δ in terms of μ, λ, ν seems available. This limits the practical utility of interpreting D as departure process for a method of moments estimation, and makes the alternative interpretation as Poisson process modulated by the diffusion X (for which the autocorrelation structure is known) a more natural and convenient choice.

7.2 Assessing the fit of a point process model

We now survey three classical results related to the “residual analysis” of point processes popularised by Ogata (1988). These are generic results inasmuch as they apply to any point process with stochastic intensity, and have been widely used in the applied statistics literature. Let $M = (M_t; 0 \leq t \leq T)$ denote such a point process with intensity process $\lambda = (\lambda_t; 0 \leq t \leq T)$; and let

$$\Lambda_t = \int_0^t \lambda_s \, ds$$

denote the integrated intensity process.

It is known (Papangelou, 1972) that the (random) time change $\tau(t)$, defined by

$$\Lambda_{\tau(t)} = \int_0^{\tau(t)} \lambda_s \, ds = t,$$

makes $\widetilde{M}_t = M_{\tau(t)}$ a standard Poisson process of intensity 1. This suggests that, if an estimate $\hat{\lambda}$ is a close approximation to the true intensity λ and (T_i) denote the event times of M , then the transformed times $\hat{\tau}_i = \hat{\tau}(T_i)$ defined by

$$\int_0^{\hat{\tau}(t)} \hat{\lambda}_s \, ds = t$$

should have very similar, if not the same statistical properties as the event times in a unit rate Poisson process.

We fix some approximate intensity $\hat{\lambda}$, and write $\hat{\tau}_i = \tau_i$ for brevity. Whether or not the τ_i conform with event times from a standard Poisson process can be tested by several means. We mention three popular techniques that yield easy to interpret graphical diagnostics:

- (i) Ogata (1988) plots the cumulative number of events $\#\{\tau_j; \tau_j \leq \tau_i\}$ against the transformed times τ_i , and then normalises the counts by M_T and the times by τ_{M_T} . This should result in a straight line with unit slope, as the arrival times of a Poisson process are uniformly distributed. Error bounds based on the Kolmogorov-Smirnov statistic can be added to identify significant aberrations from this line.
- (ii) Whether the inter-arrival times $(\tau_i - \tau_{i-1})$ are distributed according to an exponential distribution can be checked by plotting the empirical log-survivor function, or by checking whether the sequence of

$$U_i = 1 - \exp\{-(\tau_i - \tau_{i-1})\}$$

are uniform random variates on $[0, 1]$ (proposed by Berman (1983)). The latter can be visualised by the same technique as suggested above for the transformed times,

using the same error bounds.

- (iii) Berman also proposed to test for serial correlation between the inter-arrival times by plotting U_{i+1} against U_i . The resulting scatter plot should not exhibit any area of concentration in $[0, 1]^2$.

We give a straightforward illustration of these techniques when $\lambda(t) = X_t$ and $M_t = N_t$, based on the same simulated data as in Section 7.1.1 with $T = 200$ and $N_T = 1610$. We use the method of moments estimates given in Table 7.2 for $\Delta = 1$ to initialise and run the filter discussed in Section 6.1; and we use the posterior mean of the filtering distribution as intensity approximation, that is, $\tilde{\mu}_t \approx X_t$ in the notation of Section 6.1.

The results are shown in Figure 7.3. While the deviation from the straight line in the second and third plot remains just within a 1% error bound, this has to be weighed against the error that is introduced by using the filtering mean $\tilde{\mu}$ as approximation to the true intensity process X . This error can be quantified by studying the posterior variance of the filter (discussed in Section 6.1 as well), and should be taken into account when a serious violation of the 1% error bound occurs.

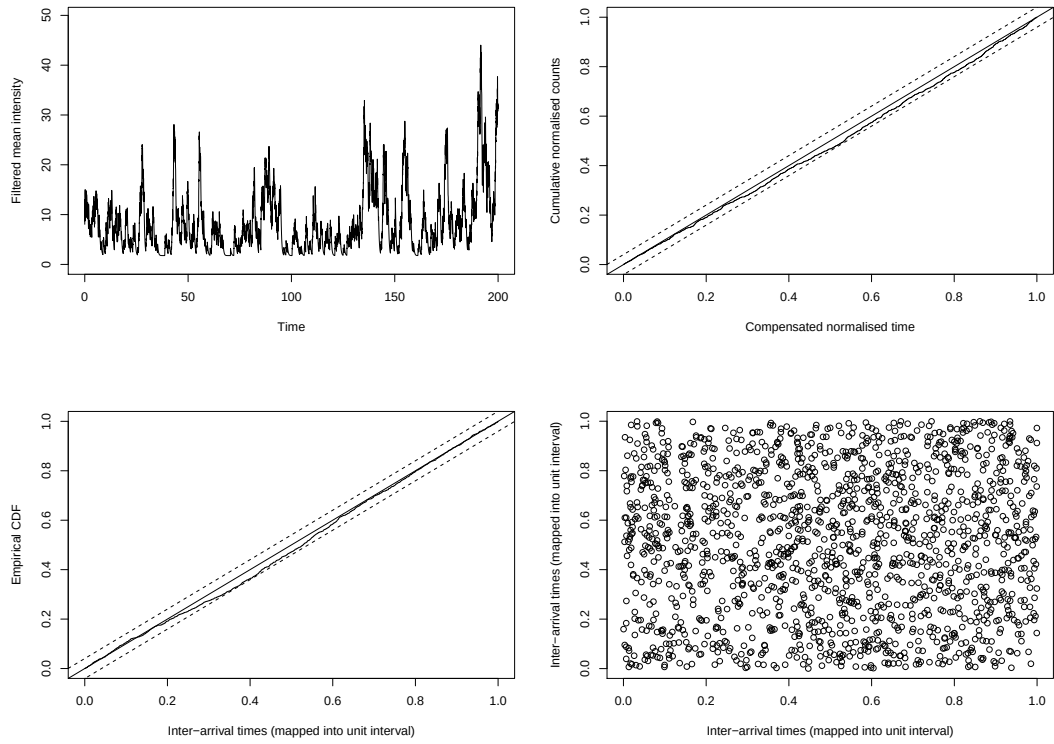


Figure 7.3 From left to right and top to bottom: the first plot shows the filtered mean intensity $\tilde{\mu}$ that we use as approximation to the true intensity process. The second plot shows the cumulative count of events on a compensated time scale (technique (i) above). The third plot shows the empirical cumulative distribution function of the U_i defined in technique (ii) above. The fourth plot shows the scatter plot of U_{i+1} against U_i mentioned in (iii). The error bounds given in the second and third plot correspond to the 1% confidence levels of the Kolmogorov-Smirnov statistic.

8 Event times in a limit order book

The limit order book of a financial asset presents a rich source of interacting point processes. Quote data can change on either side of the book, either by cancellation or arrival of new order volume, or by removal of price levels in aggressive trading. Trade events, whether price changing or not, are generally regarded as useful characteristics of market activity. Together, these events evolve as a complex multivariate marked point process.

This chapter focusses specifically on trade events as recorded in the EBS Live dataset. (The event times of sell trades exhibit the same statistical properties; the choice of the buy times over the sell times is arbitrary.) We start by assessing whether these events can realistically be modelled as a diffusion driven point process. Attention is restricted to a three-parameter CIR diffusion X . This is a flexible family of diffusions that can be adapted to have a given mean, variance and autocorrelation.

The estimation of the unknown parameters related to X is described in Section 8.1. We will see that the maximum likelihood estimates do not yield stable, usable results, and that method of moments estimates need to be used instead. We show in Section 8.2 that nevertheless the model is inadequate, failing various tests of goodness of fit.

We will see that sudden, “self-exciting” spikes in the trading activity are the cause for this. Several adjustments or alternatives to a CIR driven Poisson process model are then discussed in Section 8.3.

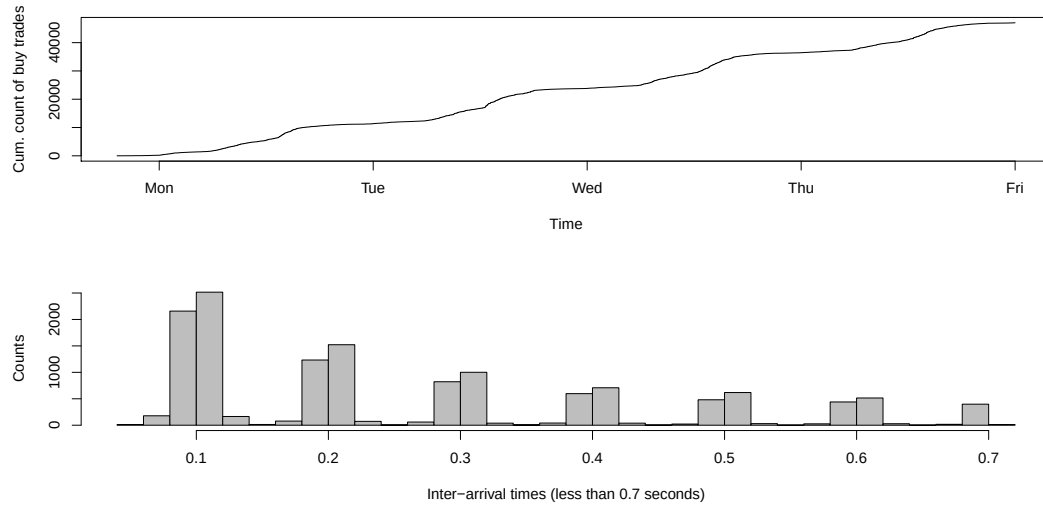


Figure 8.1 The top plot shows the cumulative counts of the point process N of buy events. The range of inter-arrival times (in seconds) is 0.047 to 2091.2, with median 2.30 seconds. The bottom plot hints at the economic and technological factors that influence the dataset: the distribution of inter-arrival times less than 0.7 seconds shows clusters around multiples of 100 milliseconds.

8.1 Parameter estimation

8.1.1 Using the maximum likelihood estimates

We consider as point process N the buy events in the EBS Live dataset. There are around 47000 such buy trades spread over a time window of about 100 hours; and while electronic trading is operated continuously throughout the day, there are clear “intra-day seasonalities” (such as the opening and closing hours of physical exchanges in London, New York or Chicago) that need to be taken into account. An illustration of this pattern is shown in Figure 8.1. This means that estimates for $\theta = (\alpha, \mu_0, V)$ need to be derived at various time points throughout Monday to Thursday, and that only trade events “close” to these points should be used for the estimation. There are two choices to make this requirement precise:

- (i) We derive rolling estimates that are *equidistant in event time*; that is, the MLE for θ at time t is based on a fixed number (we use 1500 as indicated by the results in Section 7.1.1) of buy events prior to t . This results in estimates at points irregularly

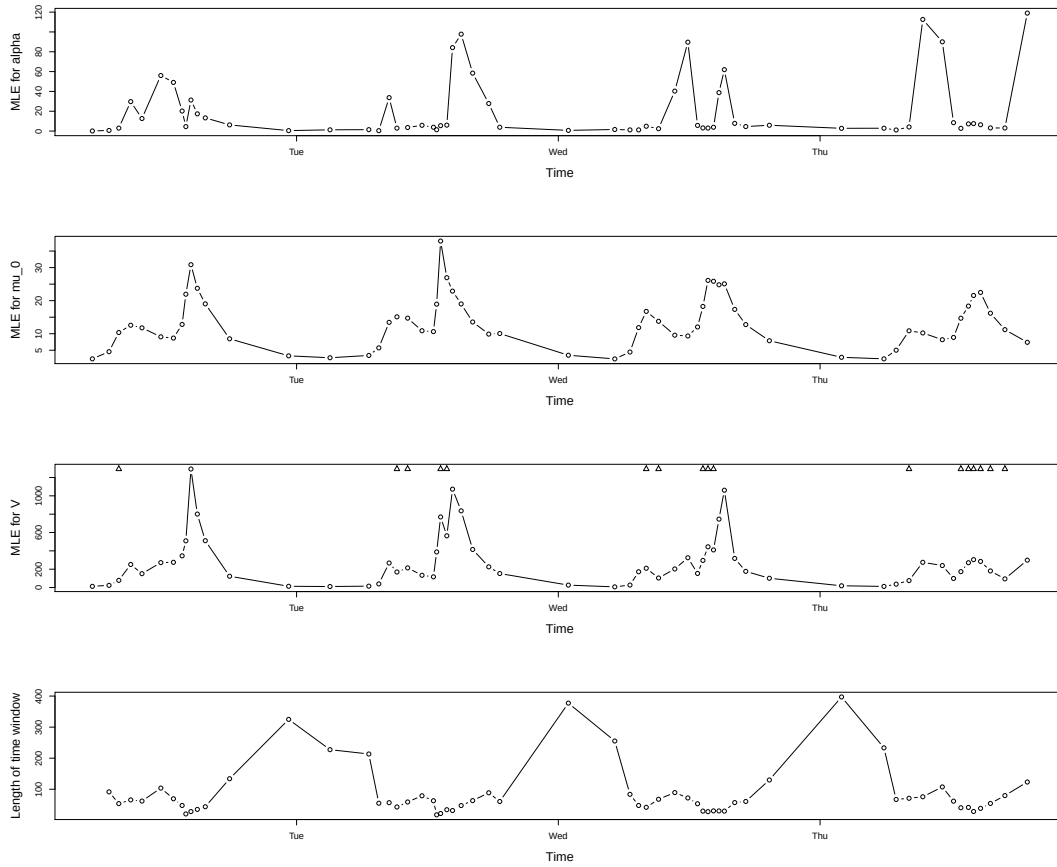


Figure 8.2 The MLEs for θ calculated at 61 points equidistant in event time; the unit of time is one minute. The calculations are effected on overlapping intervals: for two consecutive estimates, the last half of the event times for the first estimate is identical to the first half of the event times for the second estimate. Each estimate uses 1500 event times. Parameter estimates that result in a stationary point process are marked with a triangle in the third plot; there are 17 such estimates, mostly occurring during periods with high trading intensity. The fourth plot shows the length of the physical time window used for the corresponding estimation. Tick marks on the time axis correspond to midnight of the indicated day.

spaced in physical time: if trading activity is high, the MLEs are recalculated much more often than during periods with low trading activity.

- (ii) We use rolling estimates *equidistant in physical time*; that is, the MLE is recalculated after a fixed time period has passed (we use one hour), using all events that occurred in that time window. Parameter estimates for periods with low trading activity are hence based on much less events than estimates for periods with high trading activity, and may be less accurate.

The results for these two procedures are shown in Figures 8.2 and 8.3 respectively. A first

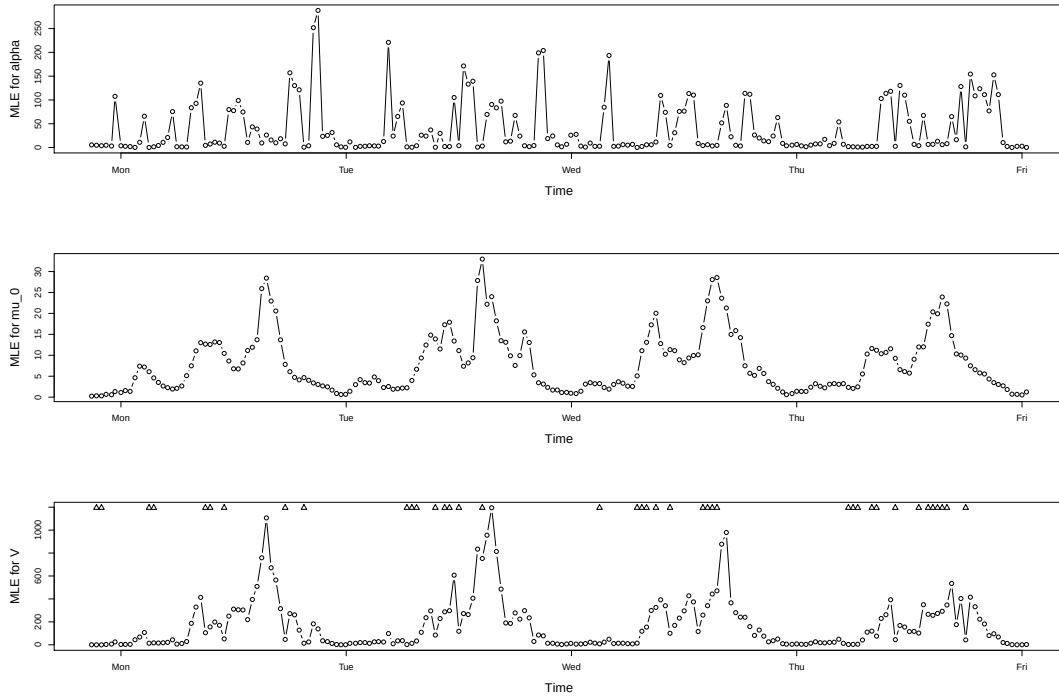


Figure 8.3 The MLEs for θ calculated at 200 points equidistant in physical time; the unit of time is one minute. Each estimate uses the event times of the previous 60 minutes. The calculations are effected on overlapping intervals: for two consecutive estimates, the last 30 minutes for the first estimate are identical to the first 30 minutes for the second estimate. Parameter estimates that result in a stationary point process are marked with a triangle in the last plot; there are 40 such estimates, again mostly occurring during periods with high trading intensity. Tick marks on the time axis correspond to midnight of the indicated day. The pattern of the number of observations within each hour follows the pattern displayed in the plot for $\hat{\mu}_0$.

worrying discovery is that most of these estimates do not correspond to a *stationary* point process N . Most stationary estimates occur for periods with high trading activity (the converse is not true, ie. there are periods with high activity where estimates do not give a stationary point process). Secondly, while the mean parameter μ_0 and the variance V seem to largely agree for the two choices of estimation times, there is substantial disagreement with respect to the mean-reversion parameter α . The results in equidistant physical time are clearly distorted by the unequal number of events in the one hour time windows used; the results in equidistant event time seem more regular but take implausibly high values during midday trading.

Neither of these results is reliable. We consider the time window from 1pm to 2pm on Tuesday, September 15 2009. (All calendar times reported are in GMT.) Within this hour, 1334 buy trades occur, giving an MLE of $\hat{\theta} = (\hat{\alpha} = 69.68, \hat{\mu}_0 = 22.20, \hat{V} = 951.62)$. The inverse of the observed information matrix is

$$j(\hat{\theta})^{-1} = \begin{pmatrix} 189.40 & -1.43 & 930.48 \\ -1.43 & 0.82 & 70.57 \\ 930.48 & 70.57 & 21148.15 \end{pmatrix},$$

demonstrating the uncertainty with respect to the estimates of α and V . The positive correlation between these two estimates may offer an explanation for the implausibly high estimates for α : should the point process N exhibit overdispersion relative to a true CIR driven process (N could for instance be the superposition of such a true CIR driven process and another point process), then this additional variance not only feeds into the maximum likelihood estimate for V but also drives up the estimate of α . This is corroborated by the contour plot of the profile log-likelihood for this time period, displayed in Figure 8.4.

8.1.2 Using the method of moments

We now investigate whether the method of moments technique described in Section 7.1.2 offers a more robust estimation of the parameters of N . We use the same 61 points

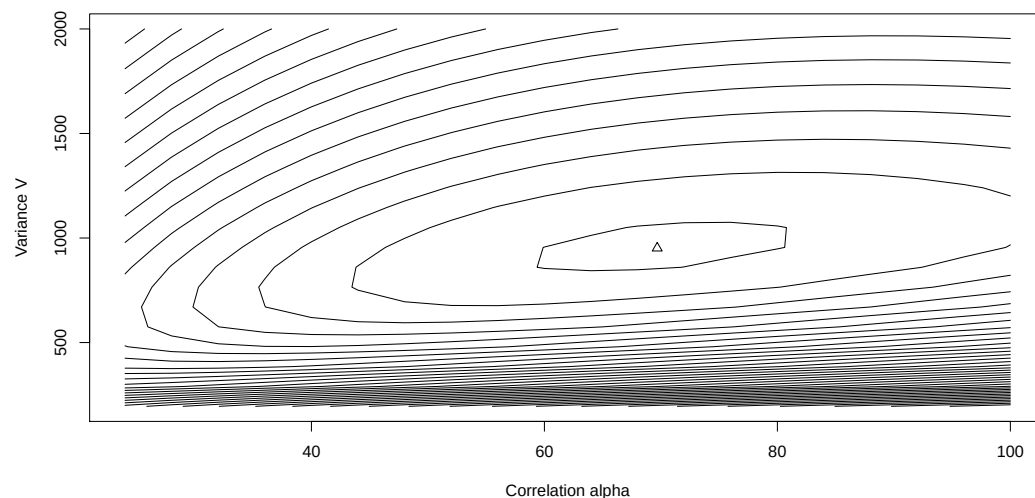


Figure 8.4 Contour plot for the profile log-likelihood function based on the buy trades 1pm to 2pm on Tuesday, September 15 2009. The MLE is indicated by a triangle.

equidistant in event time as for the MLE calculation used for Figure 8.5; a calculation equidistant in physical time is impractical due to very low counts of buy events during the night hours. The results are displayed in Figure 8.5 and clearly improve upon the maximum likelihood estimates. The estimates for the mean intensity remain unchanged as expected; the pattern for the variance is retained but on a much more reasonable scale. The coefficient α is again hardest to estimate; but the estimates now all remain within a range of 0.07 and 1.45, and the increase in this mean-reversion parameter during the busy mid-day trading conforms with economic intuition. The number of parameter estimates that correspond to a stationary point process is now 52 out of 61; and all of the non-stationary estimates occur outside the core trading hours of 8am to 5pm.

8.2 Assessing the model fit

In this section, we investigate whether a CIR driven point process with parameter estimates obtained by the method of moments fits the observed sequence of buy trades sufficiently well. As first step towards this, we calculate the arithmetic mean of the parameter estimates shown in Figure 8.5 across Monday to Thursday in the hope of obtaining

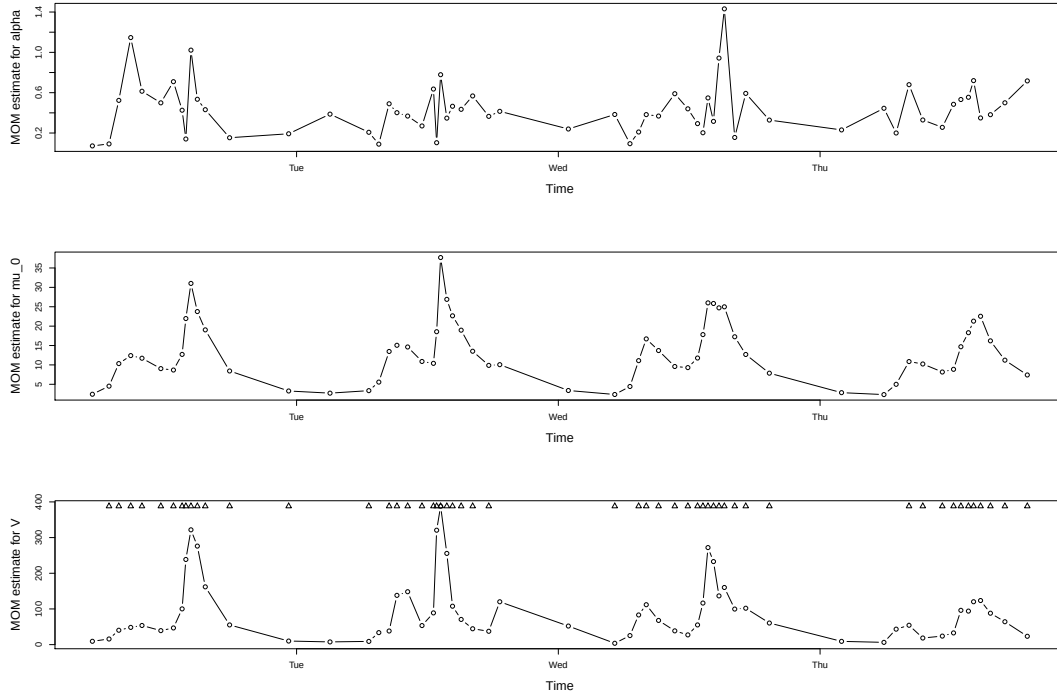


Figure 8.5 The method of moment estimates for θ . Explanations of the plots shown are as for Figure 8.2.

smoother estimates that still retain the intra-day variation due to the opening and closing of physical exchanges across the different time zones and financial centres. The results are shown in Figure 8.6. The interpretations for the mean intensity and variance are indeed straightforward and meet our expectations: the increase at about 8am corresponds to the start of trading in London, and is followed by a distinct but not too sharp slow-down during the rest of the morning. The increase at around 1pm is due to the start of trading in New York and Chicago. The averaged values for α seem less stable, but the increase in mean-reversion during the busy trading hours (with trade numbers and sizes fluctuating massively) does not contravene economic intuition.

We consider again the sequence of buy trades between 1pm and 2pm on Tuesday, 15 September 2009. The averaged parameter estimates are $\hat{\theta} = (0.63, 25.20, 195.62)$; and the results for the diagnostic tests described in Section 7.2 are shown in Figure 8.7. We can see that there is very clear deviation from the straight line in the third plot; this indicates that the inter-arrival times (even on the *compensated* timescale) do not come from an

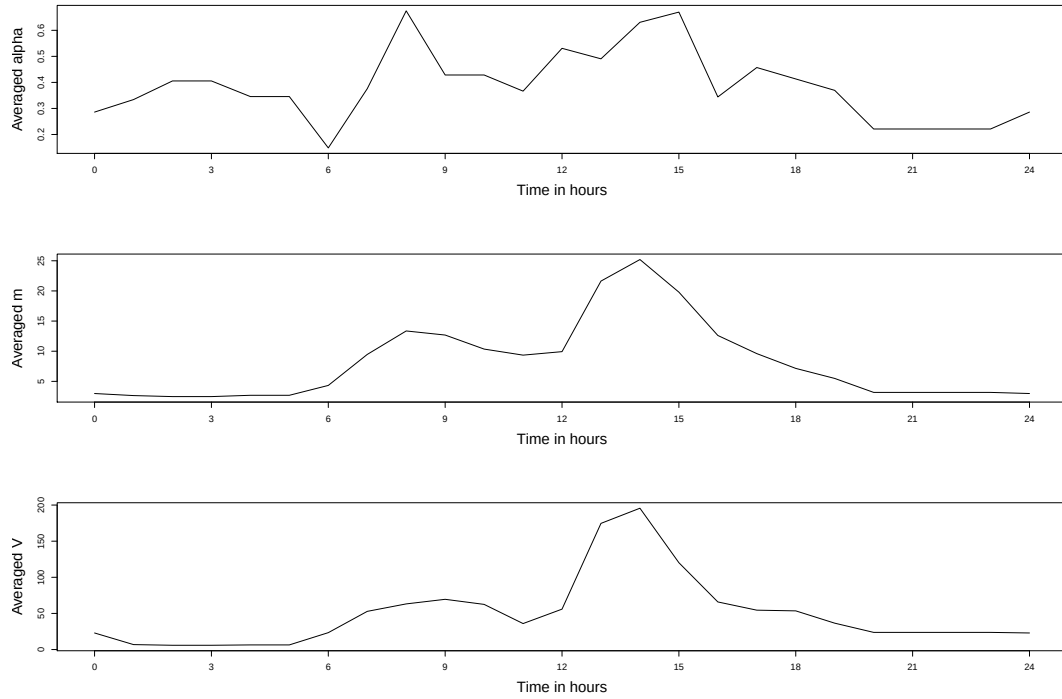


Figure 8.6 The averaged method of moment estimates for θ during a trading day starting and ending at midnight.

exponential distribution. The fourth plot indeed shows that there is dependence between consecutive inter-arrival times: a very short inter-arrival time is likely to be followed by another very short inter-arrival time, or in other words, event times still form clusters on a compensated time scale. The cause for this behaviour is the sudden spike in trading activity at around 1:35pm (see the first plot): the increase is so sudden and strong that it cannot be explained by the diffusive dynamics of X alone; and the filtered mean intensity does not adapt rapidly enough, causing the clustering on the compensated time scale. This holds even when we take the posterior variance of the filtering mean into account: the deviation from the straight line in the third plot for inter-arrival times less than 0.2 even occurs when we add two standard deviations to the posterior mean (increasing the maximum mean intensity from 71 to 98).

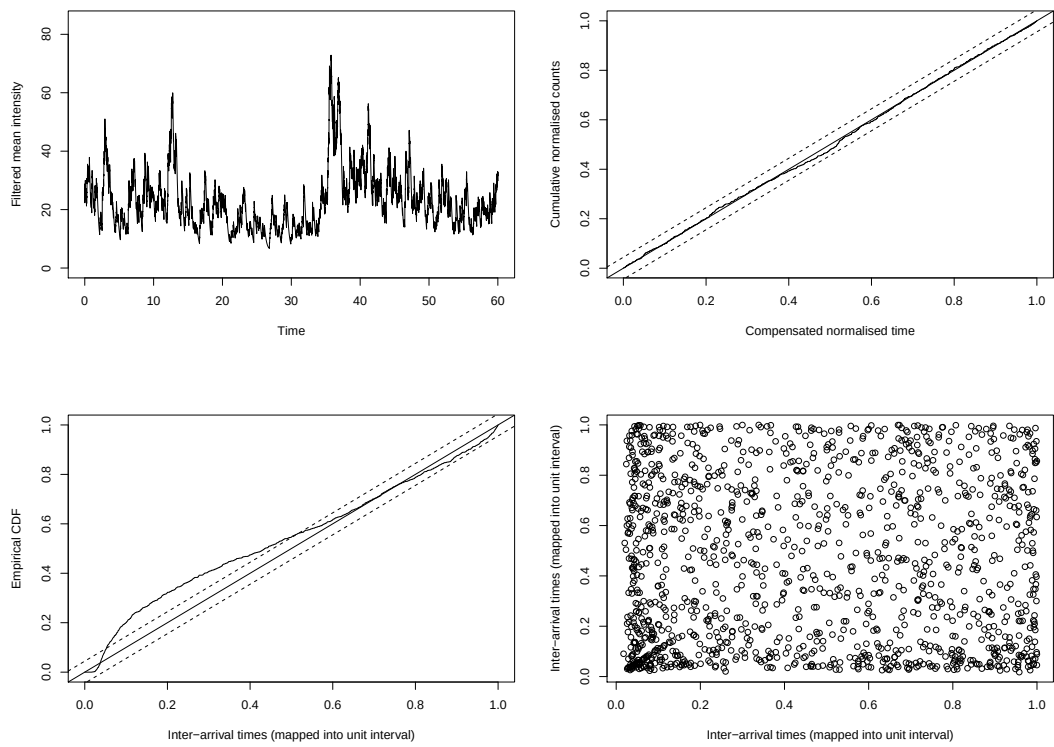


Figure 8.7 Results for assessing the fit of a CIR driven point process model for the buy trades between 1pm and 2pm on Tuesday, 15 September 2009. The first plot shows the filtered mean intensity, obtained by the approximate technique from Section 6.1 and using averaged method of moments estimates $\hat{\theta} = (0.63, 25.20, 195.62)$. The explanation of the remaining plots is as in Figure 7.3.

8.3 Alternative models and adjustments

There are several approaches to address and incorporate the above findings into a more appropriate model for the trade flow data:

- (i) If an endogenous explanation of the arrivals of trades (including the sudden bursts of activity) is desired, a bi-directional interaction between the point process N and its intensity process λ seems necessary. Two such choices are discussed below, namely the Hawkes process with a self-excitation term, and a diffusive intensity process that runs on a time scale that speeds up and down dependent on the arrival of trades.
- (ii) Alternative models of the trade flow can be based on interpreting sudden bursts of activity as irregular behaviour that needs to be treated as exogenous factor in addition to a more regularly varying base-intensity. The statistical modelling of such a situation comprises the detection of anomalous trade flow, and a technique of separation between “regular events” on which estimation of the base-intensity is founded, and “irregular events” that are due to such a sudden burst of activity. We describe a simple approach based on the idea of a *temporarily unresponsive detector* below.

An extension of any model for the point process of *buy events* is to also include information given by the *sell events*. With a temporary change of notation, we could write

$$d\lambda_t^A = dX_t + a_t dN_t^A + b_t dN_t^B$$

for the trading intensity λ_t^A at the ask price; where X is a CIR diffusion, N^A the point process of buy events (at the Ask), N^B the point process of sell events (at the Bid), and a_t and b_t coefficients that might vary in time or stochastically. Parameter estimation and filtering is of course much more complicated in the presence of such modifications. Progress has only been achieved for the multi-dimensional Hawkes process (for which we omit the dX_t term and assume constant coefficients $a_t = a$ and $b = b_t$; see Bowsher (2007) or Toke (2011)), as this again reduces to maximum likelihood estimation.

The Hawkes process. We now consider in more detail a linear Hawkes process with exponential kernel as intensity for the point process N of buy trades, with event times denoted by T_i . Fitting such a Hawkes process reduces to estimating the parameters λ_0 , α , β for the intensity process

$$\lambda_t = \lambda_0 + \int_{-\infty}^t \alpha e^{-\beta(t-s)} dN_s = \lambda_0 + \alpha \sum_{T_i < t} e^{-\beta(t-T_i)}.$$

As mentioned, this can be effected easily by maximum likelihood as the log-likelihood function is known explicitly (Ogata, 1981). The resulting fitted intensity process and the model fit diagnostics are shown in Figure 8.8. We see that despite allowing for a self-exciting interaction term, the compensated inter-arrival times still do not come from an exponential distribution and continue to exhibit positive auto-correlation. Furthermore, the estimated intensity process λ (first plot) compares unfavourably to the corresponding CIR estimate in Figure 8.7, as it displays larger variation and excessive spiking throughout the estimation period, including the comparatively calm period from 1:20pm to 1:35pm.

Using a time-changed diffusion. Another way of allowing for sharp and sudden increases in the intensity process has been described in Section 3.3.3. We recall that this approach is based on modelling N as point process with time-changed diffusive intensity process

$$\lambda_t = \varphi'(t - t^*) X(\tau(t)), \quad (8.1)$$

where X is a CIR diffusion. We have shown that inference about λ from N is equivalent to inference about X from M , where M is related to N by a time-change that rescales all inter-arrival times by the function φ .

The choice of φ is of course crucial for this approach. An appropriate choice in our situation is $\varphi(x) = \log(1 + cx)$ with $c > 1$: on the one hand, we wish to increase very small inter-arrival times and hence require a function that grows super-linearly for small arguments; on the other hand, we wish to avoid an unbounded derivative term in Equation

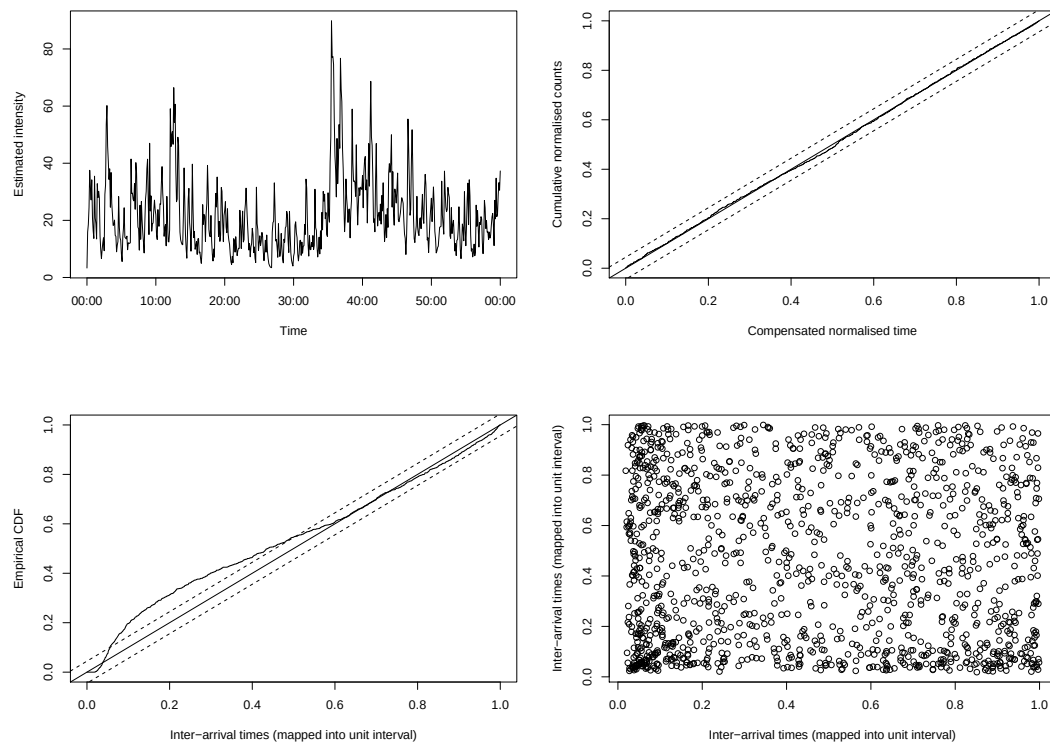


Figure 8.8 Results for assessing the fit of a Hawkes point process model for the buy trades between 1pm and 2pm on Tuesday, 15 September 2009. The model parameters were estimated using the event times between 11:30am and 1pm on the same day, resulting in $\hat{\lambda}_0 = 3.31$, $\hat{\alpha} = 5.69$, $\hat{\beta} = 6.71$. The first plot shows the Hawkes intensity process using these estimates. The explanation of the remaining plots is as in Figure 7.3.

(8.1) for $(t - t^*) \rightarrow 0$.

The effect of this transformation on the inter-arrival times is indeed a regularisation towards the mean inter-arrival time, as small such times are increased and large ones decreased. This results in a more robust parameter estimation for the CIR diffusion X driving M (for instance, we now have agreement of the MLE and method of moment estimates), and also yields a less variable conditional expectation $\hat{X}_t$. However, applying the speed adjustment φ' to recover the intensity λ for N (our real object of interest) implies that at *each* event time, $\hat{X}_t$ is multiplied by $c > 1$, but decreased by a factor less than 1 (and tending to 0) as soon as an inter-arrival time exceeds a certain threshold. This introduces large “bouncing” fluctuations into the filtering estimate $\hat{\lambda}$, making this approach unsuitable for practical applications of intensity estimation.

Using an unresponsive detector. We present a simple approach to separate between “regular events” on which estimation of a trading intensity is founded, and “irregular events” that are due to sudden spikes of activity (which arguably need to be modelled separately). The technique is motivated by applications in electrical engineering, where a counting device is considered that becomes (unavoidably, and for technical reasons) unresponsive for a small time period after an event has been recorded. Applying this idea to the trade flow data, we introduce a time window of “intentional unresponsiveness” following a (recorded) trade event; trade events that occur during this time window are simply discarded and left unrecorded. This will contract clusters of many trade events into a single or a few more widely spaced event points. Based on the distribution of small inter-arrival times illustrated in the second plot in Figure 8.1, we have chosen this time window as 0.15 seconds.

It is tempting to allow this window to vary according to some current base intensity, which in turn more generally be formulated as a technique of thinning our dataset (that is, ignoring or deleting events) at a variable rate. This approach would however introduce difficulties concerning the estimation of the model parameters (because the filtering and

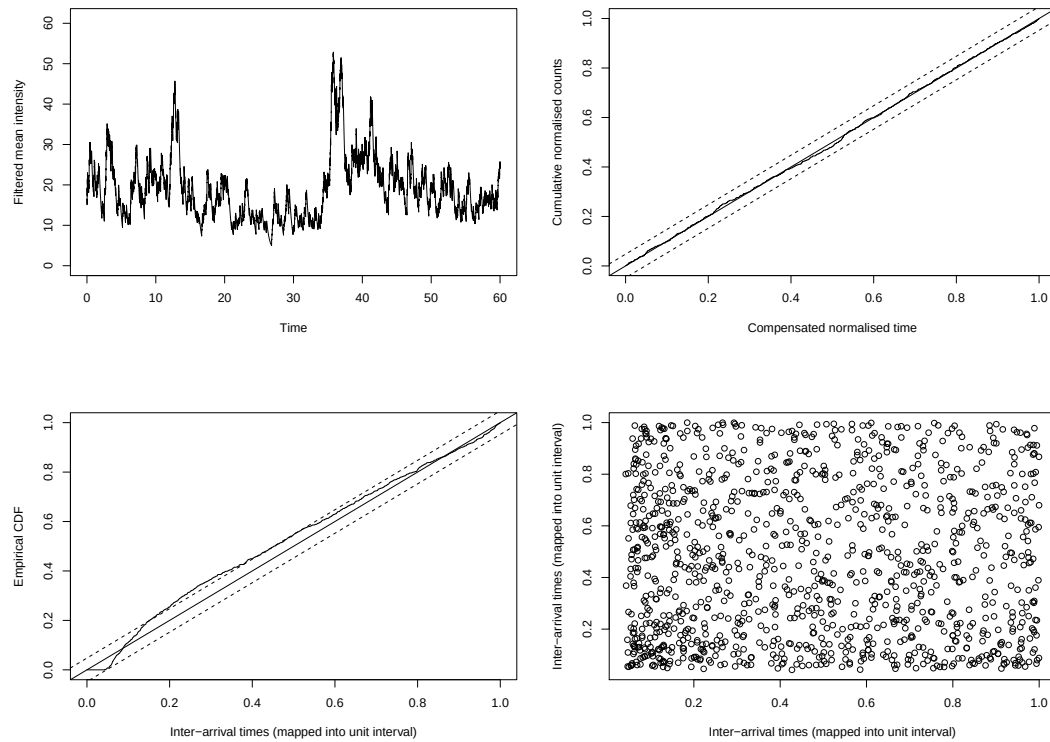


Figure 8.9 Results for assessing the fit of an unresponsive detector model for the buy trades between 1pm and 2pm on Tuesday, 15 September 2009.

thinning may now impact each other), and we have hence not pursued this idea further.

13% of the buy events between 1pm and 2pm occurred during a 0.15 second period of unresponsiveness and were hence discarded. We assume that the remaining event times were generated by a CIR driven Poisson process. The parameters of the underlying diffusion were obtained by method of moments estimates averaged over the entire trading week as described in Section 8.2; these parameter estimates are similar to Figure 8.6 but with (unsurprisingly) generally smaller values for the mean and variance.

Figure 8.9 shows the results of this approach. The first plot shows a mean filtered intensity that is qualitatively similar to the intensity displayed in Figure 8.7 but with much less pronounced spikes. The plots for the inter-arrival times and autocorrelation now indicate largely satisfactory behaviour. Note that the exactly horizontal start of the empirical CDF in the third plot is generally due to a minimum threshold for inter-arrival times rooted in

technological factors (and as such also visible in Figures 8.7 and 8.8); this effect is now amplified as we further increase this minimum threshold.

Conclusion. Neither the CIR model in Section 8.2 nor the Hawkes process or the time-changed diffusion model in this section offer a fully satisfactory statistical explanation of the trade event data. This is by and large an unsurprising finding in the light of the notoriously intricate and often obscure nature of high-frequency financial data. Far too many different factors generally affect the structure of such datasets, ranging from technological issues (such as network speed between exchanges and financial institutions) to economical matters (such as news releases) and even psychological considerations (such as intentionally misleading behaviour by a trader). A statistical model that attempts to uncover structure in the trade flow of an asset purely on basis of a sequence of indiscriminate trade events, and without taking external information into account, seems hopelessly disadvantaged in the light of these complexities and the wealth of additional information available.

The results for using an unresponsive detector can be seen as tentative indication that a fair amount of progress in this direction can be made with a limited increase in complexity. Indeed, a CIR diffusion seems to offer sufficient flexibility to explain the *regular* variations in the trade event data, and appears to improve upon a Hawkes process formulation in this respect. A more adequate treatment of the sudden bursts of activity now seems to offer the greatest reward for any additional modelling effort. The throttled recording of trade events suggested above is a very practical but only a first step in this direction, and does not even begin to address the many choices that are available to detect or even predict sudden increases in trading activity. This may present a promising area of further research.

9 Backtesting of a market making strategy

The function of a financial institution or trader that *makes a market* for some asset is to quote and publicly display bid and offer prices throughout the trading day, and to buy or sell the asset at the quoted prices from or to other market participants. A market maker hence supplies other traders with price information and most importantly *liquidity*. This is a valuable service as the (competitive) quoting of two-way prices ensures that other market participants can monitor the value of their portfolio and make adjustments to it by buying or selling the asset under acceptable transaction costs. The compensation for this service comes in terms of the *bid-ask spread*: the market maker hopes to sell the asset at the offer price to one market participant and to buy it back from some other participant at the bid price. This is called *making the spread*, and the difference between the offer and bid price is the market maker's profit.

Before the advent of electronic trading, this activity was effected by specialist firms (often just called "specialists") that were under a contractual obligation with the exchange to perform this role. This arrangement continues on some stock exchanges, where such firms are now called *Designated Market Maker (DMM)* and sometimes receive favourable treatment by the exchanges in return for the provision of liquidity. These advantages can be in terms of market data (more detailed or quicker access to the order book), but may also include rebates on transaction costs or lighter rules and regulations. Many of the electronic exchanges in the foreign exchange markets have evolved in a different fashion, and by means of limit orders now allow virtually any professional trader to act

as a voluntary provider of liquidity. We focus on this situation and ignore the specific and complex rules on exchanges with DMMs. For a recent discussion of modern (equity) trading, see Angel et al. (2011), and O’Hara (1995) for a less recent but influential general discussion of the theory of market microstructure.

The objective of a market maker is usually to earn the bid-ask spread as often as possible without being overly exposed to unpredictable or unfavourable variations in the price of the asset. How can either objective in this risk-reward trade-off be improved? (i) To increase the number of negotiated trades, a market maker will often try to improve the priority position of his quotes in the order book. This can be achieved by either tightening the bid-ask spread (in other words, to improve upon competitors’ prices and be the first to quote at a new price inside the spread), or by maintaining quotes far away from both sides of the mid price (in order to be in a favourable position should the mid price eventually move closer to these quotes). (ii) To limit or reduce “risk” presupposes a definition of this concept. A natural choice is *inventory risk* (or *market risk*), which relates to the profit or loss that will be incurred when a long or short position in the asset is affected by a large movement of the price. This risk is most effectively limited by controlling the size of the inventory. *Adverse selection risk* (“being picked off”) is another form of risk that relates to an unfavourable price move immediately after a quote is taken off by an informed trader. This can be countered by submitting less aggressive quotes away from the best bid and ask, and by quick cancellation of existing quotes should the price move unexpectedly.

Mathematical models for the optimal control of such a situation necessarily need to make simplifying assumptions of how trading in limit order books is operated, and need to agree on formal definitions of risk and reward that only approximately capture the complexities of modern trading environments.

Contributions to this area of research are comparatively recent and often very diverse in nature. One of the earliest formulations was developed by Ho and Stoll (1981), where a monopolistic dealer selects bid and ask prices around a subjective “true price” of the asset in order to maximise the utility of terminal wealth. This wealth accumulates both

by making the spread (modelled by Poisson point processes of buy and sell events) and by changes in the value of the inventory (using a mid price driven by Brownian motion). Avellaneda and Stoikov (2008) connect a similar model to modern limit order books and empirical findings in market microstructure. In their model, a (non-monopolistic) market maker selects bid and ask prices such that he is indifferent between his current inventory and his current inventory plus one unit of the asset. Of crucial importance are the (Poisson) rates at which the trader buys and sells the asset; it is assumed these can be modelled as functions that decrease in the distance of the chosen quotes to the mid price. In a very recent contribution, Guilbaud and Pham (2013) have shown how to incorporate a bid-ask spread that takes *discrete values*, in multiples of the tick size, and how to obtain numerical approximations to the optimal control for two explicit choices of the utility function.

9.1 Optimal market making as an intensity control problem

A common crucial element of the contributions cited above is to relate both the reward and the risk to the (stochastic) arrival of trade events. The inclusion of a stochastic behaviour for the mid price adds to the realistic nature of these models, but hinders a succinct description of this interpretation of optimal market making as an *intensity control problem*. (We refer to (Brémaud, 1981, Chapter VII) for the mathematical background material of point process control.)

In this section, we state a value function for the market maker's control problem under the assumption of constant bid and ask prices. We will not pursue this further to find an optimal control, as the purpose of the derivation is to illustrate the above mentioned link between risk and reward on the one hand and intensity controls on the other, and to indicate how involved the mathematical formulation of a fully realistic model, let alone its solution would be.

We place ourselves in the position of a trader (the market maker) that sells and buys some financial asset. Let $A = (A_t; t \geq 0)$ denote the counting process for how often this trader

sells the asset (at the ask price); similarly, let $B = (B_t; t \geq 0)$ denote the counting process for how often the trader has bought the asset (at the bid price). We write

$$N = \begin{pmatrix} N(1) \\ N(2) \end{pmatrix} = \begin{pmatrix} A \\ B \end{pmatrix}$$

for the joint process.

On the reward side, we suppose that the trader makes a constant profit (denoted by b) for each trade effected. This is of course only adequate if a trade of one gender (sell or buy) is in fact offset by a trade of the opposite gender, and both transactions are agreed at constant bid and ask prices. This can be a legitimate approximation for situations of low price volatility and high trading volume (such as highly liquid interest rate futures), but is arguably not satisfactory in general.

On the risk side, we assume that the trader faces *inventory risk*. Sale and purchase transactions of the asset do not occur in lockstep, and this implies that the trader will most often find himself with a non-zero net position or *inventory*

$$I_t = (-1, 1) \cdot N_t = B_t - A_t,$$

where a dot denotes the inner product. Running a non-zero inventory exposes the trader to fluctuations in the price of the asset, and the severity of this increases in the size and the duration of this imbalance. We let $c: \mathbb{Z} \rightarrow \mathbb{R}$ denote a function to penalise the imbalance of the inventory; that is, the penalty at time t is $c(I_t)$. We assume c is symmetric, that is $c(I_t) = c(-I_t)$.

We assume the trader is able to control the intensities for the point processes A and B , and can hence modify at which rate he buys and sells the asset. In the simplest form, we let $U = (U_t; t \geq 0)$ denote the *control process* with values in the set of controls

$$\mathcal{U} = \left\{ \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \end{pmatrix}, \begin{pmatrix} 0 \\ 0 \end{pmatrix} \right\}.$$

We assume N_t admits the *controlled intensity*

$$\lambda(U_t) = \left(\lambda_1 U_t(1), \lambda_2 U_t(2) \right)^t,$$

where λ_1 and λ_2 denote constant hit and lift rates for the bid and offer side of the book.

A control U_t hence simply indicates whether the market maker is willing to sell the asset ($U_t(1) = 1$ if so, $U_t(1) = 0$ if not), and whether he is willing to buy (analogously by $U_t(2)$).

Note that this choice of model and controls ignores both the question of *how many* units the market maker wishes to quote, and *at which prices* these quotes are placed. Both are important questions in practice. We also ignore the requirement to unwind (i.e. bring to zero) the inventory at the end of the trading day. How such a liquidation of a portfolio is done optimally is usually only of interest for transaction sizes large enough to introduce detrimental market impact.

We instead consider an “infinite horizon” for our control problem. This leads to a value function that assigns the expected instantaneous profit to a control U :

$$J(U) = \lim_{T \rightarrow \infty} T^{-1} \mathbb{E} \left[\int_0^T b \, dI_s - \int_0^T c(I_s) \, ds \right],$$

where dI_s is shorthand for $dA_s + dB_s$. The first integral within the expectation adds a profit of size b (which could be half the spread between bid and ask price) whenever the inventory changes, and 0 otherwise. The second expectation subtracts continuously from this by penalising an imbalanced inventory. We can use the alternative characterisation of point processes admitting an intensity from Section 3.1, Equation (3.2), to rewrite the value function as

$$J(U) = \lim_{T \rightarrow \infty} T^{-1} \mathbb{E} \left[\int_0^T b(1, 1) \cdot \lambda(U_s) - c((-1, 1) \cdot N_s(U_s)) \, ds \right]. \quad (9.1)$$

This accentuates the dependence of the average wealth on U and the controlled intensities. It is intuitively clear that an optimal control for the above value function is either $U_t =$

$(1, 1)^t$ or $U_t = (0, 0)^t$, dependent on the choice of λ_1, λ_2 and the penalty function c .

This trivial solution is not necessarily a limiting factor for the usefulness of this model – more interesting situations can indeed be introduced easily by making quantities such as λ_1, λ_2 or b time-varying. The difficulty of a realistic specification of the penalty function c is a much more significant shortcoming: While the benefit derived from executing a trade admits a natural description in terms of monetary profit, few immediate, objective specifications of inventory risk come to mind. A possible attempt could be to set $c(x) = \gamma(\sigma)|x|$, where σ is the volatility of the asset and γ some function; but no simple specification of either quantity seems immediately available.

9.2 Description of the backtesting framework

The key insight gained from the mathematical specification of the value function in Equation (9.1) is that a market maker needs to continuously adjust the arrival rates of his buy and sell transactions in such a way as to avoid or compensate for the risk of having an imbalanced inventory.

The functional form of these controls in realistic scenarios can be complicated and dependent on the price and volatility of the asset, and further exogenous variables. We want to investigate empirically the special case where only the trade flow (measured as the intensity of the arrivals of buy and sell trades, respectively) and a constant maximum inventory size J are used to determine these controls.

A characteristic of such an empirical proceeding is that the performance of a market making strategy can be evaluated in terms of its *profits and losses* (*PnL*) without explicit modelling of the price variable. This allows to adopt the approach popular amongst practitioners to identify risk as the variance of terminal wealth. An indication of the performance of a trading strategy is then the magnitude of terminal wealth weighted by its variance; this reward-to-variability ratio is known as the *Sharpe ratio* (Sharpe, 1966),

which we define rigorously in Equation (9.4). A high such ratio indicates that a trading strategy accrues profits in a smooth and consistent fashion.

We now describe and define the determinants of our empirical backtesting study.

Underlying data. We consider a time window $W = [T_0; T_1]$ that is contained within the four days from 14 September 2009 to 17 September 2009. The EBS Live dataset contains information about the first three levels on either side of the order book of the EUR/USD currency pair at (ascending) times $\{t_m : 1 \leq m \leq M\}$ that lie within W . We place ourselves in the situation of a market maker with immediate access to this dataset (i.e. the time it takes for the updates to travel from the exchange to the market maker is assumed zero), and the ability to adjust his quotes at each time t_m at no cost.

Admissible prices for quoting. We assume the market maker can submit quotes at up to three different prices on either side of the visible range (as defined in Subsection 2.2.2). Should the mid price be a tradable price, ie. be a full pip, then we do not allow to post either bid or offer quotes at this price. To illustrate this rule, if the order book displayed at t_1 is

$$\begin{array}{cccccc} 1.4997 & 1.4998 & 1.4999 & 1.5002 & 1.5003 & 1.5004 \\ -10 & -8 & -2 & 1 & 5 & 8 \end{array},$$

then the visible range of prices is

$$\mathcal{R} = \{1.4997, 1.4998, 1.4999, 1.5000, 1.5001, 1.5002, 1.5003, 1.5004\}$$

with mid price 1.50005; the first four of these can be used for bid quotes, and the last four for offer quotes. If the order book displayed however is

$$\begin{array}{cccccc} 1.4997 & 1.4998 & 1.4999 & 1.5001 & 1.5003 & 1.5004 \\ -10 & -8 & -2 & 1 & 5 & 8 \end{array},$$

then the visible range remains as above, but the mid price is now 1.5000, and bid quotes can only be submitted at $\mathcal{R}^B = \{1.4997, 1.4998, 1.4999\}$, and offer quotes at $\mathcal{R}^O = \{1.5001, 1.5002, 1.5003, 1.5004\}$. The interpretation of this rule is that the market maker has no information about the future direction of the prices, and hence does not want to touch the “disputed” mid price between the bid and offer territory.

Queue spots and the impact of trades and cancellations. We assume that any addition to the book is submitted fast enough to end up directly on top of the visible queues. That is, if at t_1 we wish to offer two units at a price of 1.5003, we assume we occupy the queue spots 6 and 7.

As long as a particular price level remains within the visible range, the queue spot claimed by us will improve whenever quotes submitted earlier are removed due to trading or cancellation. The Life Cycle dataset gives information about all these changes and hence allows to update the queue position from t_m to t_{m+1} . Furthermore, if we hold a quote long enough to arrive at the bottom of the queue, any trade occurring at this price can be attributed to us. The intricate interplay of the Life Cycle dataset and EBS Live dataset (see Subsection 2.2.3 on the mismatch of time) adds technical difficulties to this improvement of the queue spot and the simulation of trades. To continue the example above, suppose we hold queue spots 6 and 7 at the offer price 1.5003. It may then happen that the stream of messages in the Life Cycle dataset between t_1 and t_2 contains quote deletion messages that effect the cancellation of more than five units at 1.5003, effectively assigning a negative queue spot for at least one of our offers. Similarly, trade messages may occur at prices worse than our highest bid or lowest offer. Appropriate corrections for these cases were put in place (namely to stop processing cancellations of prior quotes when our quotes have reached the front spot; and to assign trades to our account that occur at prices worse than our most competitive ones). It should be noted that these corrections remain exceptional as they are only necessary for a small fraction of the times t_m .

Once a price level disappears from the visible range, we cancel any quotes at that particular price. This is put into place for several reasons: (i) Our belief about our current queue

spot will be based on the interpretation of the messages in the Life Cycle dataset. The issues encountered in the reconstruction of the order book (see 2.2.4 above) then become relevant, and render our belief the more questionable the longer we do not see any update in the EBS Live dataset. (ii) The more quotes we maintain, the more time-consuming the backtesting will be. This is a serious issue given the size of the dataset. (iii) The restriction to a limited number of outstanding quotes is likely to put us at a disadvantage in terms of queue spots during trading periods with fluctuating prices, as we may be forced to relinquish a price before enough cancellations have occurred to put us into the front spot. It however also provides a safeguard against “accidental trades” that may arise when a sharp price change occurs and outstanding orders cannot be cancelled fast enough. This trade-off between obtaining good queue spots and operational risks is a very realistic business decision; and our implementation, although partially driven by practical concerns, may not be too different from what actual market makers have chosen to implement.

The volume quoted and its distribution over permissible prices. We assume that the inventory may only grow up to a constant size J , and that the market maker needs to quote carefully enough to never violate this limit. To illustrate this, assume the inventory at time t_1 is $I_{t_1} = -2$ (that is, the number of units sold within $[T_0, t_1]$ exceeds the number of units bought in the same time window by 2); and that $J = 8$, i.e. the inventory is allowed to be in $\{-8, \dots, 0, \dots, 8\}$. The market maker may then submit quotes on the bid side for up to 10 units, but only up to 5 units on the offer side.

We assume there is a *market making signal*

$$S = (S_t; T_0 \leq t \leq T_1)$$

with values in $[0; 1]$ that indicates how “aggressive” a market maker should conduct his business (higher values represent more aggressive market making). Concrete proposals will be presented later; but the defining requirement is that S is a random variable measurable with respect to the bivariate point process given by the buy and sell events in the EBS Live dataset (these are the entries for which either the “Last Sell” or the “Last Buy” is

non-zero). In other words, the market making signal S does not depend on the inventory process $I = (I_t; T_0 \leq t \leq T_1)$, the profits and losses (PnL) that have accrued up to time t , or any other variable such as the price of the asset or its volatility.

We let $\mathcal{V}^B(S_t, J, I_t)$ and $\mathcal{V}^O(S_t, J, I_t)$ denote the total volume to be quoted on the bid and offer side at time t . We set

$$\mathcal{V}^B(S_t, J, I_t) = ([S_t J] - I_t)^+, \quad \mathcal{V}^O(S_t, J, I_t) = ([S_t J] + I_t)^+, \quad (9.2)$$

where square brackets denote rounding to the nearest integer. This specification ensures that we always quote at least as much volume as to rebalance the inventory, irrespective of how small S_t is.

The distribution of $\mathcal{V}^B$ over the range of permissible prices $\mathcal{R}^B$ on the bid side (and analogously of $\mathcal{V}^O$ over $\mathcal{R}^O$ on the offer side) can be specified with almost arbitrary sophistication and complexity. A concentration of the volume at a single price can however result in erratic and undesirable behaviour of the market making strategy: if this price disappears from the visible range, the entire volume needs to be resubmitted at lower priority queue positions, and the proportion of trades we negotiate afterwards drops unnecessarily.

We hence favour a simple and robust approach that distributes the volume more broadly: the first unit is submitted at the second best price of $\mathcal{R}^B$ and $\mathcal{R}^O$, respectively. We then proceed to quote one unit at the third best price, and so forth until we reach the end of the range. The next unit is then quoted at the best price, after which we again work towards the end of the range. This process is stopped as soon as the total volume to be quoted is exhausted.

PnL and Sharpe ratio as performance measures. We now define $V = (V_t; T_0 \leq t \leq T_1)$ as the process of *profits and losses* (PnL). We set $V_{T_0} = 0$, and let $P_{t_m}^B$ and $P_{t_m}^O$ denote the best bid and best offer price for each time step t_m . The *liquidation value* L of

the inventory at time t_m is then defined as

$$L_{t_m} = \begin{cases} I_{t_m} P_{t_m}^A & \text{if } I_{t_m} \leq 0, \\ I_{t_m} P_{t_m}^B & \text{if } I_{t_m} > 0. \end{cases}$$

Let $u = (I_{t_{m+1}} - I_{t_m})$ denote the inventory increment at time t_m : If $u > 0$ the market maker buys, if $u < 0$ he sells, and if $u = 0$ he does not trade at all. We let P_{t_m} denote the price at which the trade has been negotiated. The PnL process is then defined as

$$V_{t_m} = L_{t_m} - u P_{t_m}, \quad (9.3)$$

for all $t_m > T_0$. It can be extended to all $T_0 \leq t \leq T_1$ by constant interpolation between times t_m , continuous from the left.

To illustrate Equation (9.3) we observe that, should we for instance sell two units of the asset at a price of 1.4600, we receive $-(-2) \cdot 1.46 = 2.92$ as an “immediate cash reward”. The liquidation value however also immediately diminishes by that exact amount; and a positive or negative change in the PnL will only appear later, when these two units are bought back at a different price or the liquidation value changes due to market price fluctuations.

The Sharpe ratio for the PnL can now be defined as

$$R = \frac{V_{T_1}}{\sqrt{\sum_{m=1}^{M-1} (V_{t_{m+1}} - V_{t_m})^2}}. \quad (9.4)$$

The example of constant signal market making. We give a simple example to illustrate the above description. We set T_0 as Monday, 14th November 2009, 13:00:00, and $T_1 = T_0 + 3600$; the unit of time being one second (ie. T_1 is 2pm on the same day). We set $S_t = 1$ for all $t \in [T_0; T_1]$, and $J = 8$. There are $M = 19625$ updates contained in the EBS Live dataset between T_0 and T_1 , amongst which are 1544 buy events and 1690 sell events.

Figure 9.1 shows several graphical diagnostics and summaries for the resulting market making. The first plot shows the mid price for the currency pair. The second plot shows the PnL (with Sharpe ratio equal to -0.27); the third plot displays the inventory. The fourth plot is more complex: In red, it displays the number of quotes on the bid side that have priority of execution over our best bid quote. For instance, if the best bid price is 1.4600 with three units quoted, the second best bid price is 1.4599 with twelve units quoted, and our best bid has queue spot 5 at this price, then the number shown is 7. The corresponding quantity for the offer side is displayed in blue.

We can see that the asset appreciates in price from 13:20 to 13:45, and that the price remains steady afterwards but with higher volatility than at the beginning of the trading hour. This coincides with an increase in the trading intensity, which can be read off from the changes in inventory or from the distances to the front of the queue at the best bid and ask in the fourth plot. We can also see from this plot how much fiercer the competition for the front spot has become after 13:20, and that the losses we incur are due to predominantly negative inventory positions in times of an increasing price.

9.3 Market making with a signal informed by the trade flow

The most successful trading period in the above example has indeed been the time from 13:00 to 13:20, where trading activity was moderately high but consistent, and there was little price volatility. It seems fairly uncontroversial to suspect a link between the strength of trading activity on the bid and offer side and the volatility of the asset price, but a precise statistical analysis of this conjecture is beyond the scope of this thesis. We have to make do with a rough classification of the market making signal S in different situations of trade flow:

Similarly high trading intensities on both sides. If the trading intensities at both the bid and ask price are high and of similar magnitude (two qualities yet to be made

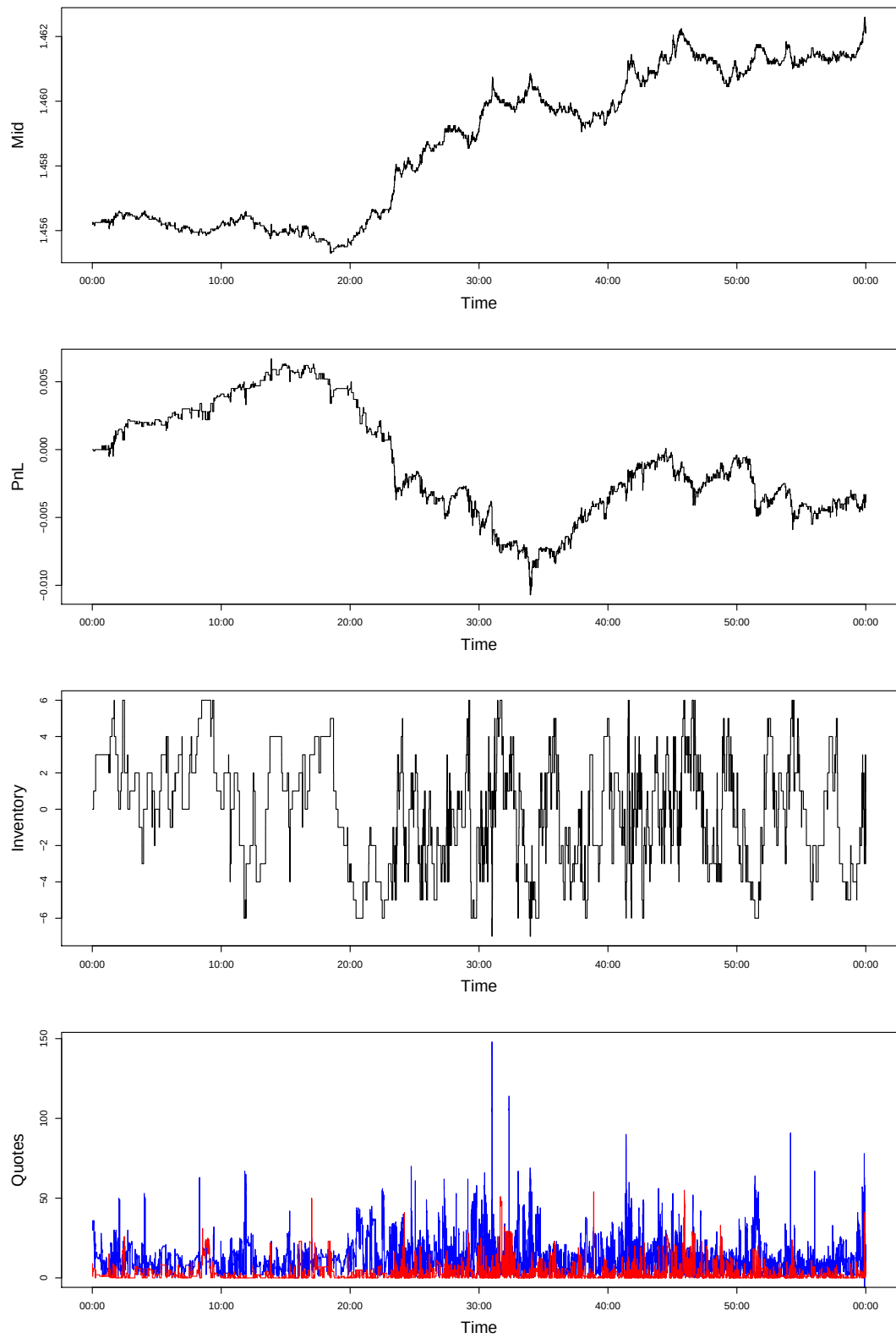


Figure 9.1 Graphical diagnostics and summaries for a market making with constant signal $S = 1$.

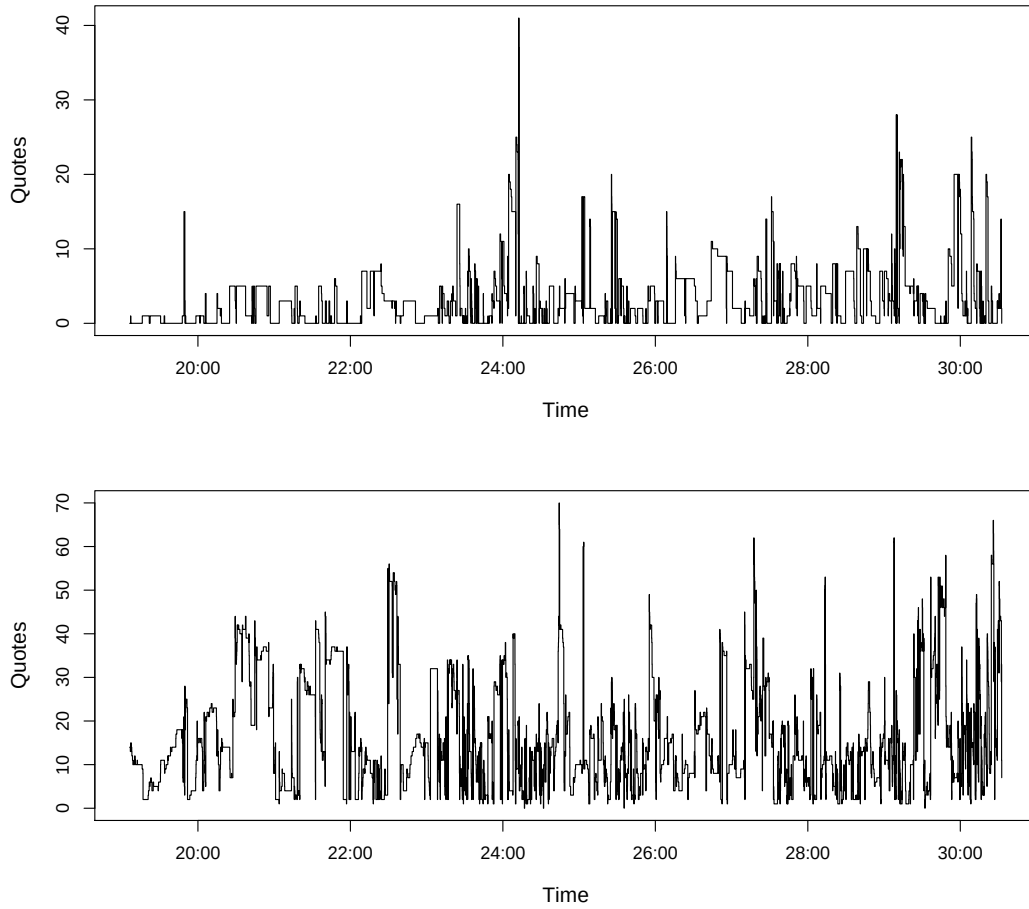


Figure 9.2 The number of third-party quotes that have priority over our best quote. Top plot for the bid side, bottom plot for the offer side.

precise), the market maker needs to quote aggressively on both sides to ensure that he is awarded a sizeable portion of the trades in the light of increased competition.

It is tempting to argue that such behaviour is justified from a risk perspective as the inventory can now be rebalanced more quickly, and that the market maker should hence be willing to exhaust his inventory constraints and to quote very close to or even at the best bid and best ask.

The premise of this reasoning is however controversial if not simply incorrect in such generality. Our argument against it is based on the example of market making with constant signal discussed above. Figure 9.2 shows the queue spots for the time window between 13:20 and 13:30; this is a time period with very high trading activity where most

of our losses accumulate. We first notice that we have consistently good queue positions on the bid side. This can be explained easily: as the price is increasing, new bid prices become available for quoting; and we occupy these virtually immediately ahead of the competition. Some trades on the bid side are hence negotiated by us, which is a desirable outcome as our inventory is negative and the price is increasing. The situation on the offer side is more intricate: most of the time, we are far away from the offer spot with highest priority, but we do occasionally and very abruptly reach this position. As a consequence, we manage to attract trades on this side of the book as well (see the change of inventory in Figure 9.1).

This is however much less desirable: as the mid price increases rather rapidly, it approaches quotes on the offer side that have been placed there much earlier, possibly for strategic reasons. (This explains our generally unfavourable queue spot positions on the offer side.) However, these quotes are usually cancelled just before they can be traded against; the reason for this could be that whoever put these quotes onto the book is no longer interested in execution as the price is now moving too rapidly. (These cancellations explain why our queue position suddenly improves so much.) As a result, we negotiate trades that do not help to rebalance our inventory. Even worse, the cancellations by other traders indicate that our offer quotes are in fact “picked off”, and that we are supplying the market with offer side liquidity at a time when most other traders prefer not to.

Periods with very high intensities of trading should hence be treated with more caution than a first intuitive reasoning may suggest, as such strong activity is often accompanied by strong price movements. Unless the market maker has information with respect to the direction of such price movements, a reduction of the inventory size (for both bid and offer) seems advisable in such situations. For periods with moderate but not excessive trading activity, the intuitive reasoning may however be applicable.

Dissimilar trading intensities. If the trading intensity is high on one side of the book but not on the other, an imbalanced inventory can be much harder to unwind. The market risk resulting from such periods of imbalance should be contained by a more careful

quoting behaviour: the trader will only use a fraction of the allowed maximum inventory size, and possibly also quote somewhat more aggressively on that side of the book where intensity is low in order to improve the price and stimulate demand.

Dormant markets. If the trading intensities on both sides of the book are low, the market maker may decide not to quote at all: being exposed to market risk and operational hazards may not be sufficiently compensated for by the few trades negotiated during dormant trading periods.

9.3.1 Using the filtering results for CIR-driven point processes

We now set T_0 as Monday, 14th November 2009, 00:00:00, and T_1 as Thursday, 17th November 2009, 18:00:00; and we assume a maximum inventory size of $J = 10$. We wish to derive an alternative to using a constant signal $S = (S_t, T_0 \leq t \leq T_1)$ for the determination of the total volumes described in Equation (9.2). Its derivation is guided by the findings discussed above: during periods with very high or dissimilar trading intensities we may decide to keep buying and selling the asset, but only within now more restrictive inventory requirements. We do however quote aggressively during times with normal and balanced trading intensities. Dormant periods (say overnight hours) are skipped entirely.

To begin with, we set $S_t = 0$ for all times t that are outside the core trading hours of 8am and 5:30pm on each of the four trading days. This should ensure that we will always be in a situation with enough trading activity to perform market making profitably.

To characterise the different levels of trade flow, we interpret the buy and sell trades as point processes driven by two independent CIR diffusions. Using the numerical results from Section 6.1, we calculate the conditional expectation of both the buy and sell intensity. We denote these by $\hat{\lambda}^B = (\hat{\lambda}_t^B; T_0 \leq t \leq T_1)$ for the bid side and $\hat{\lambda}^O = (\hat{\lambda}_t^O; T_0 \leq t \leq T_1)$ for the offer side.

We now require a precise notion of what it means for these two intensities to have “high” or “similar” magnitude.

(i) To characterise time periods when the two trading intensities are of similar magnitude or *in agreement*, we use an exponentially weighted moving average of the absolute differences $|\hat{\lambda}_t^O - \hat{\lambda}_t^B|$. A normalised version $S^A = (S_t^A; T_0 \leq t \leq T_1)$ of this agreement signal with values between 0 and 1, derived by using the inverse quantile function, is shown in Figure 9.3.

(ii) To characterise the time periods when overall trading activity is high, we have used an exponentially weighted moving average of the sum $(\hat{\lambda}_t^O + \hat{\lambda}_t^B)$. A normalised version of this signal of strength $S^S = (S_t^S; T_0 \leq t \leq T_1)$ with values between 0 and 1, derived by using the inverse quantile function, is shown in Figure 9.3.

We combine these two measures to reflect our preferences when to quote aggressively and when not to. For this purpose, we set

$$S_t = \frac{1 + S_t^A - S_t^S}{2} \quad (9.5)$$

as our market making signal. The interpretation in terms of $\mathcal{V}^B$ and $\mathcal{V}^O$, defined in Equation (9.2), is simple: we exhaust our inventory constraints during core trading periods when trading intensities are comparatively low on both sides; during periods where these become higher or more unequal, we voluntarily tighten our inventory constraints. The resulting signal for Monday, 14 September 2009 is shown in Figure 9.4 along with the asset mid price during that trading day. As surmised at the beginning of this section, there is tentative indication for a link between the nature and strength of trading activity and the volatility of the asset price.

We now use this signal to inform our market making. The outcome for Monday, 14 September 2009, is displayed in Figure 9.5. The result is satisfactory. During the calm morning period leading up to 1pm, we are generally prepared to accept a more imbalanced inventory; and the profit obtained from making the spread during this period is not negatively

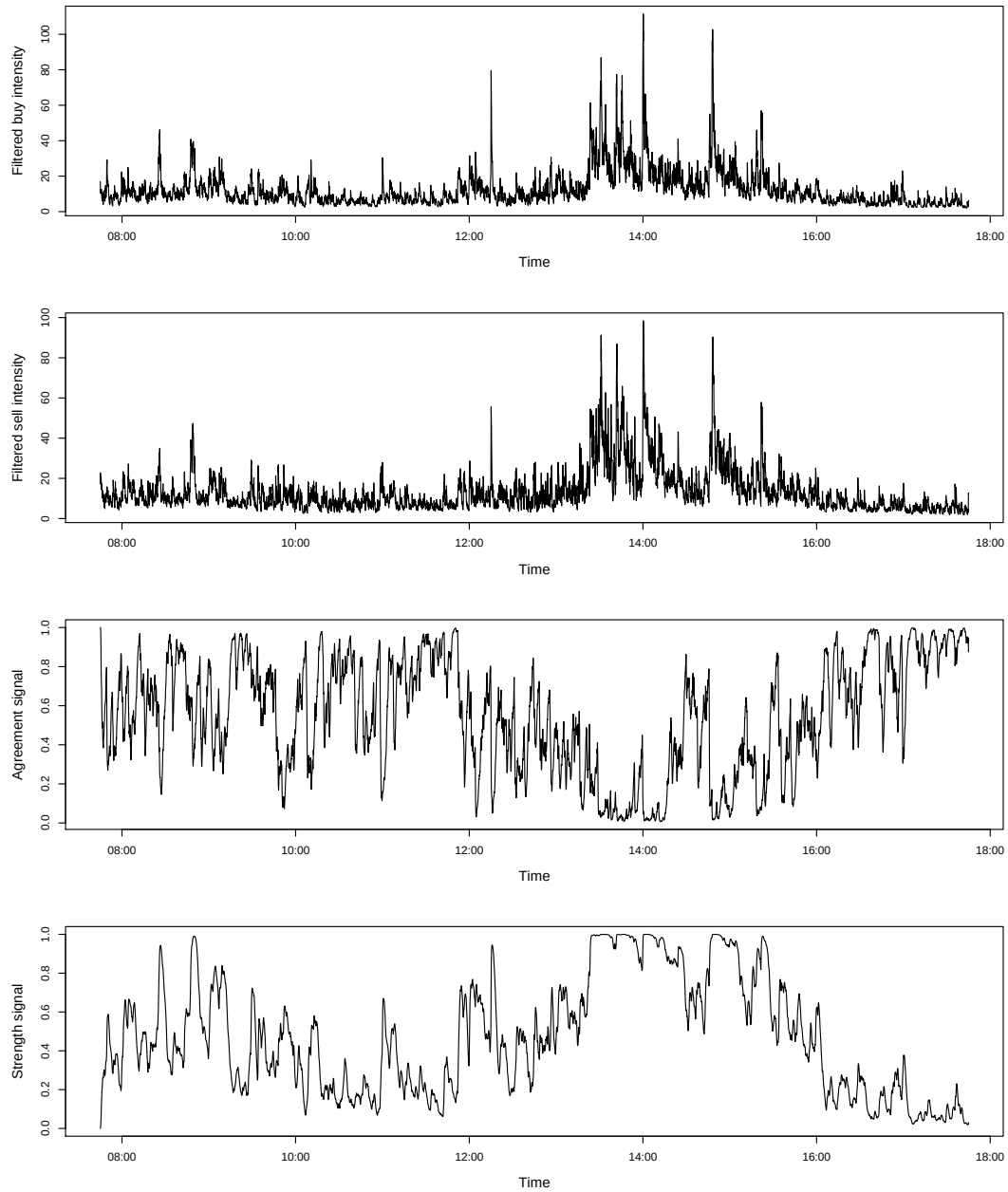


Figure 9.3 The first two plots show the filtered buy and sell intensity on Monday, 14 September 2009. The third plot shows the (normalised) agreement signal S^A ; the fourth plot the (normalised) strength signal S^S .

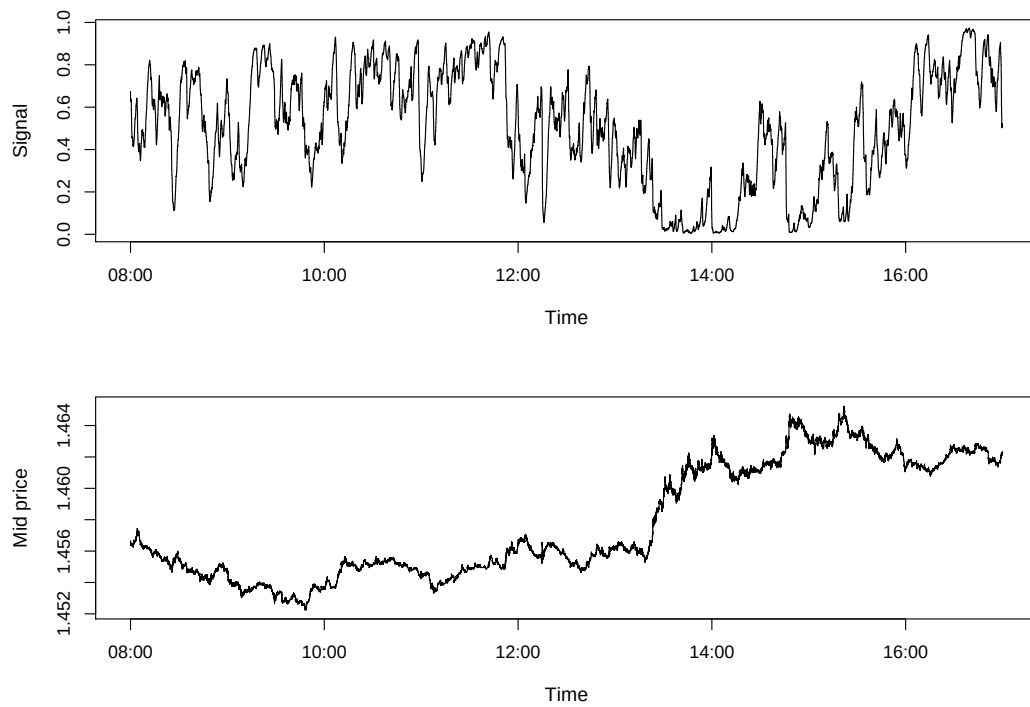


Figure 9.4 The top plot shows the signal S for Monday, 14 September 2009. The bottom plot shows the mid price of the currency pair in the same time window.

affected by market price fluctuations. From 1pm onwards, we begin to actively tighten our inventory constraints, and effectively stop trading at around 2pm. This coincides with a period of large price volatility. Trading is resumed more aggressively at around 3pm; the Sharpe ratio for this day's PnL is 0.76.

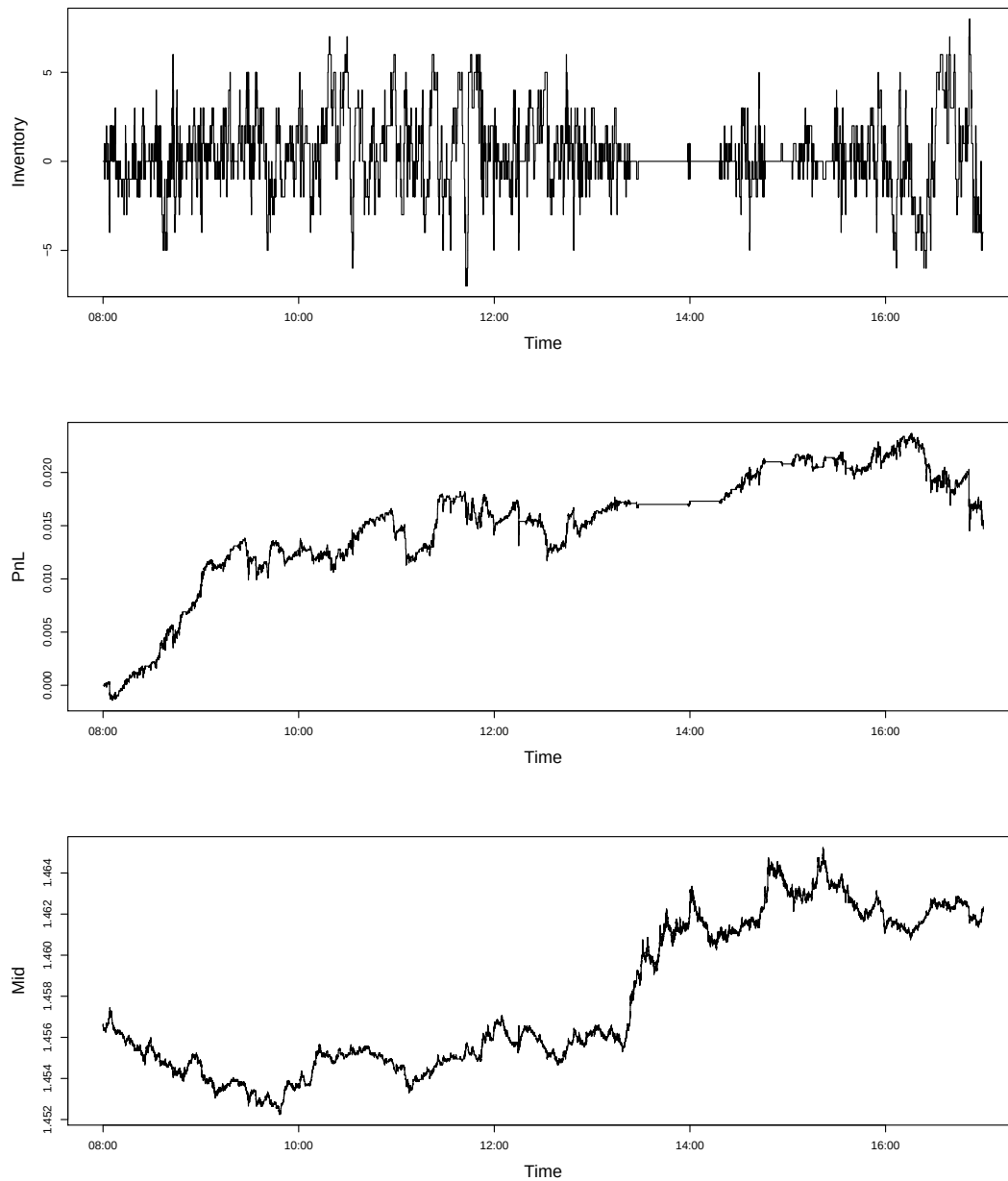


Figure 9.5 Graphical summaries for the market making simulation on Monday, 14 September 2009. The top plot shows the inventory, the middle plot the PnL, the bottom plot displays the mid price.

9.3.2 Comparison of different choices for the market making signal

We have investigated alternative specifications of the volume signal S in Equation (9.5):

- (i) A Hawkes process interpretation can be adopted for the sequence of buy and sell trades. Consequently, the filtering estimates $\hat{\lambda}^B$ and $\hat{\lambda}^O$ are replaced by the estimated Hawkes processes; these estimates can be obtained as described in Section 8.3. The derivation of the signal S then proceeds as before: we first derive (smoothed) signals of agreement and strength, and then obtain S using Equation (9.5).
- (ii) We have used a constant signal $S_t = 1$ to determine how a market maker fares that is ignorant of the trade flow.
- (iii) Based on S as defined in Equation (9.5) (using the CIR interpretation of the buy and sell processes), we have used the “inverse signal” $(1 - S_t)$ to inform the market making. We include this scenario to corroborate our intuitive reasoning at the beginning of this section: Using the inverse signal, we should predominantly be exposed to periods of high price volatility, and in the absence of information with respect to the direction of the price and no transaction costs, our expected profit from maintaining an inventory during these periods should be zero.

The Sharpe ratio results for market making simulations for Monday to Thursday, based on all four different choices of the signal S , are summarised in Table 9.1.

	Mon	Tue	Wed	Thu	Avg
CIR	0.76	0.73	0.60	0.20	0.57
Hawkes	0.93	0.54	0.87	-0.24	0.52
Constant	-0.07	-0.01	-0.52	-0.03	-0.16
Inverse CIR	-0.10	-0.22	-0.50	0.45	-0.09

Table 9.1 Sharpe ratios for the first four tradings, using four different choices for the volume signal S .

9.4 Conclusion

The mathematical modelling of (optimal) market making in a limit order book is challenging and complex. We have argued that one of its key components can be identified and formulated as an intensity control problem for the inventory that the market maker accumulates during trading.

Many factors can inform such controls. We have focussed exclusively on the flow of trades on the bid and offer side of the order book. Intuitive reasoning and empirical findings for a market making strategy oblivious to buy and sell events have lead us to relate the trade flow (represented by the estimated intensities of two point processes) to measures of agreement and strength of the overall trading activity. The results obtained in our backtesting study indicate that a market maker should trade aggressively when trading activity is of similar, moderately high magnitude on both sides of the book; and that he should contain risk during other market periods by limiting the size of his inventory.

This conclusion hints at a link between the strength and nature of the flow of trades and the price volatility of the traded asset. A characteristic of the popular class of stochastic volatility models (mentioned in Section 4.1) is that inference with respect to the latent volatility process is often solely based on the observed fluctuations of the *price of the asset*. It may be of great interest to study how such models can be improved by also taking into account the information represented by the times of trades on the bid and offer side, and whether those trades had an impact on the mid price or not. This question in fact reiterates what has been formulated in Section 2.1: to which extent can microscopic order book information be used to predict “macroscopic” characteristics of the asset price?

Such an analysis may also illustrate which additional information *not related* to price volatility is contained in trade event data. A more refined mathematical model of market making in limit order books is then required to decide how this information may benefit a trader, and which statistical model for the point processes of buy and sell events is the most appropriate. The results obtained within our current framework (collected in

Table 9.1) are only a first approximation and cannot provide a conclusive answer to this question. This is not a cause of the limited number of trading days analysed, but rooted in the simplifications and assumptions that were necessitated by the challenging data format and the abundance of parameters and choices for the design of the backtesting.

Bibliography

- Aalen, O. (1978). Nonparametric inference for a family of counting processes. *The Annals of Statistics*, 6(4):701–726.
- Angel, J. J., Harris, L. E., and Spatt, C. S. (2011). Equity trading in the 21st century. *The Quarterly Journal of Finance*, 1(01):1–53.
- Avellaneda, M. and Stoikov, S. (2008). High-frequency trading in a limit order book. *Quantitative Finance*, 8(3):217–224.
- Berman, M. (1983). Comment on ‘Likelihood analysis of point processes and its applications to seismological data’ by Y. Ogata. *Bulletin Internatl. Stat. Instit*, 50:412–418.
- Boel, R. and Beneš, V. (1980). Recursive nonlinear estimation of a diffusion acting as the rate of an observed Poisson process. *IEEE Transactions on Information Theory*, 26(5):561–575.
- Boel, R., Varaiya, P., and Wong, E. (1975). Martingales on jump processes. II: Applications. *SIAM Journal on Control*, 13(5):1022–1061.
- Bouchaud, J.-P., Mézard, M., and Potters, M. (2002). Statistical properties of stock order books: empirical results and models. *Quantitative Finance*, 2(4):251–256.
- Bowsher, C. (2007). Modelling security market events in continuous time: Intensity based, multivariate point process models. *Journal of Econometrics*, 141(2):876–912.

- Brémaud, P. (1972). *A Martingale Approach to Point Processes*. PhD thesis, University of California, Berkeley.
- Brémaud, P. (1981). *Point Processes and Queues, Martingale Dynamics*. Springer Verlag, New York.
- Cappé, O. (2001). Ten years of HMMs. www.tsi.enst.fr/~cappe/docs/hmmbib.html.
- Cappé, O., Moulines, E., and Rydén, T. (2005). *Inference in Hidden Markov Models*. Springer Series in Statistics. Springer, New York. With Randal Douc's contributions to Chapter 9 and Christian P. Robert's to Chapters 6, 7 and 13, With Chapter 14 by Gersende Fort, Philippe Soulier and Moulines, and Chapter 15 by Stéphane Boucheron and Elisabeth Gassiat.
- Chen, Z. (2003). Bayesian filtering: From Kalman filters to particle filters, and beyond. *Adaptive Systems Lab., McMaster University, Hamilton, Ontario, Canada*.
- Clifford, P. and Wei, G. (1993). The equivalence of the Cox process with squared radial Ornstein-Uhlenbeck intensity and the death process in a simple population model. *The Annals of Applied Probability*, 3(3):863–873.
- Cont, R. and De Larrard, A. (2013). Price dynamics in a Markovian limit order market. *SIAM Journal on Financial Mathematics*, 4(1):1–25.
- Cont, R., Stoikov, S., and Talreja, R. (2010). A stochastic model for order book dynamics. *Operations Research*, 58(3):549–563.
- Cox, D. (1955). Some statistical methods connected with series of events. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 129–164.
- Cox, D. (1963). Some models for series of events. *Bull. ISI*, 40:737–746.
- Cox, D. and Isham, V. (1980). *Point Processes*, volume 12. Chapman and Hall, London.

- Cox, D. and Miller, H. (1965). *The Theory of Stochastic Processes*, volume 134. Chapman and Hall, London.
- Cox, D. R. (1972). The statistical analysis of dependencies in point processes. In *Stochastic Point Processes*, volume 55. Wiley, New York.
- Cox, D. R. and Lewis, P. A. W. (1978). *The Statistical Analysis of Series of Events*. Chapman and Hall, London.
- Cox, J., Ingersoll Jr, J., and Ross, S. (1985). A theory of the term structure of interest rates. *Econometrica: Journal of the Econometric Society*, pages 385–407.
- Daley, D. and Vere-Jones, D. (2007). *An Introduction to the Theory of Point Processes: General Theory and Structure*, volume 2. Springer Verlag, New York.
- Davis, M. (1977). *Linear Estimation and Stochastic Control*. Chapman and Hall, London.
- Davison, A. (2008). *Statistical models*. Cambridge University Press, first paperback edition.
- Engle, R. and Russell, J. (1998). Autoregressive conditional duration: A new model for irregularly spaced transaction data. *Econometrica*, pages 1127–1162.
- Erlang, A. K. (1920). Telefon-Ventetider. Et Stykke Sandsynlighedsregning. *Matematisk Tidsskrift B*, 31:25–42.
- Fearnhead, P. (1998). *Sequential Monte Carlo Methods in Filter Theory*. PhD thesis, University of Oxford.
- Fearnhead, P. and Sherlock, C. (2006). An exact Gibbs sampler for the Markov-modulated Poisson process. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(5):767–784.
- Feller, W. (1939). Die Grundlagen der Volterra’schen Theorie des Kampfes ums Dasein in wahrscheinlichkeitstheoretischer Behandlung. *Acta Biotheoretica*, 5(1):11–40.

- Feller, W. (1950). *An Introduction to Probability Theory and its Applications*, volume 1. John Wiley & Sons, New York.
- Feller, W. (1951). Two singular diffusion problems. *Annals of mathematics*, pages 173–182.
- Fischer, W. and Meier-Hellstern, K. (1993). The Markov-modulated Poisson process (MMPP) cookbook. *Performance Evaluation*, 18(2):149–171.
- Gordon, N., Salmond, D., and Smith, A. (1993). Novel approach to nonlinear/non-Gaussian Bayesian state estimation. In *IEE Proceedings*, volume 140, pages 107–113.
- Gould, M., Porter, M., Williams, S., McDonald, M., Fenn, D., and Howison, S. (2012). Limit order books. *Available at SSRN 1970185*.
- Grandell, J. (1976). *Doubly Stochastic Poisson Processes*. Springer Verlag, New York.
- Guilbaud, F. and Pham, H. (2013). Optimal high-frequency trading with limit and market orders. *Quantitative Finance*, 13(1):79–94.
- Hall, A. D. and Hautsch, N. (2008). Order aggressiveness and order book dynamics. In *High Frequency Financial Econometrics*, pages 133–165. Springer.
- Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica: Journal of the Econometric Society*, pages 357–384.
- Hawkes, A. (1971). Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90.
- Ho, T. and Stoll, H. R. (1981). Optimal dealer pricing under transactions and return uncertainty. *Journal of Financial economics*, 9(1):47–73.
- Iacus, S. (2008). *Simulation and Inference for Stochastic Differential Equations: with R Examples*. Springer Verlag, New York.

- ICAP (2008). *EBS Market Data: Data Format and Message Flow*. ICAP plc., EBS Dealing Resources, Inc., Technology and Operations, One Upper Pond Road, Building F, 3rd Floor, Parsippany, New Jersey, USA 07054.
- Karlin, S. and Taylor, H. (1999). *A Second Course in Stochastic Processes*. Academic Press, New York.
- Kass, R. E. and Ventura, V. (2001). A spike-train probability model. *Neural computation*, 13(8):1713–1720.
- Kendall, D. G. (1953). Stochastic processes occurring in the theory of queues and their analysis by the method of the imbedded Markov chain. *The Annals of Mathematical Statistics*, pages 338–354.
- Kim, C.-J. and Nelson, C. R. (2003). *State-space models with regime switching: classical and Gibbs-sampling approaches with applications*, volume 1. MIT Press Books.
- Large, J. (2007). Measuring the resiliency of an electronic limit order book. *Journal of Financial Markets*, 10(1):1–25.
- Le Cam, L. (1947). Un instrument d’étude des fonctions aléatoires, la fonctionnelle caractéristique. *CR Acad. Sci., Paris*, 224:710–711.
- Leijdekker, V. and Spreij, P. (2011). Explicit computations for a filtering problem with point process observations with applications to credit risk. *Probability in the Engineering and Informational Sciences*, 25(03):393–418.
- Ogata, Y. (1981). On Lewis’ simulation method for point processes. *Information Theory, IEEE Transactions on*, 27(1):23–31.
- Ogata, Y. (1988). Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical Association*, 83(401):9–27.
- O’Hara, M. (1995). *Market Microstructure Theory*, volume 108. Blackwell Cambridge.

- Øksendal, B. (2003). *Stochastic Differential Equations: an Introduction with Applications*. Springer Verlag, Heidelberg, 5th edition.
- Papangelou, F. (1972). Integrability of expected increments of point processes and a related random change of scale. *Trans. Amer. Math. Soc.*, 165(3):483–506.
- Petris, G., Petrone, S., and Campagnoli, P. (2009). *Dynamic Linear Models*. Springer Verlag, Dordrecht.
- Potters, M. and Bouchaud, J.-P. (2003). More statistical properties of order books and price impact. *Physica A: Statistical Mechanics and its Applications*, 324(1):133–140.
- Rudemo, M. (1972). Doubly stochastic Poisson processes and process control. *Advances in Applied Probability*, pages 318–338.
- Segall, A. (1973). *A Martingale Approach to Modeling, Estimation and Detection of Jump Processes*,. PhD thesis, Stanford University, California.
- Segall, A. and Kailath, T. (1975). The modeling of randomly modulated jump processes. *Information Theory, IEEE Transactions on*, 21(2):135–143.
- Sharpe, W. F. (1966). Mutual fund performance. *The Journal of Business*, 39(1):119–138.
- Shephard, N. (2005). Stochastic volatility. *Economics Group, Nuffield College, University of Oxford, Economics Papers*.
- Smith, E., Farmer, J. D., Gillemot, L., and Krishnamurthy, S. (2003). Statistical theory of the continuous double auction. *Quantitative finance*, 3(6):481–514.
- Smith, J. (1979). A generalization of the Bayesian steady forecasting model. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 375–387.
- Snyder, D. (1972). Filtering and detection for doubly stochastic Poisson processes. *Information Theory, IEEE Transactions on*, 18(1):91–102.

- Snyder, D. and Miller, M. (1991). *Random Point Processes in Time and Space*. Springer Verlag, New York.
- Toke, I. M. (2011). An introduction to Hawkes processes with applications to finance. Presentation, http://fiquant.mas.ecp.fr/ioane_files/HawkesCourseSlides.pdf.
- Toth, B., Lemperiere, Y., Deremble, C., De Lataillade, J., Kockelkoren, J., and Bouchaud, J. (2011). Anomalous price impact and the critical nature of liquidity in financial markets. *Physical Review X*, 1(2):021006.
- Tsay, R. (2005). *Analysis of Financial Time Series*. Wiley-Interscience.
- Van Schuppen, J. (1973). *Estimation Theory for Continuous Time Processes: a Martingale Approach*. Electronics Research Laboratory, University of California.
- Wei, G. (1995). *Modelling and Inference for a Class of Doubly Stochastic Point Processes*. PhD thesis, University of Oxford.
- Wei, G., Clifford, P., and Feng, J. (2002). Population death sequences and Cox processes driven by interacting Feller diffusions. *Journal of Physics A: Mathematical and General*, 35:9309.
- Yashin, A. (1970). Filtering of jump processes. *Automat. Remote Control*, 31:725–730.
- Zakai, M. (1969). On the optimal filtering of diffusion processes. *Probability Theory and Related Fields*, 11(3):230–243.
- Zhang, T. and Kou, S. (2010). Nonparametric inference of doubly stochastic Poisson process data via the kernel method. *The Annals of Applied Statistics*, 4(4):1913.