Camilla Damian, Zehra Eksi, and Rüdiger Frey

EM Algorithm for Markov Chains Observed via Gaussian Noise and Point Process Information: Theory and Case Studies

Received June 23, 2017

Abstract: In this paper we study parameter estimation via the Expectation Maximization (EM) algorithm for a continuous-time hidden Markov model with diffusion and point process observation. Inference problems of this type arise for instance in credit risk modelling. A key step in the application of the EM algorithm is the derivation of finite-dimensional filters for the quantities that are needed in the E-step of the algorithm. In this context we obtain exact, unnormalized and robust filters, and we discuss their numerical implementation. Moreover, we propose several goodness-of-fit tests for hidden Markov models with Gaussian noise and point process observation. We run an extensive simulation study to test speed and accuracy of our methodology. The paper closes with an application to credit risk: we estimate the parameters of a hidden Markov model for credit quality where the observations consist of rating transitions and credit spreads for US corporations.

Keywords: Expectation maximization (EM) Algorithm, hidden Markov models, point processes, non-linear filtering, goodness-of-fit tests, credit risk ratings

1 Introduction

Continuous-time hidden Markov models (models where key variables are affected by an unobservable finite state Markov chain) are commonly used in finance, insurance and economics. Examples include portfolio optimization in models with unobservable Markov-modulated drift such as Sass and Haussmann [2004] or Rieder and Bäuerle [2005]; dynamic credit risk modelling such as Frey and

Camilla Damian, Institute for Statistics and Mathematics, Vienna University of Economics and Business (WU),camilla.damian@wu.ac.at
Zehra Eksi, Institute for Statistics and Mathematics, WU, zehra.eksi@wu.ac.at
Rüdiger Frey, Institute for Statistics and Mathematics, WU,ruediger.frey@wu.ac.at

Schmidt [2012]; Markov modulated risk processes in insurance such as Asmussen [1989]; or high frequency data in finance, see for instance Frey and Runggaldier [2001]. For further examples of hidden Markov models in finance we refer to the interesting collections and R. J. Elliott [2007] and and R. J. Elliott [2014].

Statistical inference for hidden Markov models is thus an important issue. This problem is frequently addressed via the Expectation Maximization (EM) algorithm (Dempster et al. [1977]). In the so-called E-Step of the algorithm one needs to solve a complicated non-linear filtering problem, so that a substantial effort is needed to tailor the method to a given model setting. We now list a few important contributions in that regard. Dembo and Zeitouni [1986] provide general results on the EM algorithm for continuous-time stochastic processes that are observed in Gaussian noise. In an important paper Elliott [1993] specializes these results to a hidden Markov model observed in Gaussian noise. In particular, he obtains finite-dimensional filters and smoothers for the quantities needed for the E-Step, see also the textbook and L. Aggoun and J. B. Moore [1995]. Elliott and Malcolm [2008] finally study the EM algorithm for a Poisson process whose intensity is modulated by a finite-state Markov chain.

In this paper we generalize the model for the observation process and consider a hidden Markov model where the state process is observed simultaneously via diffusive and point processes information. In recent years there has been an increasing interest in such models in finance and insurance. In particular, these models are relevant in the analysis of credit risk, see for instance Frey and Schmidt [2012] and the example we provide in Section 2.2. We make the following contributions. First, we derive the exact normalized and unnormalized recursive filters that are needed for the EM algorithm in our setup. This is a nontrivial extension of the work of Elliott [1993] and of Elliott and Malcolm [2008]. In the practical implementation of the algorithm it is important to work with a version of the filters that depends continuously on the observations (so-called robust filters, see Clark [1978] or James et al. [1996]). Our second contribution is therefore the derivation of robust filters for our framework. Third, we develop goodness-of-fit tests for the hypothesis that the hidden Markov model parametrized in terms of an estimated parameter vector describes the observed data well. We are not aware of any prior use of such tests in the context of hidden Markov models, so that this is also a contribution to the "classical" case of hidden Markov models with only diffusion or point process observation. Fourth we perform an extensive simulation analysis that tests the speed and accuracy of the algorithm and of the proposed goodness-of-fit tests. This analysis suggests that the method yield satisfactory results in the sense that the EM estimates converge to the corresponding full-information MLE estimates. Furthermore, we observe that the use of robust filters improves the stability of the algorithm, especially when working on a

coarser time grid, and we give examples where the proposed tests are able to distinguish the correctly estimated model from a misspecified one. Finally we give an example with real data and apply the methodology to rating transitions and credit spreads for US corporations. The unobservable 'true' credit quality is modeled as a finite state Markov chain; the point process observation is generated by defaults and rating changes and the diffusive observation is generated by observed credit spreads. We obtain reasonable estimates for the parameters of the model and we find that the filter estimate for the unobservable credit quality balances the spread- and the rating information in a plausible fashion. A discrete-time hidden Markov model for rating transitions was estimated via the EM approach by Korolkiewicz and Elliott [2008]; spread data were not considered in their analysis.

The remainder of this paper is structured as follows. In Section 2 we introduce the notation and the setting, we give a motivating example related to credit risk modeling and we discuss the main steps of the EM algorithm. In Section 3 we study in detail the filtering problems arising in the E-Step of the algorithm; this is the mathematical core of the paper. Goodness-of-fit tests are discussed in Section 4. In Section 5 we present the results of our simulation study; the application to credit data is discussed in Section 6.

Acknowledgments.

C. Damian and R. Frey are grateful for the support by the Vienna Science and Technology Fund (WWTF) through project MA14-031.

2 EM Algorithm for Diffusion and Point Process Information

In this section we introduce our setup and provide a motivating example. Moreover, we derive the form that the EM-algorithm takes in our setting.

2.1 The setup

We consider a finite time interval $[0, T]$ and a continuous-time finite-state Markov chain X defined on the filtered probability space $(\Omega, \mathcal{G}, \mathbb{G}, \mathbb{P})$ where $\mathbb{G} = (\mathcal{G}_t)_{0 \leq t \leq T}$ satisfies the usual conditions. All processes we consider are

$\mathbb{G}$-adapted, that is $\mathbb{G}$ is the *global* filtration. The chain X has the state space $S = \{e_1, e_2, ..., e_K\}$, where, without loss of generality, we assume that e_k is the kth basis column vector of $\mathbb{R}^K$. The initial distribution of X is denoted by $p = (p^1, \ldots, p^K)'$, and the matrix $A = (a^{jk})$, $1 \leq j, k \leq K$, represents the transpose of the generator matrix of X. Hence the process M^X with

$$M_t^X = X_t - X_0 - \int_0^t AX_s ds, \quad t \leq T, \tag{1}$$

is a $\mathbb{G}$-martingale.

Information.

We assume that X is not directly observable. Instead, we consider the continuous noisy observation $\tilde{Z}$ with $\tilde{Z}_t = \int_0^t \tilde{g}(X_s)ds + \sigma_Z W_t$. Here $\sigma_Z > 0$ measures the amount of noise in the continuous observation of X and W is a standard $\mathbb{P}$-Brownian motion with respect to the filtration $\mathbb{G}$, independent of X. We note that the extension of our results to an arbitrary vector observation process is straightforward. We introduce the normalized observation process $Z_t = \tilde{Z}_t/\sigma_Z$ and we let $g = \tilde{g}(\cdot)/\sigma_Z$. Then Z has dynamics

$$Z_t = \int_0^t g(X_s)ds + W_t, \quad t \geq 0, \tag{2}$$

so that during the theoretical analysis we assume without loss of generality that $\sigma_Z = 1$. In theory, the value of σ_Z is equal to $[Z]_t/t$ where $[Z]$ is the quadratic variation of Z and is thus observable. However, in practice the observations are typically not continuous and the value of σ_Z has to be estimated. We will discuss this problem in Section 5.

The second source of information stems from a univariate point process D; the extension to multivariate point process observation is straightforward. We assume that D admits the $\mathbb{G}$-intensity $h_t(D)\lambda(X_t)$; here $h_t(\cdot)$ is a bounded functional that depends on the left-continuous version $(D_{s-})_{0 \leq s \leq t}$ of the past trajectory of D. For instance, the choice $h_t(D) = 1_{\{D_{t-}=0\}}$ models the case where we can observe only one jump. Alternatively, h can be used to model the case where D is self-exciting. To this we might set $h_t(D) = 1 + M \wedge \int_0^{t-} e^{-\kappa(t-s)} dD_s$ for some (large) threshold M and some decay rate κ. It is in principle possible to estimate parameters of the function h such as κ using the EM methodology. However, details depend very much on the specific functional form of h. For this reason we assume in the present paper that all parameters of h are known. Since

h is bounded the process

$$M_t^D = D_t - \int_0^t h_s(D,c)\lambda(X_s)ds \quad t \geq 0, \tag{3}$$

is a $\mathbb{G}$-martingale (see for example [Brémaud, 1981, Section II, T8]).

The information available to the observer of the system is carried by the filtration $\mathbb{F}$ which is generated by the noisy diffusion information $\mathbb{F}^{\mathbb{Z}}$ and the point process information $\mathbb{F}^{\mathbb{D}}$, that is, $\mathbb{F} = \mathbb{F}^{\mathbb{D}} \vee \mathbb{F}^{\mathbb{Z}}$. Note that $\mathcal{F}_t \subset \mathcal{G}_t$ for all $t \leq T$. For a generic integrable process Y we denote its $\mathbb{F}$-optional projection by $\widehat{Y}$, in particular, $\widehat{Y}_t = \mathbb{E}[Y_t|\mathcal{F}_t]$ for all $t \leq T$.

Remark 2.1. In practical applications one is often dealing with noisy observations arising discretely in time, say, at time points $t_n = n\Delta$ for a step size $\Delta > 0$. If one works on a fine time scale, that is with a small Δ, it is still reasonable to use continuous time models. We now explain how this situation can be embedded in our setting. Suppose we have a noisy observation of the form $z_n = \widetilde{g}(X_{t_n}) + \epsilon_n$ for an i.i.d. sequence of noise variables with mean zero and variance σ_ϵ^2. Define the *scaled cumulative observations process* by

$$\widetilde{Z}_t := \Delta \sum_{t_n \leq t} z_n = \sum_{t_n \leq t} \Delta \widetilde{g}(X_{t_n}) + \Delta \sum_{t_n \leq t} \epsilon_n \,. \tag{4}$$

For small Δ the first term on the right side is an approximation of $\int_0^t \widetilde{g}(X_s)ds$ and the second term is an approximation of $\sigma_\epsilon \sqrt{\Delta} W_t$ for a standard Brownian motion W (by Donsker's invariance theorem). It is therefore natural to apply the continuous-time filtering formulas derived in Section 3.2 to the observation process $\widetilde{Z}$ from (4). This immediately raises the issue of robust filtering: one seeks filters that perform well even if the dynamics of Z are not exactly of the form (2), see Section 3.3 below.

2.2 Examples

2.2.1 A Hidden Markov Model for Credit Quality

In this section we introduce an example from the field of credit risk modelling that fits into our framework. We consider a sample of m firms indexed by $i = 1, \ldots, m$; all of these firms are rated by some rating agency and have CDS contracts outstanding. The corresponding credit ratings (including default) and CDS spreads are observable and constitute the available information.

State Process.
Let X_t^i denote the *true* credit quality of firm i, modeled as a finite-state Markov chain with state space $S = \{e_1, \dots, e_K\}$ and generator matrix $A^\top$; following the literature, we assume that this matrix is identical for all firms. Here the state e_1 represents the *best* credit quality, while e_K represents the *worst* non-default state. In the sequel, we write for two elements of S $e_l \geq e_k$ whenever $l > k$, so that states are ordered according to credit quality.

Observation Process.
We have two sources of information available to the observer of the system. First, there is the point-process information which stems from observable ratings and defaults. Second, there is the continuous information provided by the time series of CDS spreads.

Point process observation. We denote by $R_t^i \in S$ the observed rating of firm i at time t. In order to model the dynamics of the process R^i in a simple way, we assume that there are only three types of events possible. Suppose that the current state of R_t is e_l. First, there may be an *upgrading* of firm i, that is a transitions of R_t^i to the state e_{l-1}; second there may be a *downgrading* of firm i, that is a transitions of R_t^i to the state e_{l+1}; third, firm i might default. Note that an upgrading is only possible if $l > 1$, that is if the observed rating of the firm is not yet in the best rating category; similarly, a downgrading is only possible if $l < K$. Hence the dynamics of R_t^i can be described in terms of the following three point processes:

i) $D_t^{+,i}$, the number of upgradings of firm i up to time t;
ii) $D_t^{-,i}$, the number of downgradings of firm i up to time t;
iii) $D_t^{d,i}$, the default indicator of firm i (a point process that jumps to one at the default time of firm i).

Note that for simplicity we do not consider upgradings or downgradings of size larger than one; if real rating data exhibit an upgrading (downgrading) by more than one category we will treat this as several upgradings (downgradings) of size one.

We denote by λ^+, λ^- and by λ^d the intensities of D^+, D^- and of D^d, respectively. We assume that these intensities are identical across firms. We propose the following parametrization: if $R_t^i > e_1$ we let

$$\lambda^+(X_t^i, R_t^i) = \lambda_1^+ 1_{\{X_t^i < R_t^i\}} + \lambda_2^+ 1_{\{X_t^i = R_t^i\}} + \lambda_3^+ 1_{\{X_t^i > R_t^i\}};$$

moreover, $\lambda^+(X_t^i, e_1) \equiv 0$. This parametrization is motivated by the idea that the observed rating follows the true credit quality, albeit with some rating error.

In particular we expect that $\lambda_1^+ > \lambda_2^+ > \lambda_3^+$, that is, an upgrading is most likely when the true credit quality is lower than the observed rating. Similarly, for $R_t^i < e_K$ we let

$$\lambda^-(X_t^i, R_t^i) = \lambda_1^- 1_{\{X_t^i < R_t^i\}} + \lambda_2^- 1_{\{X_t^i = R_t^i\}} + \lambda_3^- 1_{\{X_t^i > R_t^i\}}.$$

Moreover, $\lambda^-(X_t^i, e_K) \equiv 0$. In this case, we expect to have $\lambda_1^- < \lambda_2^- < \lambda_3^-$. Finally, concerning the default process we take

$$\lambda^d(X_t^i) = \langle \lambda^d, X_t^i \rangle, \quad \text{for} \quad \lambda_1^d < \cdots < \lambda_K^d.$$

Continuous observation. The diffusion information stems from observed CDS spreads as we explain next. Let $z_n^i = \log(CDS_{t_n}^i)$, that is, z_n^i is the logarithm of the observed CDS spread of firm i at time t_n, $n \in \{1, \ldots, N\}$, $t_N = T$. We assume that

$$z_n^i = \tilde{g}(X_{t_n}^i) + \epsilon_n^i, \tag{5}$$

where ϵ_n^i, $n \in \{1, \ldots, N\}$, $1 \le i \le m$ are independent noise variables with mean zero and some variance σ_ϵ. The relation (5) is motivated by empirical work for corporate credit markets such as Berndt et al. [2008], which shows that there is a reasonably stable regression-type relation between observed logarithmic CDS spreads and credit quality as measured by short-term default probabilities or by ratings. Identifying (5) with a continuous model as in Remark 2.1 gives the observation process $\widetilde{Z}_t^i$.

2.2.2 Other applications

Parameter estimation problems for hidden Markov models with point process information arise also in other areas of finance and insurance and we now give a few examples. In insurance one considers frequently Markov modulated risk processes where the arrival intensity of claims is driven by an unobservable Markov chain. Some authors such as Schmidli [1995] consider perturbed risk processes, where a Brownian component is added to the risk process in order to model fluctuations caused by small claims, investment returns or other sources of randomness. This line of modelling gives rise to a hidden Markov model with diffusion and point process information. Another interesting area of application for our methodology is high frequency data in finance. It is well-known that on very fine time-scales asset prices follow a pure jump process since in reality quoted prices are constant between trades and jump only when new orders arrive. Moreover, there are good reasons for introducing an unobservable regime

switching factor in the price dynamics: this helps to reproduce the clustering in inter-event durations, and a hidden Markov chain can be used to model the feedback effect from the trading activity of the rest of the market, see for instance Cont [2011] or Colaneri et al. [2017] for details. Hence it makes sense to consider hidden Markov models with point process information in the analysis of high frequency data.

2.3 The EM Algorithm

Note that for a generic function $f\colon S \to \mathbb{R}$ it holds that $f(X_t) = \langle X_t, f\rangle$ where $\langle\ ,\ \rangle$ denotes the scalar product on $\mathbb{R}^K$ and $f_k = f(e_k)$, $1 \le k \le K$, so that functions of the Markov chain can be identified with K-vectors. Hence parameters to be estimated are given by the parameter vector

$$\theta = \left(a^{jk}, g^j, \lambda^j,\ j,k \in \{1,\dots,K\}, j \neq k\right);$$

the set of admissible parameter vectors is denoted by Θ.

We use the EM algorithm to estimate the model parameters and to infer the unobserved realization of the state process X. Denote by $\mathbb{P}_\theta$ the probability measure corresponding to the parameter vector $\theta \in \Theta$. In order to describe the algorithm, we define the *full-information log-likelihood* by

$$L(\theta,\theta') := \log \frac{d\mathbb{P}_\theta}{d\mathbb{P}_{\theta'}}\Big|_{\mathcal{G}_T},\ \text{for all } \theta,\theta' \in \Theta\,. \tag{6}$$

Of course, in making this definition we implicitly assume that $\mathbb{P}_\theta$ and $\mathbb{P}_{\theta'}$ are equivalent on $\mathcal{G}_T$ which is stronger than requiring equivalence of these measures on the observation σ-field $\mathcal{F}_T$.

The EM algorithm is an iterative procedure that leads to a sequence $\{\theta^m\}_{m\ge1}$ of parameter estimates such that the likelihood of the observations increases in each step. Schematically, given the optimal parameter vector θ^m after the m^{th} iteration of the algorithm, iteration $m+1$ of the algorithm consists of the following two steps:
Expectation (E) : Compute the estimate $\widehat{L(\theta,\theta^m)} = E_{\theta^m}\left[L(\theta,\theta^m) \mid \mathcal{F}_T\right]$.
Maximization (M) : Find $\theta^{m+1} \in \operatorname{argmax}_{\theta\in\Theta} \widehat{L(\theta,\theta^m)}$.

In our setup the steps of the EM algorithm are as follows.

E-Step.

In order to write the full-information log likelihood in compact form we introduce a couple of stochastic processes related to the Markov chain X. Let for $t \leq T$ and $1 \leq j, k \leq K, j \neq k$,

$$N_t^{jk} = \sum_{0<s\leq t} 1_{\{X_{s-}=e_j\}} 1_{\{X_s=e_k\}} \text{ (number of jumps from state } j \text{ to state } k), \tag{7}$$

$$G_t^j = \int_0^t 1_{\{X_{s-}=e_j\}} dZ_s, \text{ (level integral for state } j), \tag{8}$$

$$J_t^j = \int_0^t 1_{\{X_{s-}=e_j\}} ds \text{ (occupation time for state } j), \tag{9}$$

$$B_t^j = \int_0^t 1_{\{X_{s-}=e_j\}} dD_s \text{ (jump level integral for state } j), \tag{10}$$

$$C_t^j = \int_0^t 1_{\{X_{s-}=e_j\}} h_s(D) ds \text{ (modified occupation time for state } j). \tag{11}$$

Combining the Girsanov theorem for point processes (see [Brémaud, 1981, pg 166, Theorem T3]) with the likelihood function given in Elliott [1993] gives that the full-information log-likelihood equals

$$\begin{aligned} L(\theta, \theta') = &\sum_{j,k=1 j\neq k}^{K} \left(N_T^{jk} \log a^{kj} - a^{kj} J_T^j \right) + \sum_{j=1}^{K} \left(g^j G_T^j - \frac{1}{2}(g^j)^2 J_T^j \right) \\ &+ \sum_{j=1}^{K} \left(\log(\lambda^j) B_T^j - \lambda^j C_T^j \right) + R(\theta'), \end{aligned}$$

where $R(\theta')$ is independent of θ. This gives

$$\begin{aligned} \widehat{L(\theta, \theta^m)} = E_{\theta^m} \left[\log \frac{d\mathbb{P}_\theta}{d\mathbb{P}_{\theta^m}} \middle| \mathcal{F}_T \right] = &\sum_{j,k=1, j\neq k}^{K} \left(\widehat{N_T^{jk}} \log a^{kj} - a^{kj} \widehat{J_T^j} \right) \\ &+ \sum_{j=1}^{K} \left(g^j \widehat{G_T^j} - \frac{1}{2}(g^j)^2 \widehat{J_T^j} \right) + \sum_{j=1}^{K} \left(\log\left(\lambda^j\right) \widehat{B_T^j} - \lambda^j \widehat{C_T^j} \right) + \widehat{R}(\theta^m). \end{aligned} \tag{12}$$

M-Step.

Since $\widehat{L(\cdot,\theta^m)}$ is concave, the new parameter vector θ^{m+1} is given by equating the partial derivatives of (12) to zero. We thus obtain

$$(a^{kj})^{m+1} = \frac{\widehat{N_T^{jk}}}{\widehat{J_T^j}}, \quad (g^j)^{m+1} = \frac{\widehat{G_T^j}}{\widehat{J_T^j}} \text{ and } (\lambda^j)^{m+1} = \frac{\widehat{B_T^j}}{\widehat{C_T^j}}. \tag{13}$$

3 Filtering

To perform the EM algorithm, one has to obtain the filtered estimates of the quantities in (12). This is a non-linear filtering problem with diffusion and point process information. Following Frey and Schmidt [2012], in Section 3.1 we address this problem via the innovations approach to nonlinear filtering. In Section 3.2 we derive unnormalized filters and the Zakai equation; Section 3.3 is concerned with robust filtering.

In the following we consider θ^m given and fixed and we simply write $\mathbb{P}$ and $\mathbb{E}$ instead of $\mathbb{P}_{\theta^m}$ and $\mathbb{E}_{\theta^m}$. Moreover, we always denote $\mathbb{G}$ martingales by upper-case letters and $\mathbb{F}$ martingales by lower-case letters.

3.1 Filtering via the Innovations Approach

The innovations Brownian motion (the martingale part in the $\mathbb{F}$-semimartingale decomposition of Z) is given by

$$w_t = Z_t - \int_0^t \langle \widehat{X}_s, g\rangle ds, \tag{14}$$

and the $\mathbb{F}$-martingale part of the point process D is given by

$$m_t^D = D_t - \int_0^t h_s(D)\langle \widehat{X}_s, \lambda\rangle ds. \tag{15}$$

The next theorem gives the first filtering result.

Theorem 3.1. *Consider a scalar process H of the form*

$$H_t = H_0 + \int_0^t \alpha_s^H ds + \int_0^t \gamma_s^H dW_s + \int_0^t \left(\beta_s^H\right)^\top dM_s^X + \int_0^t \delta_s^H dM_s^D, \tag{16}$$

where α^H, γ^H and δ^H are $\mathbb{G}$-predictable scalar processes and β^H is a K-dimensional vector process that is $\mathbb{G}$-predictable. Moreover, suppose that

(A2) $\mathbb{E}\left[\int_0^T \left(|\alpha_s^H| + |\delta_s^H| + \left(\gamma_s^H\right)^2\right) ds\right] + \mathbb{E}\left[\int_0^T \sum_{i=1}^K |\left(\beta_s^H\right)^i| ds\right] < \infty.$

Then, $\widehat{H}$ has the dynamics $\widehat{H}_t = \widehat{H}_0 + \int_0^t \widehat{\alpha^H}_s ds + \int_0^t \mu_s^H dw_s + \int_0^t \kappa_s^H dm_s^D$, where

$$\mu_s^H = \widehat{\gamma_s^H} + (\widehat{\langle X, g\rangle H})_s - \langle \widehat{X}_s, g\rangle \widehat{H}_s, \tag{17}$$

$$\kappa_s^H = \frac{1}{\langle \widehat{X}_{s-}, \lambda\rangle} \left((\widehat{\langle X, \lambda\rangle \delta^H})_{s-} + (\widehat{\langle X, \lambda\rangle H})_{s-} - \langle \widehat{X}_{s-}, \lambda\rangle \widehat{H}_{s-} \right). \tag{18}$$

Proof. During the proof we will frequently make use of the following facts:

(F1) For every true $\mathbb{G}$-martingale M, the projection $\widehat{M}$ is $\mathbb{F}$-martingale.

(F2) For a $\mathbb{G}$-adapted, integrable process α, the process $(\widehat{\int_0^t \alpha_s ds} - \int_0^t \widehat{\alpha}_s ds)_{0\le t\le T}$ is an $\mathbb{F}$-martingale.

(F3) For every $\mathbb{F}$-martingale m, there exists a $\mathbb{F}$-adapted process δ and an integrable, $\mathbb{F}$-predictable process ν such that m has the representation $m_t = \int_0^t \delta_s dw_s + \int_0^t \nu_s dm_s^D$.

(F1) and (F2) are standard in the nonlinear-filtering literature, for a proof of (F3) we refer to Frey and Schmidt [2012]. Using (F1) and (F2) we first write $\widehat{H}_t = \widehat{H}_0 + \int_0^t \widehat{\alpha_s^H} ds + m_t^H$ for some $\mathbb{F}$ martingale m^H. Using (F3) therefore gives that

$$\widehat{H}_t = \widehat{H}_0 + \int_0^t \widehat{\alpha_s^H} ds + \int_0^t \mu_s^H dw_s + \int_0^t \kappa_s^H dm_s^D. \tag{19}$$

It remains to identify the integrands μ^H and κ^H. Define for some arbitrary bounded, $\mathbb{F}$-predictable process ζ the $\mathbb{F}$-adapted process ρ by

$$\rho_t := \int_0^t \zeta_s dD_s\,.$$

In order to identify κ^H we will compare two different representations for $\widehat{\rho H}$. On the one hand we get from Ito's product formula that $H_t\rho_t = \int_0^t H_{s-} d\rho_s + \int_0^t \rho_{s-} dH_s + [\rho, H]_t$. As $[\rho, H]_t = \int_0^t \delta_s^H \zeta_s dD_s$ we have for $t \ge 0$

$$H_t\rho_t = \int_0^t \rho_s \alpha_s^H ds + \int_0^t H_r \zeta_s h_s(D)\langle X_s, \lambda\rangle ds + \int_0^t \delta_s^H \zeta_s h_s(D)\langle X_s, \lambda\rangle ds + M_t,$$

where M is a $\mathbb{G}$-martingale. Then, using (F1) and (F2), we get the following representation for $(\widehat{H\rho})_t$

$$(\widehat{H\rho})_t = \int_0^t \rho_s \widehat{\alpha_s^H} ds + \int_0^t \zeta_s h_s(D)(\widehat{\langle X, \lambda\rangle H})_s ds + \int_0^t \zeta_s h_s(D)(\widehat{\langle X, \lambda\rangle \delta^H})_s ds + m_t, \tag{20}$$

where m is an $\mathbb{F}$-martingale.

On the other hand, it holds that $(\widehat{H\rho})_t = \widehat{H}_t\rho_t$, as ρ is $\mathbb{F}$-adapted. Moreover, $\widehat{H}_t\rho_t = \int_0^t \widehat{H}_{s-} d\rho_s + \int_0^t \rho_{s-} d\widehat{H}_s + [\rho, \widehat{H}]_t$, and $[\rho, \widehat{H}]_t = \int_0^t \zeta_s \kappa_s^H dD_s$. Hence we obtain

$$\widehat{H}_t\rho_t = \int_0^t \rho_s \widehat{\alpha_s^H} ds + \int_0^t h_s(D)\langle \widehat{X}_s, \lambda\rangle \widehat{H}_s \zeta_s ds + \int_0^t h_s(D)\langle \widehat{X}_s, \lambda\rangle \kappa_s^H \zeta_s ds + \tilde{m}_t\,, \tag{21}$$

for an $\mathbb{F}$-martingale $\tilde{m}$. Now, $\widehat{H\rho}$ is a special semimartingale and hence has a unique decomposition (see, e.g, Protter [2013][Chapter 7, Thm. 34]). This implies that the martingale and finite variation parts in (20) and (21) must be equal. Comparing the two equations, we get

$$0 = \int_0^t \zeta_s h_s(D) \left((\widehat{\langle X, \lambda\rangle \delta^H})_s + (\widehat{\langle X, \lambda\rangle H})_s - \langle \widehat{X}_s, \lambda\rangle \widehat{H}_s - \kappa_s^H \langle \widehat{X}_s, \lambda\rangle \right) ds. \tag{22}$$

Moreover, the integrands in (22) are continuous in s for almost all s (they jump only at the jump times of D). Hence it also holds that

$$0 = \int_0^t \zeta_s h_s(D) \left((\widehat{\langle X, \lambda\rangle \delta^H})_{s-} + (\widehat{\langle X, \lambda\rangle H})_{s-} - \langle \widehat{X}_{s-}, \lambda\rangle \widehat{H}_{s-} - \kappa_s^H \langle \widehat{X}_{s-}, \lambda\rangle \right) ds. \tag{23}$$

As ζ is arbitrary and as κ^H is predictable, (23) yields (18).

In order to determine μ^H one follows same strategy and compares two different representations for $\widehat{HZ}$, we omit the details. □

Note that the integrands in (17) involve the filtered estimate of the product term $H_t\langle X_t, g\rangle$. This is inconvenient for practical purposes as the resulting filters are not recursive. As a remedy Elliott [1993] proposes to derive filters for the product H_tX_t. One has $H_tX_t = \sum_{i=1}^K H_t\langle X_t, e_i\rangle e_i$ and hence

$$(\widehat{HX})_t = \sum_{i=1}^K (\widehat{H\langle X, e_i\rangle})_t e_i.$$

Let $\mathbf{1} = (1, \cdots, 1)^\top \in \mathbb{R}^K$. Given $(\widehat{HX})_t$, the filter for H can then be obtained from the relation

$$\mathbb{E}[H_t \mid \mathcal{F}_t] = \mathbb{E}\Big[H_t \sum_{i=1}^K \langle X_t, e_i\rangle \mid \mathcal{F}_t\Big] = \sum_{i=1}^K (\widehat{H\langle X, e_i\rangle})_t = \langle (\widehat{HX})_t, \mathbf{1}\rangle.$$

Theorem 3.2. *Consider a $\mathbb{G}$-adapted process Y of the form*

$$Y_t = Y_0 + \int_0^t \alpha_s^Y\, ds + \int_0^t \gamma_s^Y\, dW_s + \int_0^t \left(\beta_s^Y\right)^\top dM_s^X + \int_0^t \delta_s^Y\, dM_s^D\,. \tag{24}$$

Define the diagonal matrices $\Gamma = \mathrm{diag}(g)$ and $\Lambda = \mathrm{diag}(\lambda)$ Then we get with $\beta = \beta^Y$

$$\begin{aligned}(\widehat{YX})_t = (\widehat{YX})_0 &+ \int_0^t (\widehat{\alpha^Y X})_s + A(\widehat{YX})_s ds + \int_0^t \mu_s^{YX} dw_s + \int_0^t \kappa_s^{YX} dm_s^D \\ &+ \sum_{i,j=1}^K \int_0^t \big\langle (\widehat{\beta^j X})_s - (\widehat{\beta^i X})_s, e_i\big\rangle a^{ji} ds (e_j - e_i),\end{aligned} \tag{25}$$

where

$$\mu_s^{YX} = (\widehat{\gamma^Y X})_s + \Gamma(\widehat{YX})_s - \langle \widehat{X}_s, g\rangle (\widehat{YX})_s, \tag{26}$$

$$\kappa_s^{YX} = \frac{1}{\langle \widehat{X}_{s-}, \lambda\rangle}\left(\Lambda\left((\widehat{\delta^Y X})_{s-} + (\widehat{YX})_{s-}\right) - \langle \widehat{X}_{s-}, \lambda\rangle (\widehat{YX})_{s-}\right). \tag{27}$$

Proof. In order to reduce the claim to Theorem 3.1 we need to find the $\mathbb{G}$-semimartingale decomposition of $H = YX$. Note first that $[Y, X]_t = \sum_{0\le s\le t} (\beta_s^\top \Delta X_s)\Delta X_s$. Hence we get from the Itô product formula that

$$\begin{aligned}Y_t X_t = Y_0 X_0 &+ \int_0^t (Y_s A X_s + \alpha_s^Y X_s) ds + \int_0^t \gamma_s^Y X_s dW_s + \int_0^t (Y_{s-} + X_{s-}\beta_s^\top) dM_s^X \\ &+ \int_0^t X_{s-}\delta_s^Y dM_s^D + \sum_{0\le s\le t} (\beta_s^\top \Delta X_s)\Delta X_s.\end{aligned}$$

It is shown in the proof of [Elliott, 1993, Theorem 2] that

$$\sum_{0\le s\le t} (\beta_s^\top \Delta X_s)\Delta X_s = \int_0^t \sum_{i,j=1}^K \langle \beta_s^j X_s - \beta_s^i X_s, e_i\rangle a^{ji}(e_j - e_i) + M_t$$

for some $\mathbb{G}$-martingale M. Hence we may write $H = YX$ in the form (16) with

$$\alpha_s^H = AY_sX_s + \alpha_s^Y X_s + \sum_{i,j=1}^K \langle \beta_s^j X_s - \beta_s^i X_s, e_i\rangle a^{ji}(e_j - e_i),$$
$$\delta_s^H = \delta_s^Y X_{s-} \text{ and } \gamma_s^H = \gamma_s^Y X_{s-}.$$

The claim follows by substituting these identities in Theorem 3.1. In order to illustrate the computational tricks involved we now explain in detail the derivation of κ_s^{YX}, the integrand in the stochastic integral with respect to the compensated point process m^D. With $H = YX$ and hence $\delta_t^H = \delta_t^Y X_t$ we get from Theorem 3.1 that

$$\kappa_t^{YX} = \frac{1}{\langle \widehat{X}_t, \lambda\rangle}\left((\widehat{\langle X_t, \lambda\rangle \delta^H})_t + (\widehat{\langle X, \lambda\rangle H})_t - \langle \widehat{X}_t, \lambda\rangle \widehat{H}_t\right). \tag{28}$$

Moreover, $\langle X_t, \lambda\rangle \delta_t^Y X_t = \sum_{i=1}^K \lambda_i \delta_t^Y \langle X_t, e_i\rangle e_i$ so that

$$(\widehat{\langle X, \lambda\rangle \delta^Y X})_t = \sum_{i=1}^K \lambda_i (\widehat{\delta^Y \langle X, e_i\rangle})_t e_i = \Lambda(\widehat{\delta^Y X})_t\,.$$

Similarly, one gets that $(\widehat{\langle X, \lambda\rangle YX})_t = \sum_{i=1}^K \lambda_i (\widehat{Y\langle X, e_i\rangle})_t e_i = \Lambda(\widehat{YX})_t$, and the form of κ^{YX} follows by plugging these identities in (28). □

In the following, we compute the filters for the quantities needed for the E-Step of the EM algorithm. We begin with the state filter.

Corollary 3.3. *The filtered estimate of the unobserved process X is given by*

$$\widehat{X}_t = \widehat{X}_0 + \int_0^t A\widehat{X}_s ds + \int_0^t \left(\Gamma \widehat{X}_s - \langle \widehat{X}_s, g\rangle \widehat{X}_s\right) dw_s + \int_0^t \left(\frac{\Lambda \widehat{X}_{s-}}{\langle \widehat{X}_{s-}, \lambda\rangle} - \widehat{X}_{s-}\right) dm_s^D.$$

Proof. The result follows from Theorem 3.2 with $Y_t = Y_0 = 1$ and hence with $\alpha^Y = \gamma^Y = \beta^Y = \delta^Y = 0$. □

Next we consider the number of jumps N^{ij} defined in (7). Fix two states $i \neq j$. Since a^{ji} gives the transition intensity from state i to state j and since $1_{\{X_{s-}=e_i\}} 1_{\{X_s = e_j\}} = \langle X_{s-}, e_i\rangle \langle \Delta X_s, e_j\rangle$, the semimartingale decomposition of N^{ij} is given by

$$N_t^{ij} = \int_0^t \langle X_{s-}, e_i\rangle \langle dX_s, e_j\rangle = \int_0^t \langle X_{s-}, e_i\rangle \langle dM_s^X, e_j\rangle + \int_0^t \langle X_{s-}, e_i\rangle a^{ji} ds.$$

This gives the following.

Corollary 3.4. *The filtered estimate for N^{ij} is given by $\widehat{N_t^{ij}} = \langle(\widehat{N^{ij}X})_t, \mathbf{1}\rangle$ where*

$$(\widehat{N^{ij}X})_t = \int_0^t \langle \widehat{X}_s, e_i\rangle a^{ji} e_j ds + \int_0^t A(\widehat{N^{ij}X})_s ds + \int_0^t \left(\frac{\Lambda(\widehat{N^{ij}X})_{s-}}{\langle \widehat{X}_{s-}, \lambda\rangle} - (\widehat{N^{ij}X})_{s-} \right) dm_s^D$$
$$+ \int_0^t \left(\Gamma(\widehat{N^{ij}X})_s - \langle \widehat{X}_s, g\rangle (\widehat{N^{ij}X})_s \right) dw_s.$$

Proof. The result follows if we take in Theorem 3.2 $Y = N^{ij}$ and hence $Y_0 = 0$, $\alpha_s^Y = \langle X_s, e_i\rangle a^{ji}$, $\beta_s^Y = \langle X_s, e_i\rangle e_j$, $\gamma^Y = 0$, $\delta^Y = 0$. To identify the drift of $(\widehat{N^{ij}X})$ we argue as follows: it holds that $\beta_s^\ell = \langle X_s, e_i\rangle \delta_{\ell,j}$ so that $\beta^j X_s = \langle X_s, e_i\rangle e_i$ and $\beta^\ell X_s = 0 \in \mathbb{R}^K$ for $\ell \neq j$. Hence

$$\sum_{k,\ell=1}^K \langle (\widehat{\beta^\ell X})_s, e_k\rangle a^{\ell k} = \langle \widehat{X}_s, e_i\rangle a^{ji}$$

and $\sum_{k,\ell=1}^K \langle (\widehat{\beta^k X})_s, e_k\rangle a^{\ell k} = 0$ so that

$$\sum_{k,\ell=1}^K \int_0^t \left\langle (\widehat{\beta^\ell X})_s - (\widehat{\beta^k X})_s, e_k \right\rangle a^{\ell k} ds (e_\ell - e_k) = \int_0^t \langle \widehat{X}_s, e_i\rangle a^{ji} ds (e_j - e_i)\,.$$

Moreover, $\alpha_s^Y X_s = \langle X_s, e_i\rangle a^{ji} e_i$ and hence $(\widehat{\alpha^Y X})_s = \langle \widehat{X}_s, e_i\rangle a^{ji} e_i$. Plugging these identities into Theorem 3.2 gives the result.

□

Next we consider the occupation time J^i defined in (9).

Corollary 3.5. *The filtered estimate of the occupation time of state e_i is given by $\widehat{J_t^i} = \langle (\widehat{J^i X})_t, \mathbf{1}\rangle$, where*

$$(\widehat{J^i X})_t = \int_0^t \langle \widehat{X}_s e_i\rangle e_i ds + \int_0^t A(\widehat{J^i X})_s ds + \int_0^t \left(\Gamma(\widehat{J^i X})_s - \langle \widehat{X}_s, g\rangle (\widehat{J^i X})_s \right) dw_s$$
$$+ \int_0^t \left(\frac{\Lambda(\widehat{J^i X})_{s-}}{\langle \widehat{X}_{s-}, \lambda\rangle} - (\widehat{J^i X})_{s-} \right) dm_s^D.$$

Proof. Substituting $Y = J^i$ and hence $Y_0 = 0$, $\alpha_s^Y = \langle X_s, e_i\rangle$, $\beta_s^Y = 0 \in \mathbb{R}^K$, $\gamma^Y = 0$ and $\delta_s^Y = 0$ in Theorem 3.2 yields the result. □

Next we turn to the level integral defined in (8).

Corollary 3.6. *The filtered estimate of G_t^i is given by $\widehat{G_t^i} = \langle(\widehat{G^iX})_t, \mathbf{1}\rangle$, where*

$$(\widehat{G^iX})_t = g^i \int_0^t \langle \widehat{X}_s, e_i\rangle e_i ds + \int_0^t A(\widehat{G^iX})_s ds + \int_0^t \left(\frac{\Lambda(\widehat{G^iX})_{s-}}{\langle \widehat{X}_{s-}, \lambda\rangle} - (\widehat{G^iX})_{s-} \right) dm_s^D + \int_0^t \left(\langle \widehat{X}_s, e_i\rangle e_i + \Gamma(\widehat{G^iX})_s - \langle \widehat{X}_s, g\rangle (\widehat{G^iX})_s \right) dw_s.$$

Proof. Note that $G_t^i = g^i \int_0^t \langle X_s, e_i\rangle ds + \int_0^t \langle X_s, e_i\rangle dW_s$. Hence the claim follows from Theorem 3.2 if we let $Y_t = G_t^i$ and hence $Y_0 = 0$, $\alpha_s^Y = g^i\langle X_s, e_i\rangle$, $\gamma_s^Y = \langle X_s, e_i\rangle$, $\beta^Y = 0 \in \mathbb{R}^K$ and $\delta_s^Y = 0$. □

Finally, we consider the jump level integral $B_t^i := \int_0^t \langle X_s, e_i\rangle dD_s$ and the modified occupation time C^i that were introduced in (10) and (11).

Corollary 3.7. *The filtered estimate of C_t^i and B_t^i are given by $\widehat{C_t^i} = \langle(\widehat{C^iX})_t, \mathbf{1}\rangle$ and $\widehat{B_t^i} = \langle(\widehat{B^iX})_t, \mathbf{1}\rangle$, where*

$$(\widehat{C^iX})_t = \int_0^t h_s(D)\langle \widehat{X}_s, e_i\rangle e_i ds + \int_0^t A(\widehat{C^iX})_s ds + \int_0^t \Gamma(\widehat{C^iX})_s - \langle \widehat{X}_s, g\rangle (\widehat{C^iX})_s dw_s + \int_0^t \frac{1}{\langle \widehat{X}_{s-}, \lambda\rangle} \left(\Lambda(\widehat{C^iX})_{s-} - \langle \widehat{X}_{s-}, \lambda\rangle (\widehat{C^iX})_{s-} \right) dm_s^D,$$

$$(\widehat{B^iX})_t = \int_0^t h_s(D)\lambda^i\langle \widehat{X}_s, e_i\rangle e_i ds + \int_0^t A(\widehat{B^iX})_s ds + \int_0^t \Gamma(\widehat{B^iX})_s - \langle \widehat{X}_s, g\rangle (\widehat{B^iX})_s dw_s + \int_0^t \frac{1}{\langle \widehat{X_{s-}}, \lambda\rangle} \left(\Lambda\big(\langle \widehat{X}_{s-} e_i\rangle e_i + (\widehat{B^iX})_{s-}\big) - \langle \widehat{X}_{s-}, \lambda\rangle (\widehat{B^iX})_{s-} \right) dm_s^D.$$

Proof. In order to compute $(\widehat{B^iX})_t$, we take $Y_t = B_t^i$, $Y_0 = 0$, $\alpha_s^Y = h_s(D)\lambda^i\langle X_s, e_i\rangle$, $\gamma_s^Y = 0$, $\beta^Y = 0 \in \mathbb{R}^K$ and $\delta_s^Y = \langle X_s, e_i\rangle$ and apply Theorem 3.2. To obtain $(\widehat{C^iX})_t$ we take $Y_t = C_t^i$, $Y_0 = 0$, $\alpha_s^Y = h_s(D)\langle X_s, e_i\rangle$, $\gamma_s^Y = 0$, $\beta^Y = 0 \in \mathbb{R}^K$ and $\delta_s^Y = 0$ and apply Theorem 3.2. □

3.2 Unnormalized Filters

In this section we derive so-called unnormalized filters for the quantities arising in the E-Step of the EM algorithm. The resulting filtering equations are linear and driven directly by the observation processes Z and D. Moreover, unnormalized filters are needed for the derivation of robust filters in Section 3.3 below.

Denote by $\mathbb{P}^*$ the so-called *reference probability measure* on $(\Omega, \mathcal{G})$. That is, under $\mathbb{P}^*$, Z is a Brownian motion and D is a Poisson process with unit intensity, independent of X. Let

$$\frac{d\mathbb{P}}{d\mathbb{P}^*}\Big|_{\mathcal{G}_t} = L_t = 1 + \int_0^t L_s g(X_s) Z_s + \int_0^t L_{s-}(\lambda(X_{s-})h_{s-}(D) - 1)\,(dD_s - ds).$$

It follows from the Girsanov theorem that under $\mathbb{P}$, Z and D have the correct joint law. For any $\mathbb{G}$-adapted and integrable process Y we denote the unnormalized conditional expectation by

$$\sigma(Y)_t = \mathbb{E}^*[L_t Y_t | \mathcal{F}_t]. \tag{29}$$

From Bayes' rule, we have $\widehat{Y}_t = \sigma_t(Y)/\sigma_t(1)$. In what follows our objective is to derive the Zakai equation (the dynamics of the unnormalized conditional expectation (29)). The first step towards this goal is to derive $\sigma_t(1) = \mathbb{E}^*[L_t|\mathcal{F}_t]$.

Lemma 3.8. *The dynamics of $\sigma_t(1)$ are given by*

$$\sigma_t(1) = 1 + \int_0^t \sigma_s(1)\langle g, \widehat{X}_s\rangle dZ_s + \int_0^t \sigma_{s-}(1)\big(\langle \lambda, \widehat{X}_{s-}\rangle h_s(D) - 1\big) d(D_s - s). \tag{30}$$

Proof. The proof follows similar arguments as in [Elliott, 1993, Thm 3]: we use the fact that the process L is a $(\mathbb{P}^*, \mathbb{G})$ martingale so that a version of Theorem 3.1 applies with $Y_r = L_r$ $\alpha^Y = 0$, $\gamma_r^Y = L_r\langle g, X_r\rangle$, $\beta^Y = 0$, $\delta_r^Y = L_r(\langle \lambda, X_r\rangle h_r(D) - 1)$, Z a Brownian motion and $D_t - t$ a martingale. Then we use Bayes' rule and obtain the result. □

Now we are ready to prove the main theorem of this section.

Theorem 3.9. *Consider a $\mathbb{G}$-adapted process Y of the form (24). Then, with $\beta = \beta^Y$, it holds*

$$\begin{aligned}\sigma_t(YX) = \sigma_0(YX) &+ \int_0^t \sigma_s(\alpha^Y X)ds + \int_0^t A\sigma_s(YX)ds \\ &+ \sum_{i,j=1}^K \int_0^t \left\langle \sigma_s(\beta^j X) - \sigma_s(\beta^i X), e_i \right\rangle a^{ji} ds(e_j - e_i) \\ &+ \int_0^t \sigma_s(\gamma^Y X) + \Gamma\sigma_s(YX)dZ_s \\ &+ \int_0^t \left(h_s(D)\Lambda\sigma_{s-}(\delta^Y X) + (h_s(D)\Lambda - I)\sigma_{s-}(YX)\right)(dD_s - ds).\end{aligned}$$

Proof. It follows from Bayes' formula that $\sigma_t(YX) = \widehat{YX}_t\sigma_t(1)$. Hence, we apply Itô's product rule for jump diffusions and write

$$d\sigma_t(YX) = \sigma_{t-}(1)d(\widehat{YX})_t + (\widehat{YX})_{t-}d\sigma_t(1) + d[\sigma(1), \widehat{YX}]_t. \tag{31}$$

Then, inserting (25) and (30) in (31), using Bayes' formula and making the necessary cancellations, we obtain the result; the details are omitted. □

Using unnormalized filters, $\widehat{Y}_t$ can be computed from $\sigma_t(YX)$ and $\sigma_t(X)$ as follows:

$$\widehat{Y}_t = \langle(\widehat{YX})_t, \mathbf{1}\rangle = \frac{\langle\sigma_t(YX), \mathbf{1}\rangle}{\sigma_t(1)} = \frac{\langle\sigma_t(YX), \mathbf{1}\rangle}{\langle\sigma_t(X), \mathbf{1}\rangle}. \tag{32}$$

Next we compute the unnormalized filters for the various quantities of interest. In what follows, we use the simpler notation $q_t = \sigma(X)_t$.

Corollary 3.10. *The dynamics of the unnormalized filter for X (the Zakai equation) are given by*

$$q_t = q_0 + \int_0^t Aq_s ds + \int_0^t \Gamma q_s dZ_s + \int_0^t (h_s(D)\Lambda - I)q_{s-}(dD_s - ds). \tag{33}$$

Proof. We set $Y_t = 1$, $\alpha_r^Y = \beta_r^Y = \gamma_r^Y = \delta_r^Y = 0$ and we apply Theorem 3.9. □

Corollary 3.11. *We have the following unnormalized filters:*

$$\begin{aligned}\sigma_t(N^{ij}X) &= \int_0^t \langle q_s, e_i\rangle a^{ji} e_j \, ds + \int_0^t A\sigma_s(N^{ij}X)\, ds + \int_0^t \Gamma\sigma_s(N^{ij}X)\, dZ_s \\ &\quad + \int_0^t (h_s(D)\Lambda - I)\sigma_{s-}(N^{ij}X)\,(dD_s - ds) \\ \sigma_t(J^iX) &= \int_0^t \langle q_s, e_i\rangle e_i ds + \int_0^t A\sigma_s(J^iX) ds + \int_0^t \Gamma\sigma_s(J^iX) dZ_s \\ &\quad + \int_0^t (h_s(D)\Lambda - I)\sigma_{s-}(J^iX)\,(dD_s - ds),\end{aligned}$$

$$\begin{aligned}\sigma_t(G^iX) &= g^i\int_0^t \langle q_s, e_i\rangle e_i \, ds + \int_0^t A\sigma_s(G^iX)\, ds + \int_0^t (\Gamma\sigma_s(G^iX) + \langle q_s, e_i\rangle e_i)\, dZ_s \\ &\quad + \int_0^t (h_s(D)\Lambda - I)\sigma_{s-}(G^iX)\,(dD_s - ds), \\ \sigma_t(B^iX) &= \lambda^i \int_0^t h_s(D)\langle q_s, e_i\rangle e_i \, ds + \int_0^t A\sigma_s(B^iX)\, ds + \int_0^t \Gamma\sigma_s(B^iX)\, dZ_s \\ &\quad + \int_0^t \left((h_s(D)\Lambda - I)\sigma_{s-}(B^iX) + h_s(D)\Lambda\langle q_{s-}, e_i\rangle e_i\right)\,(dD_s - ds), \\ \sigma_t(C^iX) &= \int_0^t h_s(D)\langle q_s, e_i\rangle e_i ds + \int_0^t A\sigma_s(C^iX) ds + \int_0^t \Gamma\sigma_s(C^iX) dZ_s \\ &\quad + \int_0^t (h_s(D)\Lambda - I)\sigma_{s-}(C^iX)(dD_s - ds)\,.\end{aligned}$$

Proof. In order to obtain these results, we use an analogous reasoning as in the proof of Corollary 3.10. □

3.3 Robust Filters and discretization

In this section, our objective is to derive *robust* filters in the sense of Clark [1978] and James et al. [1996]. These filters are Lipschitz continuous "in the observation process", so that they perform well if applied to a situation where Z and D are only approximately of the form (2) and (3); a case in point is that of discrete observations, see Remark 2.1. In order to derive these filters one transforms the filter dynamics in such a way that they involve a minimal number of stochastic integrals.

Robust filters

Throughout this section we assume that $h_t(D) > 0$ for all t; see however Remark 3.12, point 3, below. Following the approach of James et al. [1996] and of Elliott and Malcolm [2008] we first define

$$\Pi_t^i = \exp\left\{ g^i Z_t - \frac{1}{2}(g^i)^2 t + (1 - h_t(D)\lambda^i)t + D_t \log(h_t(D)\lambda^i) \right\}$$

and we let

$$\Pi_t = \operatorname{diag}\left\{\Pi_t^1, \dots, \Pi_t^K\right\} = \exp\left\{ \Gamma Z_t - \frac{1}{2}\Gamma^2 t + (I - h_t(D)\Lambda)t + D_t \Lambda_t^L \right\},$$

where $\Lambda_t^L = \operatorname{diag}\left\{\log(h_t(D)\lambda^1), \dots, \log(h_t(D)\lambda^K)\right\}$. The corresponding dynamics are $d\Pi_t = \Pi_t \Gamma dZ_t + \Pi_t(h_t(D)\Lambda - I)(dD_t - dt)$, and the Itô formula yields

$$\begin{aligned} d\Pi_t^{-1} &- \Pi_t^{-1}\Gamma dZ_t + \Pi_t^{-1}\Gamma^2 dt \\ &- \Pi_t^{-1}\left(I - \frac{1}{h_t(D)}\Lambda^{-1} \right) dD_t + \Pi_t^{-1}(h_t(D)\Lambda - I)dt. \end{aligned}$$

For any $\mathbb{G}$-adapted, integrable process Y we define

$$\overline{\sigma}_t(YX) = \Pi_t^{-1}\sigma_t(YX). \tag{34}$$

It follows from Itô's product formula that $\overline{q}_t := \overline{\sigma}(X_t)$ has the dynamics

$$\frac{d}{dt}\overline{q}_t = \Pi_t^{-1} A \Pi_t \overline{q}_t; \tag{35}$$

in particular, $\overline{q}_t$ is a process of finite variation. The unnormalized filter is then given by $\sigma_t(X) = \Pi_t \overline{q}_t$. Note that in order to compute $\sigma_t(X)$ in this way we only have to discretize the ODE (35) and to evaluate Z and D at given time points;

it is not necessary to approximate a stochastic integral driven by these processes (as one would have to do in a naive discretization of the Zakai equation (33)).

In a similar vein $\overline{\sigma}_t(YX)$ can be computed for the other quantities needed for the EM algorithm. We obtain

$$d\overline{\sigma}_t(J^iX) = \langle \overline{q}_t, e_i\rangle e_i\,dt + \Pi_t^{-1}A\Pi_t\overline{\sigma}_t(J^iX)\,dt. \tag{36}$$
$$d\overline{\sigma}_t(N^{ij}X) = \langle \overline{q}_t, e_i\rangle\langle Ae_i, e_j\rangle e_j\,dt + \Pi_t^{-1}A\Pi_t\overline{\sigma}_t(N^{ij}X)\,dt, \tag{37}$$
$$d\overline{\sigma}_t(G^iX) = \langle \overline{q}_t, e_i\rangle e_i\,dZ_t + \Pi_t^{-1}A\Pi_t\overline{\sigma}_t(G^iX)\,dt, \tag{38}$$
$$d\overline{\sigma}_t(B^iX) = \langle \overline{q}_t, e_i\rangle e_i\,dD_t + \Pi_t^{-1}A\Pi_t\overline{\sigma}_t(B^iX)\,dt, \tag{39}$$
$$d\overline{\sigma}_t(C^iX) = h_t(D)\langle \overline{q}_t, e_i\rangle e_i\,dt + \Pi_t^{-1}A\Pi_t\overline{\sigma}_t(C^iX)\,dt. \tag{40}$$

Discretization.

For the numerical implementation we need to discretize the filter equations. We now explain how to do this for the robust filters derived above. In what follows we will consider the partition $0 = t_0 < t_1 < \cdots < t_N = T$ on the interval $[0, T]$, and we let $\Delta_n = t_n - t_{n-1}$. The easiest approach to discretize the ODE for $\overline{\sigma}_t$ is to use a simple explicit Euler scheme, see also Elliott and Malcolm [2008][Section D]. If we apply this to the ODE (35) for $\overline{q}_t$ we obtain

$$\overline{q}_{t_n} \approx \overline{q}_{t_{n-1}} + \Pi_{t_{n-1}}^{-1}A\Pi_{t_{n-1}}\Delta_n\overline{q}_{t_{n-1}}.$$

In order to obtain the unnormalized filter we multiply both sides with Π_{t_n}:

$$q_{t_n} \approx \Pi_{t_n}\Pi_{t_{n-1}}^{-1}(I + A\Delta_n)q_{t_{n-1}}. \tag{41}$$

Hence, (41) gives a way for the recursive estimation of the unnormalized state probabilities. Now we define $z_n^\Delta = Z_{t_n} - Z_{t_{n-1}}$, apply the same procedure and obtain the following recursions for the remaining filters:

$$\begin{aligned}
\sigma_{t_n}(G^iX) &\approx \Pi_{t_n}\Pi_{t_{n-1}}^{-1}\left((I + A\Delta_n)\sigma_{t_{n-1}}(G^iX) + \langle q_{t_{n-1}}, e_i\rangle e_i z_n^\Delta\right),\\
\sigma_{t_n}(J^iX) &\approx \Pi_{t_n}\Pi_{t_{n-1}}^{-1}\left((I + A\Delta_n)\sigma_{t_{n-1}}(J^iX) + \langle q_{t_{n-1}}, e_i\rangle e_i\Delta_n\right),\\
\sigma_{t_n}(N^{ij}X) &\approx \Pi_{t_n}\Pi_{t_{n-1}}^{-1}\left((I + A\Delta_n)\sigma_{t_{n-1}}(N^{ij}X) + \langle q_{t_{n-1}}, e_i\rangle\langle Ae_i, e_j\rangle e_j\Delta_n\right),\\
\sigma_{t_n}(B^iX) &\approx \Pi_{t_n}\Pi_{t_{n-1}}^{-1}\left((I + A\Delta_n)\sigma_{t_{n-1}}(B^iX) + \langle q_{t_{n-1}}, e_i\rangle e_i(D_{t_n} - D_{t_{n-1}})\right),\\
\sigma_{t_n}(C^iX) &\approx \Pi_{t_n}\Pi_{t_{n-1}}^{-1}\left((I + A\Delta_n)\sigma_{t_{n-1}}(C^iX) + h_t(D)\langle q_{t_{n-1}}, e_i\rangle e_i\Delta_n\right).
\end{aligned}$$

Remark 3.12. 1) Note that when the filters are represented in integral form, it is possible to make use of the integration by parts formula and convert the

stochastic integrals with respect to the processes D and Z into integrals with respect to time. For example,

$$\overline{\sigma}_t(B^i X) = \langle \overline{q}_t, e_i\rangle e_i D_t - \int_0^t D_s \langle d\overline{q}_s, e_i\rangle e_i + \int_0^t \Pi_s^{-1} A \Pi_s \overline{\sigma}_s(B^i X)\, ds. \tag{42}$$

This can be used to obtain a robust filter for B^i that does not contain a stochastic integral with repect to D. In a similar vein, partial integration can be used to eliminate the stochastic integral with respect to Z in the robust filter for G^i.

2) There are other ways to discretize the ODE part in the robust filter equations (35) and (36) – (40) that can be advantageous if the generator A' of X has been constructed as a discrete approximation of a diffusion process. In that case Clark [1978] recommends an implicit discretization of the ODEs.

3) If $h_t(D) \equiv 0$ on some stochastic interval (τ_1, τ_2) for $\mathbb{F}$-stopping times τ_1 and τ_2 one works with the solution of the matrix SDE $d\Pi_t = \Gamma \Pi_t dZ_t$ instead; this leads to the robust filters derived in James et al. [1996].

4 Goodness-of-fit Tests

In this section we propose several statistical tests for the hypothesis that the hidden Markov model from Section 2.1, parameterized in terms of a (estimated) parameter vector θ^*, models the observed data (Z, D) well. These tests are based on two observations: first,

$$w_t = Z_t - \int_0^t \langle g^*, \widehat{X}_s\rangle ds \text{ is a } \mathbb{F}\text{-Brownian motion;} \tag{43}$$

second, D is a point process with $\mathbb{F}$-intensity $\lambda_t^* := h_t(D)\langle \lambda^*, \widehat{X}_t\rangle dt$ (both under $\mathbb{P}_{\theta^*}$). Define the time change $\mathcal{T}(t) = \int_0^t \lambda_s^* ds$ and suppose that $\lim_{t\to\infty} \mathcal{T}(t) = \infty$ a.s. Denote by $\mathcal{T}^{-1}(t) = \inf\{s \geq 0, \mathcal{T}(s) \geq t\}$ the inverse transform. Then it holds that the process $\widetilde{D}$ defined by

$$\widetilde{D}_t = D \circ \mathcal{T}^{-1}(t),\ 0 \leq t \leq \mathcal{T}(T),\ \text{is a standard Poisson process.} \tag{44}$$

The hypotheses (43) and (44) can be tested in various ways; this leads to a number of goodness-of-fit tests for our setup. Note that in (43) and (44) the filter $\widehat{X}_t$ and the time change $\mathcal{T}$ and $\mathcal{T}^{-1}$ are computed under $\mathbb{P}_{\theta^*}$, so that all components of θ^* enter into the testing procedure.

Testing the Brownian-motion hypothesis

Fix some time interval $\bar{\Delta}$ and let $t_k = k\bar{\Delta}$, $k = 0, 1, \ldots, \kappa := \lfloor T/\bar{\Delta} \rfloor$. Define the random variables

$$V_k = w_{t_k} - w_{t_{k-1}} \approx Z_{t_k} - Z_{t_{k-1}} - \langle g^*, \widehat{X}_{t_{k-1}} \rangle \bar{\Delta}.$$

Under $\mathbb{P}_{\theta^*}$, $\{V_k\}_{k \geq 1}$ is a sequence of independent, $N(0, \bar{\Delta})$-distributed random variables. All of this can be tested: a standard t-test can be performed to test the hypothesis that the V_k have mean zero, that is that the drift estimate $\langle g^*, \widehat{X}_t \rangle$ is correct 'on average'. The normality assumption can be tested graphically via a QQ-plot or numerically using for instance a Kolmogorov Smirnov goodness-of-fit test. The independence assumption (which implies zero autocorrelation at all lags) can be assessed graphically via correlograms or numerically using for instance a Ljung-Box test. Tests of the independence assumption are in fact particularly important, as this provides a check if the model captures the dynamics of Z well.

Testing the Poisson-process hypothesis

If $\widetilde{D}$ is standard Poisson the random variables $U_k = \widetilde{D}_{t_k} - \widetilde{D}_{t_{k-1}}$, $k = 1, \ldots, \kappa$, are iid Poisson with parameter $\bar{\Delta}$. This implies that the random variables $\widetilde{U}_k = U_k \wedge 1$, $k = 1, \ldots, \kappa$ are Bernoulli distributed with parameter $p = 1 - \exp(-\bar{\Delta})$, which can be tested with a standard binomial test. Moreover, one can test if the inter-arrival times of the jumps of $\widetilde{D}$ follow a standard exponential distribution (as they should under the Poisson hypothesis). This can be tested graphically via a QQ-plot or numerically using Kolmogorov Smirnov. Testing the exponentiality of the inter-arrival times of $\widetilde{D}$ is very useful to check if the model is able to 'decluster' the jumps of D and hence to capture the dynamics of D reasonably well. Note that the tests of the Poisson hypothesis are closely related to tests for the accuracy of Value at Risk models in market risk management, see for instance Section 9.3. of McNeil et al. [2015]. For numerical illustrations of the proposed goodness-of-fit tests we refer to the example in Section 5.

5 Simulation Analysis

In this section we present the results from a simulation study that tests the speed, efficiency and accuracy of the various algorithms and methods introduced so far. This analysis is crucial for the fine-tuning of the methods and it serves as

a bridge between theoretical results and practical implementation. We discuss the performance of the EM algorithm (in Section 5.1), the advantage of robust filtering algorithms (Section 5.2) and the performance of the goodness-of-fit tests (Section 5.3).

Throughout we consider a 3-state Markov chain and we use the parameter vector from Table 1 to generate data sets. The stepsize is taken as $\Delta_n = \frac{1}{500}$, and we use $N = 20\,000$ observations.

a^{12}	a^{13}	a^{21}	a^{23}	a^{31}	a^{32}	$\tilde{g}^1$	$\tilde{g}^2$	$\tilde{g}^3$	λ^1	λ^2	λ^3	σ_Z
0.2	**0.3**	**0.3**	**0.2**	**0.2**	**0.3**	**-0.5**	**0**	**0.5**	**0.6**	**1**	**4**	**0.05**

Table 1: Parameters used in simulation study.

5.1 EM Algorithm

In this section we illustrate the performance of the EM algorithm. For this we fix a parameter vector θ, an initial distribution p, some noise variance σ_Z^2 and we generate trajectories of size N with step size Δ_n for the Markov chain X, the continuous observation Z and the point process D. Given these data and an initial parameter vector θ^0 we run several iterations of the algorithm and we stop as soon as the relative change in the parameter values is lower than a given tolerance level. Here the following issue arises. Recall that the theoretical results are derived assuming that the Brownian motion W has unit variance. When this is not the case, but the noise variance has a known value, it suffices to normalize Gaussian observations before initiating the EM algorithm. In practice, though, this value is typically unknown. However, as explained in James et al. [1996][Section VI-B], in a discretized setting it is possible to obtain an MLE estimate of the noise variance. Following this, throughout this section we assume that the noise variance is unknown and we estimate it accordingly.

In Table 2 we present a typical outcome of this analysis. The table contains a comparison between the true parameters, the parameters estimated via the EM algorithm and the MLE estimates (the estimates computed in the hypothetical case where the trajectory of the chain is observable). For the EM algorithm the termination tolerance is set to 1% and the starting values are 50% of the true ones. The final estimate for the volatility is given by $\sigma_Z^{\mathrm{MLE}} = 0.05054923$. The evolution of the estimates for the generator of X (in dependence of the number of

iterations m of the EM algorithm) is shown in Figure 1. Our analysis shows that in this case the performance of the algorithm is reasonable. Only the estimates for the generator matrix of the chain are somewhat off from their true values. However, these parameters are difficult to estimate as can be seen from the fact that the MLE estimates also deviate from the true parameter value by a similar amount.

Parameters	a^{12}	a^{13}	a^{21}	a^{23}	a^{31}	a^{32}
True	**0.20000**	**0.30000**	**0.30000**	**0.20000**	**0.20000**	**0.30000**
EM	**0.10275**	**0.18030**	**0.21879**	**0.14359**	**0.39011**	**0.22059**
MLE	**0.09608**	**0.14972**	**0.34329**	**0.14972**	**0.34329**	**0.09608**
Parameters	$\tilde{g}^1$	$\tilde{g}^2$	$\tilde{g}^3$	λ^1	λ^2	λ^3
True	**-0.50000**	**0.00000**	**0.50000**	**0.60000**	**1.00000**	**4.00000**
EM	**-0.48445**	**-0.03426**	**0.48943**	**0.52917**	**1.00920**	**3.45954**
MLE	**-0.49259**	**-0.03862**	**0.48687**	**0.51493**	**1.00884**	**3.44363**

Table 2: True parameters, EM estimates and MLE estimates (estimates for the hypothetical case where X is observable).

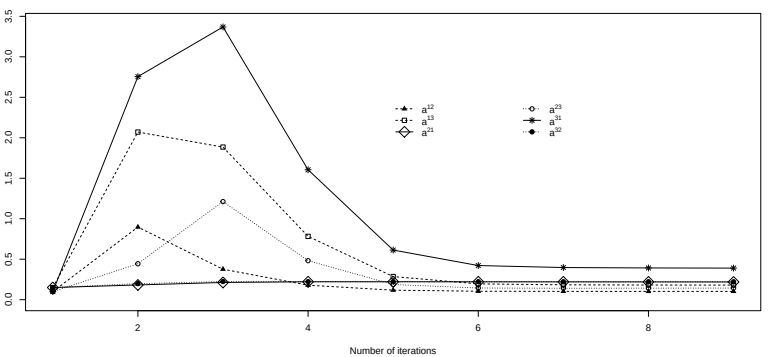

Fig. 1: Evolution of parameter estimates for matrix A.

5.2 Effect of Robust Discretization

Performing a robust discretization of the filters allows one to obtain much smoother estimates of the quantities of interest in compared to those obtained when directly discretizing the exact filters using the Euler-Maruyama method. Naturally, the larger the discretization step size (Δ_n in our notation), the more evident this effect. The robust discretization effect is illustrated in Figure 2, where we focused on state filters for the data set generated with the parameters given in Table 1. For visualization purposes, only the first 5 000 points are shown.

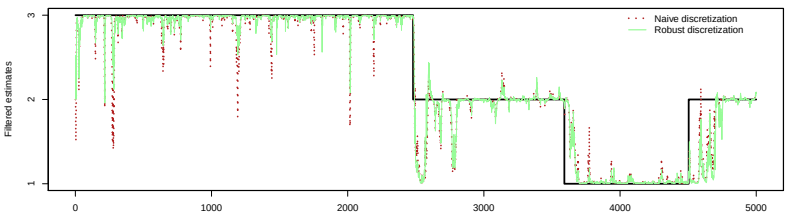

Fig. 2: Naive discretization of exact state filters vs. robust discretization, discretization step $\Delta_n = \frac{1}{500}$

5.3 Tests for Model Validation

In this section we provide a numerical illustration of the goodness-of-fit tests proposed in Section 4. For this we compare two cases. Case 1 corresponds to the situation where all parameters have been estimated correctly. In Case 2 we assume that the generator matrix of X has been estimated correctly while the estimates for λ and g are constant across states and given by $\lambda^{*,j} = \langle \lambda, \pi \rangle$ for all j and $g^{*,j} = \langle g, \pi \rangle$ for all j, where π is the stationary distribution of the Markov chain X. This parameter choice implies that the increments of w and of the time transformed process $\widetilde{D}$ have the correct mean. However, the model misses the autocorrelation caused by the randomness in the drift of w and the clustering in the jump times of D. Hence we expect that the t test and the binomial test do not reject Case 2, but the tests for independence respectively for the exponentiality of the inter-arrival times should lead to a rejection.

To test this conjecture we again use a data set generated with the parameters given in Table 1. The left plot in Figure 3 shows the correlogram of the increments of w for Case 1 (correct parameters); the right plot shows the correlogram for Case 2. We see that working erroneously with a constant g induces significant autocorrelation at all lags. Next we turn to the Poisson hypothesis. Here the null hypothesis of the Kolmogorov-Smirnov test for exponentiality is rejected in Case 2 (p-value: 0.02201), but not in Case 1 (p-value: 0.8934). Figure 4 provides a graphical illustration: the left QQ-plot corresponds to the correctly estimated model, the right one to Case 2. The more erratic behavior in the latter is evident. The null hypothesis of the standard binomial test is not rejected in both cases (the p-values are 0.5484364 and 0.4388406, respectively).

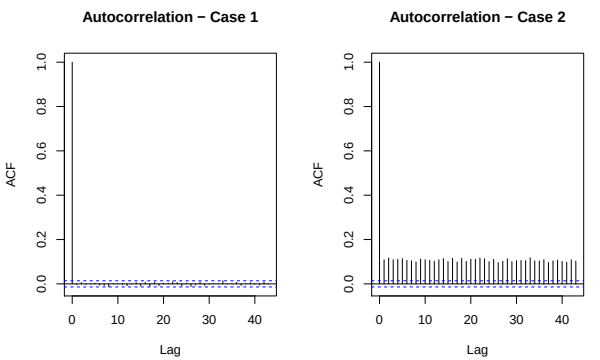

Fig. 3: Comparison of correlograms: Case 1 (left) and Case 2 (right).

6 Application to credit risk

Implementation

We applied the HMM for credit quality described in Example 2.2 to a real data set consisting of five US corporations, considering seven non-default rating categories. We assume that the model parameters are identical for all firms, but that signal and observation for different firms are independent. This implies that the log-likelihood of the observations is the sum of the likelihoods for the different firms, and the EM parameter updates are easily computed. In fact, this

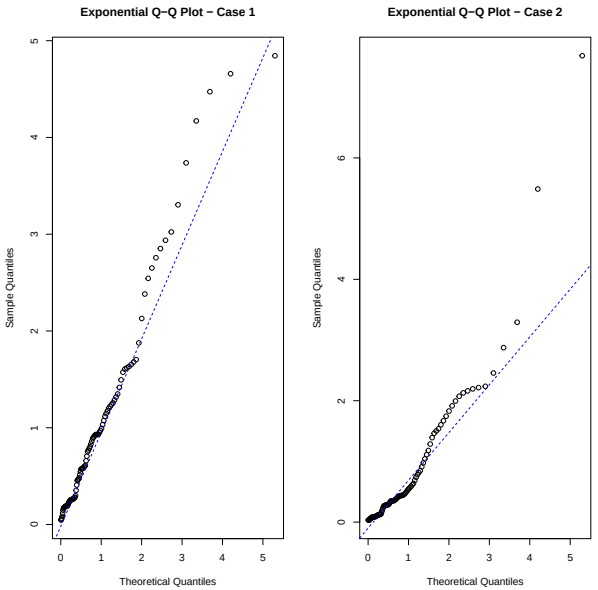

Fig. 4: Comparison of QQ-plots for inter-arrival times: Case 1 (left) and Case 2 (right).

assumption leads to the *filter-based cohort approach* proposed in Korolkiewicz and Elliott [2008]. We use normalized continuous observations as input, where the normalizing factor is given by an average of the volatility estimates for each of the five companies.

We introduce a couple of restrictions on the parameters: first, we assume for simplicity that the Markov chain X that models the true credit quality can jump only to neighboring states; second, since we observe only one default in our data set, we do not estimate λ^d but instead we keep it fixed and take it equal to

$$\lambda^d = (0.00005, 0.00020, 0.00060, 0.00180, 0.01116, 0.04134, 0.17972)^\top.$$

These values are derived from estimated default rates of different rating classes, as given in McNeil et al. [2015], Table 10.2. Since the intensity of the up-and downgrade processes depends on the observable ratings, we need a slight extension of the EM methodology developed in Section 2 and 3; details are discussed in Appendix A.

Results

Since the state space S of the observed rating R^i and of the hidden true credit quality X are taken identical, we identify S with the rating categories we consider (AAA, AA, A, BBB, BB, B, CCC-C). The estimates for the *transition rates* between the different states of X are given in Table 3. Note that transition intensities to non-neighbor states are zero by assumption. Overall the estimates appear reasonable; the large transition rates between the highest categories (labelled AAA, AA and A) are probably due to the fact that given the limited amount of data the algorithm is not able to distinguish clearly between the three classes.

	AAA	AA	A	BBB	BB	B	CCC-C
AAA	-5.9814	5.9814	0	0	0	0	0
AA	6.6849	-15.5480	8.8630	0	0	0	0
A	0	18.6123	-20.9233	2.3109	0	0	0
BBB	0	0	0.3700	-0.6641	0.2943	0	0
BB	0	0	0	0.3492	-0.3493	0.0000	0
B	0	0	0	0	0.1122	-0.1122	0.0000
CCC-C	0	0	0	0	0	0.0000	0.0000

Table 3: Estimated generator matrix $A^\top$.

The estimates for the *drift coefficients* g are given in Table 4. These estimates are monotonously increasing, in line with the stylized fact that credit spreads are higher when the credit quality of a firm is worse.

g_{AAA}	g_{AA}	g_A	g_{BBB}	g_{BB}	g_B	g_{CCC-C}
33.97179	34.44579	35.06211	39.61831	48.75934	64.17487	99.36986

Table 4: Estimated drift coefficients

The estimates for the *up- and downgrade intensities* λ^+ and λ^- are given in Table 5. Note that the elements of the estimated vectors λ^+ and λ^- respect the expected ordering (decreasing for λ^+ and increasing for λ^-), vindicating the intuition that observed ratings follow the true credit quality albeit with some rating error.

λ_1^+	λ_2^+	λ_3^+	λ_1^-	λ_2^-	λ_3^-
0.47943	**0.20457**	**0.00042**	**0.05020**	**0.12984**	**0.27643**

Table 5: Estimates for λ^+ and λ^-

We also applied the *goodness-of-fit* tests described in Section 4. The results for the Poisson process hypothesis were quite satisfactory: the p value of the Kolmogorov-Smirnov test for the exponentiality of the interarrival times of the time transformed processes $\widetilde{D}^+$ and $\widetilde{D}^-$ was 0.2125 for upgrades and 0.5685 for downgrades D^-; the p value of the binomial test was 0.1645 for upgrades and 0.255 for downgrades. The tests for the Brownian motion hypothesis were a bit more problematic, essentially because the observed log-credit spreads show a very strong degree of autocorrelation; we omit the details.

Finally we use the model to compute a filter estimate for the unobservable true credit quality of a given firm. In Figure 5 and in Figure 6 we graph the estimated credit quality for Medtronic and Abbott, together with the observed (logarithmic) CDS spread and the observed rating. The analysis shows that the estimated credit quality balances the impact of both sources of information (ratings and CDS spreads).

7 Conclusion

In this work we study an EM algorithm for the setting where the state variable follows a Markov chain observed via diffusive *and* point processes information. On the theoretical side, we derived the dynamics for the exact and the unnormalized filters, and we computed discretized, robust versions of the filters in the sense of Clark [1978]. Moreover, we proposed several goodness-of-fit tests for hidden Markov models with Gaussian noise and point process observation. On the applied side we carried out a simulation analysis to test the performance of our methodology, and we considered an application to credit risk: we estimate the parameters of a HMM for credit quality where the observations consist of rating transitions and credit spreads for five US corporations. Our work opens interesting avenues for future empirical work such as an analysis of sovereign credit spreads and of contagion effects between sovereigns or parameter estimation in hidden Markov models for high frequency data in finance.

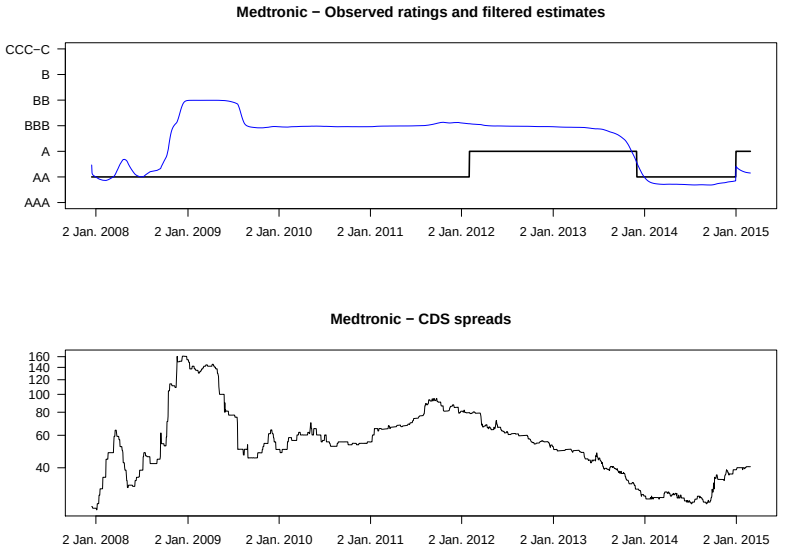

Fig. 5: Medtronic - Observed ratings and filtered estimate for credit quality (top) and observed credit spread on log-scale (bottom).

A EM algorithm for Example 2.2

In what follows we are going to provide the steps of the EM algorithm corresponding to Example 2.2. To this, we need to define new processes due to the dependence of the variates λ_t^+ and λ_t^- on the rating observation R_t^i. Namely, we define the processes C^{jk} and B^{jk}, $1 \leq j, k \leq K$ with the following:

$$B_t^{+,jk} = \int_0^t 1_{\{R_s = e_j\}} \langle X_s, e_k \rangle dD_s^+ \quad \text{and} \quad C_t^{jk} = \int_0^t 1_{\{R_s = e_j\}} \langle X_s, e_k \rangle ds.$$

We can define $B^{-,jk}$ in a similar fashion. Now we have the following likelihood function

$$L(\theta, \theta') = \cdots + \int_0^t \log(\lambda^+(X_s, R_s))\, dD_s^+ - \int_0^t \lambda^+(X_s, R_s)\, ds \tag{45}$$

$$+ \int_0^t \log(\lambda^-(X_s, R_s))\, dD_s^- - \int_0^t \lambda^-(X_s, R_s)\, ds + R(\theta'). \tag{46}$$

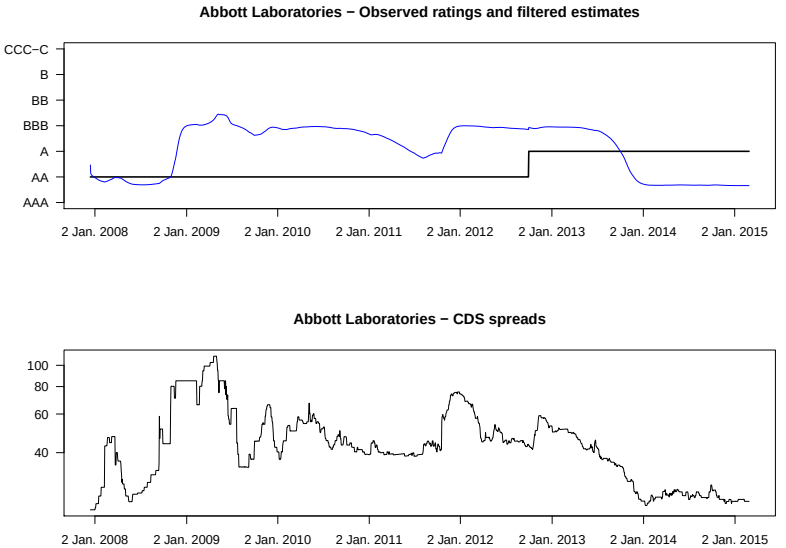

Fig. 6: Abbott - Observed ratings and filtered estimate for credit quality(top) and observed credit spread on log-scale (bottom).

Note that we can write $\lambda^+(X_s, R_s) = \sum_{i=1}^K \sum_{j=1}^K \lambda^{+,ij} 1_{\{R_s=e_j\}} \langle X_s, e_i \rangle$. Hence

$$L(\lambda^+, \lambda^{+\prime}) = \int_0^t \sum_{k=1}^K \sum_{j=1}^K \log(\lambda^{+,jk}) 1_{\{R_s=e_j\}} \langle X_s, e_k \rangle \, dD_s^+ - \int_0^t \sum_{k=1}^K \sum_{j=1}^K \lambda^{+,jk} 1_{\{R_s=e_j\}} \langle X_s, e_k \rangle \, ds + R(\lambda^{+\prime}).$$

Thus we have

$$L(\lambda^+, \lambda^{+\prime}) = \sum_{k=1}^K \sum_{j=1}^K \log(\lambda^{+,jk}) B_t^{+,jk} - \sum_{k=1}^K \sum_{j=1}^K \lambda^{+,jk} C_t^{jk} + R(\lambda^{+\prime}). \tag{47}$$

Next we write the filtered estimate of the log-likelihood function:

$$L(\widehat{\lambda^+, \lambda^{+\prime}}) = \sum_{k=1}^K \sum_{j=1}^K \log(\lambda^{+,jk}) \widehat{B_t^{+,jk}} - \sum_{k=1}^K \sum_{j=1}^K \lambda^{+,jk} \widehat{C_t^{jk}} + R(\lambda^{+\prime}). \tag{48}$$

Hence, we have what is needed for the E-step. Let us now use the parametrization

$$\lambda^{+,jk} = \lambda_1^+ 1_{\{k<j\}} + \lambda_2^+ 1_{\{k=j\}} + \lambda_3^+ 1_{\{k>j\}}, \quad 1 \le j, k \le K, k > 1. \tag{49}$$

Hence

$$L(\widehat{\lambda^+,\lambda^{+\prime}}) = \sum_{k=2}^{K}\sum_{j=1}^{K} \log(\lambda_1^+ 1_{\{k<j\}} + \lambda_2^+ 1_{\{k=j\}} + \lambda_3^+ 1_{\{k>j\}})\widehat{B_t^{+,jk}}$$
$$- \sum_{k=2}^{K}\sum_{j=1}^{K} (\lambda_1^+ 1_{\{k<j\}} + \lambda_2^+ 1_{\{k=j\}} + \lambda_3^+ 1_{\{k>j\}})\widehat{C_t^{jk}} + R(\lambda^{+\prime}).$$

From the first order conditions we then obtain the following estimates

$$\widehat{\lambda_1^+} = \frac{\sum_{j=1}^{K}\sum_{1<k<j}^{K}\widehat{B_t^{+,jk}}}{\sum_{j=1}^{K}\sum_{1<k<j}^{K}\widehat{C_t^{jk}}}, \quad \widehat{\lambda_2^+} = \frac{\sum_{k=2}^{K}\widehat{B_t^{+,kk}}}{\sum_{k=2}^{K}\widehat{C_t^{kk}}}.$$
$$\widehat{\lambda_3^+} = \frac{\sum_{j=1}^{K}\sum_{k>j}^{K}\widehat{B_t^{+,jk}}}{\sum_{j=1}^{K}\sum_{k>j}^{K}\widehat{C_t^{jk}}}.$$

To apply the algorithm we need to obtain the filtered estimates for the quantities C_t^{jk} and B_t^{jk}, and their robust version. These are computed exactly as in Section 3.

References

R. J. Elliott and L. Aggoun and J. B. Moore. *Hidden Markov Models: Estimation and Control*. Springer, New York, 1995.

R. S. Mamon and R. J. Elliott, editor. *Hidden Markov Models in Finance*, International Series in Operations Research and Management Science, 2007. Springer.

R. S. Mamon and R. J. Elliott, editor. *Hidden Markov Models in Finance: Further Developments and Applications, Volume II*, International Series in Operations Research and Management Science, 2014. Springer.

S. Asmussen. Risk theory in a Markovian environment. *Scandinavian Actuarial Journal*, pages 69–100, 1989.

A. Berndt, R. Douglas, D. Duffie, F. Ferguson, and D. Schranz. Measuring Default Risk Premia from Default Swap rates and EDFs, 2008.

P. Brémaud. *Point Processes and Queues, Martingale Dynamics*. Springer, 1981.

J. M. C. Clark. The design of robust approximations to the stochastic differential equations of nonlinear filtering. *Communication systems and random process theory*, 25:721–734, 1978.

K. Colaneri, Z. Eksi, R. Frey, and M. Szoelgyenyi. Shall I sell or shall I wait: Optimal liquidation under partial information with market impact. preprint, Vienna University of Economics and Business, 2017.

R. Cont. Statistical modeling of high-frequency financial data. *IEEE Signal Processing Magazine*, 28(5):16–25, 2011.

A. Dembo and O. Zeitouni. Parameter estimation of partially observed continuous time stochastic processes via the EM algorithm. *Stochastic Processes and Their Applications*, 23(1):91–113, 1986.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, pages 1–38, 1977.

R. J. Elliott. New finite-dimensional filters and smoothers for noisily observed Markov chains. *Information Theory, IEEE Transactions on*, 39(1):265–271, 1993.

R. J. Elliott and W. P. Malcolm. Discrete-time expectation maximization algorithms for Markov-modulated poisson processes. *IEEE Transactions on Automatic Control*, 53(1): 247–256, 2008.

R. Frey and W. Runggaldier. A nonlinear filtering approach to volatility estimation with a view towards high frequency data. *International Journal of theoretical and Applied Finance*, 4:199–210, 2001.

R. Frey and T. Schmidt. Pricing and hedging of credit derivatives via the innovations approach to nonlinear filtering. *Finance and Stochastics*, 16(1):105–133, 2012.

M. R. James, V. Krishnamurthy, and F. Le Gland. Time discretization of continuous-time filters and smoothers for HMM parameter estimation. *IEEE Transactions on Information Theory*, 42(2):593–605, 1996.

M. W. Korolkiewicz and R. J. Elliott. A hidden Markov model of credit quality. *Journal of Economic Dynamics and Control*, 32(12):3807–3819, 2008.

A. J. McNeil, R. Frey, and P. Embrechts. *Quantitative Risk Management: Concepts, Techniques and Tools*. Princeton University Press, 2nd edition, 2015.

P. E. Protter. *Stochastic Integration and Differential Equations*. Springer, 2013.

U. Rieder and N. Bäuerle. Portfolio optimization with unobservable Markov-modulated drift process. *Journal of Applied Probability*, 43:362–378, 2005.

J. Sass and U. G. Haussmann. Optimizing the terminal wealth under partial information: The drift process as a continuous time Markov chain. *Finance and Stochastics*, 8(4):553–577, 2004.

H.P. Schmidli. Cramér Lundberg approximations for ruin probabilities of risk processes perturbed by diffusion. *Insurance: Mathematics and Economics*, 16:135–149, 1995.