EM algorithm for Markov chains under mixed observations and applications to credit risk

Zehra Eksi   Camilla Damian   Rüdiger Frey

Stochastic Dynamical Models in Mathematical Finance,
Econometrics, and Actuarial Sciences,
Lausanne, May 2017

Introduction

- Partial observation models are frequently used in finance and insurance $\Rightarrow$ parameter estimation in these models is of high relevance.
- Examples:
  - Credit risk: unobservable default intensity or credit quality of obligors (corporates or sovereigns)
  - Insurance: unobservable claims-arrival intensity or mortality rate
  - High frequency data: unobservable 'state of the market' that is affected by trading activity of others
- EM algorithm is a possible approach for parameter estimation under partial information; particularly useful if unobservable state variable can be approximated by a finite state Markov chain

Our contributions

- Extending Elliott [1993] and Elliott and Malcolm [2008], we obtain an EM algorithm for a setting in which the state variable follows a finite-state Markov chain observed via diffusive and point process observation; this is quite relevant for the applications mentioned beforehand.
- In such setting, we derive the corresponding exact, unnormalized and *robust* (in the sense of Clark [1978] and James et al. [1996]) filters needed in the E step.
- We propose *goodness of fit tests* and we run an extensive *simulation study*.
- We present a *case study* with rating data

Markov Chain

- We consider a finite time interval $[0, T]$ and a continuous-time finite-state Markov chain X defined on $(\Omega, \mathcal{G}, \mathbb{G}, \mathbb{P})$.
- X has the state space $S = \{e_1, e_2, ..., e_K\}$ where, without loss of generality, we assume e_k is the basis column vector of $\mathbb{R}^K$.
- The initial distribution is $\pi = (\pi^1, \ldots, \pi^K)$.
- The transpose of the infinitesimal generator is $A = (a^{ij})$, $i, j \in \{1, \ldots, K\}$.
- Accordingly, we define

$$M_t^X := X_t - X_0 - \int_0^t AX_s ds.$$

Clearly, M^X is a $\mathbb{G}$-martingale.

X is not directly observable !

Observation Processes

Diffusion information. We consider the noisy observation process

$$Z_t = \int_0^t g(X_s)ds + W_t. \quad (1)$$

Often Z is constructed from discrete observations on timescale Δ.
Consider $z_n = \widetilde{g}(X_{t_n}) + \epsilon_n$ for $\{\epsilon_n\}$ iid mean zero with variance σ_ϵ^2.
Define *scaled cumulative observations process*

$$\widetilde{Z}_t := \Delta \sum_{t_n \le t} z_n = \sum_{t_n \le t} \Delta \widetilde{g}(X_{t_n}) + \Delta \sum_{t_n \le t} \epsilon_n \,. \quad (2)$$

Then $\tilde{Z}_t \approx \int_0^t \tilde{g}(X_s)ds + \sigma_\epsilon \sqrt{\Delta} W_t$ (as in (1) after normalisation).

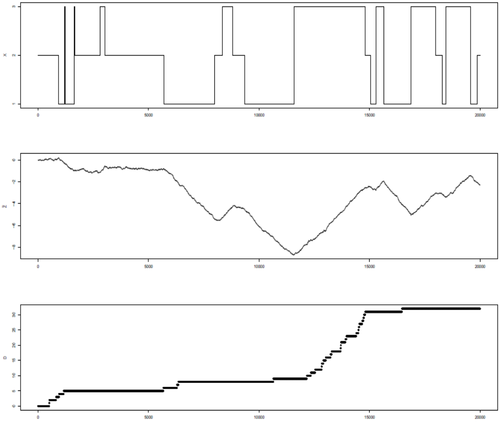
Point process. Second source of information is a point process D with $\mathbb{G}$-intensity $\lambda(X_t)$. Hence we have the $\mathbb{G}$-martingale

$$M_t^D = D_t - \int_0^t \lambda(X_s)ds, \quad t \le T.$$

Graphical illustration

Parameter set: $N = 20000$, $\Delta_n = \frac{1}{500}$, $\sigma_z = 0.2$, $\tilde{g} = (-1, 0, 1)^\top$, $\lambda = (0.2, 1, 3)^\top$, $(a^{12}, a^{13}, a^{21}, a^{23}, a^{31}, a^{32}) = (0.3, 0.1, 0.1, 0.2, 0.2, 0.2)$ and $h \equiv 1$.

Figure: Markov chain, Gaussian observation, point process observation



Information and estimation problem

- The information available to the observer of the system is carried by $\mathbb{F} = \mathbb{F}^{\mathbb{Z}} \vee \mathbb{F}^{\mathbb{D}}$. Note that $\mathcal{F}_t \subset \mathcal{G}_t$.
- For an integrable and measurable process Y, $\widehat{Y}_t$ denotes the $\mathbb{F}$-optional projection, that is $\widehat{Y}_t = \mathbb{E}[Y_t|\mathcal{F}_t]$, for every t. T
- For a generic function f it holds that $f(X_t) = \langle X_t, \mathbf{f} \rangle$ where $\langle \ , \ \rangle$ denotes the scalar product and $f_k = f(e_k)$, $1 \leq k \leq K$.
- Hence, the unobserved parameters to be estimated are given by the vector

$$\theta = \{a_{jk}, g_j, \lambda_j, \quad j,k \in \{1,\dots,K\}\}.$$

EM Methodology: General Description

- Assume that measures corresponding to different parameters θ, θ' are equivalent on $\mathcal{G}_T$ (full information!). Define the corresponding full-information log-likelihood:

$$L(\theta, \theta') := \log \frac{d\mathbb{P}_\theta}{d\mathbb{P}_{\theta'}}\Big|_{\mathcal{G}_T}.$$

EM Methodology: General Description

- Assume that measures corresponding to different parameters θ, θ' are equivalent on $\mathcal{G}_T$ (full information!). Define the corresponding full-information log-likelihood:

$$L(\theta, \theta') := \log \frac{d\mathbb{P}_\theta}{d\mathbb{P}_{\theta'}}\Big|_{\mathcal{G}_T}.$$

- Let θ^m be the optimal set of parameters after the m^{th} iteration of the algorithm. Then, iteration $m+1$ of the EM algorithm consists of the following two main steps:
  - **Expectation (E)**: compute filtered estimate

$$\widehat{L(\theta, \theta^m)} = \mathbb{E}_{\theta^m}\left[\log \frac{d\mathbb{P}_\theta}{d\mathbb{P}_{\theta^m}}\Big|\mathcal{F}_T\right].$$

  - **Maximization (M)**: find $\theta^{m+1} \in \underset{\theta}{\operatorname{argmax}}\, \widehat{L(\theta, \theta^m)}$.

EM Methodology: Current Setting

Define for $i, j \in \{1, \ldots, K\}$, the following quantities:

- $N_t^{ij} = \sum_{0<s\leq t} 1_{\{X_{s-}=e_i\}} 1_{\{X_s=e_j\}}$, *(number of jumps)*
- $G_t^i = \int_0^t \langle X_s, e_i \rangle dZ_s$, *(level integral)*
- $J_t^i = \int_0^t \langle X_s, e_i \rangle ds$, *(occupation time)*
- $B_t^i = \int_0^t \langle X_s, e_i \rangle dD_s$, *(jump level integral)*

EM Methodology: Current Setting

Define for $i, j \in \{1, \ldots, K\}$, the following quantities:

- ▷ $N_t^{ij} = \sum_{0<s\leq t} 1_{\{X_{s-}=e_i\}} 1_{\{X_s=e_j\}}$, *(number of jumps)*
- ▷ $G_t^i = \int_0^t \langle X_s, e_i \rangle dZ_s$, *(level integral)*
- ▷ $J_t^i = \int_0^t \langle X_s, e_i \rangle ds$, *(occupation time)*
- ▷ $B_t^i = \int_0^t \langle X_s, e_i \rangle dD_s$, *(jump level integral)*

E Step. By Girsanov, the filtered estimate for the log-likelihood is

$$\widehat{L(\theta, \theta^m)} = \mathbb{E}_{\theta^m}\left[\log \frac{d\mathbb{P}_\theta}{d\mathbb{P}_{\theta^m}} \Big| \mathcal{F}_T\right] = \sum_{i,j=1, i\neq j}^{K} \left(\widehat{N_T^{ij}} \log a^{ji} - a^{ji}\widehat{J_T^i}\right) + \sum_{i=1}^{K}\left(g^i \widehat{G_T^i} - \frac{1}{2}(g^i)^2 \widehat{J_T^i}\right)$$

$$+ \sum_{i=1}^{K}\left(\log(\lambda^i)\widehat{B_T^i} - \lambda^i \widehat{J_T^i}\right) + \widehat{R}(\theta^m).$$

EM Methodology: Current Setting

Define for $i, j \in \{1, \dots, K\}$, the following quantities:

- ▷ $N_t^{ij} = \sum_{0<s\leq t} 1_{\{X_{s-}=e_i\}} 1_{\{X_s=e_j\}}$, *(number of jumps)*
- ▷ $G_t^i = \int_0^t \langle X_s, e_i \rangle dZ_s$, *(level integral)*
- ▷ $J_t^i = \int_0^t \langle X_s, e_i \rangle ds$, *(occupation time)*
- ▷ $B_t^i = \int_0^t \langle X_s, e_i \rangle dD_s$, *(jump level integral)*

E Step. By Girsanov, the filtered estimate for the log-likelihood is

$$\widehat{L(\theta, \theta^m)} = \mathbb{E}_{\theta^m}\left[\log \frac{d\mathbb{P}_\theta}{d\mathbb{P}_{\theta^m}} \middle| \mathcal{F}_T\right] = \sum_{i,j=1, i\neq j}^{K} \left(\widehat{N_T^{ij}} \log a^{ji} - a^{ji} \widehat{J_T^i}\right) + \sum_{i=1}^{K} \left(g^i \widehat{G_T^i} - \frac{1}{2}(g^i)^2 \widehat{J_T^i}\right)$$

$$+ \sum_{i=1}^{K} \left(\log(\lambda^i) \widehat{B_T^i} - \lambda^i \widehat{J_T^i}\right) + \widehat{R}(\theta^m).$$

M-Step. FOCs give new parameter set:

$$(a^{ji})^{m+1} = \frac{\widehat{N_T^{ij}}}{\widehat{J_T^i}}, \quad (g^i)^{m+1} = \frac{\widehat{G_T^i}}{\widehat{J_T^i}}, \quad (\lambda^i)^{m+1} = \frac{\widehat{B_T^i}}{\widehat{J_T^i}}.$$

Unnormalized Filters

Reference Probability Measure

- We work under the so-called *reference probability measure* $\mathbb{P}^*$ on $(\Omega, \mathbb{G})$. Under $\mathbb{P}^*$,
  - Z is a Brownian motion,
  - D is a Poisson process with unit intensity, independent of X.
- $P << P^*$ with density $\frac{d\mathbb{P}}{d\mathbb{P}^*}\Big|_{\mathcal{G}_t}$
- For any $\mathbb{G}$-adapted, integrable process Y define the *unnormalized conditional expectation* by

$$\sigma_t(Y) = \mathbb{E}^*[Y_t \mid \mathcal{F}_t];$$

by Bayes it holds that $\widehat{Y}_t = \frac{\sigma_t(Y)}{\sigma_t(1)}$

Unnormalized Filters

Theorem 3.1 (Main Result)

Consider a $\mathbb{G}$-adapted process Y of the form

$$Y_t = Y_0 + \int_0^t \alpha_s^Y ds + \int_0^t \gamma_s^Y dW_s + \int_0^t \left(\beta_s^Y\right)^\top dM_s^X + \int_0^t \delta_s^Y d(D_s - s).$$

Let $\Gamma = diag(g)$, $\Lambda = diag(\lambda)$ and I the identity matrix. Then

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

On Theorem3.1

Comments

- The resulting filtering equations are linear and (for appropriate $\alpha^Y, \beta^Y, \gamma^Y, \delta^Y$) recursive.
- They are driven by observation processes Z and D.

On Theorem3.1

Comments

- The resulting filtering equations are linear and (for appropriate $\alpha^Y, \beta^Y, \gamma^Y, \delta^Y$) recursive.
- They are driven by observation processes Z and D.

Corollary 3.1 (**Zakai equation**)

The unnormalized filter for the unobserved state of the Markov chain is given by ($q_t := \sigma_t(X)$)

$$q_t = q_0 + \int_0^t A q_s ds + \int_0^t \Gamma q_s dZ_s + \int_0^t (\Lambda - I) q_{s-} d(D_s - s)$$

Similar expressions for other quantities of interest

Robust Filters

Goal. Derive versions of the unnormalized filters that depend *continuously* on observations. $\Rightarrow$ transform the filter dynamics such that the resulting expressions involve a minimal number of stochastic integrals. (Clark [1978])

Robust Filters

Goal. Derive versions of the unnormalized filters that depend *continuously* on observations. $\Rightarrow$ transform the filter dynamics such that the resulting expressions involve a minimal number of stochastic integrals. (Clark [1978])

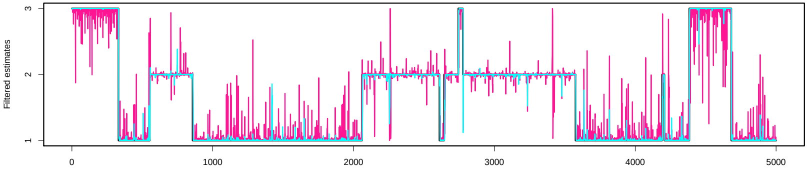


Figure: Naive discretization of exact state filters (pink) vs. robust discretization (cyan), using $\Delta_n = \frac{1}{100}$

Goodness of Fit Tests

Goal. Find tests for the hypothesis that the model, parameterized in terms of $\theta^* = \left(a^{jk,*}, g^{j,*}, \lambda^{j,*}, j,k \in \{1,\ldots,K\}, j \neq K\right)$, models the observed data (Z, D) well.

Two testable observations.

1. $w_t = Z_t - \int_0^t \langle g^*, \widehat{X}_s \rangle ds$ is a $\mathbb{P}_{\theta^*}$-Brownian motion;
2. Define $\mathcal{T}(t) := \int_0^t \lambda^*(\hat{X}_s) ds$. Then the process $\widetilde{D}$ with

$$\widetilde{D}_t := D \circ \mathcal{T}^{-1}(t),\ 0 \leq t \leq \mathcal{T}(T),$$

is a standard Poisson process under $\mathbb{P}_{\theta^*}$.

Potential Tests

Brownian motion hypothesis:

- QQ-plot and Kolmogorov-Smirnov for normality.
- Correlograms.

Potential Tests

Brownian motion hypothesis:

- QQ-plot and Kolmogorov-Smirnov for normality.
- Correlograms.

Poisson process hypothesis:

- QQ-plot and Kolmogorov-Smirnov for exponentiality of inter-arrival times.

Potential Tests

Brownian motion hypothesis:

- QQ-plot and Kolmogorov-Smirnov for normality.
- Correlograms.

Poisson process hypothesis:

- QQ-plot and Kolmogorov-Smirnov for exponentiality of inter-arrival times.
- If $\widetilde{D}$ is standard Poisson,
  - → $U_k := \widetilde{D}_{t_k} - \widetilde{D}_{t_{k-1}}$, $k = 1, \ldots, \kappa$, are i.i.d. Poisson with parameter $\bar{\Delta}$.
  - → Hence, the rvs $\widetilde{U}_k = U_k \wedge 1$, $k = 1, \ldots, \kappa$ are Bernoulli with parameter $p = 1 - \exp(-\bar{\Delta})$.
  - → To check this, one can employ a standard Binomial test.

Simulation Procedure

Algorithm

1. Fix a parameter set θ, an initial distribution π, some noise variance σ_z^2 and generate trajectories of size N with step size Δ_n for the Markov chain X, Brownian motion W and the point process D. Obtain the corresponding observation series $(\tilde{Z}, D)$.
2. Run the EM algorithm and obtain estimates for the hidden states, as well as for the parameters:
   - **I:** Initialize the algorithm with some parameter set θ^0 and σ_z^0.
   - **N:** Normalize the data by σ_z^m.
   - **E:** Obtain the filtered estimates of the quantities of interest.
   - **M:** Compute θ^{m+1} and σ_z^{m+1}.
   - **T:** Terminate if $\frac{|\theta^{m+1}-\theta^m|}{\theta^m}$ and $\frac{|\sigma_z^{m+1}-\sigma_z^m|}{\sigma_z^m}$ are below the termination tolerance; else return to step N.

Performance of EM

- Parameter set: $N = 20000$, $\Delta_n = \frac{1}{100}$, $\sigma_z = 0.1$, $\tilde{g} = (0.01, 0.8, 1.3)^\top$, $\lambda = (2, 8, 10)^\top$, $(a^{12}, a^{13}, a^{21}, a^{23}, a^{31}, a^{32}) = (0.2, 0.7, 0.3, 0.2, 0.2, 0.2)$.

	J_T^1	J_T^2	J_T^3	G_T^1	G_T^2
Actual	92.86	73.13	34.01	9.39	58.63
Filters	94.54	70.95	34.52	10.17	57.88
Relative error %	1.80	2.98	1.50	8.30	1.28
	G_T^3	B_T^1	B_T^2	B_T^3	N_T^{12}
Actual	43.70	180.00	521.00	265.00	26.00
Filters	45.02	169.49	530.53	265.98	22.95
Relative error %	3.03	5.84	1.83	0.37	11.73
	N_T^{21}	N_T^{13}	N_T^{31}	N_T^{23}	N_T^{32}
Actual	17.00	16.00	26.00	18.00	9.00
Filters	16.95	15.60	22.59	19.77	13.78
Relative error %	0.27	2.51	13.10	9.84	53.07

- The final parameter estimates read:

$$(a^{12}, a^{13}, a^{21}, a^{23}, a^{31}, a^{32})^{EM} = (0.239, 0.655, 0.243, 0.399, 0.165, 0.279),$$
$$(g^1, g^2, g^3)^{EM} = (1.075, 8.158, 13.042)^\top,$$
$$(\lambda^1, \lambda^2, \lambda^3)^{EM} = (1.793, 7.478, 7.705)^\top,$$

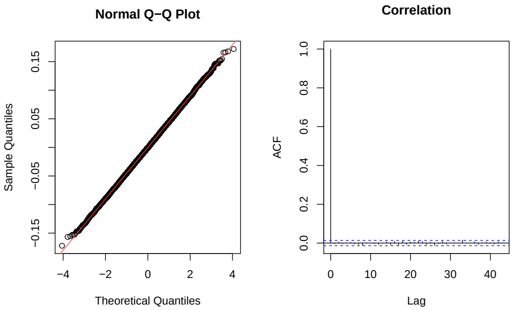
Approach works well under favorable circumstances!

Simulation Analysis: Tests

Comparison of 2 models

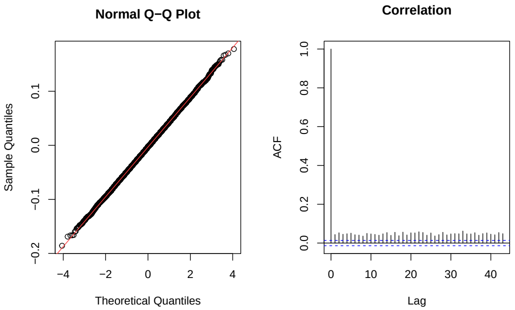
- *Correctly* specified model
- *Weighted average case*. Here generator A is specified correctly, but $\lambda^{*,j} = \langle \lambda, \pi \rangle \quad \forall j$ and $g^{*,j} = \langle g, \pi \rangle \quad \forall j$, π the *stationary distribution* of X. (constant accross states)
- Weighted average case has on average correct drift and correct number of jumps, but misspecifies *dependence structure* of data

Diffusion tests, correct model



Parameter set: $N = 20000$, $\Delta_n = \frac{1}{500}$, $\sigma_z = 0.1$, $\tilde{g} = (0.01, 0.8, 1.3)^\top$, $\lambda = (0.6, 1, 4)^\top$, $(a^{12}, a^{13}, a^{21}, a^{23}, a^{31}, a^{32}) = (0.1, 0.1, 0.1, 0.1, 0.1, 0.1)$.

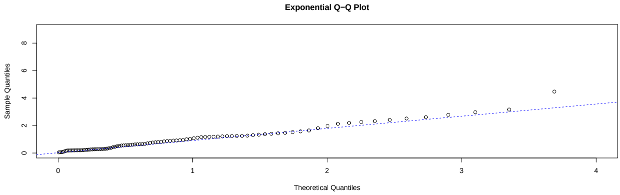
Diffusion tests, weighted average case

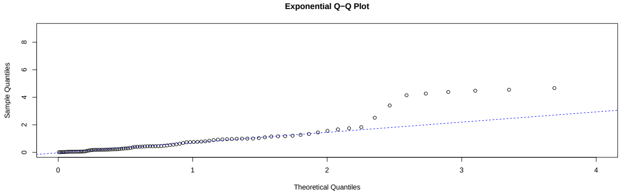


Parameter set: $N = 20000$, $\Delta_n = \frac{1}{500}$, $\sigma_z = 0.1$, $\tilde{g} = (0.01, 0.8, 1.3)^\top$, $\lambda = (0.6, 1, 4)^\top$, $(a^{12}, a^{13}, a^{21}, a^{23}, a^{31}, a^{32}) = (0.1, 0.1, 0.1, 0.1, 0.1, 0.1)$.

Test for Point Process

Correctly Estimated Model vs. 'Weighted Average' Case





Corresponding Kolmogorov-Smirnov p-values: 0.8122 and 0.0001351.

A Hidden Markov Model for Credit Quality

We consider m rated firms with CDS contracts. Ratings and CDS spreads form available information.

State Process. X_t^i is *true* credit quality of firm i at t;

- Finite state Markov chain; generator matrix A^t identical accross firms;
- e_1 is *best* credit quality, e_K is the *worst* (non-default) state; $e_l \geq e_k$ whenever $l > k$

Continuous observation (CDS spreads). Let $z_n^i = \log(CDS_{t_n}^i)$, (log of CDS spread of firm i at t_n).Assume that

$$z_n^i = \tilde{g}(X_{t_n}^i) + \epsilon_n^i, \text{ for } \epsilon_n^i, 1 \leq n \leq N, 1 \leq i \leq m, \text{ iid} \qquad (3)$$

Identifying (3) with a continuous model gives Z^i.

Point Process observation (ratings)

- $R_t^i \in S$ *observed* rating of firm i at time t.
- For simplicity only three types of events possible: upgrading (by one category); downgrading; default.
- $\Rightarrow$ Dynamics of R^i described by three point processes:
  - $D_t^{+,i}$ (number of upgradings of firm i up to time t)
  - $D_t^{-,i}$ (number of downgradings of firm i up to time t)
  - $D_t^{d,i}$ (default indicator of firm i)
- Intensities. Idea: observed rating tracks 'true' credit quality, possibly with rating error. We take

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

we expect $\lambda_1^+ > \lambda_2^+ > \lambda_3^+$ and $\lambda_1^- < \lambda_2^- < \lambda_3^-$.

Estimation

Methodology

- We considered 5 rating categories, 7 American firms
- Assumption: parameters identical across firms but signal and observation are independent across firms.
- For simplicity we imposed next neighbour dynamics for X,
- Slight extension of previous methodology necessary since intensities depend on observable rating.

Estimated generator $Q = A'$.

	AAA-A	BBB	BB	B	CCC-C
AAA-A	-0.04390	0.04390	0	0	0
BBB	0.43039	-1.06006	0.62967	0	0
BB	0	0.80777	-0.80777	0.00000	0
B	0	0	0.11317	-0.11317	0.00000
CCC-C	0	0	0	0.00000	0.00000

Results ctd

Estimated drifts g.

g_{AAA-A}	g_{BBB}	g_{BB}	g_{B}	g_{CCC-C}
35.52313	41.83595	50.60478	64.23868	99.37381

Estimated λ^+.

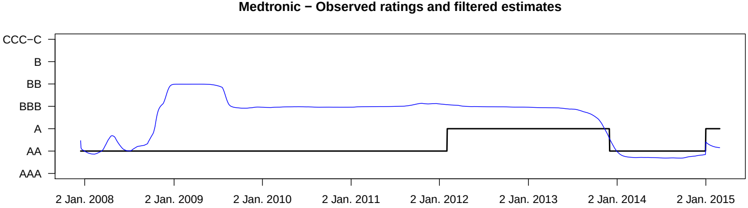
λ_1^+	λ_2^+	λ_3^+
1.00008	0.08778	0.00000

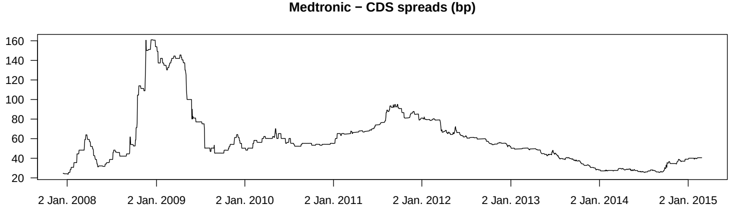
Estimated λ^-.

λ_1^-	λ_2^-	λ_3^-
0.00000	0.04386	0.29266

Parameter estimates seem reasonable!

Observed and estimated credit quality for Medtronic





References

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

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

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

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

Thank you!