Filtered Likelihood for Point Processes

Kay Giesecke and Gustavo Schwenkler

July 12, 2016†

Abstract

Point processes are used to model the timing of defaults, market transactions, births, unemployment and many other events. We develop and study likelihood estimators of the parameters of a marked point process and of incompletely observed explanatory factors that influence the arrival intensity and mark distribution. We establish an approximation to the likelihood and analyze the convergence and large-sample properties of the associated estimators. Numerical results highlight the computational efficiency of our estimators, and show that they can outperform EM Algorithm estimators.

*Giesecke is at the Department of Management Science & Engineering, Stanford University. Schwenkler is at the Department of Finance, Boston University Questrom School of Business. Schwenkler is corresponding author: Phone (617) 358-6266, email: gas@bu.edu, web: http://people.bu.edu/gas/.

†We thank Darrell Duffie, Peter Glynn, Tze Leung Lai, Marcel Rindisbacher, Yang Song, and Stefan Weber, as well as the participants at the 2013 INFORMS Annual Meeting, the 2014 Joint Mathematics Meeting, the 2015 SoFiE Annual Conference, and the 2015 World Congress of the Econometric Society for insightful comments. This work was completed while Schwenkler was a PhD student in the Department of Management Science and Engineering at Stanford University. Schwenkler acknowledges support from a Mayfield Fellowship and a Lieberman Fellowship.
1 Introduction

Point processes are stochastic models for the timing of discrete events. They are widely used in finance and economics to model the timing of defaults, market transactions, births, and other events of interest. In many cases, the point process intensity, which represents the conditional event arrival rate, is a function of explanatory factors. In practice, however, the factor data available to the econometrician are often incomplete. This paper addresses the inference problem arising in this situation. It develops and analyzes likelihood estimators of the parameters of the point process and of the explanatory factors. The results provide a rigorous statistical foundation for the applications of event timing models in the presence of incomplete factor data.

More specifically, we consider a marked point process whose intensity and mark distribution are parametric functions of a vector $V$ of explanatory factors. The factors are allowed to follow stochastic processes whose dynamics are specified by a parameter. The data include the event times and marks observed during a sample period. They also include certain observations of $V$. Some elements of $V$ are observed at any time, others are sampled only on certain dates, and the remaining ones are frailties that are never observed. The values between the sampling times of the partially observed factors, and the values of the frailty factors are not available for inference.

The model and data structure we study is motivated by a range of applications of event timing models in which incomplete information about $V$ is a key feature. An important example is the prediction of corporate defaults, see Azizpour, Giesecke & Schwenkler (2015), Duffie, Eckner, Horel & Saita (2009), Koopman, Lucas & Monteiro (2008), Koopman, Lucas & Schwaab (2011), Lando & Nielsen (2010) and others. Here, the event times of the point process represent default dates. A mark variable might describe the financial loss at default. The vector $V$ might include stochastically varying factors, such as stock returns, which are observed at any time during the sample period. It might also include macro-economic factors, such as industrial production, which are published only monthly or quarterly. Moreover, there is strong evidence of the presence of a dynamically varying, latent frailty factor influencing bankruptcy timing. Similar model and data structures are also common in the study of market transactions, unemployment spells, mortgage prepayments and delinquencies, the timing of births, and the duration of strikes. Section 2.3 provides several concrete examples.

We use a filtering approach to address the parameter inference problem with incomplete factor data. We begin by developing a representation of the likelihood of the data as a product of two terms, one addressing the event data and the other the available factor data. The decomposition is based on a change of probability measure that resolves the interaction between the point process and the factors influencing the arrival intensity.

---

The likelihood representation generalizes to incomplete factor data the representations of the complete data likelihood studied by Andersen, Borgan, Gill & Keiding (1993), Ogata (1978), Puri & Tuan (1986), and others. Our filtering approach treats the partial observation of explanatory factors exactly, and eliminates the need to interpolate between discrete observations of factors values. Interpolation between discrete observations is a standard practice in the literature to construct a data series with values for every time in the sample period. However, the implications of this practice for the properties of the estimators are not well understood.

One of the terms governing the likelihood is a point process filter whose practical computation may be difficult. We propose an approximation to this filter that eliminates the need to perform simulations to evaluate the likelihood estimator, which may be computationally burdensome for longer sample periods and cause a loss of estimation efficiency. The approximation is based on a quadrature method; an efficient algorithm reduces its computation to a sequential multiplication of matrices. Our approach is an alternative to the development and numerical solution of a Kushner-Stratonovich equation for the filter (see, e.g., Kurtz & Ocone (1988), Kliemann, Koch & Marchetti (1990), Kurtz (1998)).

We provide conditions guaranteeing the uniform convergence of the approximate likelihood to the exact likelihood. We also show that the parameter maximizing the approximate likelihood converges to the parameter at which the exact likelihood attains a global optimum and that it inherits the large-sample properties of the estimator maximizing the exact likelihood. The approximation does not generate additional asymptotic variance, so there is no loss of estimation efficiency when maximizing the approximate instead of the exact likelihood. Our approach to developing the properties of an estimator obtained from an approximate likelihood is an alternative to the approach of Dupacova & Wets (1988), who directly addressed the behavior of that estimator by viewing it as a solution of a stochastic optimization problem with partial information.

Numerical results illustrate the behavior of our estimators and provide a comparison with alternative simulation-based estimators. We analyze an extension of the classical self-exciting point process of Hawkes (1971), which is widely used in empirical applications (see Aït-Sahalia, Cacho-Diaz & Laeven (2015), Aït-Sahalia, Laeven & Pelizzon (2014), Azizpour et al. (2015), Bowsher (2007), Lando & Nielsen (2010), and others). The arrival intensity is a function of three factors, the first following a jump-diffusion that is observed monthly, the second following an unobservable geometric Brownian motion, and the third given by a functional of the point process path. We verify the convergence of the approximate to the true likelihood. Our approximation of the likelihood achieves the same rate of convergence as unbiased Monte Carlo simulation (which is intractable in our setting). It converges at a faster rate than an approximation that is based on standard

\[2\]
A similar quadrature method is used by Brillinger & Preisler (1983) to approximate the likelihood in a regression model with latent variables.

\[3\]
Zeng (2003) uses this approach to perform Bayesian least-squares estimation for a point process model that is a special case of our formulation.
Euler discretization. We compare our parameter estimates with those of the stochastic EM Algorithm (see Dempster, Laird & Rubin (1977), Wei & Tanner (1990), Celeux & Diebolt (1992), and others), which is often used to treat the inference problem with incomplete data. We find that our estimates are more accurate and less susceptible to the initial values for the majority of the model parameters. Our estimates also imply significantly better model fits to simulated data than the EM Algorithm parameters. Turning to large-sample properties (see Nielsen (2000)), we find that the EM Algorithm generates asymptotic standard errors in excess of those implied by our method for the majority of the parameters. Our estimates are found to be highly statistically efficient.

Prior research has considered inference for event timing models. Ogata (1978) develops and analyzes likelihood estimators of stationary and ergodic point processes with complete data. Andersen & Gill (1982) examine the proportional hazards model of Cox (1972) with complete data, in which the intensity is the product of an exponential function of factors and a deterministic, semi-parametric baseline hazard function. Nielsen, Gill, Andersen & Sorensen (1992), Murphy (1994), Murphy (1995) and Parner (1998) study a frailty model in which the intensity is the product of a semi-parametric baseline hazard function, an observable factor, and an unobservable frailty represented by an independent random variable. We focus on parametric intensity models and treat more general model and data structures. For example, we allow the frailty to follow a stochastic process that may be correlated with the observable factors. We also allow for arbitrary link functions, going beyond the proportional hazard formulation. Moreover, we allow for partial observation of explanatory factors. These features are common in the applied event timing literature; see the examples in Section 2.3. Finally, our results are related to the results of Le Cam & Yang (1988), who analyze likelihood estimators with incomplete data. Unlike these authors, we allow for some factors to be observed at all times. As a result, our data cannot be viewed as the product of repeated independent experiments and our likelihood cannot be decomposed into a product of conditional densities.

The rest of this paper is organized as follows. Section 2 formulates the point process model with incomplete factor data and provides examples. Section 3 develops a representation of the likelihood. Section 4 introduces and analyzes an approximation to the likelihood, and examines the asymptotic properties of the associated estimator. Section 5 provides an efficient algorithm to evaluate the approximate likelihood. Section 6 numerically illustrates the behavior of our estimators. Section 7 concludes. There are several appendices. Appendices A and B provide additional results on the exact and approximate estimators. Appendix C provides the proofs of our results.

The EM estimator is guaranteed to converge to a local optimum of the likelihood only; see Wu (1983) and Dembo & Zeitouni (1986). Our estimator is guaranteed to converge to a global optimum.

We have implemented our estimators in the R package. The code can be downloaded at http://people.bu.edu/gas. It can be easily customized to treat alternative models specifications.
Marked point process

Fix a complete probability space \((\Omega, \mathcal{F}, \mathbb{P})\) and a right-continuous and complete filtration \((\mathcal{F}_t)_{t \geq 0}\). Consider a marked point process \((T_n, U_n)_{n \geq 1}\). The \(T_n\) are strictly positive and strictly increasing stopping times that represent the arrival times of events. The \(U_n\) are \(\mathcal{F}_{T_n}\)-measurable mark variables that take values in a subset \(R_U\) of Euclidean space. They provide additional information about the events.

2.1 Model

Suppose the counting process \(N\) given by \(N_t = \sum_{n \geq 1} 1\{T_n \leq t\}\) has intensity \(\lambda\) (see II.D7 in Brémaud (1980)). The intensity represents the conditional event arrival rate. We take

\[
\lambda_t = \Lambda(V_t; \alpha)
\]

for a function \(\Lambda : R_V \to (0, \infty)\) specified by a parameter \(\alpha \in \Theta_N\), and an explanatory factor process \(V = (X, Y, Z) \in R_V \subset \mathbb{R}^{d_X} \times \mathbb{R}^{d_Y} \times \mathbb{R}^{d_Z}\), where \(d_X, d_Y\) and \(d_Z\) are natural numbers. The dynamics of \(V\) are specified by a parameter \(\gamma \in \Theta_U\). The conditional distribution of the mark variables is described by a probability transition kernel \(\Pi_t(du)\) from \(\Omega \times [0, \infty)\) into \(R_U\) (see VIII.D5 in Brémaud (1980)). We take

\[
\Pi_t(du) = \pi(u, V_t^-; \beta)du
\]

for a density function \(\pi : R_U \times R_V \to (0, \infty)\) specified by a parameter \(\beta \in \Theta_U\). Technical conditions need to be imposed on \(\pi, \Lambda\), and the dynamics of \(V\) in order for the point process to be non-explosive (see, for example, Protter (2004) and Gjessing, Røysland, Pena & Aalen (2010)).

2.2 Data

The goal is to estimate the parameter \(\theta = (\alpha, \beta, \gamma)\) governing the intensity, the distribution of the marks, and the dynamics of the explanatory factor \(V\). The data available for inference include the event times and marks observed during a sample period \([0, \tau]\), where \(\tau > 0\). The data also include certain observations of \(V\) during the sample period. More precisely, the data are given by \(D_\tau = (N_\tau, U_\tau, X_\tau, Y_\tau)\), where

- \(N_\tau = (N_t : 0 \leq t \leq \tau)\) represents the observations of the \(T_n\);
- \(U_\tau = (\sum_{n \leq N_t} U_n : 0 \leq t \leq \tau)\) represents the observations of the \(U_n\);
- \(X_\tau = (X_t : 0 \leq t \leq \tau)\) represents the observations of \(X\);
- \(Y_\tau = (Y^j_t : t \leq \tau, t \in S^j, j = 1, \ldots, d_Y)\) represents the observations of \(Y\), where the \(S^j\) are discrete subsets of \([0, \infty)\) with \(S = \bigcap_{j=1}^{d_Y} S^j \neq \emptyset\).
The structure of the data addresses the fact that in practice, the data on the explanatory factor \( V = (X, Y, Z) \) may be incomplete. While the data include the entire history of \( X \) for the sample period, they include the values of \( Y \) only on certain dates. The data do not include any values of \( Z \); this element of \( V \) is an unobservable frailty.

2.3 Examples

The point process model and data structure we study have a wide scope. We illustrate with a number of examples from the empirical literature.

**Example 2.1** (Timing of Corporate Bankruptcies). Azizpour et al. (2015) examine the timing of bankruptcies in the United States between 1970 and 2012. An event time \( T_n \) represents a bankruptcy date. A mark \( U_n \in \mathbb{R}_+ \) represents the debt outstanding at default. The density \( \pi \) is parameter independent and modeled by the empirical distribution in the sample. The intensity function is \( \Lambda(v; \alpha) = \exp(\alpha_1 \cdot v_1) + \alpha_2 v_2 + \alpha_3 v_3 \) for \( v = (v_1, v_2, v_3) \) and \( \alpha = (\alpha_1, \alpha_2, \alpha_3) \). The factor vector is \( V = (V^1, \ldots, V^6) \), where

\[
V^1_t = \sum_{n \leq N_t} e^{-\gamma_1(t - T_n)} \log U_n
\]

generates self-exciting behavior. The element \( V^2 \) is an unobservable frailty governed by the SDE

\[
dV^2_t = \gamma_3 (\gamma_4 - V^2_t) dt + \sqrt{V^2_t} dW_t
\]

where \( W \) is a Brownian motion and \( \gamma = (\gamma_1, \ldots, \gamma_4) \) with \( \gamma_1 > 0 \). The initial value \( V^2_0 \) is sampled from its stationary gamma distribution with shape parameter \( 2\gamma_3\gamma_4 \) and scale parameter \( \frac{1}{2\gamma_3} \). The element \( V^3 \) follows a vector autoregressive model and represents several interest rates, stock index returns and volatilities that are sampled daily. The element \( V^4 \) follows an autoregressive model and represents credit spreads that are sampled weekly. The element \( V^5 \) follows an autoregressive moving-average model, represents the industrial production growth rate, and is sampled monthly. The element \( V^6 \) follows a moving-average model, represents the growth rate of the gross domestic product, and is sampled quarterly. We have \( V = (X, Y, Z) \) with \( X = V^1 \), \( Y = V^2 \) and \( Z = V^3 \), \( dX = 1 \), \( dY = 4 \), \( dZ = 1 \), \( S^1 = \{k/365 : k \in \mathbb{N}\} \), \( S^2 = \{k/52 : k \in \mathbb{N}\} \), \( S^3 = \{k/12 : k \in \mathbb{N}\} \), and \( S^4 = \{k/4 : k \in \mathbb{N}\} \). Related models are analyzed by Duffie, Saita & Wang (2007), Duffie et al. (2009), Koopman et al. (2008), Koopman et al. (2011), and Lando & Nielsen (2010).

**Example 2.2** (Timing of Births). Newman & McCulloch (1984) explore the factors driving the timing of births for 4724 women between the ages of 20 and 49 in Costa Rica. A unit of time is one month, and time is measured starting at the 13-th birthday of a woman. The event time \( T_n \) corresponds to the time of first birth for the \( n \)-th woman. The event intensity is \( \Lambda(v; \alpha) = \exp(\alpha \cdot v) \) for \( v = (v_1, \ldots, v_4) \) and \( \alpha = (\alpha_1, \ldots, \alpha_4) \). The explanatory factor \( V_t = (V^1_t, \ldots, V^4_t) \) satisfies \( V^2_t = t \), \( V^3_t = \max\{t - 48, 0\} \), and \( V^4_t = \max\{t - 144, 0\} \). The factor \( V^1 \) is a 6-dimensional vector that represents personal
and regional characteristics such as schooling level and mortality rates, and is observed monthly. Overall, \( d_X = 3, d_Y = 6, d_Z = 0 \), \( X = (V^2, V^3, V^4) \), \( Y = V^1 \), and \( S^j = \mathbb{N}_0 \) for \( 1 \leq j \leq 6 \). Similar models are analyzed by Heckman & Walker (1990), Hotz, Klerman & Willis (1997), and Todd, Winters & Stecklov (2012).

**Example 2.3** (Timing of Market Transactions). Engle & Russell (1998) analyze IBM stock transactions data during regular trading hours for the period between November 1, 1990, and January 31, 1991. The event time \( T_n \) represents the time of the \( n \)-th transaction. The event intensity satisfies \( \Lambda(v; \alpha) = \Lambda_0(v_0/v_1; \alpha)v_1^{-1} \) for \( v = (v_0, v_1) \). The vector of explanatory factors is \( V_t = (V^0_t, V^1_t) \), where \( V^0_t = t - T_N \) and \( V^1_t = V^1_t \) for

\[
V^1_n = \gamma_0 + \sum_{j=0}^{m} \gamma_{j+1} (T_{n-j} - T_{n-j-1}) + \sum_{j=0}^{q} \gamma_{m+2+j} V^1_{n-j}
\]

for \( \gamma = (\gamma_0, \gamma_1, \ldots, \gamma_{m+q+2}) \) and \( m, q \in \mathbb{N} \). This model is called the autoregressive conditional duration model because the conditional distribution of the time until the next event depends on the realization of past inter-event times. The data include complete observations of \( V^0 \) and \( V^1 \). Thus, \( d_X = 2, d_Y = d_Z = 0 \), and \( V = X \). Related models of inter-transaction times are analyzed by Dufour & Engle (2000) and Engle (2000).

**Example 2.4** (Unemployment Timing). Lancaster (1979) studies the duration of unemployment among 479 British individuals. The event time \( T_n \) represents the arrival time of the \( n \)-th job offer. The \( n \)-th mark is \( U_n = (w_n, a_n) \in \mathbb{R}_+ \times \{0, 1\} \), where \( w_n \) indicates the wage offered for the \( n \)-th job and \( a_n \) indicates acceptance of an offer. The density \( \pi \) is parameter-independent and modeled by the empirical distribution in the sample. The intensity \( \Lambda(v; \alpha) = v_2 \exp(\alpha \cdot v_1) \) for \( v = (v_1, v_2) \) and \( V_t = (V^1_t, V^2_t) \). The three-dimensional observable explanatory factor component \( V^1 \) represents the age of the individual, the unemployment rate in the UK, and social security benefits. The component \( V^2 \) is an unobservable frailty that has a gamma distribution with unit mean and variance \( \gamma \). Thus, \( d_X = 3, d_Y = 0, d_Z = 1 \) and \( V = (X, Z) \) with \( X = V^1 \) and \( Z = V^2 \). Lancaster & Nickell (1980), Meyer (1990) and Dolton & O’Neill (1996) examine related models.

**Example 2.5** (Mortgage Delinquencies). Deng et al. (2000) investigate home mortgage default and prepayment events in the United States using quarterly data from Freddie Mac on single-family mortgage loans issued between 1976 and 1983. The event time \( T_n \) represents the first time of either default or prepayment of the \( n \)-th loan. The event intensity is \( \Lambda(v; \alpha) = v_{00} \exp(v_2 + v_3 + \alpha_d \cdot v_6) + v_1 \exp(v_4 + v_5 + \alpha_p \cdot v_6) \) for \( v = (v_0, \ldots, v_6) \) and \( \alpha = (\alpha_d, \alpha_p) \). The vector of explanatory factors is \( V_t = (V^0_t, \ldots, V^6_t) \). Here, \( V^0_t = V^0 \) and \( V^1_t = V^1 \) are time-invariant one-dimensional frailty factors that take values in discrete sets and characterize the borrower type, \( V^2_t \) and \( V^3_t \) are one-dimensional baseline functions, \( V^3_t \) and \( V^4_t \) are two-dimensional measures of the moneyness of the default and prepayment options held by the borrower, respectively, and \( V^6_t \) is a two-dimensional vector of macro-factors such as employment and divorce rates. The data includes continuous
observations of \((V^2, V^4)\) and quarterly observations of \((V^3, V^5, V^6)\). Therefore, \(d_X = 2\), \(d_Y = 6\), \(d_Z = 2\), and \(V = (X, Y, Z)\) for \(X = (V^2, V^4)\), \(Y = (V^3, V^5, V^6)\), \(Z = (V^0, V^1)\), and \(S^j = N_0\) for \(1 \leq j \leq 6\). Related models are analyzed by Demyanyk & Van Hemert (2011), Deng & Gabriel (2006), Foote, Gerardi & Willen (2008), and Pennington-Cross & Ho (2010).

Our formulation facilitates the analysis of multiple influences on event timing and various data structures. The arrival intensity \((1)\) is a non-negative function of a vector \(V\) of explanatory factors. The elements of \(V\) may include time, allowing for time-inhomogeneity as in Examples 2.2 and 2.3. They may be represented by time-invariant random variables (Examples 2.4 and 2.5), or by stochastic processes (Examples 2.1 and 2.3). The elements of \(V\) may be correlated (Examples 2.1 and 2.5) and need not be Markovian (Example 2.3). The point process may be an element of \(V\) or have an impact on the dynamics of other elements of \(V\), to generate self-exciting behavior as in Examples 2.1 and 2.3. Some elements of \(V\) may be completely unobservable frailties (Examples 2.4, 2.5, and 2.1). Others may be observable only on certain dates (Examples 2.1, 2.2, and 2.5) or even at any time (Examples 2.1 and 2.3). Our formulation allows one to address all of these features in a common model and data framework.

3 Likelihood inference

We analyze likelihood estimators of the parameter \(\theta = (\alpha, \beta, \gamma) \in \Theta = \Theta_N \times \Theta_U \times \Theta_V\). We write \(P_{\theta}\) and \(E_{\theta}\) for the probability and the expectation operators when the underlying parameter is \(\theta\). The true but unknown parameter \(\theta_0\) belongs to \(\Theta^0\), the interior of \(\Theta\). The observed data is a realization of \(D_\tau\) under \(P_{\theta_0}\).

The likelihood \(L_\tau(\theta)\) of the data \(D_\tau\) at the parameter \(\theta \in \Theta\) is taken as the Radon-Nikodym derivative of the \(P_{\theta}\)-distribution of the data with respect to the true \(P_{\theta_0}\)-distribution; see Appendix A.1 for a precise definition. With complete factor data \((d_Y = d_Z = 0)\), the likelihood is given by \(dP_{\theta}/dP_{\theta_0}\). In the general case, the likelihood is given by the projection of \(dP_{\theta}/dP_{\theta_0}\) onto the observed data:

\[
L_\tau(\theta) = E_{\theta_0}\left[\frac{dP_{\theta}}{dP_{\theta_0}}\bigg| D_\tau\right].
\] (3)

Unfortunately, the representation (3) is impractical because the true parameter \(\theta_0\) is unknown. In order to obtain an alternative representation, we introduce an auxiliary probability measure. Let \(\pi^*\) be a strictly positive probability density on \(R_U\). If the variable

\[
M_\tau(\theta) = \prod_{n \leq N_\tau} \frac{\pi^*(U_n)}{\pi(U_n, V_{T_n-}; \beta)} \times \exp \left( -\int_0^\tau \log(\Lambda(V_s-; \alpha))dN_s - \int_0^\tau (1 - \Lambda(V_s; \alpha))ds \right)
\] (4)

is again a strictly positive probability density on \(R_U\), then

\[
L_\tau(\theta) = E_{\theta_0}\left[\frac{dP_{\theta}}{dP_{\theta_0}}\bigg| D_\tau\right] = E_{\theta_0}\left[\frac{dP_{\theta}}{dP_{\theta_0}}\bigg| D_\tau\right] = \int_0^\tau \pi^*(U_n)dN_n
\]

for all \(\theta \in \Theta\).
has expected value equal to one, then it induces an equivalent probability measure \( P^* \) on \( \mathcal{F}_\tau \). Under this measure, the event counting process \( N \) is a standard Poisson process on the interval \([0, \tau]\) and a mark \( U_n \) has density \( \pi^* \). The intensity and the mark distribution relative to \( P^* \) do neither depend on the factor \( V \) nor the parameter \( \theta \). This property facilitates a decomposition of the likelihood (3) into a product of two terms, one addressing the event data \((N_\tau, U_\tau)\) and the other the factor data \((X_\tau, Y_\tau)\). Each of these terms is given by an expectation under the auxiliary measure.

**Theorem 3.1.** Suppose the following conditions hold:

(A1) The expectation \( \mathbb{E}_\theta[M_\tau(\theta)] = 1 \).

(A2) The conditional likelihood \( \mathcal{L}^*_F(\theta; \tau) \) of the factor data \((X_\tau, Y_\tau)\) given the event data \((N_\tau, U_\tau)\) exists under \( P^*_\theta \).

Then the likelihood satisfies

\[
\mathcal{L}_\tau(\theta) \propto \mathbb{E}_\theta^* \left[ \frac{1}{M_\tau(\theta)} \Bigg| D_\tau \right] \mathcal{L}^*_F(\theta; \tau). \tag{5}
\]

Condition (A1) guarantees that the measure \( P^*_\theta \) is well-defined. Blanchet & Ruf (2013) provide sufficient conditions for (A1). Condition (A2) requires the absolute continuity of the conditional \( P^*_\theta \)-law of the factor data \((X_\tau, Y_\tau)\) given the event data \((N_\tau, U_\tau)\) relative to the corresponding conditional law under \( P^*_\theta_0 \); see Appendix A for details. Sufficient conditions for this property depend on the model and data structure specified for \( V \). For a jump-diffusion model of \( V \) with discrete observation dates, for example, see Komatsu & Takeuchi (2001) and Takeuchi (2002).

Theorem 3.1 generalizes to incomplete factor data the representations of the complete data point process likelihood studied by Andersen et al. (1993), Ogata (1978), Puri & Tuan (1986) and others. Indeed, if \( d_Y = d_Z = 0 \), the data \( D_\tau \) include all values of the factor \( V = X \) required to evaluate \( M_\tau(\theta) \). In this case the conditional expectation in (5) is trivial so

\[
\mathcal{L}_\tau(\theta) \propto \frac{\mathcal{L}^*_F(\theta; \tau)}{M_\tau(\theta)}. \tag{6}
\]

With incomplete factor data \((d_Y > 0 \text{ or } d_Z > 0)\), the data \( D_\tau \) are insufficient to evaluate \( M_\tau(\theta) \). In this case, the conditional expectation in (5) represents a posterior mean of \( 1/M_\tau(\theta) \) given the data.

From Theorem 3.1, ignoring an additive term that is independent of \( \theta \), the log-likelihood takes the form

\[
L_\tau(\theta) = \log \mathcal{L}^*_E(\theta; \tau) + \log \mathcal{L}^*_F(\theta; \tau), \tag{7}
\]

where \( \mathcal{L}^*_E(\theta; \tau) = \mathbb{E}_\theta^*[1/M_\tau(\theta) \big| D_\tau] \). A maximum likelihood estimator (MLE) \( \hat{\theta}_\tau \in \Theta^o \) of the parameter \( \theta \) is a solution to

\[
0 = \nabla L_\tau(\theta). \tag{8}
\]
Appendix A.2 studies the asymptotic properties of a MLE. Identifiability, smoothness, boundedness, and non-singularity conditions ensure that \( \hat{\theta}_\tau \to \theta_0 \) in \( P_{\theta_0} \)-probability and \( \sqrt{\tau}(\hat{\theta}_\tau - \theta_0) \to N(0, \Sigma_0) \) in \( P_{\theta_0} \)-distribution as \( \tau \to \infty \). The asymptotic variance-covariance matrix \( \Sigma_0 \in \mathbb{R}^{p \times p} \) is given by

\[
\Sigma_0 = I(\theta_0)^{-1} + 2\zeta I(\theta_0)^{-1} \Sigma_{EF}(\theta_0) I(\theta_0)^{-1},
\]

where \( \zeta = \lim_{\tau \to \infty} \frac{|S \cap [0, \tau]|}{\tau} \) is the mean number of observations of \( Y \) per unit of time, \( I(\theta_0) \) is the Fisher information matrix for the estimation of the true parameter \( \theta_0 \), and \( \Sigma_{EF}(\theta_0) \) is a matrix that addresses the filtering of the unobserved factors. According to Ogata (1978), the asymptotic variance-covariance matrix is \( \Sigma_0 = I(\theta_0)^{-1} \) in the case of complete factor data. As a result, the filtering calls for a correction term.

4 Approximate likelihood estimators

Theorem 3.1 states that the likelihood is governed by two terms. The term \( \mathcal{L}_E^*(\theta; \tau) \) addresses the available factor data; its computation is standard for a range of models of \( V \), see Section 6 as well as Aït-Sahalia (2008), Giesecke & Schwenkler (2015), Lo (1988), and others. The term \( \mathcal{L}^*_E(\theta; \tau) = \mathbb{E}_\theta[1/M_\tau(\theta) \mid D_\tau] \) addresses the event data. Depending on the model specification, its practical computation may be difficult.

This section develops and analyzes an approximation to the point process filter \( \mathcal{L}_E^*(\theta; \tau) \). It also examines the convergence and asymptotic properties of the estimators maximizing the approximate likelihood. We focus on the case that at least one of the factors is an unobserved frailty \( (dZ > 0) \).

4.1 Filter approximation

We consider filters of the form

\[
\mathcal{E}(\theta; \tau, g) = \mathbb{E}_\theta^*[g(V_\tau, \tau)/M_\tau(\theta) \mid D_\tau]
\]

for functions \( g \) on \( R_V \times [0, \infty) \). The term \( \mathcal{L}_E^*(\theta; \tau) \) can be computed by taking \( g(v, t) = 1 \). While we treat \( \mathcal{E}(\theta; t, g) \) at the horizon \( t = \tau \), the extension to dates \( t < \tau \) is straightforward.

A standard approach to computing (10) if \( d_Y = 0 \) is to develop and numerically solve the associated Kushner-Stratonovic equation (see Kliemann et al. (1990)). To our knowledge, however, the KS equation has not yet been worked out for the case that \( d_Y > 0 \), i.e., when some factors are observed at fixed dates only.

We propose to approximate \( \mathcal{E}(\theta; \tau, g) \) using a quadrature-based method. Fix a number \( m_S \in \mathbb{N} \) and choose points \( v^1, \ldots, v^{m_S} \in R_V \), together with compact sets \( \mathcal{A}^1, \ldots, \mathcal{A}^{m_S} \subseteq R_V \) that are pairwise disjoint such that \( \mathcal{A}^k \) is a neighborhood of \( v^k \) with non-empty interior. Additionally, fix another number \( m_T \in \mathbb{N} \), and choose times \( T^0, T^1, \ldots, \)
$T^{m_T} \in [0, \tau]$ with $0 = T^0 < T^1 < \ldots < T^{m_T} = \tau$. Our approximation of $E(\theta; \tau, g)$ is given by

$$E^T(\theta; \tau, g) = \sum_{i_0=1}^{m_S} \cdots \sum_{i_{m_T}=1}^{m_S} g(v^{m_T}, \tau) F^T_{\theta}(i_1, \ldots, i_{m_T}) P^T_{\theta}(i_0, i_1, \ldots, i_{m_T})$$

where $I = \{m_S, m_T, v, A, T\}$ is the implementation of the approximation with $v = \{v^1, \ldots, v^{m_S}\}$, $A = \{A^1, \ldots, A^{m_S}\}$ and $T = \{T^1, \ldots, T^{m_T}\}$. Further, writing $\Lambda(\cdot; \theta)$ and $\pi(\cdot; \theta)$ rather than $\Lambda(\cdot; \alpha)$ and $\pi(\cdot; \beta)$, let

$$F^T_{\theta}(i_1, \ldots, i_{m_T}) = \exp \left\{ \sum_{j=1}^{m_T} \sum_{T_n \in \{T^j-1, T^j\}} \log \frac{\Lambda(v^{i_j}; \theta) \pi(U_n, v^{i_j}; \theta)}{\pi(U_n)} \right\}$$

$$+ \sum_{j=1}^{m_T} (1 - \Lambda(v^{i_j}; \theta)) \left( T^j - T^{j-1} \right)$$

$$P^T_{\theta}(i_0, \ldots, i_{m_T}) = \mathbb{P}^*_{\theta}[V_0 \in A^{i_0}, V_t \in A^{i_j} (1 \leq j \leq m_T, t \in (T^{j-1}, T^j)] | D_r].$$

The approximation $E^T(\theta; \tau, g)$ of $E(\theta; \tau, g)$ is a weighted average of certain paths of the explanatory factors $V$. The paths are those for which $V_0 \in \{v^1, \ldots, v^{m_S}\}$ and $V_t$ is constant in the interval $(T^{j-1}, T^j]$ for $1 \leq j \leq m_T$, taking values in the set $\{v^1, \ldots, v^{m_S}\} \subset R_V$. The weights assigned to these paths are the conditional $\mathbb{P}^*_\theta$-probabilities, given the data $D_r$, that $V$ is in the set $A^{i_j}$ during the interval $(T^{j-1}, T^j]$ for $1 \leq j \leq m_T$ and $i_j \in \{1, \ldots, m_S\}$. Thus, $E^T(\theta; \tau, g)$ is based on a multidimensional rectangular quadrature approximation to the Lebesgue integral in $E(\theta; \tau, g)$, taken with respect to the measure $\mathbb{P}^*_\theta$ conditional on the data $D_r$. Section 5 provides an algorithm for evaluating the approximate filter $E^T(\theta; \tau, g)$.

### 4.2 Convergence

We examine the convergence of the approximate filter $E^T(\theta; \tau, g)$ and of the estimators based on $E^T(\theta; \tau, g)$. The idea is to choose sequences of compact sets $\{A^{n,j}\}_{1 \leq j \leq m^n_S}$ and of times $\{T^{n,j}\}_{1 \leq j \leq m^n_T}$, such that, in the limit as $n \to \infty$, the entire range of $V$ is covered by the compact sets while both the sets $A^{n,j}$ and the time intervals $(T^{n,j-1}, T^{n,j}]$ become smaller. Let $\partial_v$ denote the partial derivative with respect to $v$.

**Theorem 4.1** (Convergence of approximate filter). Let $g$ be a continuous or bounded function. Suppose the following conditions hold.

**\(B1\)** The partial derivative $\partial_v \Lambda(v; \theta)$ of the intensity function $\Lambda$ exists. Both $\Lambda(v; \theta)$ and $\partial_v \Lambda(v; \theta)$ are continuous in $\theta \in \Theta$.

**\(B2\)** The partial derivative $\partial_v \pi(u, v; \theta)$ of the $\mathbb{P}$-density function $\pi$ of the marks exists. The functions $\pi(u, v; \theta)$ and $\partial_v \pi(u, v; \theta)$ are continuous in $\theta \in \Theta$. 

11
(B3) The space of admissible parameters $\Theta$ is a compact subset of an Euclidean space and has non-empty, convex interior.

(B4) $\mathbb{E}_\theta[1/M_r(\theta)^2] < \infty$ for any $\theta \in \Theta$.

Let $\mathcal{T}^n = \{m^n_S, m^n_T, v^n, A^n, T^n\}$ with $v^n = (v^{n,j})_{1 \leq j \leq m^n_S}$, $A^n = (A^{n,j})_{1 \leq j \leq m^n_S}$, and $T^n = (T^{n,i})_{1 \leq i \leq m^n_T}$ be a sequence of data-dependent implementations that satisfies the following conditions $\mathbb{P}$-almost surely:

(B5) $m^n_S, m^n_T < \infty$ for all $n \in \mathbb{N}$.

(B6) $A^n$ is a sequence of compact and convex $\mathbb{R}^d$-sets with non-empty, pairwise disjoint interiors, such that $A^n = \bigcup_{j=1}^{m^n_S} A^{n,j} \xrightarrow{n \to \infty} R_V$.

(B7) $v^{n,j} \in R_V$ for each $n \in \mathbb{N}$ and $1 \leq j \leq m^n_S$, and the interior of $A^{n,j}$ is an open neighborhood of $v^{n,j}$.

(B8) $0 = T^{n,0} < T^{n,1} < \ldots < T^{n,m^n_T} = \tau$.

(B9) In the limit as $n \to \infty$, it holds $\mathbb{P}$-almost surely that

$$
\psi^{n,1}_r = \max_{1 \leq j \leq m^n_S} \text{vol}(A^{n,j}) \left[ \tau \max_{\theta \in \Theta} \max_{1 \leq j \leq m^n_S} \left| \partial_v \Lambda(v^{j}; \theta) \right| + N_r \max_{\theta \in \Theta} \max_{1 \leq k \leq N_r} \max_{1 \leq j \leq m^n_S} \left| \partial_v \log \frac{\Lambda(v^{j}; \theta) \pi(U_k, v^{j}; \theta)}{\pi^*(U_k)} \right| \right] \to 0.
$$

(B10) In the limit as $n \to \infty$,

$$
\psi^{n,2}_r = \sup_{\theta \in \Theta} \mathbb{P}_\theta^* \left[ \exists t \in [0, \tau] : V_t \notin A^n \big| D_r \right] \to 0.
$$

Then, for fixed $\tau > 0$, the absolute error of $E_T^n(\cdot; \tau, g)$ satisfies

$$
|E_T^n(\theta; \tau, g) - \mathcal{E}(\theta; \tau, g)| = O \left( \psi^{n,1}_r + \sqrt{\psi^{n,2}_r} \right), \quad (11)
$$

and $E_T^n(\cdot; \tau, g)$ converges uniformly in the space of admissible parameters $\Theta$ to $\mathcal{E}(\cdot; \tau, g)$ as $n \to \infty$, $\mathbb{P}$-almost surely.

Theorem 4.1 states that if one can construct implementations such that Conditions (B5)–(B9) are satisfied, then the approximation $E_T^n(\theta; \tau, g)$ converges uniformly in the parameter $\theta \in \Theta$ as quickly as $O(\psi^{n,1}_r + \sqrt{\psi^{n,2}_r})$. To achieve this, the sequence of compact sets $A^{n,j}$ has to be chosen such that the volume of each set becomes smaller faster than $|\partial_v \log \Lambda(v; \theta) \pi(u, v; \theta)/\pi^*(u)|$ and $|\partial_\theta \Lambda(v; \theta)|$ grow, while their union covers the range $R_V$ of the explanatory factor $V$. The state space discretization determines the rate of convergence of the algorithm by controlling for extreme cases of the intensity function and
the mark distribution. The time discretization grid \( \{ T^1, \ldots, T^{m_T} \} \) can be chosen so as to facilitate the calculation of the probabilities \( P^T_{\theta} (i_0, i_1, \ldots, i_{m_T^n}) \); we address this choice in Section 5. Assumptions (B1)–(B4) are standard. They guarantee that Assumptions (B5)–(B9) are well-posed. A sufficient condition for the existence of an implementation \( I^n \) as in Theorem 4.1 is that there exists a sequence of compact and convex sets \( A_n \in \sigma(R_V) \) with non-empty interior, \( A_n \neq R_V \), such that \( P^* \)-almost surely

\[
\sup_{\theta \in \Theta} \mathbb{P}_\theta \left[ \exists t \in [0, \tau] : V_t \notin A^n \mid D_\tau \right] \to 0. \tag{12}
\]

Theorem 4.1 guarantees that our scheme approximates the likelihood function to the same degree of accuracy for any \( \theta \in \Theta \) and converges uniformly. Next, we analyze the convergence of the estimators obtained from the approximate likelihood.

**Proposition 4.2** (Convergence of estimators). Suppose the conditions of Theorem 4.1 apply and, in addition, \( \sup_{\theta \in \Theta} \mathcal{L}_\tau^* (\theta; \tau) < \infty, \mathbb{P}\)-almost surely. Assume that \( \hat{\theta}_\tau \in \Theta \) maximizes the likelihood

\[
\mathcal{L}_\tau (\theta) = \mathcal{E}(\theta; \tau, 1) \mathcal{L}_\tau^* (\theta; \tau)
\]

and that \( \hat{\theta}_n^\tau \) for \( n \in \mathbb{N} \) maximizes the approximate likelihood

\[
\mathcal{L}_n^\tau (\theta) = \mathcal{E}^T (\theta; \tau, 1) \mathcal{L}_\tau^* (\theta; \tau).
\]

Then for fixed \( \tau > 0, \mathbb{P}\)-a.s.,

\[
\lim_{n \to \infty} \mathcal{L}_n^\tau (\hat{\theta}_n^\tau) = \mathcal{L}_\tau (\hat{\theta}_\tau)
\]

and \( \theta^n \) converges to a global optimum of the likelihood for any realization of the data \( D_\tau \).

Assuming smoothness of \( \mathcal{L}_\tau (\cdot) \), since \( \mathcal{L}_\tau (\hat{\theta}_\tau) < \infty \) and the parameter set \( \Theta \) is compact, the sequence of estimators \( \hat{\theta}_n^\tau \) obtained by the uniformly convergent implementation of our approximation converges to some parameter \( \hat{\theta}_\tau^\infty \in \Theta \). Proposition 4.2 states that \( \mathcal{L}_\tau (\hat{\theta}_\tau^\infty) = \mathcal{L}_\tau (\hat{\theta}_\tau) \), that is, both \( \hat{\theta}_\tau^\infty \) and \( \hat{\theta}_\tau \) achieve the same likelihood, making \( \hat{\theta}_\tau^\infty \) a global maximizer of the likelihood function \( \mathcal{L}_\tau (\cdot) \). When \( \hat{\theta}_\tau \) is the unique MLE, then \( \hat{\theta}_\tau^\infty = \hat{\theta}_\tau \) and hence \( \hat{\theta}_n^\tau \to \hat{\theta}_\tau \) as \( n \to \infty \).

### 4.3 Asymptotic properties of estimators

The uniform convergence of the approximate likelihood function established above entails additional results. Following the notation in Proposition 4.2, we write

\[
\sqrt{\tau} (\hat{\theta}_n^\tau - \theta_0) = \sqrt{\tau} (\hat{\theta}_n^\tau - \hat{\theta}_\tau^\infty) + \sqrt{\tau} (\hat{\theta}_\tau^\infty - \theta_0). \tag{13}
\]

The assumptions in Theorem 4.1 guarantee that, for any given degree of accuracy, one can construct an implementation of our approximation achieving that accuracy. If \( n \) goes to \( \infty \) fast enough as \( \tau \to \infty \), then \( \sqrt{\tau} (\hat{\theta}_n^\tau - \hat{\theta}_\tau^\infty) \) converges to zero \( \mathbb{P}\)-almost surely. This observation leads to the following result.
Proposition 4.3 (Asymptotic properties of estimators). Suppose that the conditions of Theorem 4.1 and Proposition 4.2 are satisfied. In addition, assume that the following conditions also hold.

(B11) Any exact maximum likelihood estimator is consistent and asymptotically normal; i.e., \( \hat{\theta}_\tau \to \theta_0 \) in \( P_{\theta_0} \)-probability and \( \sqrt{\tau}(\hat{\theta}_\tau - \theta_0) \to N(0, \Sigma_0) \) in \( P_{\theta_0} \)-distribution as \( \tau \to \infty \) for \( \Sigma_0 \) as in (9).

(B12) The likelihood \( L_\tau(\cdot) \) does not attain optima on the boundary of \( \Theta \).

(B13) For any \( n \geq 1 \), the Hessian matrix \( \nabla^2 L^\tau_n \) of the approximate likelihood is positive definite in a neighborhood of \( \hat{\theta}_\tau^n \).

Define the sequence \( \tau \mapsto n(\tau) \) such that

\[
\psi_{\tau}^{n(\tau),1} + \sqrt{\psi_{\tau}^{n(\tau),2}} = O \left( \tau^{-q} \right),
\]

for some \( q > \frac{1}{2} \). Then the sequence of parameter estimators \( (\hat{\theta}_\tau^n(\tau))_{\tau > 0} \) derived from the implementations \( I^n(\tau) \) with \( \tau \mapsto n(\tau) \) as defined in (14) satisfies

\[
\hat{\theta}_\tau^n(\tau) \to \theta_0
\]

as \( \tau \to \infty \) in \( P_{\theta_0} \)-probability, and

\[
\sqrt{\tau}(\hat{\theta}_\tau^n(\tau) - \theta_0) \to N(0, \Sigma_0)
\]

as \( \tau \to \infty \) in \( P_{\theta_0} \)-distribution, where \( \Sigma_0 \) is the same as in (9).

Proposition 4.3 gives implementations of our approximation that generate consistent and asymptotically normal estimators. The asymptotic variance-covariance matrix of the estimators is the same as that of the estimator \( \hat{\theta}_\tau \) obtained from the exact likelihood. Thus, the approximations do not generate additional asymptotic variance. There is no loss of efficiency when maximizing the approximate likelihood instead of the exact (but uncomputable) one. In order to achieve this result, the approximation has to converge to the true likelihood sufficiently fast as \( \tau \to \infty \). This is accomplished by picking a sequence \( (I^n(\tau))_{\tau \geq 0} \) of implementations of our approximation so that condition (14) is satisfied for some \( q > \frac{1}{2} \).

Appendix A.2 provides conditions ensuring consistency and asymptotic normality of exact maximum likelihood estimators (Assumption (B11)). Further, Appendix B indicates how an approximation to the asymptotic variance-covariance matrix \( \Sigma_0 \) can be constructed using our approximate filter.
5 Filter computation

We provide a convenient algorithm for evaluating the filter approximation $E^T(\theta; \tau, g)$. If the process $V$ is Markovian, then the probabilities $P^T_{\theta}(i_0, \ldots, i_{m_T})$ can be written as

$$\mathbb{E}_\theta \left[ 1_{A_0}(V_0) \mathbb{P}_\theta^* \left[ \forall j \in \{1, \ldots, m_T\} \forall t \in (T^{j-1}, T^j] : V_t \in A^j \mid D_\tau, V_0 \right] D_\tau \right].$$

We propose to approximate $P^T_{\theta}(i_0, \ldots, i_{m_T})$ by the following interpolation:

$$\hat{P}^T_{\theta}(i_0, \ldots, i_{m_T}) = \mathbb{P}_\theta^* \left[ V_0 \in A^{i_0} \mid D_\tau \right] \prod_{j=1}^{m_T} \mathbb{P}_\theta^* \left[ V_{T^j} \in A^{i_j} \mid D_\tau, V_{T^{j-1}} = v^{j-1} \right].$$

If $\max_{1 \leq j \leq m_T |A^i|} \approx 0$, then $\hat{P}^T_{\theta}(i_0, \ldots, i_{m_T}) \approx P^T_{\theta}(i_0, \ldots, i_{m_T})$. The following proposition formalizes this statement.

**Proposition 5.1 (Estimation of transition probabilities).** Suppose the conditions of Theorem 4.1 apply and assume that $V$ is Markovian. Let $(\mathcal{T}_n)_{n \geq 1}$ be a sequence of implementations guaranteeing the convergence of the approximate filter. If $m_n^T \to \infty$ as $n \to \infty$, then

$$|\hat{P}^{T_n}_{\theta}(i_0, \ldots, i_{m_n^T}) - P^{T_n}_{\theta}(i_0, \ldots, i_{m_n^T})| \to 0$$

as $n \to \infty$, $\mathbb{P}_\theta$-almost surely.

Proposition 5.1 implies that the quantities $\hat{P}^T_{\theta}(i_0, \ldots, i_{m_T})$ are consistent estimators of the probabilities $P^T_{\theta}(i_0, \ldots, i_{m_T})$. An immediate consequence is that the general convergence results of Section 4.2 are not altered if we compute the filter approximation $E^T(\theta; \tau, g)$ using $\hat{P}^T_{\theta}(i_0, \ldots, i_{m_T})$ instead of $P^T_{\theta}(i_0, \ldots, i_{m_T})$. Let

$$F^T_{\theta}(i) = e^{\sum_{n \in (T^{j-1}, T^j]} \log (\Lambda(v^j, \theta) \pi(v^j, i, \theta))} \pi(v^j, i, \theta) \mathbb{P}_\theta^*[V_{T^j} = v^j | D_\tau],$$

for $1 \leq j \leq m_T$ and $1 \leq i \leq m_S$. Then $F^T_{\theta}(i_1, \ldots, i_{m_T})$ can be rewritten as $\prod_{j=1}^{m_T} F^T_{\theta}(i_j)$.

Similarly, define the matrices $\hat{p}^T_{\theta}(i)$ for $1 \leq j \leq m_T$ with components

$$\hat{p}^T_{\theta}(i, j) = \mathbb{P}_\theta^*[V_{T^j} = v^{|j|} | D_\tau],$$

as well as the vector

$$\hat{p}^T_{\theta}(i) = \mathbb{P}_\theta^*[V_0 \in A^i | D_\tau]$$

for $1 \leq k, l \leq m_S$. We can write $\hat{P}^T_{\theta}(i_0, \ldots, i_{m_T}) = \prod_{j=1}^{m_T} \hat{p}^T_{\theta}(i_j) \hat{p}^T_{\theta}(i_0)$. It follows that the approximation

$$\hat{E}^T(\theta; \tau, g) = \sum_{i_0=1}^{m_S} \cdots \sum_{i_{m_T}=1}^{m_S} g(v^{m_T}, \tau) \prod_{j=1}^{m_T} F^T_{\theta}(i_j) \hat{p}^T_{\theta}(i_j) \hat{p}^T_{\theta}(i_0)$$

shares the same convergence properties as $E^T(\theta; \tau, g)$ in Theorem 4.1. Conveniently, $\hat{E}^T(\theta; \tau, g)$ can be calculated by a series of matrix multiplications.

We have implemented our estimator in R. The code can be downloaded at [http://people.bu.edu/gas](http://people.bu.edu/gas). We summarize the steps.
Algorithm 5.2. For a given $\theta \in \Theta$ do

1. Initialization: Let $\xi \in \mathbb{R}^{m_S}$ with $\xi(i) = \hat{p}_{\theta}^{I,0}(i)$ for $1 \leq i \leq m_S$.
2. Iteration: For $j = 1, \ldots, m_T - 1$ do
   a. Define $\hat{Q}_j \in \mathbb{R}^{m_S \times m_S}$ with elements $F_{\theta}^{I,j}(k) \hat{p}_{\theta}^{I,j}(k,l)$
   b. Update the vector $\xi$ to $\xi \leftarrow \hat{Q}_j \cdot \xi$
3. Termination:
   a. Define $\hat{Q}_{m_T} \in \mathbb{R}^{m_S \times m_S}$ with elements $g(v^k, \tau) F_{\theta}^{I,m_T}(k) \hat{p}_{\theta}^{I,m_T}(k,l)$
   b. Update the vector $\xi$ to $\xi \leftarrow \hat{Q}_{m_T} \cdot \xi$
   c. Compute $\hat{E}^I(\theta; \tau, g) = \sum_{i=1}^{m_S} \xi(i)$.

5.1 Choice of implementation

The implementation $I^n = \{m^n_S, m^n_T, v^n, A^n, T^n\}$ of our approximation determines its rate of convergence as indicated in Theorem 4.1. It also determines the computational expenses for evaluating our approximation according to Algorithm 5.2. The choice of the approximation implementation should balance error versus computational requirements.

We adopt the computational efficiency concept of Glynn & Whitt (1992) to make this choice.\(^6\) Let $c_n$ denote the computational costs for evaluating our approximation based on the implementation $I^n$, and $R_n$ denote its squared error. Suppose that there exist constants $r, v > 0$ such that

$$\lim_{n \to \infty} c_n^r R_n \to \frac{1}{v}.$$ \hspace{1cm} (15)

Then $r$ gives the asymptotic efficiency rate and $v$ the asymptotic efficiency value of the approximation based on $I^n$. Computationally efficient implementations $I^n$ are those that have large asymptotic efficiency rates and values.

Algorithm 5.2 involves $m^n_T$ multiplications of $m^n_S \times m^n_S$-matrices, $m^n_T (m^n_S)^2$ simple multiplications, and one addition of length $m^n_S$ in order to compute $\hat{E}^T(\theta; \tau, g)$. Therefore, the computational effort is of order $c_n = O(m^n_T (m^n_S)^3)$. The computational costs grow cubically with the fineness of the state space discretization, and linearly with the fineness of the time discretization. The squared error of the approximation $\hat{E}^T(\theta; \tau, g)$ is of order $R_n = O((\psi^n_{\tau,1})^2 + \psi^n_{\tau,2})$. These observations lead to the following proposition that indicates how the implementation $I^n$ has to be chosen to achieve a predetermined asymptotic efficiency rate $r > 0$.

\(^6\)Even though the computational efficiency concept of Glynn & Whitt (1992) was originally introduced for Monte Carlo estimators, this concept is also applicable here if we replace the risk function $R$ in Glynn & Whitt (1992) with the squared error of the approximation.
Proposition 5.3. Let \( r > 0 \) be given. Suppose that the implementation \( \mathcal{I}^n \) is set up such that \( \psi_{\tau}^{n,1} = O \left( \sqrt{\psi_{\tau}^{n,2}} \right) \). If

\[
m_T^n (m_S^n)^3 = O \left( (\psi_{\tau}^{n,1})^{-\frac{2}{r}} \right),
\]

then \( r \) is the asymptotic efficiency rate of the implementation \( \mathcal{I}^n \), and \( c^n_r R_n \rightarrow \frac{1}{v} \) as \( n \rightarrow \infty \) for some asymptotic efficiency value \( v > 0 \).

The above proposition provides guidance on how to choose the implementation \( \mathcal{I}^n \). If one wants to achieve an asymptotic efficiency rate of \( r > 0 \), then the computational costs of the implementation \( \mathcal{I}^n \) have to grow at a slower rate than the \( r \)-root of the inverse squared error. This imposes restrictions on what asymptotic efficiency rates can be achieved by our approximation. The range of asymptotic efficiency rates that can be achieved will depend on the formulation of the intensity function \( \Lambda \) and on the model specification for the covariates \( V \) given the definition of \( \psi_{\tau}^{n,1} \) and \( \psi_{\tau}^{n,2} \) in (B9)-(B10). Section 6 illustrates the maximum possible asymptotic efficiency rate that can be achieved by our approximation in a numerical example of empirical relevance.

6 Numerical analysis

This section illustrates our estimators and contrasts them with alternatives within a credit risk modeling framework. We take the intensity (1) as

\[
\Lambda(v; \alpha) = \alpha_1 v_1 + \alpha_2 v_2 + \alpha_3 e^{-\alpha_4 v_5} v_3
\]

for \( v = (v_1, v_2, v_3, v_4, v_5) \in R_V = R^5_+ \) and \( \alpha = (\alpha_1, \alpha_2, \alpha_3, \alpha_4) \in \Theta_N = R^4_+ \). The vector \( V = (V^1, \ldots, V^5) \) of explanatory factors satisfies the stochastic differential equation

\[
dV_t = \mu(V_t; \gamma)dt + \sigma(V_t; \gamma)dW_t + R(V_t; \gamma)dN_t
\]

for a two-dimensional standard Brownian motion \( W = (W_1, W_2) \), \( v = (v_1, \ldots, v_5) \in R_V \), \( \gamma = (\gamma_1, \ldots, \gamma_8) \in \Theta_V = R \times R^3_+ \times R \times R^3_+ \), and

\[
\mu(v; \gamma) = \begin{pmatrix}
(\gamma_1 + \frac{\gamma_2}{2})(v_1 - \gamma_3 e^{-\gamma_4 v_5} v_4) - \gamma_3 \gamma_4 e^{-\gamma_4 v_5} v_4 \\
(\gamma_5 + \frac{\gamma_6}{2}) v_2 \\
0 \\
0 \\
1
\end{pmatrix}
\]  (17)

\[
\sigma(v; \gamma) = \begin{pmatrix}
\gamma_2 (v_1 - \gamma_3 e^{-\gamma_4 v_5} v_4) & 0 \\
0 & \gamma_6 v_2 \\
0 & 0 \\
0 & 0 \\
0 & 0
\end{pmatrix}
\]  (18)
The marks are irrelevant; to fully specify the Radon-Nikodym derivative (4) we take \( R_U = (0, 1) \) and \( \pi(u, v; \beta) = \pi^*(u) = 1 \). Theorem V.3.7 in Protter (2004) guarantees the existence of a unique, non-explosive solution \((V, N)\) to the system specified above.\(^7\) We take \( \gamma_7 = \alpha_4 \) and \( \gamma_8 = \gamma_4 \), and see that \( V \) takes the form
\[
V_t^1 = V_0^1 \exp \left( \gamma_1 t + \gamma_2 W_t^1 \right) + \gamma_3 \int_0^t e^{-\gamma_4(t-u)} dN_u,
\]
\[
V_t^2 = V_0^2 \exp \left( \gamma_5 t + \gamma_6 W_t^2 \right),
\]
\[
V_t^3 = \int_0^t e^{\alpha_4 u} dN_u, \quad V_t^4 = \int_0^t e^{\gamma_4 u} dN_u, \quad V_t^5 = t.
\]

The factor \( V^1 \) follows a jump-diffusion with jumps driven by \( N \); the impact of a jump decays exponentially with time at rate \( \gamma_4 \). The factor \( V^2 \) follows a geometric Brownian motion. The factors \( V^3 \) and \( V^4 \) are pure-jump processes driven by \( N \), and \( V^5 \) is time. The data structure is as follows. The factors \( V^3, V^4, \) and \( V^5 \) are observed at any time. The factor \( V^1 \) is observed only at times \( S = \{ \frac{k}{12} : k \in \mathbb{N} \} \), while \( V^2 \) is a completely unobserved frailty. We have \( dX = 3, \; dY = dZ = 1, \; X = (V^3, V^4, V^5), \; Y = V^1, \) and \( Z = V^2 \).

Equations (16)-(22) specify a reduced-form model of default events. The event intensity \( \lambda_t = \Lambda(V_t, \alpha) \) is influenced by observable idiosyncratic and systematic risk factors, past events, as well as by a latent frailty variable. This formulation is similar to or extends the models analyzed by Azizpour et al. (2015), Duffie et al. (2009), Duffie et al. (2007), Lando & Nielsen (2010), and others.\(^8\)

We take \( \theta_0 = (\alpha_{1,0}, \alpha_{2,0}, \alpha_{3,0}, \alpha_{4,0}, \gamma_{1,0}, \gamma_{2,0}, \gamma_{3,0}, \gamma_{4,0}, \gamma_{5,0}, \gamma_{6,0}) \) with \( \alpha_{1,0} = 3, \alpha_{2,0} = 3, \alpha_{3,0} = 2, \alpha_{4,0} = 2.5, \gamma_{1,0} = -0.5, \gamma_{2,0} = 1, \gamma_{3,0} = 0.2, \gamma_{4,0} = 5, \gamma_{5,0} = -0.125, \) and \( \gamma_{6,0} = 0.5 \). For simplicity, we take \( V_0 = (2.8, 1, 0, 0, 0) \), such that the starting value \( V_0 \) of the explanatory factor \( V \) is non-random. These values are roughly consistent with the fitted parameter values in Azizpour et al. (2015).\(^9\) We take \( \Theta = [0, 20]^4 \times [-20, 20] \times [0, 20]^3 \times [-20, 20] \times [0, 20] \subset \Theta_N \times \Theta_V \).

Time is measured in years. We generate realizations of the counting process \( N \) under the model (16)-(22) parametrized by \( \theta_0 \) using the discretization method described and analyzed by Giesecke & Teng (2012). We take 200,000 equidistant time steps during the time interval \([0, 40]\). The numerical results reported below are based on an implementation

\(^7\)To see this, consider the SDE for \((V, N)\), which is driven by the semi-martingales \( W \) and \( N - \int_0^t \lambda_s ds \). The coefficients of this SDE are functional Lipschitz in the sense of the definition in Section V.3 of Protter (2004). Thus, Theorem V.3.7 in Protter (2004) applies.

\(^8\)Azizpour et al. (2015) assume that \( \gamma_3 = 0 \), and that the frailty \( V^2 \) follows a Cox-Ingersoll-Ross process. Duffie et al. (2009) assume that \( \Lambda \) is log-linear in the factors \( v_1 \) and \( v_2 \), and that the frailty \( V^2 \) follows an Ornstein-Uhlenbeck process. Lando & Nielsen (2010) assume that there is no frailty \( (V^2 \equiv 0) \), while Duffie et al. (2007) assume that there is neither frailty \((V^2 \equiv 0)\) nor self-excitation \((\gamma_3 = 0 \) and \( V^3 \equiv 0)\).

\(^9\)We have considered alternative values, but found no significant difference in the results.
We also fix the rate of convergence in (24). We fix the computational requirements to evaluate our approximation; it does not have an effect on the accuracy of the approximation. For a given \( n \in \mathbb{N} \), we can discretize the state spaces of \( Y \) and \( Z \) separately of each other. We set \( A_{n,j}^1 = A_{Y}^{n,j} \times A_{X}^{n,j} \) and discretize the intensity process \( \lambda_t \) as

\[
\Lambda((X_t, y^{n,j}, z^{n,j}); \theta),
\]

where \( y^{n,j} \in A_{Y}^{n,j} \) constitutes the discretization of \( Y \), and \( z^{n,j} \in A_{Z}^{n,j} \) constitutes the discretization of \( Z \). We implement equidistant discretizations. This means that

\[
\text{vol}(A_{n,j}^1) = \frac{\max_{1 \leq j \leq m^S_y} y^{n,j} - \min_{1 \leq j \leq m^S_y} y^{n,j}}{m^S_y} \times \frac{\max_{1 \leq j \leq m^S_z} z^{n,j} - \min_{1 \leq j \leq m^S_z} z^{n,j}}{m^S_z}
\]

We have \( |\partial_y \Lambda((X_t, y^{n,j}, z^{n,j}); \theta)| = O(1) \) and

\[
|\partial_y \log \Lambda((X_t, y^{n,j}, z^{n,j}); \theta)| = O\left(\frac{1}{\min_{1 \leq j \leq m^S_y} y^{n,j}}\right).
\]

The analogous result holds for the partial derivatives with respect to \( z \). Therefore,

\[
\psi^{n,1}_r = O\left(\sqrt{\frac{\max_{1 \leq j \leq m^S_y} y^{n,j} \max_{1 \leq j \leq m^S_z} z^{n,j}}{(m^S_y)^2 \min_{1 \leq j \leq m^S_y} \{y^{n,j}, z^{n,j}\}}\right).
\]

For a given \( n \in \mathbb{N} \), we fix

\[
\max_{1 \leq j \leq m^S_y} y^{n,j} = O(q^n_Y), \quad \min_{1 \leq j \leq m^S_y} y^{n,j} = O(q^n_{Y}),
\]

\[
\max_{1 \leq j \leq m^S_z} z^{n,j} = O(q^n_Z), \quad \min_{1 \leq j \leq m^S_z} z^{n,j} = O(q^n_{Z}),
\]

where \( Q^n_\tau (Q^n_{Y}) \) and \( q^n_\tau (q^n_{Y}) \) are the theoretical \((1 - \frac{1}{2n})\)-th and \( \frac{1}{2n} \)-th quantiles of \( Y_\tau (Z_\tau) \). We also fix

\[
m^S = O\left(\sqrt{n} \sqrt{\frac{Q^n_Y Q^n_Z}{\min\{q^n_Y, q^n_Z\}}}\right).
\]

These choices imply that

\[
\psi^{n,2}_r = O\left((m^S)^{-3} n^{-\frac{1}{2}}\right) \quad \text{and} \quad (\psi^{n,1}_r)^2 = O\left((m^S)^{-3} n^{-\frac{1}{2}}\right).
\]

Next, we fix the time discretization. The time discretization only impacts the computational requirements to evaluate our approximation; it does not have an effect on the rate of convergence in (24). We fix \( m^T = O(n^{1/2}) \) and \( T^{n,j} = i \frac{\tau}{m^T} \). These choices yield

\[
m^T (m^S)^3 = O\left((\psi^{n,1}_r)^{-2}\right).
\]
Proposition 5.3 implies that the asymptotic efficiency rate of the implementation is \( r = 1 \). This is the same asymptotic efficiency rate as that associated with unbiased Monte Carlo estimation. It can be shown that \( r = 1 \) is the highest asymptotic efficiency rate that can be achieved by an implementation of our algorithm that consists of equidistant discretizations of \( Y \) and \( Z \).

The implementation described above naturally satisfies Conditions (B5) and (B7)-(B9) of Theorem 4.1. Given that \( Q^n_Y, Q^n_Z \to \infty \) and \( q^n_Y, q^n_Z \to 0 \) as \( n \to \infty \), Conditions (B6) and (B10) are also satisfied. The model (16)-(22) satisfies Conditions (B1)-(B3). We verify that Assumption (B4) is satisfied by estimating \( E^*_{\theta}[1/M_r(\theta)^2] \) via Monte Carlo simulation with \( 10^7 \) Monte Carlo samples. It follows that Theorem 4.1 holds, and

\[
|E^{\tau^n}(\theta; \tau, g) - E(\theta; \tau, g)| = O\left(n^{-1} \left( \frac{Q^n_Y Q^n_Z}{\min\{q^n_Y, q^n_Z\}} \right)^{-\frac{3}{4}} \right). \tag{25}
\]

Thus, our approximation converges at a faster than linear rate to the true filter as \( n \to \infty \).

6.2 Accuracy of filter approximation

We begin by examining the convergence (25) at the parameter \( \theta_0 \) and the function \( g \equiv 1 \). We first compute a proxy of the true filter \( E(\theta_0; \tau, 1) \) via Monte Carlo simulation with \( 10^7 \) samples using an Euler discretization based on the square-root rule of Duffie & Glynn (1995). The true filter is very large (in the order of \( e^{10^4} \)). We therefore only work with the logarithm of the filter. Standard Taylor expansion tells us that the same rate of convergence as in (25) also applies for the log-filter approximation \( \log E^{\tau^n}(\theta_0; \tau, 1) \). Figure 1 confirms this intuition. The error of our log-filter approximation is large for coarse discretizations (small \( n \)), but quickly converges to zero at a faster than linear rate as the discretization becomes finer (\( n \to \infty \)).

A side effect of the high rate of convergence of our approximation is that the computational costs for evaluating our approximation quickly increase as \( n \) grows large. Figure 2 illustrates this phenomenon, and shows that the computational costs grow at a faster than linear rate when \( n \to \infty \). Nonetheless, this is intentional given that the implementation of our algorithm was chosen so as to achieve an asymptotic efficiency rate of \( r = 1 \).

We next evaluate the computational efficiency of our approximation. Figure 3 shows the relative squared error of our log-filter approximation versus the computational cost to evaluate the approximation for different values of \( n \). It also shows a fitted linear regression of the logarithm of the relative squared error of our approximation on the logarithm of the computational cost. Proposition 5.3 and the implementation choice of Section 6.1 ensure that

\[
\log R_n \approx -\log v - r \log c_n
\]
for \( r = 1 \) when \( n \) is large. A slope of \(-0.77\) of the fitted regression in Figure 3 is consistent with this result.

For comparison, Figure 3 shows the relative mean squared error of a standard Euler discretization-based approximation of the filter \( E(\theta_0; \tau, 1) \).\(^{10}\) The figure also displays the fitted linear regression of the logarithm of the relative mean squared error of the Euler discretization approximation on the logarithm of the corresponding computational costs. It is well-known that Euler discretization has a smaller asymptotic efficiency rate than the unbiased Monte Carlo estimator due to the presence of bias (see Rhee & Glynn (2015) for a recent discussion). Given that our approximation has the same asymptotic efficiency rate as the unbiased Monte Carlo estimator, we expect that the Euler discretization error decreases at a slower rate than the error of our approximation as the computational budget grows large. Figure 3 confirms this intuition. The fitted regression of the Euler discretization error has a slope of \(-0.34\), implying a much lower asymptotic efficiency rate than that of our estimator.\(^{11}\)

6.3 Accuracy of parameter estimates

Next, we examine the accuracy of our parameter estimates. We first generate 10 alternative realizations of the data \( D_\tau \) from the model \( \theta_0 \) with \( \tau = 40 \). We then maximize the approximate log-likelihood function \( \log \hat{E}\mathbb{I}_n(\theta; \tau, 1) + \log \mathcal{L}_F^*(\theta; \tau) \) for data subsamples with varying time horizons and varying time and state space grids. Since \( Y \) has the explicit solution (20), the process \( Y_t - \gamma_3 \int_0^t e^{-\gamma_4(t-u)} dN_u \) has log-normal transition densities conditional on \((N_\tau, U_\tau)\) under \( P^*_\theta \), and the corresponding factor log-likelihood is given by

\[
\log \mathcal{L}_F^*(\theta; \tau) = -\sum_{i=1}^{\lfloor T \rfloor} \left( \log \frac{Y_{s_i} - \gamma_3 \int_0^{s_i} e^{-\gamma_4(s_i-u)} dN_u}{Y_{s_{i-1}} - \gamma_3 \int_0^{s_{i-1}} e^{-\gamma_4(s_{i-1}-u)} dN_u} \right)^2 2\gamma_2^2(s_i - s_{i-1}) \\
- \sum_{i=1}^{\lfloor T \rfloor} \log \left( \frac{Y_{s_i} - \gamma_3 \int_0^{s_i} e^{-\gamma_4(s_i-u)} dN_u}{Y_{s_{i-1}} - \gamma_3 \int_0^{s_{i-1}} e^{-\gamma_4(s_{i-1}-u)} dN_u} \right) \gamma_2 \sqrt{s_i - s_{i-1}}.
\]

Proposition 4.3 guarantees the consistency of the parameter estimator that maximizes \( \log \hat{E}\mathbb{I}_n(\theta; \tau, 1) + \log \mathcal{L}_F^*(\theta; \tau) \). We can thus expect that our estimates converge to \( \theta_0 \) as \( \tau \) increases. Figure 4 plots the errors of our estimates for different values of \( \tau \), averaged over the 10 realizations of \( D_\tau \) that we have sampled. To ensure that the estimates are as accurate as possible, we initialize the numerical maximization of the approximate likelihood at the true parameters.\(^{12}\) We fix \( n = n(\tau) = 2\sqrt{T} \) in order to satisfy the conditions of Proposition 4.3. Consistent with our theoretical results, we see that our estimates

\(^{10}\) The number of time steps is chosen according to the square-root rule of Duffie & Glynn (1995).

\(^{11}\) We attribute the underperformance of the Euler estimator here to the large \( D_\tau \)-conditional variance of the Radon-Nikodym density \( 1/M_\tau(\theta_0) \) under \( P^*_\theta \).

\(^{12}\) Below we also consider alternative choices for the initial parameters.
become more accurate as more data becomes available. However, convergence to the true parameter is slow and noisy for some estimates.

We compare our estimates with the parameter estimates generated by the standard stochastic EM Algorithm of Celeux & Diebolt (1992). In the E-step we sample 250τ Euler paths of Y and Z based on their conditional law given the data Dτ, and fix the number of equidistant Euler steps according to the square-root rule of Duffie & Glynn (1995). These choices imply that the approximation of the likelihood in the E-step of the EM Algorithm is at least as accurate as our approximation of the likelihood and that both approximations take about the same time to compute, as suggested by Figure 3. They also imply that the resulting EM parameter estimators are consistent (see Nielsen (2000)). In the M-step, the complete-data partial likelihood functions of V and N are maximized separately. Since the gradient and the Hessian matrix of the complete data likelihood are available analytically, we implement the optimization using the BFGS quasi-Newton method in R. The EM Algorithm is also initialized at the true parameters. We stop it when the M-step increases the likelihood by less than 0.5; a criterion consistent with, e.g., Chan & Ledolter (1995). The resulting average errors of the parameter estimates are shown in Figure 4 by the red lines. We have considered alternative implementations, including performing a larger number of simulation trials and using a tighter stopping criterion, but found no significant differences in the results.

As for our estimates and in accordance with the theory of Nielsen (2000), the EM parameter estimates also converge to the true parameters as more data becomes available. The EM estimates for the intensity and frailty parameters tend to be more accurate than our estimates. In constrast, our estimates of the parameters of the factor Y tend to be more accurate. Despite these differences, the errors of our estimates and the EM estimates are comparable for most parameters when the time horizon is large (τ = 40).

At first sight, the EM Algorithm appears to perform at least as well if not better than our methodology in Figure 4. However, this observation may be misleading because in the above analysis we implicitly assumed that the econometrician knows the true parameters and can start the optimization routines at the true parameters. Such an assumption is naturally infeasible. A more realistic assumption is that the optimization routines are initialized at some noisy guess of the true parameters. Next, we analyze the impact of noisy initial parameters on the accuracy of our estimates. For each data sample Dτ, we randomly select an initial parameter of the form θ0(1 + ϵ) for a standard normally distributed ϵ. If that value lies outside of Θ, we take its orthogonal projection onto the boundary of Θ.13 We then initialize the maximization of our approximate log-likelihood and the EM Algorithm at the same parameters, and evaluate the parameter errors. In the E-step of the EM Algorithm we sample 10,000 Euler paths of Y and Z based on their conditional law given the data Dτ and fix the number of equidistant Euler steps according to the square-root rule of Duffie & Glynn (1995). This ensures that the approximation of

13Below we also examine an alternative, uniform sampling rule for the initial parameters.
the likelihood in the E-step is about as accurate and computationally expensive as our log-likelihood approximation with \( n = 200 \) (see Figure 3).

Figure 5 shows the errors of our estimates for \( \tau = 40 \), averaged over 10 alternative realizations of \( D_\tau \), as a function of \( n \) parametrizing the fineness and range of the state space and time discretizations. It also shows the corresponding average errors of the EM estimates, which are independent of \( n \) because we fixed the Euler discretization in the E-step to match the accuracy and computational expenses of our approximation with \( n = 200 \), the largest value of \( n \) we consider in this analysis. To provide some perspective, Figure 5 also displays the average errors of the initial parameters at which the optimization routines are initialized, and of proxies of true maximum likelihood estimates (MLE). The MLE proxies we use here are the maximizers our approximate log-likelihood with \( n = 400 \), where we initialize the optimization routine at the true parameters.

The errors of our estimates tend to become smaller as \( n \) increases, although there is noise in some of the estimates due to the finite nature of the data and the randomized initial parameters. All fitted values lie in the interior of the parameter space. Confirming the results of Proposition 4.2, the errors of our estimates of the intensity parameters \((\alpha_1, \alpha_2, \alpha_3, \alpha_4)\) and of the factor parameters \((\gamma_1, \gamma_2, \gamma_3, \gamma_4)\) are similar to the errors of the true MLE. For many of these parameters, the errors of our estimates are smaller than the errors of the corresponding EM parameters. On the other hand, the EM estimates of the frailty parameters \((\gamma_5, \gamma_6)\) are more accurate than our estimates.

A close inspection of Figure 5 reveals some subtle issues with the estimates derived from the stochastic EM Algorithm. Since the EM Algorithm maximizes the complete data partial likelihood of the frailty separately from the likelihoods of \( N \) and \( Y \), the mean and the volatility parameters of the frailty are estimated by matching the first two sample moments. Because of this, the estimation of these parameters is especially susceptible to the initial parameters. This is indicated by the location of the dotted line in Figure 5, which represents the average absolute error of the initial parameter with respect to \( \theta_0 \). The EM estimates of the frailty parameters hardly deviate from the respective initial values, consistent with the fact that the EM estimator is guaranteed to converge to a local optimum of the likelihood only (see Wu (1983) and Dembo & Zeitouni (1986)). This may be an issue in practice when \( \theta_0 \) is unknown. We have investigated this further by analyzing the estimates obtained from initial parameter values sampled from a uniform distribution on the parameter space \( \Theta \). Unlike our method, the EM Algorithm was found to converge extremely slowly when the initial parameters were very different from \( \theta_0 \), often resulting in inappropriate estimates that lie on the boundary of the parameter space. Boundary optima are an issue for conventional asymptotic results, which apply only to optima in the interior of the parameter space.

Another issue of the separate maximization of the complete data partial likelihood functions of \( N \) and \( V \) in the EM Algorithm is that the interaction of the events and the factor is ignored. As a result, the EM Algorithm faces difficulties in differentiating between diffusive movements and jumps of \( Y \). It tends to converge to boundary solutions
that imply an event sensitivity $\gamma_3$ of $Y$ equal or close to zero, a decay rate $\gamma_4$ equal or close to 20, and too high estimates of the mean rate $\gamma_1$ and the volatility $\gamma_2$. In order to compensate for these biases, the EM Algorithm tends to overstate the role of the factor $Y$ and understate the role of the frailty $Z$ in the intensity. This interpretation is consistent with the large errors of the EM estimates for the factor and frailty sensitivities of the intensity in Figure 5. In these cases, the errors are larger than the errors of the initial values and of true MLE.

In order to shed more light on these issues, we analyze how the two methods address the filtering problem. We compare the fitted filtered factor $E_{\theta}[Y_t|D_t]$ to the corresponding monthly observations, and the fitted filtered intensity $h_t = E_{\theta}[\lambda_t|D_t]$ to the actual number of events per year for $0 \leq t \leq \tau = 40$. For a given realization of the data, Figure 6 shows the approximations to these filters, given by $E_{I_n}(\theta; t, g) = (1, 0, 0, 0, 0) \cdot v$ and $g(v, t) = \Lambda(v; \theta)$, respectively, evaluated at the corresponding estimator $\theta = \hat{\theta}_\tau$. For that realization, Figure 6 also shows Monte Carlo approximations to these filters, obtained using 10,000 trials and evaluated at the EM estimator of $\theta$. The confidence bands were calculated using subsampling with group size 100.

As expected, in the case of our method the chosen value of $n$ has a large impact on the accuracy of the filter. In accordance with Theorem 4.1, the larger $n$, the finer and more extensive the state space grid, and the closer the filters track the observed values of the factor and the event counts. If the state space discretization is too coarse, then the approximation to the filtered factor and filtered intensity fluctuate only between few values. As $n$ increases, the filtered paths track the observations better. The Monte Carlo approximation to the factor filter, which is based on EM parameter estimates, tracks the factor well. Nevertheless, the corresponding Monte Carlo approximation to the filtered intensity $h$ severely understates the actual event counts. We have found similar issues for all other realizations of the data $D_{\tau}$ that we have considered.

We examine the implications of the estimation errors for model fit. Proposition D.1 of Azizpour et al. (2015) implies that, after a change of time given by the $P_{\theta}$-compensator $\int_0^t h_s ds$ relative to the right-continuous and complete filtration generated by $D_t$, the counting process $N$ is a standard $P$-Poisson process in the time-changed filtration. We test whether the fitted filtered intensity $h$ time-scales the realized $N$ correctly. Figure 7 provides a QQ-plot of the time-changed inter-arrival times for each of several fitting methods (95% confidence bands of the order statistics of the exponential distribution are included). We see that, as $n$ increases, the model estimated using our scheme fits the event data better. Already with $n = 200$ we obtain an acceptable fit that is comparable to that achieved by the true MLE. The model estimated using the EM Algorithm, for which the filtered intensity $h$ is approximated by Monte Carlo simulation, fails the test.

Overall, the results of this section suggest that the parameter estimators derived from our log-likelihood approximation outperform estimators derived from the EM Algorithm for model (16)-(22). This result holds even when matching the computational effort and the accuracy of our log-likelihood approximation with that of the likelihood approximation.
constructed in the E-step of the EM Algorithm.

### 6.4 Asymptotic properties of parameter estimates

To conclude we examine the standard errors of our parameter estimators. In the setting of the previous analyses, we calculate the approximate asymptotic variance-covariance matrix \( \hat{\Sigma}_0 \) of our estimators according to (31)-(33) at the true parameter vector \( \theta_0 \). In order to satisfy the conditions of Proposition 4.3, for the implementation we let \( n \) vary with \( \tau \) as in \( n = n(\tau) = 2\sqrt{\tau} \).

Figure 8 shows the asymptotic standard errors for each parameter, for a given realization of \( D_\tau \). As \( \tau \) increases and more data is included in the estimation, the errors decrease for most parameters, albeit slowly in some cases. The errors behave similarly for the other realizations of \( D_\tau \) we have considered (the results for these are available upon request). The standard errors implied by our likelihood approximation are small when compared to the true parameter values, especially when the time horizon is large.

For comparison, Figure 8 also indicates the asymptotic standard errors of the EM estimates. These are calculated using the following approximation of the theoretical maximum likelihood asymptotic variance-covariance matrix:

\[
\frac{\zeta^{EM}_{\tau}}{\tau} \left( \nabla^2 Q(\theta, \theta_0) |_{\theta = \theta_0} \right)^{-1} \nabla Q(\theta, \theta_0) |_{\theta = \theta_0} \nabla Q(\theta, \theta_0) |_{\theta = \theta_0}^\top \left( \nabla^2 Q(\theta, \theta_0) |_{\theta = \theta_0} \right)^{-1}
\]

where \( Q(\theta, \theta_0) \) is the complete data likelihood evaluated at \( \theta \) through simulations with respect to the conditional \( \mathbb{P}_{\theta_0} \)-distribution given \( D_\tau \), and \( \zeta^{EM}_{\tau} \) is an approximation of the fraction of missing data as in Nielsen (2000). In order to achieve convergence to the true asymptotic distribution, in the E-step of the EM Algorithm we sample 250\( \tau \) paths of \( Y \) and \( Z \) based on their conditional law given the data \( D_\tau \) and fix the number of Euler steps per path according to the square-root rule of Duffie & Glynn (1995).

For most parameters and observation periods, the standard errors of the EM estimators are larger than the standard errors of our estimators. Only for the factor mean rate and the frailty parameters we find that the standard errors of the EM estimators and of our estimators are comparable. For many of the intensity parameters, the standard errors of the EM estimators are excessively large when compared to the true parameter values.

The large standard errors of the EM estimators provide an explanation for the issues highlighted in Section 6.3. It appears that the likelihood approximated by the stochastic EM Algorithm is flat and noisy, making identification difficult. This yields inaccurate estimators for model (16)-(22), and may lead to misguided conclusions in empirical applications. Our parameter estimators do not seem to suffer under this issue. Indeed, in many cases, the loss in efficiency due to simulation in the stochastic EM Algorithm appears to exceed the loss of efficiency due to filtering in our method. The loss of efficiency due to filtering, indicated in Proposition A.2, appears to be small in the cases considered. This renders our estimates highly efficient.
7 Conclusion

We develop and study likelihood estimators of the parameters of a marked point process and of incompletely observed explanatory factors that influence the arrival intensity and mark distribution. We establish a quadrature-based approximation to the likelihood, and show that our approximation converges to the true likelihood uniformly in the parameter space under mild conditions. We also show that the parameter estimators that maximize our approximate likelihood converge to true maximum likelihood estimators, and that they inherit the same asymptotic properties of true maximum likelihood estimators. These results provide a rigorous statistical foundation for empirical applications of event timing models in the presence of incomplete factor data.

A numerical analysis highlights our theoretical results. We find that our likelihood approximation is highly accurate and computationally efficient. It has a higher rate of convergence than a standard Euler discretization-based approximation. Turning to parameter inference, we find that the parameter estimators that result from maximizing our approximate likelihood are also highly accurate and statistically efficient. They outperform estimators derived from the stochastic EM Algorithm, a standard alternative for parameter inference in incomplete data cases as those considered in this paper.

A Exact likelihood

This appendix complements Section 3. It provides a rigorous definition of the exact likelihood $L_{\tau}(\theta)$ of the observed data $D_{\tau}$ and the conditional likelihood $L^*_{\tau}(\theta; \tau)$ of the factor data. It also derives conditions that ensure consistency and asymptotic normality of an exact maximum likelihood estimator $\hat{\theta}_{\tau}$, and introduces notation that will be used throughout the proofs given in Appendix C.

A.1 Likelihood

Let $R_{D,\tau}$ denote the range of the date $D_\tau$ and $\mathcal{D}_\tau$ the Borel $\sigma$-algebra of $R_{D,\tau}$. We have $R_{D,\tau} = D([0, \tau]; \mathbb{N}) \times D([0, \tau]; R_U) \times D([0, \tau]; \mathbb{R}^d) \times D_{Y,\tau}$, where $D(A; B)$ is the Skorokhod space of càdlàg functions mapping the set $A$ onto $B$, and $R_{Y,\tau} = \{(y_{i,j} : y_{i,j} \in \mathbb{R} \text{ for } i = 1, \ldots, d_Y, j = 1, \ldots, |S_i \cap [0, \tau]|)\}$ is the range of the data $Y_\tau$ of the partially observed factor element. Write $D_{\theta,\tau}$ for the $\mathbb{P}_{\theta}$-law of the data $D_\tau$ on $(R_{D,\tau}, \mathcal{D}_\tau)$. The likelihood of the data is given by the Radon-Nikodym derivative

$$L_{\tau}(\theta) = \frac{dD_{\theta,\tau}}{dD_{\theta_0,\tau}},$$

evaluated at the observed data $D_{\tau}$.

We now give a precise definition of the conditional likelihood $L^*_{\tau}(\theta; \tau)$ appearing in Theorem 3.1. Let $D^*_{\theta,\tau}$ denote the $\mathbb{P}^*$-law of the data $D_\tau$ on $(R_{D,\tau}, \mathcal{D}_\tau)$. Write $D^*_{E,\theta,\tau}$ for
the \( \mathbb{P}^\ast \)-law of the event data \((N_\tau, U_\tau)\) on \((D([0, \tau]; N) \times D([0, \tau]; R_U)), \sigma(D([0, \tau]; N) \times D([0, \tau]; R_U)))\), and \(D^\ast_{F,\theta,\tau}\) for the conditional \(\mathbb{P}^\ast_{\theta}\)-law of the factor data \((X_\tau, Y_\tau)\) given the event data \((N_\tau, U_\tau)\) on \((D([0, \tau]; R^d X) \times R^d Y, \sigma(D([0, \tau]; R^d X) \times R^d Y, \sigma))\). The conditional likelihood \(L^\ast_F(\theta; \tau)\) of the factor data \((X_\tau, Y_\tau)\) given the event data \((N_\tau, U_\tau)\) under \(\mathbb{P}^\ast\) is the Radon-Nikodym derivative

\[
L^\ast_F(\theta; \tau) = \frac{dD^\ast_{F,\theta,\tau}}{dD^\ast_{F,\theta_0,\tau}},
\]

evaluated at the observed data.

### A.2 Asymptotic properties of exact likelihood estimators

The representation of the likelihood obtained in Theorem 3.1 facilitates the analysis of the asymptotic properties of an exact MLE \(\hat{\theta}_\tau\) as \(\tau \to \infty\).

**Proposition A.1** (Consistency). Suppose Assumptions (A1)-(A2) hold. In addition, suppose that:

(A3) For any \(\theta \in \Theta\), it holds \(\mathbb{P}_\theta\)-almost surely that \(L^\ast_F(\theta; \tau) > 0\).

(A4) The true parameter \(\theta_0\) is the unique maximizer of the asymptotic likelihood, i.e.,

\[
\arg \sup_{\theta \in \Theta} \lim_{\tau \to \infty} \frac{1}{\tau} L_\tau(\theta) = \{\theta_0\}.
\]

Then every \(\hat{\theta}_\tau\) solving (8) is consistent: \(\hat{\theta}_\tau \to \theta_0\) as \(\tau \to \infty\) in \(\mathbb{P}_{\theta_0}\)-probability.

Consistency follows from Lemma 4.1 in Section 1.4.3 of Ibragimov & Has’minskii (1981) after establishing that the filtering in (5) guarantees the smoothness of the likelihood function.

Condition (A4) is a standard identifiability condition. Condition (A3) is mild and implies both the equivalence of the laws of the data \(D_\tau\) under \(\mathbb{P}_\theta\) and \(\mathbb{P}_{\bar{\theta}}\) for any \(\theta, \bar{\theta} \in \Theta\), as well as the stochastic equicontinuity of the likelihood function. It is a restriction on the parameter space \(\Theta\), since it requires that the likelihood of the observed factor data be strictly positive at any parameter \(\theta \in \Theta\). Thus, consistency is guaranteed as long as the parameter space is chosen adequately.

We now consider asymptotic normality. We make some assumptions to simplify the exposition. Suppose that \(S^i = S^j\) and \(0 \in S^i\) for all \(1 \leq i, j \leq d_Y\). Hence, all partially observed factors are observed on the same dates, including at time 0. Suppose \(S_\tau = S \cap [0, \tau]\) takes the form \(\{s_0, s_1, \ldots, s_{m(\tau)}\}\) for \(0 = s_0 < \cdots < s_{m(\tau)} = \tau\). Suppose the limit \(\zeta = \lim_{\tau \to \infty} m(\tau)/\tau\) exists and is strictly positive. Define the random functions

\[
\theta \in \Theta \mapsto L^\ast_{E,i}(\theta) = L^\ast_E(\theta; s_i)/L^\ast_E(\theta; s_{i-1})
\]

(27)
\[ \theta \in \Theta \implies L_{F,i}^*(\theta) = L_F^*(\theta; s_i) / L_F^*(\theta; s_{i-1}) \]

for \( i = 1, \ldots, m(\tau) \), as well as
\[ \theta \in \Theta \implies L_{E,0}^*(\theta) = 1 \]
\[ \theta \in \Theta \implies L_{F,0}^*(\theta) = L_F^*(\theta; 0). \]

It follows that \( L_E^*(\theta; \tau) = \prod_{i=0}^{m(\tau)} L_{E,i}^*(\theta) \) and \( L_F^*(\theta; \tau) = \prod_{i=0}^{m(\tau)} L_{F,i}^*(\theta) \). Write \( \theta = (\theta_1, \ldots, \theta_p) \) for some \( p \in \mathbb{N} \), \( \partial_v = \frac{\partial}{\partial \theta_v} \), \( \partial_{v,w} = \frac{\partial^2}{\partial \theta_v \partial \theta_w} \), and \( \partial_{u,v,w} = \frac{\partial^3}{\partial \theta_u \partial \theta_v \partial \theta_w} \) for any \( 1 \leq u, v, w \leq p \), and let \( \nabla, \nabla^2 \) denote the gradient and the Hessian matrix, respectively, with respect to \( \theta \).

**Proposition A.2** (Asymptotic normality). In addition to (A1)-(A4), suppose that the following conditions hold:

(A5) The functions \( L_{E,i}^*(\theta) \) and \( L_{F,i}^*(\theta) \) for \( i = 0, \ldots, m(\tau) \) are three times continuously differentiable at \( \theta \), \( \mathbb{P}_\theta \)-almost surely, for any \( \theta \in \Theta^0 \).

(A6) For any \( \theta \in \Theta^0 \), \( 1 \leq u, v \leq p \), and \( 0 \leq i \leq m(\tau) \), it holds \( \mathbb{P}_\theta \)-a.s. that
\[
0 = \mathbb{E}_\theta \left[ \partial_u \log L_{E,i}^*(\theta) \middle| D_{s_{i-1}} \right] = \mathbb{E}_\theta \left[ \partial_u \log L_{F,i}^*(\theta) \middle| D_{s_{i-1}} \right],
\]
\[
0 = \mathbb{E}_\theta \left[ \frac{\partial^2_{u,v} L_{E,i}^*(\theta)}{L_{E,i}^*(\theta)} \middle| D_{s_{i-1}} \right] = \mathbb{E}_\theta \left[ \frac{\partial^2_{u,v} L_{F,i}^*(\theta)}{L_{F,i}^*(\theta)} \middle| D_{s_{i-1}} \right].
\]

(A7) There exist positive constants \( K_1, K_2, K_3, \) and \( K_4 \) such that for all \( 1 \leq u, v \leq p \) and \( 0 \leq i \leq m(\tau) \)
\[
\mathbb{E}_\theta \left[ (\partial_u \log L_{E,i}^*(\theta))^2 \right] \leq K_1, \quad \mathbb{E}_\theta \left[ \left( \frac{\partial_{u,v} L_{E,i}^*(\theta)}{L_{E,i}^*(\theta)} \right)^2 \right] \leq K_2,
\]
\[
\mathbb{E}_\theta \left[ (\partial_u \log L_{F,i}^*(\theta))^2 \right] \leq K_3, \quad \mathbb{E}_\theta \left[ \left( \frac{\partial_{u,v} L_{F,i}^*(\theta)}{L_{F,i}^*(\theta)} \right)^2 \right] \leq K_4.
\]

(A8) For any \( \theta \in \Theta^0 \), the following limits of matrices exist in \( \mathbb{P}_\theta \)-probability and are deterministic:
\[
\Sigma_E(\theta) = \lim_{m \to \infty} \frac{1}{m} \sum_{i=0}^{m} \nabla \log L_{E,i}^*(\theta) \nabla \log L_{E,i}^*(\theta) \in \mathbb{R}^{p \times p},
\]
\[
\Sigma_F(\theta) = \lim_{m \to \infty} \frac{1}{m} \sum_{i=0}^{m} \nabla \log L_{F,i}^*(\theta) \nabla \log L_{F,i}^*(\theta) \in \mathbb{R}^{p \times p},
\]
\[
\Sigma_{EF}(\theta) = \lim_{m \to \infty} \frac{1}{m} \sum_{i=0}^{m} \nabla \log L_{E,i}^*(\theta) \nabla \log L_{F,i}^*(\theta) \in \mathbb{R}^{p \times p}.
\]
(A9) For any $\theta \in \Theta^\circ$, the matrix $\Sigma_E(\theta) + \Sigma_F(\theta)$ is non-singular.

(A10) For any $\theta \in \Theta^\circ$, there exists a neighborhood $U(\theta)$ of $\theta$ and finite constants $G^E = G^E(\theta)$, $G^F = G^F(\theta)$, such that for any $\tilde{\theta} \in U(\theta)$ and $1 \leq u, v, w \leq p$

$$\lim_{m \to \infty} \mathbb{P}_\theta \left[ \frac{1}{m} \sum_{i=0}^{m} \partial_{u,v,w} \log \mathcal{L}_{E,i}^*(\tilde{\theta}) < G^E \right] = 1,$$

$$\lim_{m \to \infty} \mathbb{P}_\theta \left[ \frac{1}{m} \sum_{i=0}^{m} \partial_{u,v,w} \log \mathcal{L}_{F,i}^*(\tilde{\theta}) < G^F \right] = 1.$$

Then every $\hat{\theta}_\tau$ solving (8) is asymptotically normal: $\sqrt{\tau}(\hat{\theta}_\tau - \theta_0) \to \mathcal{N}(0, \Sigma_0)$ as $\tau \to \infty$ in $\mathbb{P}_{\theta_0}$-distribution. Letting $\Sigma(\theta_0) = \Sigma_E(\theta_0) + \Sigma_F(\theta_0)$, the asymptotic variance-covariance matrix is given by

$$\Sigma_0 = \frac{1}{\zeta} \Sigma(\theta_0)^{-1} + \frac{2}{\zeta} \Sigma(\theta_0)^{-1} \Sigma_{EF}(\theta_0) \Sigma(\theta_0)^{-1}.$$

Proposition A.2 is a martingale CLT for the score function. Because of the incompleteness of the factor data, the analysis is not completely standard: in our setting, the likelihood ratios are locally asymptotically quadratic but not locally asymptotically mixed normal.\textsuperscript{14} Thus, the conventional martingale CLTs of Feigin (1976), Jeganathan (1995), van der Vaart (2000) and others cannot be directly applied in our setting. Our asymptotic analysis uses an argument similar to the one behind Proposition 6.1 of Hall & Heyde (1980), after controlling for the incompleteness of the factor data.

Condition (A5) is standard. Condition (A7) is also standard; it imposes regularity on the intensity (1) and the $\mathbb{P}_\theta$-law of the factor $V$. Conditions (A8) and (A9) are necessary to obtain well-defined covariance matrices. Condition (A10) guarantees that a Taylor expansion of the score function is sufficiently accurate. Condition (A6) is an identifiability condition. It implies that $\theta_0$ maximizes the asymptotic likelihood.

Proposition A.2 highlights the influence of the partially observed data on the asymptotic variance-covariance matrix $\Sigma_0$: it is inversely proportional to $\zeta$, the average number of observations of $Y$ per unit of time. We compare with the results of Ogata (1978), who studied likelihood estimators for stationary and ergodic point processes with complete factor data. Consider the Fisher information matrix $I(\theta_0)$ for the estimation of $\theta_0 = (\alpha_0, \beta_0, \gamma_0)$, given in our setting by

$$I(\theta_0) = \zeta (\Sigma_E(\theta_0) + \Sigma_F(\theta_0)).$$

Under the assumptions of stationarity and ergodicity as well as complete factor data, Ogata (1978) shows that $\Sigma_0 = I(\theta_0)^{-1}$. We relax these assumptions and allow for incomplete factor data. We find that

$$\Sigma_0 = I(\theta_0)^{-1} + 2\zeta I(\theta_0)^{-1} \Sigma_{EF}(\theta_0) I(\theta_0)^{-1}.$$

\textsuperscript{14}See Definitions 1 and 3 of Jeganathan (1995).
The matrix $\Sigma_{EF}(\theta_0)$ corrects for the filtering of the unobserved path of the factors. If $V$ is observed perfectly, then one can show that $\Sigma_{EF}(\theta_0) = 0$, producing the result of Ogata (1978).

We end this section with alternative and more traditional representations of the matrices $\Sigma_E(\theta_0), \Sigma_F(\theta_0)$ and $\Sigma_{EF}(\theta_0)$.

**Corollary A.3.** Suppose the assumptions of Proposition A.2 hold. For any $\theta \in \Theta^0$, it holds $P_\theta$-almost surely that
\[
\zeta \Sigma_E(\theta) = - \lim_{\tau \to \infty} \frac{1}{\tau} \nabla^2 \log \mathcal{L}_E^*(\theta; \tau) = \lim_{\tau \to \infty} \frac{1}{\tau} \nabla \log \mathcal{L}_E^*(\theta; \tau) \nabla \log \mathcal{L}_E^*(\theta; \tau),
\]
\[
\zeta \Sigma_F(\theta) = - \lim_{\tau \to \infty} \frac{1}{\tau} \nabla^2 \log \mathcal{L}_F^*(\theta; \tau) = \lim_{\tau \to \infty} \frac{1}{\tau} \nabla \log \mathcal{L}_F^*(\theta; \tau) \nabla \log \mathcal{L}_F^*(\theta; \tau),
\]
\[
\zeta \Sigma_{EF}(\theta) = \lim_{\tau \to \infty} \frac{1}{\tau} \nabla \log \mathcal{L}_E^*(\theta; \tau) \nabla \log \mathcal{L}_F^*(\theta; \tau).
\]

**B Asymptotic variance-covariance matrix**

Proposition 4.3 states that the approximate MLE has the same asymptotic variance-covariance matrix $\Sigma_0$ as the exact MLE analyzed in Appendix A. Thus, there is no loss of efficiency when maximizing the approximate likelihood rather than the exact likelihood. From Proposition A.2 we know that $\Sigma_0 = \frac{1}{\zeta} \Sigma(\theta_0)^{-1} + \frac{2}{\zeta} \Sigma(\theta_0)^{-1} \Sigma_{EF}(\theta_0) \Sigma(\theta_0)^{-1}$ for $\Sigma(\theta_0) = \Sigma_E(\theta_0) + \Sigma_F(\theta_0)$. For finite $\tau$, we propose to approximate $\Sigma_0$ as follows. We approximate $\zeta$ by $\hat{\zeta} = \frac{m(\tau)}{\tau}$. For given $n$, we approximate $\Sigma_E(\theta_0), \Sigma_F(\theta_0)$, and $\Sigma_{EF}(\theta_0)$ by
\[
\hat{\Sigma}_E = \frac{1}{m(\tau)} \sum_{i=1}^{m(\tau)} \nabla \log \frac{E^{T^n}(\hat{\theta}_r^n; s_i, 1)}{E^{T^n}(\hat{\theta}_r^n; s_{i-1}, 1)} \nabla \log \frac{E^{T^n}(\hat{\theta}_r^n; s_i, 1)}{E^{T^n}(\hat{\theta}_r^n; s_{i-1}, 1)}
\]
\[
\hat{\Sigma}_F = \frac{1}{m(\tau)} \sum_{i=0}^{m(\tau)} \nabla \log \frac{L_{F,i}^*(\theta)}{L_{F,i}^*(\theta)} \nabla \log \frac{L_{F,i}^*(\theta)}{L_{F,i}^*(\theta)}
\]
\[
\hat{\Sigma}_{EF} = \frac{1}{m(\tau)} \sum_{i=1}^{m(\tau)} \nabla \log \frac{E^{T^n}(\hat{\theta}_r^n; s_i, 1)}{E^{T^n}(\hat{\theta}_r^n; s_{i-1}, 1)} \nabla \log \frac{E^{T^n}(\hat{\theta}_r^n; s_i, 1)}{E^{T^n}(\hat{\theta}_r^n; s_{i-1}, 1)}
\]
respectively. The implementation $T^n$ is chosen so as to achieve some given degree of accuracy in accordance with Theorem 4.1. Finally, we set $\hat{\Sigma} = \hat{\Sigma}_E + \hat{\Sigma}_F$ and approximate $\Sigma_0$ by $\hat{\Sigma}_0 = \frac{1}{\hat{\zeta}} \hat{\Sigma}_E^{-1} + \frac{2}{\hat{\zeta}} \hat{\Sigma}_F^{-1} \hat{\Sigma}_{EF}^{-1}$.

**C Proofs**

*Proof of Theorem 3.1.* The measure $d\mathbb{P}_\theta^* = M_r(\theta) d\mathbb{P}_\theta$ is well-defined by Condition (A1) and Theorem VIII.T10 of Brémaud (1980). We have
\[
\mathbb{D}_{\theta, \tau}(A) = \mathbb{E}_{\theta}[1_A] = \mathbb{E}_{\theta}^*[1_A \mathbb{E}_{\theta}^*[1/M_r(\theta) | D_\tau]] = \int_A \mathbb{E}_{\theta}^*[1/M_r(\theta) | D_\tau] d\mathbb{D}_{\theta, \tau}^*.
\]
for $A \in \mathcal{D}_\tau$. Consequently, the law $\mathbb{D}_{\theta,\tau}$ is absolutely continuous with respect to $\mathbb{D}_{\theta,0,\tau}$ with Radon-Nikodym derivative

$$\frac{d\mathbb{D}_{\theta,\tau}}{d\mathbb{D}_{\theta,0,\tau}} = \mathbb{E}_{\theta}^{*} \left[ \frac{1}{M_{\tau}(\theta)} \right] \mathbb{D}_{\tau}. \quad (34)$$

Bayes’ formula implies that the $\mathbb{P}_{\theta}^{*}$-law of the data is equal to the product of the $\mathbb{P}_{\theta}^{*}$-law of the event data $(N_{\tau}, U_{\tau})$ and the conditional $\mathbb{P}_{\theta}^{*}$-law of the factor data $(X, Y_{\tau})$ given $(N_{\tau}, U_{\tau})$:

$$\mathbb{D}_{\theta,\tau}^{*} = \mathbb{D}_{F,\theta,\tau}^{*} \times \mathbb{D}_{E,\theta,\tau}^{*}. \quad (35)$$

Now, the $\mathbb{P}_{\theta}^{*}$-law of the event data $(N_{\tau}, U_{\tau})$ is parameter independent by construction since the counting process $N$ has unit intensity and the event marks have parameter-independent density $\pi^{*}$ under $\mathbb{P}_{\theta}^{*}$. Hence,

$$\frac{d\mathbb{D}_{E,\theta,\tau}^{*}}{d\mathbb{D}_{E,\theta,0,\tau}^{*}} \propto 1. \quad (36)$$

In addition,

$$\frac{d\mathbb{D}_{F,\theta,\tau}^{*}}{d\mathbb{D}_{F,\theta,0,\tau}^{*}} = \mathcal{L}_{F}^{*}(\theta; \tau) \quad (37)$$

by Assumption (A2) and the definition of the conditional likelihood $\mathcal{L}_{F}^{*}(\theta; \tau)$ in (26). Because $\mathbb{D}_{\theta}^{*}$ is a product law as in (35), Fubini’s theorem together with (36)-(37) imply that

$$\frac{d\mathbb{D}_{\theta,\tau}^{*}}{d\mathbb{D}_{\theta,0,\tau}^{*}} = \frac{d\mathbb{D}_{F,\theta,\tau}^{*} \times \mathbb{D}_{E,\theta,\tau}^{*}}{d\mathbb{D}_{F,\theta,0,\tau}^{*} \times \mathbb{D}_{E,\theta,0,\tau}^{*}} = \frac{d\mathbb{D}_{F,\theta,\tau}^{*}}{d\mathbb{D}_{F,\theta,0,\tau}^{*}} \frac{d\mathbb{D}_{E,\theta,\tau}^{*}}{d\mathbb{D}_{E,\theta,0,\tau}^{*}} \propto \mathcal{L}_{F}^{*}(\theta; \tau)$$

We conclude that the likelihood satisfies

$$\mathcal{L}_{\tau}(\theta) = \frac{d\mathbb{D}_{\theta,\tau}}{d\mathbb{D}_{\theta,0,\tau}} = \frac{d\mathbb{D}_{F,\theta,\tau}}{d\mathbb{D}_{F,\theta,0,\tau}} \frac{d\mathbb{D}_{E,\theta,\tau}}{d\mathbb{D}_{E,\theta,0,\tau}} \propto \mathbb{E}_{\theta}^{*} \left[ \frac{1}{M_{\tau}(\theta)} \right] \mathbb{D}_{\tau} \mathcal{L}_{F}^{*}(\theta; \tau). \quad (38)$$

This completes the proof. \qed

**Proof of Theorem 4.1.** For simplicity, we only consider the case that $g \equiv 1$. Since $\Theta$ is compact by Assumption (B3), the reasoning below remains valid whenever the function $g$ is either continuous or bounded.

Define $C_{1}^{\tau} = \max_{\theta \in \Theta} \mathbb{E}_{\theta}^{*} \left[ \frac{1}{M_{\tau}(\theta)} \right] \mathbb{D}_{\tau}$ and $C_{2}^{\tau} = \max_{\theta \in \Theta} \mathbb{E}_{\theta}^{*} \left[ \frac{1}{M_{\tau}(\theta)} \right] \mathbb{D}_{\tau}$, which are finite by Assumptions (B3) and (B4). Write

$$\Delta_{\tau}(\theta) = |E^{\mathcal{T}_{\tau}^{n}}(\theta; \tau, 1) - \mathcal{E}(\theta; \tau, 1)|$$

31
for the absolute error generated by the approximation at the parameter \( \theta \). By Jensen’s inequality, we can bound \( \Delta_r(\theta) \) from above by

\[
E^*_{\theta} \left[ \frac{1}{M_r(\theta)} \right] ^{1_{\{3 \in [0, \tau] \forall \epsilon \in A\}} } D_r

+ \sum_{i_1=1}^{m^2_1} \ldots \sum_{i_{m^2_{\tau}}=1}^{m^2_{\tau}} E^*_{\theta} \left[ F_{\theta}^n(i_1, \ldots, i_{m^2_{\tau}}) - \frac{1}{M_r(\theta)} I_{i_1,1, \ldots, i_{m^2_{\tau}}}, D_r \right]
\]  
(39)

for \( I_{i_1,1, \ldots, i_{m^2_{\tau}}} = 1_{\{V_0 \in A^{i_0}, \forall j=1, \ldots, m^2_{\tau}, \forall t \in \{T^{n,j} \in \Theta \} : V_0 \in A^{i_0,j} \} \} \). The second summand in the right-hand side of (39) can be bounded above by

\[ C^\tau_1 \left( \exp \left( \psi^{n,1}_\tau \right) - 1 \right) \mathbb{P}^*_{\theta} \left[ \forall t \in [0, \tau] : V_t \in A^n, D_r \right] \leq C^\tau_1 \left( \exp \left( \psi^{n,1}_\tau \right) - 1 \right), \]

where

\[
\psi^{n,1}_\tau = \max_{1 \leq j \leq m^2_\tau} \text{vol} \left( A^{j,n} \right) \left[ \tau \max_{\theta \in \Theta} \max_{1 \leq j \leq m^2_\tau} \left| \partial_v \Lambda(v^j; \theta) \right| \right.

\]

\[ + N_r \max_{\theta \in \Theta} \max_{1 \leq k \leq N_r} \max_{1 \leq j \leq m^2_\tau} \left| \partial_v \log \Lambda(v^j; \theta) \pi(k, v^j; \theta) \pi^*(U_k) \right| \]

as in Assumption (B9). This follows from Taylor expansion around the points \( v^j \) and Jensen’s inequality, together with Assumptions (B1), (B2), and (B5)-(B8). We use Hölder’s inequality with Assumptions (B4) and (B10) for the first term on the right-hand side of (39) to obtain the bound \( \sqrt{C^\tau_2} \sqrt{\psi^{n,1}_\tau} \) for

\[ \psi^{n,2}_\tau = \sup_{\theta \in \Theta} \mathbb{P}^*_{\theta} \left[ \exists t \in [0, \tau] : V_t \not\in A^n, D_r \right] . \]

Thus, we have that \( \left| E^{\tau,n}(\theta; \tau, 1) - \mathcal{E}(\theta; \tau, 1) \right| \) is bounded above by

\[ \left( \exp \left( \psi^{n,1}_\tau \right) - 1 \right) C^\tau_1 + \sqrt{C^\tau_2} \sqrt{\psi^{n,2}_\tau}. \]
(40)

Note that (40) does not depend on \( \theta \). As a result, the error bound in (40) holds uniformly over the parameter space \( \Theta \). Assumptions (B3), (B4), (B10), and (B9) then lead to the claim for \( n \to \infty \).

**Proof of Proposition 4.2.** Take \( \epsilon > 0 \). By Theorem 4.1 there exists some \( n_0 \in \mathbb{N} \), depending only on \( \tau \), such that

\[ \left| E^{\tau,n}(\theta; \tau, 1) - \mathcal{E}(\theta; \tau, 1) \right| \leq \epsilon \]

for all \( n \geq n_0 \) and all \( \theta \in \Theta \). Since \( \max_{\theta \in \Theta} \mathcal{L}^*_{\tau}(\theta; \tau) < \infty \), \( \mathbb{P} \)-almost surely, it follows that

\[ \left| \mathcal{L}^*_{\tau}(\theta) - \mathcal{L}^*_{\tau}(\theta) \right| \leq \epsilon \max_{\theta \in \Theta} \mathcal{L}^*_{\tau}(\theta; \tau) < \infty , \]

\( \mathbb{P} \)-almost surely. Define \( \tilde{\epsilon} = \epsilon \max_{\theta \in \Theta} \mathcal{L}^*_{\tau}(\theta; \tau) \). Then, since \( \hat{\theta}^n_{\tau} \) is an optimizer of \( \mathcal{L}^*_{\tau} \),

\[ \hat{\theta}^n_{\tau} \in \{ \theta \in \Theta : \mathcal{L}^*_{\tau}(\hat{\theta}^n_{\tau}) - \tilde{\epsilon} \leq \mathcal{L}^*_{\tau}(\theta) \leq \mathcal{L}^*_{\tau}(\hat{\theta}^n_{\tau}) + \tilde{\epsilon} \}, \]
(41)

\( \mathbb{P}_g \)-almost surely. The claim follows for \( \epsilon \to 0 \). □

32
Proof of Proposition 4.3. For fixed $\tau > 0$, we deduce from Proposition 4.2 that there exists the limit of $\tilde{\theta}_n^\tau$ for $n \to \infty$. Let $\tilde{\theta}_n^\infty$ be this limit. Then we know that $L_\tau(\tilde{\theta}_n^\infty) = L_\tau(\hat{\theta}_n)$ for $\hat{\theta}_n = \arg \max_{\theta \in \Theta} L_\tau(\theta)$. Since $\hat{\theta}_n$ is a global maximizer and no optimum is attained on the boundary of the parameter space, necessarily $\hat{\theta}_n^\infty$ satisfies the first-order condition

$$\nabla L_\tau(\hat{\theta}_n^\infty) = 0. \quad (42)$$

Thus, we can derive a linear-quadratic Taylor expansion of $L_\tau(\hat{\theta}_n^\infty)$ around $\theta_0$ as in Proposition A.2, and conclude that

$$\sqrt{\tau} \left( \hat{\theta}_n^\infty - \theta_0 \right) \to N(0, \Sigma_0). \quad (43)$$

From the proof of Proposition 4.2 we know that for the mapping $\tau \mapsto n(\tau)$ as defined in (14) it holds that

$$\left| L_\tau^{n(\tau)}(\tilde{\theta}_n^{n(\tau)}) - L_\tau(\tilde{\theta}_n^\infty) \right| = O(\tau^{-q}).$$

Taylor expansion tells us that

$$L_\tau(\tilde{\theta}_n^\infty) - L_\tau^{n(\tau)}(\tilde{\theta}_n^{n(\tau)}) = \nabla^2 L_\tau^{n(\tau)}(\tilde{\theta})(\tilde{\theta} - \tilde{\theta}_n^\tau)$$

for $\tilde{\theta}$ in a neighborhood of $\tilde{\theta}_n^{n(\tau)}$. Assumption (B13) implies that

$$\left\| \left( \nabla^2 L_\tau^{n(\tau)}(\tilde{\theta}) \right)^{-1} \right\| < \infty$$

so that

$$\sqrt{\tau} \left| \tilde{\theta}_n^{n(\tau)} - \tilde{\theta}_n^\infty \right| \leq \sqrt{\tau} \left\| \left( \nabla^2 L_\tau^{n(\tau)}(\tilde{\theta}) \right)^{-1} \right\| \left| L_\tau^{n(\tau)}(\tilde{\theta}_n^{n(\tau)}) - L_\tau(\tilde{\theta}_n^\infty) \right| = O(\tau^{-\frac{1}{2} - q}) \to 0$$

as $\tau \to \infty$ given that $q > \frac{1}{2}$. \qed

Proof of Proposition 5.1. We proceed inductively. The claim follows immediately for $m_T^n = 0$. For $m_T^n \geq 1$, $\left| \tilde{P}_\theta^{T_n}(i_0, \ldots, i_{m_T^n}) - P_\theta^{T_n}(i_0, \ldots, i_{m_T^n}) \right|$ is bounded above by the triangle inequality:

$$\left| \tilde{P}_\theta^{T_n}(i_{m_T^n}) - P_\theta^{T_n}(i_0, \ldots, i_{m_T^n}) \right| \leq \left| \tilde{P}_\theta^{T_n}(i_{m_T^n}) - P_\theta^{T_n}(i_0, \ldots, i_{m_T^n-1}) \right|$$

$$+ \left| P_\theta^{T_n}(i_{m_T^n}) - P_\theta^{T_n}(i_0, \ldots, i_{m_T^n-1}) \right|.$$

Note that $P_\theta^{T_n}(i_{m_T^n}) \leq 1$ and $P_\theta^{T_n}(i_0, \ldots, i_{m_T^n-1}) \leq 1$. By the induction assumption we know that $\left| \tilde{P}_\theta^{T_n}(i_0, \ldots, i_{m_T^n-1}) - P_\theta^{T_n}(i_0, \ldots, i_{m_T^n-1}) \right| \to 0$ as $n \to \infty$, $P_\theta$-almost surely. Further, from Jensen’s and Hölder’s inequality we derive for $j = m_T^n$

$$\left| \tilde{P}_\theta^{T_n}(i_j) - P_\theta^{T_n}(i_j) \right| \leq \mathbb{E}_\theta^* \left[ 1_{\{V_t \notin A_n(i_j) \text{ for all } t \in (T^{n,j}_n, T^{n,j}_{n+1})\}}^{1_{\{V_{T_n,i} \notin A_n(i_j)\}}} \right]_{D_T}$$

$$\leq \max_{\theta \in \Theta} \mathbb{P}_\theta^* \left[ \forall t \in (T^{n,j-1}_n, T^{n,j}_n) : V_t \notin A_n(i_j) \right]_{D_T}.$$

Now $(T^{n,j-1}_n, T^{n,j}_n) \to \emptyset$ as $n \to \infty$ since $m_T^n \to \infty$. Hence, the monotone convergence theorem implies that $\left| \tilde{P}_\theta^{T_n}(i_{m_T^n}) - P_\theta^{T_n}(i_{m_T^n}) \right| \to 0$ as $n \to \infty$, $P_\theta$-almost surely. \qed
Proof of Proposition 5.3. This follows immediately from the definition of the computational costs \( c_n \) and the squared error \( R_n \) of our approximation.

Proof of Proposition A.1. Lemma 4.1 in Section 1.4.3 of Ibragimov & Has’minskii (1981) states that it is sufficient to show that for any \( \gamma \geq 0 \) and any \( \theta \in \Theta \),

\[
P_\theta \left[ \sup_{|\xi| \geq \gamma} \frac{1}{\tau} L_\tau(\theta + \xi) - \frac{1}{\tau} L_\tau(\theta) \geq 0 \right] \to 0 \tag{44}
\]

as \( \tau \to \infty \). Write

\[
\frac{1}{\tau} L_\tau(\theta + \xi) - \frac{1}{\tau} L_\tau(\theta) = \frac{1}{\tau} \log \frac{\mathcal{L}_\tau(\theta + \xi)}{\mathcal{L}_\tau(\theta)}.
\]

We will show that for any given \( \epsilon, \delta > 0 \), there exists \( \tau^* > 0 \), such that for all \( \tau \geq \tau^* \) and \( \theta, \tilde{\theta} \in \Theta \):

\[
P_\theta \left[ \frac{1}{\tau} \log \frac{\mathcal{L}_\tau(\tilde{\theta})}{\mathcal{L}_\tau(\theta)} > \delta \right] \leq \epsilon.
\]

This implies convergence in probability uniformly over the parameter space \( \Theta \), and

\[
P_\theta \left[ \sup_{|\xi| \geq \gamma} \frac{1}{\tau} (L_\tau(\theta + \xi) - L_\tau(\theta)) > 0 \right] \to 0. \tag{45}
\]

The inequality “\( \geq \)” in Condition (44) then follows from Assumption (A4).

Fix \( \epsilon, \delta > 0 \) and take \( \theta, \tilde{\theta} \in \Theta \). The likelihood function \( \mathcal{L}_\tau(\theta) \) satisfies \( \mathcal{L}_\tau(\tilde{\theta}) = \frac{d\mathbb{D}_{\tilde{\theta},\tau}}{d\mathbb{D}_{\theta,\tau}} \) by definition. Consequently,

\[
\mathbb{E}_{\theta_0} \left[ \mathcal{L}_\tau(\tilde{\theta}) \right] = 1. \tag{46}
\]

Theorem 3.1 states that the likelihood function satisfies

\[
\mathcal{L}_\tau(\theta) = \frac{d\mathbb{D}_{\theta,\tau}}{d\mathbb{D}_{\theta_0,\tau}} \propto \mathbb{E}_\theta^* \left[ \frac{1}{M_\tau(\theta)} \right] \mathcal{D}_\tau \mathcal{L}_F^*(\theta; \tau),
\]

which is strictly positive under Assumption (A3). Thus, \( \mathbb{D}_{\theta_0,\tau} \) also absolutely continuous with respect to \( \mathbb{D}_{\theta,\tau} \) with Radon-Nikodym density

\[
\frac{d\mathbb{D}_{\theta_0,\tau}}{d\mathbb{D}_{\theta,\tau}} = \frac{1}{\mathcal{L}_\tau(\theta)}.
\]

By construction, \( \frac{\mathcal{L}_\tau(\tilde{\theta})}{\mathcal{L}_\tau(\theta)} \) is \( \mathcal{D}_\tau \)-measurable. We conclude from the exponential Chebyshev inequality and (46) that

\[
P_\theta \left[ \frac{1}{\tau} \log \frac{\mathcal{L}_\tau(\tilde{\theta})}{\mathcal{L}_\tau(\theta)} > \delta \right] \leq e^{-\tau \delta} \mathbb{E}_\theta \left[ \frac{\mathcal{L}_\tau(\tilde{\theta})}{\mathcal{L}_\tau(\theta)} \right] = e^{-\tau \delta} \mathbb{E}_{\theta_0} \left[ \mathcal{L}_\tau(\tilde{\theta}) \right] = e^{-\tau \delta}.
\]

Choosing \( \tau \) large enough leads to the claim. \( \square \)
Proof of Proposition A.2. We will use an argument similar to the one used to prove Proposition 6.1 in Hall & Heyde (1980), after generalizing to a multivariate parameter and our incomplete data structure. For simplicity, write \( m = m(\tau) \).

By Taylor expansion and Assumption (A5),

\[
0 = \frac{1}{\tau} \nabla L_\tau(\hat{\theta}_\tau) = \frac{1}{\tau} \nabla L_\tau(\theta_0) + \frac{1}{\tau} \nabla^2 L_\tau(\theta_0) \cdot (\hat{\theta}_\tau - \theta_0) + o_P(\|\hat{\theta}_\tau - \theta_0\|) \tag{47}
\]
since the MLE \( \hat{\theta}_\tau \) is consistent by Proposition A.1 and the third derivatives of the log-likelihood function are uniformly bounded in probability in a neighborhood of \( \theta_0 \) by Assumption (A10). Reformulate equation (47) as

\[
\sqrt{\tau} \left( \hat{\theta}_\tau - \theta_0 \right) = \left( -\frac{1}{\tau} \nabla^2 L_\tau(\theta_0) + o_P(1) \right)^{-1} \sqrt{\frac{1}{\tau} \nabla L_\tau(\theta_0)}.
\]

Assumption (A5) and the definition of \( \mathcal{L}^*_E,i \) and \( \mathcal{L}^*_F,i \) in (27)-(28) imply

\[
\nabla^2 L_\tau(\theta) = \nabla^2 \sum_{i=1}^{m} \log \mathcal{L}^*_E,i(\theta) + \nabla^2 \sum_{i=1}^{m} \log \mathcal{L}^*_F,i(\theta)
\]

\[
= \sum_{i=1}^{m} \nabla^2 \mathcal{L}^*_E,i(\theta) \frac{\mathcal{L}^*_E,i(\theta)}{\mathcal{L}^*_E,i(\theta)} + \sum_{i=1}^{m} \nabla^2 \mathcal{L}^*_F,i(\theta) \frac{\mathcal{L}^*_F,i(\theta)}{\mathcal{L}^*_F,i(\theta)}
\]

\[
- \sum_{i=1}^{m} \nabla \log \mathcal{L}^*_E,i(\theta) ^\top \nabla \log \mathcal{L}^*_E,i(\theta)
\]

\[
- \sum_{i=1}^{m} \nabla \log \mathcal{L}^*_F,i(\theta) ^\top \nabla \log \mathcal{L}^*_F,i(\theta).
\]

Chebyshev’s inequality together with Assumptions (A6) and (A7) imply

\[
\lim_{m \to \infty} \frac{1}{m} \sum_{i=1}^{m} \nabla^2 \mathcal{L}^*_E,i(\theta) \mathcal{L}^*_E,i(\theta) = 0
\]

in \( \mathbb{P}_\theta \)-probability, and Assumption (A8) leads to

\[
\frac{1}{\tau} \nabla^2 L_\tau(\theta) \to -\zeta (\Sigma_E(\theta) + \Sigma_F(\theta)) \tag{48}
\]

in \( \mathbb{P}_\theta \)-probability as \( \tau \to \infty \). Thus, it suffices to show that

\[
\frac{1}{\sqrt{\tau}} \nabla L_\tau(\theta_0) \to N \left( 0, \zeta (\Sigma_E(\theta_0) + \Sigma_F(\theta_0) + 2\Sigma_{EF}(\theta_0)) \right) \tag{49}
\]

in \( \mathbb{P}_{\theta_0} \)-distribution. Asymptotic normality then follows given that \( \Sigma_E(\theta_0) + \Sigma_F(\theta_0) + 2\Sigma_{EF}(\theta_0) \) is positive semidefinite.
To see that (49) holds, note that Assumptions (A6)-(A7) together with the fact that
\[ \nabla \log L^*_E(\theta) \text{ and } \nabla \log L^*_F(\theta) \] are $D_{s_i}$-measurable imply that
\[ \nabla L_{s_k}(\theta) = \sum_{i=1}^{k} \{ \nabla \log L^*_{E,i}(\theta) + \nabla \log L^*_{F,i}(\theta) \}, \quad 1 \leq k \leq m, \]
is a $P_{\theta}$-martingale on $(D_{s_k})_{1 \leq k \leq m}$. Chebyshev’s inequality and Assumptions (A7)-(A8) lead to
\[ \Sigma_E(\theta) = \lim_{m \to \infty} \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}_{\theta} \left[ \nabla \log L^*_{E,i}(\theta)^\top \nabla \log L^*_{E,i}(\theta) \mid D_{s_{i-1}} \right] \] (50)
\[ = \lim_{m \to \infty} \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}_{\theta} \left[ \nabla \log L^*_{E,i}(\theta)^\top \nabla \log L^*_{E,i}(\theta) \right], \]
\[ \Sigma_F(\theta) = \lim_{m \to \infty} \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}_{\theta} \left[ \nabla \log L^*_{F,i}(\theta)^\top \nabla \log L^*_{F,i}(\theta) \mid D_{s_{i-1}} \right] \] (51)
\[ = \lim_{m \to \infty} \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}_{\theta} \left[ \nabla \log L^*_{F,i}(\theta)^\top \nabla \log L^*_{F,i}(\theta) \right], \]
\[ \Sigma_{EF}(\theta) = \lim_{m \to \infty} \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}_{\theta} \left[ \nabla \log L^*_{E,i}(\theta)^\top \nabla \log L^*_{F,i}(\theta) \mid D_{s_{i-1}} \right] \] (52)
\[ = \lim_{m \to \infty} \frac{1}{m} \sum_{i=1}^{m} \mathbb{E}_{\theta} \left[ \nabla \log L^*_{E,i}(\theta)^\top \nabla \log L^*_{F,i}(\theta) \right]. \]

In the notation of Hall & Heyde (1980) (HH), $I_m(\theta) = \nabla \log L_{s_m}(\theta)^\top \nabla \log L_{s_m}(\theta)$ and $J_m(\theta) = \nabla^2 \log L_{s_m}(\theta)$. We have that
\[ \lim_{m \to \infty} \frac{1}{m} I_m(\theta) = \Sigma_E(\theta) + \Sigma_F(\theta) + 2\Sigma_{EF}(\theta), \]
\[ \lim_{m \to \infty} \frac{1}{m} J_m(\theta) = - (\Sigma_E(\theta) + \Sigma_F(\theta)) \]
where the first equation follows from Assumption (A6) and Chebyshev’s inequality, and the second equation follows from (48). As a result, $I_m(\theta) \to \infty$ as $m \to \infty$. Also, $\lim_{m \to \infty} \frac{1}{m} I_m(\theta) = \lim_{m \to \infty} \frac{1}{m} \mathbb{E}_{\theta} [I_m(\theta)]$. Finally, $\lim_{m \to \infty} \frac{1}{m} J_m(\theta) = - \lim_{m \to \infty} \frac{1}{m} I_m(\theta) + 2\Sigma_{EF}(\theta)$. Given that $I_m(\theta)$ and $J_m(\theta)$ are continuous by Assumption (A5), we conclude that a slightly modified version of Assumption 1 of Proposition 6.1 of HH holds. This modified version of Assumption 1 accounts for the fact that our data is observed incompletely and there is correlation between the event data likelihood and the factor data likelihood. In addition, Assumption 2 of Proposition 6.1 of HH also holds because of our Assumptions (A5) and (A10), equation (48), and Theorem 2.1 of Billingsley (2009). Following the proof of Proposition 6.1 of HH yields the claim. 

\[ \square \]

36
Proof of Corollary A.3. The first equality for \( \Sigma_E(\theta) \) follows from (48). For the second equality note that, by construction and since \( s_0 = 0 \), \( s_m(\tau) = \tau \), we have that

\[
\Sigma_E(\theta) = \lim_{\tau \to \infty} \left\{ \frac{1}{m(\tau)} \sum_{i=1}^{m(\tau)} \nabla \log \mathcal{L}_E^*(\theta; s_i) \nabla \log \mathcal{L}_{E,i}^*(\theta) \right\} 
- \frac{1}{m(\tau)} \sum_{i=1}^{m(\tau)} \nabla \log \mathcal{L}_E^*(\theta; s_{i-1}) \nabla \log \mathcal{L}_{E,i}^*(\theta)
\]

\[
= \lim_{\tau \to \infty} \frac{1}{m(\tau)} \nabla \log \mathcal{L}_E^*(\theta; s_{m(\tau)}) \nabla \sum_{i=1}^{m(\tau)} \nabla \log \mathcal{L}_{E,i}^*(\theta)
\]

\[
= \lim_{\tau \to \infty} \frac{\tau}{m(\tau)} \nabla \log \mathcal{L}_E^*(\theta; s_{m(\tau)}) \nabla \log \mathcal{L}_E^*(\theta; s_{m(\tau)})
\]

\[
= \frac{1}{\zeta} \lim_{\tau \to \infty} \frac{\tau}{\tau} \nabla \log \mathcal{L}_E^*(\theta) \nabla \log \mathcal{L}_E^*(\theta; \tau).
\]

The analogous expressions hold for \( \Sigma_E(\theta) \) and \( \Sigma_{EF}(\theta) \). 

References


Figure 1: Relative absolute error $\frac{\log \hat{E}^{\tau_1}(\theta_0;\tau,1) - \log E(\theta_0;\tau,1)}{\log E(\theta_0;\tau,1)}$ of the approximate log-filter for one sample realization of $D_r$ generated from the model $\theta_0$. The dotted line corresponds to an error of order $O(n^{-1})$ that decreases at a linear rate.
Figure 2: Run time necessary to evaluate the approximate log-filter $\log \hat{E}^T_n(\theta_0; \tau, 1)$ for one sample realization of $D_\tau$ generated from the model $\theta_0$. The dotted line corresponds to a run time of order $O(n)$ that increases at a linear rate.
Figure 3: Relative squared error $\frac{(\log \hat{E}^{n}(\theta_0;\tau,1) - \log \mathbb{E}(\theta_0;\tau,1))^2}{\log \mathbb{E}(\theta_0;\tau,1)}$ of the approximate log-filter for one sample realization of $D_\tau$ generated from the model $\theta_0$, plotted against the time necessary to compute the log-filter approximation (black dots). The black line shows the fitted linear regression of log(relative squared error) on log(run time). We compare to the relative mean squared error (MSE) of a Monte Carlo estimator based on $K$ i.i.d. samples of an Euler discretization of $1/M_\tau(\theta_0)$ under $\mathbb{P}_{\theta_0}$ conditional on $D_\tau$ (red crosses). We implement Euler discretization following the square-root rule of Duffie & Glynn (1995), and allow for $\sqrt{K}$ Euler steps per 1-year time period. The red line shows the fitted linear regression of log(relative MSE) on log(run time) for the Euler discretization approximation.
Figure 4: Error $\hat{\theta}_n^{(\tau)} - \theta_0$ of a parameter estimate, averaged over the errors associated with 10 alternative realizations of $D_\tau$ generated from the model $\theta_0$, as a function of the time horizon $\tau$. We take $n(\tau) = 2\sqrt{\tau}$ in order to satisfy the conditions of Proposition 4.3. Also shown are the errors of the parameter estimates obtained using the stochastic EM Algorithm, averaged over the same realizations of $D_\tau$. In the E-step of the EM Algorithm, we estimate the factor likelihood as the sample average of 250 $\tau$ samples of an Euler discretization based on the square-root rule of Duffie & Glynn (1995). The dashed line marks a value of zero.
Figure 5: Parameter error $\hat{\theta}_n^n - \theta_0$, averaged over 10 alternative realizations of $D_\tau$ generated from the model $\theta_0$ with $\tau = 40$, as a function of $n$, the fineness and range of the state space and time discretizations. Also shown are the errors of the parameter estimates obtained using the stochastic EM Algorithm, averaged over the same realizations of $D_\tau$. In the E-step of the EM Algorithm we estimate the factor likelihood as the sample average of 10,000 samples of an Euler discretization based on the square-root rule of Duffie & Glynn (1995). The blue lines show the errors of proxies of true maximum likelihood estimators, also averaged over the 10 realizations of $D_\tau$. These proxies are computed as the numerical maximizers our approximate log-likelihood with $n = 400$, where the numerical optimization routines are initialized at the true parameters. The dotted lines show the average errors of the initial parameter values, and the dashed lines mark zero.
Figure 6: Fitted filtered factor and intensity vs. actual factor and event counts.
Figure 7: Goodness-of-fit tests: QQ plots of the quantiles of the time-changed inter-arrival times derived vs. theoretical standard exponential quantiles.
Figure 8: Asymptotic standard errors of our estimates and EM estimates plotted against $\tau$, the end of the observation period. The y-axes are in log-scale.