Edinburgh Research Explorer

Capturing Model Risk and Rating Momentum in the Estimation of Probabilities of Default and Credit Rating Migrations

Capturing Model Risk and Rating Momentum in the Estimation of Probabilities of Default and Credit Rating Migrations

Abstract:	We discuss the estimation of rating transition probabilities in Markov and non-Markov setups. We first estimate a continuous time Markov chain using discrete (missing) data and derive a simpler expression for the Fisher information matrix, reducing the computation time of Wald confidence intervals to less than half the current standard. We provide an efficient procedure transferring such uncertainties to the rating migrations and probabilities of default, which is of practical use. Supported by a full data-set, we propose a tractable and parsimonious self-exciting marked point processes model capturing the non-Markovian effect of rating momentum. Compared to the Markov model, the non-Markov model yields higher probabilities of default in the investment grades, but also lower default probabilities in some speculative grades. This agrees with empirical observations and has clear practical implications. We use \emph{Moody's proprietary corporate credit ratings data set}. Implementation available in the R package \emph{ctmcd}.

SCHOLARONE™
Manuscripts

Capturing Model Risk and Rating Momentum in the Estimation of Probabilities of Default and Credit Rating Migrations

G. dos Reis†‡*, M. Pfeuffer¶ and G. Smith† §

†School of Mathematics, Maxwell Institute for Mathematical Sciences, The University of
Edinburgh, James Clerk Maxwell Building, Peter Guthrie Tait Road, Edinburgh, EH9 3FD, UK
‡Centro de Matemática e Aplicações (CMA), FCT, UNL, Portugal
¶University of Erlangen-Nüremberg, Department of Statistics and Econometrics, Lange Gasse
20, 90403 Nuremberg
§Moody's Analytics, 7 Exchange Crescent Conference Square, Edinburgh EH3 8RD

We present two methodologies on the estimation of rating transition probabilities within Markov and non-Markov frameworks. We first estimate a continuous-time Markov chain using discrete (missing) data and derive a simpler expression for the Fisher information matrix, reducing the computational time needed for the Wald confidence interval by a factor of a half. We provide an efficient procedure for transferring such uncertainties from the generator matrix of the Markov chain to the corresponding rating migration probabilities and, crucially, default probabilities.

For our second contribution, we assume access to the full (continuous) data set and propose a tractable and parsimonious self-exciting marked point processes model able to capture the non-Markovian effect of rating momentum. Compared to the Markov model, the non-Markov model yields higher probabilities of default in the investment grades, but also lower default probabilities in some speculative grades. Both findings agree with empirical observations and have clear practical implications.

We use *Moody's proprietary corporate credit rating data set*. Parts of our implementation are available in the R package *ctmcd*.

1. Introduction

Credit risk modelling and financial regulations have received added attention from Mathematics and Economics disciplines since the 2008 financial crash. On January 1, 2018, and for purposes of risk assessment, the new guideline IFRS 9 took effect requiring the calculation of expected losses for the complete maturity of certain obligors' riskier contracts. Thereby, a cornerstone of credit risk modelling lies in the ability to accurately estimate probabilities of default over varying

Corresponding author. Email: G.dosReis@ed.ac.uk

time horizons. This can be either done by considering market data (e.g. bond or credit default swap prices, as well as implied probabilities of default from equities, see Bielecki *et al.* (2011)) or historical (default or rating) data. In this manuscript, we focus on the latter.

REMARK 1.1 (Obtaining Default Probabilities: Risk-neutral Vs real-world) *It is important to note the distinction between these estimation methods. Using market data such as bond prices or credit default swaps to estimate default probabilities actually gives risk neutral default probabilities. Our approach uses observed data and therefore gives real-world (physical) default probabilities. Our results can be used without any adjustment in capital requirement calculations where real-world default probabilities are needed.*

When estimating probabilities of default, it is typical that credit ratings are considered in the calculation, as they allow for more granularity. Ratings, as categorical solvency measures, might be issued by (external) rating agencies or be produced by the financial institutes themselves as part of the Pillar I internal ratings-based approach underpinning the Basel regulatory framework. Due to idiosyncratic company-level or general business-cycle changes, credit ratings vary over time, and this effect is referred to as a *rating transition* or *rating migration*. This dynamical movement is a stochastic process with a discrete state space in continuous-time. Here Markov chains are a simple, robust and tractable class to model the movement of such rating transitions. The specific models that can be used depend on the type of data available.

Most literature dealing with the modelling of credit rating transitions focuses on anonymous discrete-time data and often on an annual basis. This data is easier to use and less costly to obtain than the "full" (continuous-time company specific rating transitions) data set. In the discretely observed data case, it is not possible to follow individual obligors over the different periods which forces one to treat all companies in the same rating as equivalent. Hence, one is naturally led to a Markov chain construct (continuous or discrete). Assuming a continuous-time Markov chain (CTMC) model that has been observed only at specific discrete points, obtaining the maximum likelihood estimator (MLE) and understanding the error of this estimate is a classical problem. Many works have investigated the estimation of Generator matrices or the (intermediate) Transition Probability Matrices (TPM), see Kalbfleisch and Lawless (1985), Bladt and Sørensen (2005), Bladt and Sørensen (2009), dos Reis and Smith (2018) to mention a few and Pfeuffer (2017) for an overview and algorithm implementation in the statistical language R.

Despite these works, the problem of how to conduct statistical inference in this context or, in particular, how to derive error estimates for discretely observed Markov processes, is still an issue. Since our inference is likelihood based, Wald confidence intervals (or Wald intervals) are the natural choice for error estimation. In Kalbfleisch and Lawless (1985), the authors use numerical techniques to estimate the derivatives and use a so-called *quasi-Newton* method to obtain the MLE[1]. More recently, Bladt and Sørensen (2005) and Bladt and Sørensen (2009) consider the Expectation Maximisation (EM) algorithm and a Markov chain Monte Carlo (MCMC) algorithm to obtain the MLE. In the case of the EM, the authors provide a numerical scheme based on a formula from Oakes (1999) to obtain the error in the estimate. Following their approach dos Reis and Smith (2018) give exact expressions for errors arising from the EM algorithm. Building on these, we transfer the errors in the estimation of the CTMC's generator matrix to the estimation errors of the rating transitions probabilities themselves. As far as we are aware, such estimations have not been considered, although, they are of significant practical importance. When complete continuous-time rating transition data is available, then the computation of point estimates and Wald intervals for the parameters of a CTMC is straightforward, see e.g. Lando and Skodeberg (2002).

Concerning the second contribution of our manuscript, Lando and Skodeberg (2002) show that

---

[1] A quasi-Newton method (or scoring procedure) only requires one to estimate the first order derivative. More common approaches such as Newton-Raphson for finding the MLE would also require evaluation of the second derivative.

rating transitions exhibit non-Markovian behaviours. In particular, an obligor that has been recently downgraded into a certain rating is more likely to be downgraded further than other obligors currently in that rating. Such an effect is referred to as *(downward)-rating momentum*. A similar effect may also appear in upgrades; however, it is not as apparent. Documented (non-Markovian) effects in rating transitions include *rating drift* (or *momentum*) in Altman and Kao (1992) and Lando and Skodeberg (2002), *rating stickiness* in McNeil *et al.* (2005) and specific rating agencies' policies (see Carey and Hrycay (2001) and Løffler (2005)). Nickell *et al.* (2000) highlight non-Markovian patterns in transition probabilities for ratings and discuss their dependence regarding underlying variables like industry, domicile and business cycle. However, of these effects rating momentum is the most important to capture and what we look to model here.

The rating momentum effect has a non-negligible bearing on the risk attributed to a portfolio as it makes defaults of investment grade bonds likelier than defaults that are estimated within the standard Markov framework. (Couderc 2008, p.8) report on the temporal span of the rating drift (for a certain Standard & Poor's database) and its mean reversion. When looking to model over longer horizons, the non-Markov effects such as momentum become more pronounced, i.e. have a larger impact on transition probabilities. At a practical level, the IFRS9 regulation requires knowledge of risks on rating migrations over longer horizons where these effects can significantly change the results. When one can access the full data set (continuous-time observations), it is possible to construct tractable models that capture non-Markov effects and this is one of our contributions. The model we propose is able to capture the momentum behaviour, and we found that the purely Markov model underestimates default risk in investment grades but overestimates the risk in some speculative grades. We discuss this in more detail in Section 4 below.

For clarity, we summarize the contributions of our manuscript in the next two points.

(i) In the CTMC setting with discretely observed data, we provide a new simpler closed-form expression for the Hessian of the likelihood function, enabling faster computation of confidence intervals via the Fisher information matrix (Wald intervals). We further provide expressions allowing one to transfer confidence intervals at the level of the generator matrix to the level of rating transitions and probabilities of default, where they can be easily interpreted. Recall that in the case of discrete anonymous data one is forced to adopt the Markov assumption.

(ii) In the setting of continuously observed data, we propose a tractable and parsimonious model that captures the non-Markovian phenomenon of *rating momentum*. We provide a calibration procedure and several comparative tests based on *Moody's Corporate Credit Ratings* data set (see Section 2). Most notable is the difference between empirical, Markov (CTMC) and non-Markov (our model) estimates of probabilities of default. We observe that in several cases the Markov model under- or overestimates the probabilities of default empirically observed while the non-Markov model provides better agreement.

REMARK 1.2 (Software and R-code) *The algorithms relating to Markov Chains (in Section 3) are part of the CRAN R-package* ctmcd: Estimating the Parameters of a Continuous-Time Markov Chain from Discrete-Time Data *(see Pfeuffer (2017)) — https://CRAN.R-project.org/package=ctmcd*

Potential Non-Markov Models with Application to Rating Momentum. Different models have been introduced in the past to incorporate non-Markov phenomena. We briefly overview some of these works here. Interested readers are encouraged to consult the reference herein.

Extended State Space and Mixture Models. Christensen *et al.* (2004), attempt to take non-Markovian effects into account while preserving some Markovian structure. The idea is to extend the state space to include $+$ and $-$ states, referred to as excited states. For example, when a company downgrades from rating A to rating B, it is instead given the rating B^-, which has a higher probability of further downgrades than B. Similarly, if the company transitions from B to A, it is instead rated A^+ which has a smaller probability of downgrade than A. This construction allows

us to maintain the Markov property; however, we must calibrate many more parameters and, in real-world data, we do not observe a company belonging to the excited or non-excited state. Moreover, when successive transitions occur, it is unknown whether the company was in the excited or non-excited state. Hence calibrating an intensity between excited and non-excited states seems impossible. One could navigate around this by assuming excited states do not jump to non excited states, but this is against empirical evidence of momentum reducing over time, see (Couderc 2008, p. 35) for example.

D'Amico *et al.* (2016) apply a semi-Markov model to capture the observed effect that companies move through states not following an exponential distribution. However, they still rely on the Markov transition structure and hence they need to expand the state space in order to include momentum. Related to this approach is Frydman and Schuermann (2008), where the authors cleverly use two different time homogeneous CTMC generator matrices, however, it does not capture momentum since the jump itself is Markov.

Hidden Markov Model (HMM). A different idea is to use a hidden Markov model (HMM) (see Cappé *et al.* (2005) for a complete account). The HMM approach to credit risk can be traced back to the work of R. Elliot, see overview in Korolkiewicz (2012) and its references. Roughly, the approach considers two processes (X, Y), the observed (published) credit rating Y and the "true" credit rating X which is unobserved (or hidden). The paradigm is that the observed credit ratings are assumed to be "noisy" observations of Y and not the true representation of the credit risk. The goal is then to use Y to make inference on X. In such a setup if one considers the noisy observation and the true rating as correlated, then rating momentum can be added into the model. Although this approach has some benefits, the work appears to be constrained to the discrete time case and, from the literature point of view, the approach remains unexplored.

Hazard Rates, Point Processes and self-exciting Marked Point Processes. Let us start by discussing Hazard rates. The main work in this area for credit ratings is given in Koopman *et al.* (2008). An extensive work bringing hazard rate methodologies to the estimation of probabilities of default can be found in Couderc (2008) (and references therein). The paradigm is that each company has a corresponding hazard rate (a parameter) and in this hazard rate one can encode various factors such as momentum for example. The issue with Koopman *et al.* (2008)'s methodology is that they must calibrate parameters for each of the various transitions with the extra variables to obtain the probabilities of these transitions. This however, increases the model's complexity greatly. Our goal is to present a model as parsimonious as possible that captures rating momentum.

The approach we pursue relies on *point processes* that are dependent on their own history, so-called *self-exciting* processes (see Daley and Vere-Jones (2003), Daley and Vere-Jones (2008)). Point processes are generalizations of Markov processes and a natural choice for our model. One of the most satisfying aspects of using point processes is that one can capture rating momentum by adding only a small number of parameters (2 to 4 in our case). The most common example of a self-exciting process is the Hawkes process. These processes appear in other areas of mathematical finance, such as models for limit order books and are also used in high-frequency trading, see Bacry *et al.* (2015). However, they have not been fully utilized in credit transitions. A Hawkes process can be thought of as a counting process (similar to a Poisson process) which in one dimension has an intensity λ_t of the form (see Dassios and Zhao (2013)),

$$\lambda_t = \mu + \int_0^t \phi(t - s) \mathrm{d}N_s \,,$$

where N is a counting measure and denotes that an event has occurred (this will be a rating change in our case), μ is the baseline intensity and ϕ is the impact on the intensity and allows the intensity to depend on past events. By setting $\phi = 0$ the Hawkes process reduces to a Poisson process. A common choice for ϕ is the so-called exponential decay, namely $\phi(t - s) = \alpha\beta \exp(-\beta(t - s))$ with $\alpha, \beta > 0$. Functions of this form are useful since the event's influence on the intensity weakens as time progresses, hence we can account for momentum reducing over time (agreeing with the

findings of Couderc (2008)).

Using a Hawkes process allows us to embed past dependence in the jump times, however, in this simplistic form it is not fit for our purposes since we require different changes to intensity dependent on whether it is an upgrade or a downgrade. Further, we require the baseline intensity μ, to depend on the current state. Such extended Hawkes processes are referred to as *marked point processes*, since to each event observed one assigns a *mark* to indicate the type of event, see (Daley and Vere-Jones 2003, Chapter 6.4). We discuss this further in Section 4.

This work is organized as follows. In Section 2 we overview the data paradigms and describe the data we work with. In Section 3 we establish our closed-form expression for the Wald confidence intervals for the underlying Transition Probability Matrix (TPM) in the Markov setting. Finally, in Section 4 we analyse Moody's corporate credit rating data set, we test for non-Markovianity and calibrate the proposed non-Markov model. We also give due attention and discuss the effect of adding momentum in the estimation of default probabilities. In order to help keep this work self contained we supplement the core ideas with further discussion in the Appendix.

2.   Data description

To illustrate the statistical methods we develop in this manuscript, we use the proprietary *Moody's corporate credit rating data set*, which comprises continuous-time observations for 17097 entities (companies) in the time interval Jan 1, 1987 to Dec 31, 2017. Through the remainder of the article we refer to this set as the "*Moody's data set*". Some of the discrete data is available publicly, but the full data set is proprietary and must be purchased. Other works such as Christensen *et al.* (2004) also use the full Moody's data set.

The rating categories in Moody's data set are depicted in decreasing order of rating quality as "Aaa", "Aa1", "Aa2", "Aa3", "A1", "A2", "A3", "Baa1", "Baa2", "Baa3", "Ba1", "Ba2", "Ba3", "B1", "B2", "B3", "Caa1", "Caa2", "Caa3", "Ca", "C". We define "C" as the default category. The refinements "1", "2" and "3" shall be referred to as *modifiers* in the following. The ratings "Aaa" to "Baa3" are the so-called "Investment Grade" block while the ratings "Ba1" to "Ca" form the "Speculative Grade" block.

We employ a standard data aggregation arrangement where we aggregate all modifiers within their rating class. For instance, we group "Aa1", "Aa2", "Aa3" as "Aa" and so on to obtain the following categories in decreasing credit quality: "Aaa", "Aa", "A", "Baa", "Ba", "B", "Caa", "Ca" and "C" (Default Category). We shall use the standard aggregation unless otherwise stated.

For clarification, unlike the Standard and Poor rating classes where "C" is taken as the rating above default and "D" is used as default, in the Moody's rating system, "C" denotes default. We use the latter notation throughout our manuscript.

As described in the introduction there are two data paradigms, a *discrete* (missing) and a *continuous* (full) one. In Section 3 of the paper we construct annually discretized rating transition matrices from this data, and one is led to use a (CTMC) Markov model. In Section 4, we use the full data set and its richness allows us to expand the scope to non-Markov models.

3.   Calculating Wald Confidence Intervals for Discretely Observed Markov Processes

The working paradigm for this section is the *discrete time data* one and we work towards estimating the generator matrix $\mathbf{Q}$ of the underlying CTMC model. For this setting, it was shown in dos Reis and Smith (2018) that the Expectation-Maximization (EM) algorithm is the strongest algorithm for the estimation of $\mathbf{Q}$ (a description of the EM algorithm for this context is provided in Appendix A). The EM is built to use likelihood-based inference which has the advantage that one can obtain errors for the estimate by taking derivatives, the so-called Wald confidence intervals.

The goals of this section are to find expressions for these derivatives and then use them to obtain the corresponding intervals for the transition probabilities.

Our CTMC set up is similar to that of dos Reis and Smith (2018). We understand companies' ratings as defined on a finite state space $\{1, \ldots, h\}$, where each state corresponds to a rating. We denote Aaa as rating 1 and C (default) as rating h. Let $\mathbf{P}$ be an h-by-h stochastic matrix, which will be the corresponding TPM (at, say, time $t = 1$) and $\mathbf{Q}$ is an h-by-h generator matrix; we denote $p_{ij} := (\mathbf{P})_{ij}$, $q_{ij} := (\mathbf{Q})_{ij}$ and the intensity of state i by $q_i = \sum_{j \neq i} q_{ij}$ where $i, j \in \{1, \ldots, h\}$. A standard assumption used in credit risk modelling is that the default state is an absorbing state, hence $p_{hh} = 1$. In the data, companies are observed to withdraw (e.g. via mergers or early payment) and we treat such a withdrawn rating as a censored result.

Regarding the CTMC's generator, we work with stable generator matrices, i.e. matrices $\mathbf{Q}$ that satisfy the following definition.

DEFINITION 3.1 (Stable-Conservative infinitesimal Generator matrix of a CTMC) *We say a matrix* $\mathbf{Q}$ *is a generator matrix if the following properties are satisfied for all* $i, j \in \{1, \ldots, h\}$:
  i) $0 \leq q_{ij} < \infty$ *for* $i \neq j$;   *ii)* $q_{ii} \leq 0$;   *and iii)* $\sum_{j=1}^{h} q_{ij} = 0$.

Our quantity of interest is the time varying transition probability matrix, $\mathbf{P}(t)$, which is related to the generator matrix $\mathbf{Q}$ via,

$$\mathbf{P}(t) = e^{\mathbf{Q}t}, \qquad t \geq 0. \tag{3.1}$$

We assume throughout that $\mathbf{Q}$ is a valid generator matrix (in the sense of Definition 3.1), hence $\mathbf{P}$ is well defined. Considering the case where the CTMC is observed at times $t_0 < t_1 < \cdots < t_M$ and denote $\Delta t_u := t_u - t_{u-1}$ for $u \in \{1, \ldots, M\}$ and the transition matrix over that interval by $\mathbf{N}(u)$.

The likelihood of the discretely observed Markov process is given by,

$$L(\mathbf{Q}|\mathbf{N}) = \prod_{u=1}^{M} \prod_{s=1}^{h} \prod_{r=1}^{h} \exp(\mathbf{Q}\Delta t_u)_{sr}^{\mathbf{N}_{sr}(u)}. \tag{3.2}$$

Although this is not the full likelihood of a CTMC, it is the likelihood based on the observable data, so in effect, the EM algorithm looks to find $\mathbf{Q}$ to maximise (3.2). Therefore the Wald confidence intervals of $\mathbf{Q}$ are based on this likelihood. One can construct confidence intervals for other algorithms such as the quasi-optimization of the generator (see Kreinin and Sidelnikova (2001)) by bootstrapping, but these are computationally more expensive to calculate.

3.1. *Direct Differentiation for Gradient and Hessian of the Likelihood*

The standard procedure to derive a confidence interval is to use the variance of the estimator (in our case, the negative inverse of the Hessian H of the likelihood L in (3.2)). Since the EM algorithm deals with a missing data likelihood, these derivatives are complex to calculate, however, (Oakes 1999, Section 2) derived a simpler formula for the Hessian. This formula was used by Bladt and Sørensen (2009) and dos Reis and Smith (2018) to obtain error estimates in this setting. A formula for obtaining the Hessian is useful, however, while the second derivative can inform us about errors at the level of the generator matrix, it does not shed light on how these errors propagate to the transition probabilities (see (3.1)). For that we need to be able to take further derivatives.

Relying on first principles, it turns out that for this problem one can extract derivatives without the said formula in Oakes (1999) and derive a new closed-form solution involving matrix exponentials for the gradient and the Hessian by direct differentiation.

Similar to the situation in dos Reis and Smith (2018) the parameter space of $\mathbf{Q}$ is closed at zero and we can only differentiate in the interior of the space, hence we introduce the notion of *allowed pairs*. This concept allows one to incorporate absorbing states in the analysis.

DEFINITION 3.2 (Allowed pairs)   *Let $i, j \in \{1, \ldots, h\}$, then we say that the pair (i, j) is* allowed *if $i \neq j$ (not in the diagonal) and q_{ij} is not converging to zero under the EM algorithm.*

Essentially i, j is allowed if $q_{ij} > 0$ and thus in the interior of the parameter space of $\mathbf{Q}$. For ease of presentation we denote by $\mathbf{V_Q}$ the matrix of allowed pairs of $\mathbf{Q}$, namely: for N_a the number of allowed pairs in the estimation of $\mathbf{Q}$ we define the matrix $\mathbf{V_Q}$ as the N_a-*by*-2-dimensional matrix which records the allowed pairs of $\mathbf{Q}$.

Let $\mathbf{A}$ be an h-by-h matrix, $\alpha, \beta, s, r \in \{1, \ldots, h\}$ and $\mathbf{e}_r$ be an h-dimensional column vector with 1 at entry r and zero elsewhere. Let us further denote $a_{\alpha\beta} := (\mathbf{A})_{\alpha\beta}$ as the entries of matrix $\mathbf{A}$ and assume $a_{\alpha\beta} > 0$, then using standard properties of derivatives and integrals of matrix exponentials (see Wilcox (1967) and Van Loan (1978)) it follows that,

$$\frac{\partial \exp(\mathbf{A}t)_{sr}}{\partial a_{\alpha\beta}} = \mathbf{e}_s^{\mathsf{T}} \int_0^t \exp(\mathbf{A}v) \frac{\partial \mathbf{A}}{\partial a_{\alpha\beta}} \exp\left(\mathbf{A}(t-v)\right) \mathrm{d}v \, \mathbf{e}_r = \mathbf{e}_s^{\mathsf{T}} \exp\left(\begin{bmatrix} \mathbf{A} & \frac{\partial \mathbf{A}}{\partial a_{ij}} \\ 0 & \mathbf{A} \end{bmatrix} t\right)_{1:h, h+1:2h} \mathbf{e}_r. \quad (3.3)$$

Using (3.3), we can directly calculate the first and second derivative of the likelihood function for a discretely observed Markov process. Let (α, β) and (μ, ν) be allowed pairs for the generator $\mathbf{Q}$, then the expressions for the gradient and Hessian of the logarithm of (3.2) are as follows: for the **Gradient** we have

$$\frac{\partial \log L(\mathbf{Q}|\mathbf{N})}{\partial q_{\alpha\beta}} = \sum_{u=1}^{M} \sum_{s=1}^{h} \sum_{r=1}^{h} \mathbf{N}_{sr}(u) \frac{\exp(\mathbf{C}_\eta^{(\alpha\beta)} \Delta t_u)_{s,h+r}}{\exp(\mathbf{Q}\Delta t_u)_{s,r}} \quad \text{with} \quad \mathbf{C}_\eta^{(\alpha\beta)} = \begin{bmatrix} \mathbf{Q} & \mathbf{e}_\alpha \mathbf{e}_\beta^{\mathsf{T}} - \mathbf{e}_\alpha \mathbf{e}_\alpha^{\mathsf{T}} \\ 0 & \mathbf{Q} \end{bmatrix},$$

while for the **Hessian** we have

$$H(\mathbf{Q})_{\alpha\beta,\mu\nu} = \frac{\partial^2 \log L(\mathbf{Q}|\mathbf{N})}{\partial q_{\alpha\beta} \partial q_{\mu\nu}}$$

$$= \sum_{u=1}^{M} \sum_{s=1}^{h} \sum_{r=1}^{h} \frac{\mathbf{N}_{sr}(u)}{\exp(\mathbf{Q}\Delta t_u)_{sr}} \left[\frac{\exp(\mathbf{C}_\eta^{(\alpha\beta)} \Delta t_u)_{s,h+r} \exp(\mathbf{C}_\eta^{(\mu\nu)} \Delta t_u)_{s,h+r}}{\exp(\mathbf{Q}\Delta t_u)_{sr}} - \exp(\mathbf{C}_\xi^{(\alpha\beta,\mu\nu)} \Delta t_u)_{s,3h+r}\right],$$

$$\text{whereas} \quad \mathbf{C}_\xi^{\alpha\beta,\mu\nu} = \begin{bmatrix} \mathbf{C}_\eta^{(\alpha\beta)} & \frac{\partial \mathbf{C}_\eta^{(\alpha\beta)}}{\partial q_{\mu\nu}} \\ 0 & \mathbf{C}_\eta^{(\alpha\beta)} \end{bmatrix}.$$

These estimates are direct applications of the theory above and hence we omit the steps. Both the formula of (dos Reis and Smith 2018, p. 7) and this new one are exact expressions for the Hessian and thus for the Fisher information matrix. However, the new formula is of distinctly reduced complexity, which consequently leads to clearly shorter computing times. Since the Hessian is only defined for allowed pairs the matrix is dimension-wise smaller than $(h-1)^2$-*by*-$(h-1)^2$.

We compute the Wald confidence intervals as follows,

- recall that $\mathbf{V_Q}$ is the N_a-*by*-2 dimensional matrix recording the allowed pairs of $\mathbf{Q}$ (with N_a the number of allowed pairs in the estimated $\mathbf{Q}$).

  The ijth component of the Hessian is the differential

$$\frac{\partial^2}{\partial q_{\mathbf{V_Q}(i,1)\mathbf{V_Q}(i,2)} \partial q_{\mathbf{V_Q}(j,1)\mathbf{V_Q}(j,2)}}.$$

- The Fisher information matrix is given by $-\mathbf{H}(\cdot)$. The estimated variance of the allowed parameter q_{ab} is the ith diagonal element of $-\mathbf{H}(\cdot)^{-1}$, where $\mathbf{V_Q}(i,1) = a$ and $\mathbf{V_Q}(i,2) = b$.
- The Wald 95% confidence interval of the MLE $\hat{q}_{ab}$ is $\hat{q}_{ab} \pm 1.96\sqrt{Var(\hat{q}_{ab})}$.

A 95% confidence interval for the generator matrix estimate based on Moody's discretely observed corporate rating data is illustrated in Table 1. To obtain the Wald confidence interval, the computation time was $\approx$ 1s with the new expression compared to $\approx$ 2s for the formula of dos Reis and Smith (2018).

Aaa	Aa	A	Baa	Ba	B	Caa	Ca	C
	[0.074,0.091]	0	0	[0,0.001]	0	0	0	0
[0.009,0.012]		[0.088,0.098]	[0.001,0.004]	[0,0.001]	0	0	0	0
[0,0.001]	[0.023,0.027]		[0.061,0.067]	[0.003,0.005]	[0.001,0.002]	[0, 0.001]	0	0
[0,0.001]	[0.001,0.002]	[0.039,0.044]		[0.042,0.047]	[0.005,0.007]	[0.001,0.003]	[0,0.001]	0
0	[0,0.001]	[0.002,0.004]	[0.064,0.072]		[0.092,0.102]	[0.007,0.011]	[0.001,0.002]	0
0	[0,0.001]	[0,0.001]	[0.001,0.003]	[0.049,0.055]		[0.091,0.099]	[0.008,0.011]	0
0	0	0	[0,0.001]	[0.001,0.005]	[0.107,0.122]		[0.052,0.064]	[0.028,0.036]
0	0	0	[-0.001,0.006]	[0.003,0.018]	[0.047,0.083]	[0.127,0.181]		[0.123,0.170]
0	0	0	0	0	0	0	0	0

Table 1.   Confidence Interval (at 95 % confidence) for the entries of the Generator Matrix for Moody's Corporate Rating Discrete-Time Transition Matrix.

3.2.   *The Delta method - Confidence Intervals for probabilities*

The object we are estimating is the generator matrix $\mathbf{Q}$, thus the confidence intervals are based on the entries of this matrix. Although obtaining these confidence intervals are useful, from a practitioners standpoint it is more useful to know how this uncertainty propagates to the underlying TPM and the estimated probabilities of default. This is a classical problem in statistics where one wishes to consider how the confidence interval changes under a transformation (in this case (3.1)), the standard method to do this is known as the *Delta method*, see Lehmann and Casella (1998) for further information.

We construct confidence intervals for each individual element in $\mathbf{P}$ using the set of *allowed pairs* (Definition 3.2). We consider the confidence interval for the transition probability p_{ij} at time t as,

$$p_{ij}(\mathbf{V_Q};t) := \left(e^{\mathbf{Q}t}\right)_{ij}.$$

That is for a fixed t, $p_{ij}(\mathbf{V_Q};t)$ is a multivariate function of the allowed pairs, $\mathbf{V_Q}$, in $\mathbf{Q}$. This leads to the following result.

THEOREM 3.3   *Assume asymptotic normality holds for all allowed pairs, let $\mathbf{V}_{\hat{\mathbf{Q}}}$ denote the allowed pairs of $\hat{\mathbf{Q}}$ (our MLE estimate) and fix t. Then, for each i,j in the state space with $i \neq h$, the variance in p_{ij} is given by,*

$$Var\left(p_{ij}(\mathbf{V}_{\hat{\mathbf{Q}}};t)\right) \approx \frac{\partial p_{ij}(\mathbf{V}_{\hat{\mathbf{Q}}};t)}{\partial \mathbf{V}_{\hat{\mathbf{Q}}}} \left(-\mathbf{H}(\hat{\mathbf{Q}})^{-1}\right) \left(\frac{\partial p_{ij}(\mathbf{V}_{\hat{\mathbf{Q}}};t)}{\partial \mathbf{V}_{\hat{\mathbf{Q}}}}\right)^{\mathsf{T}}, \tag{3.4}$$

provided $\partial p_{ij}(\mathbf{V}_{\hat{\mathbf{Q}}};t)/\partial \mathbf{V}_{\hat{\mathbf{Q}}} \neq 0$, where $\frac{\partial}{\partial \mathbf{V}_{\hat{\mathbf{Q}}}}$ denotes the vector constructed by differentiating w.r.t. each element in $\mathbf{V}_{\hat{\mathbf{Q}}}$ then evaluated at $\hat{\mathbf{Q}}$, and $\mathbf{H}(\hat{\mathbf{Q}})^{-1}$ is the inverse Hessian matrix at the MLE. Moreover, for each $(\alpha,\beta) \in \mathbf{V}_{\hat{\mathbf{Q}}}$,

$$\frac{\partial p_{ij}(\mathbf{V}_{\hat{\mathbf{Q}}};t)}{\partial q_{\alpha\beta}} = \left(\exp(\mathbf{C}_{\eta}^{(\alpha\beta)}t)\right)_{i,h+j} \quad where \quad \mathbf{C}_{\eta}^{(\alpha\beta)} = \begin{bmatrix} \hat{\mathbf{Q}} & \mathbf{e}_\alpha\mathbf{e}_\beta^{\mathsf{T}} - \mathbf{e}_\alpha\mathbf{e}_\alpha^{\mathsf{T}} \\ 0 & \hat{\mathbf{Q}} \end{bmatrix}.$$

The proof of this result is given in Appendix B. The assumption that $\partial p_{ij}(\mathbf{V}_{\hat{\mathbf{Q}}};t)/\partial \mathbf{V}_{\hat{\mathbf{Q}}} \neq 0$ is extremely mild and can be easily checked once the MLE estimate is found.

At this point, we take advantage of the fact that we have already derived a closed-form expression for the Hessian. Hence we can easily compute (3.4), moreover, it is now straightforward to compute the confidence interval for the transition probabilities. This is an extremely useful result since it allows one to quantify the uncertainty at the level of the estimation of transition probabilities (instead of the generator matrix), and critically, uncertainties in the probability of default. Figures 1 and 2 show such intervals for probability of default estimates from Moody's corporate rating data 2016 and a time horizon of up to 10 years. One can see that this procedure easily allows one to quantify the error of probability of default predictions for arbitrary time horizons. This is especially interesting as this parameter is an important ingredient to the calculation of expected losses over lifetime in the IFRS 9 regulatory framework.
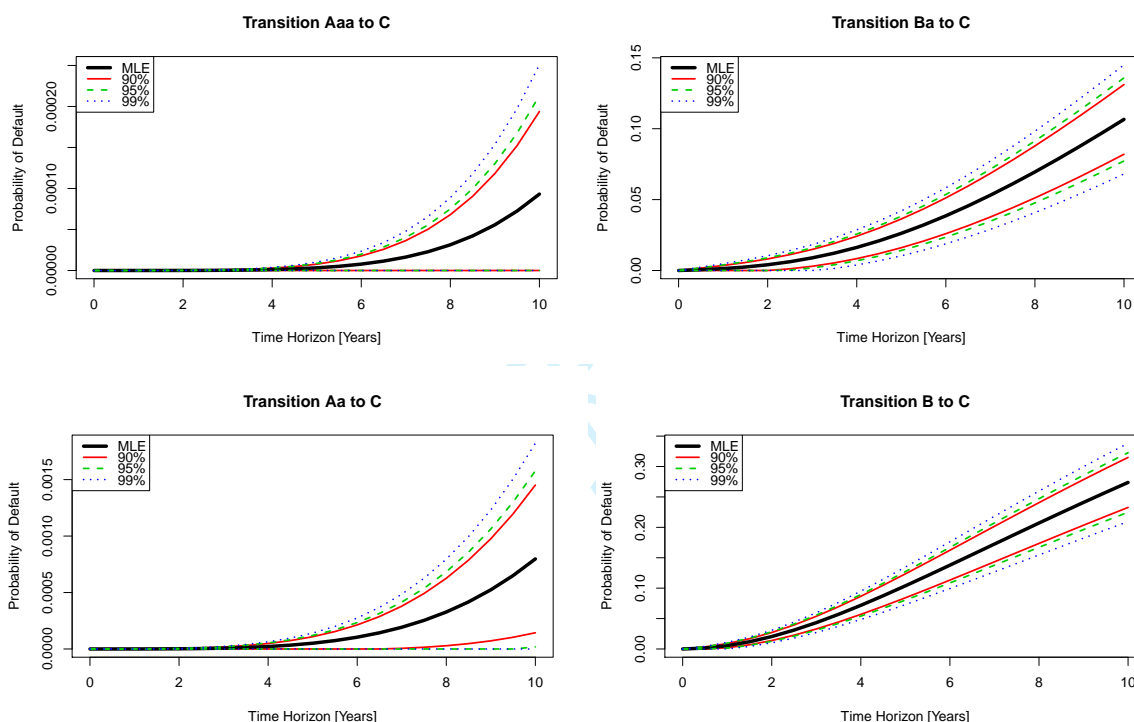

Figure 1. Confidence Intervals as maps of time for Discrete-Time Transitions into the Default Category C over 10 years-Moody's Corporate Rating Discrete-Time Transitions 2016

3.3.   *Confidence Intervals w.r.t. information*

We benchmark our analysis against (dos Reis and Smith 2018, Section 4). We consider a true generator matrix (which is the MLE Markov generator described in Section 4.5) and from that simulate multiple years worth of data which is viewed as empirical data. We then introduce the EM algorithm to increasing amounts of data and assess how the estimate and errors change. By using a known generator, we additionally assess the accuracy of the estimate and error. From a computational point of view, matrix exponentials embed highly nonlinear dependencies in the elements of $\mathbf{Q}$ and $\mathbf{P}$. Therefore, to understand the error we consider how both the error of $\mathbf{Q}$ and $\mathbf{P}$ changes as the amount of information changes.

We consider the scenario of 250 obligors per rating and simulate 50 years worth of transitions (i.e. the number of companies that made each transition). We then apply the EM algorithm using 1 year worth of data then 2 years etc up to 50 years. In the case of a company defaulting we replace
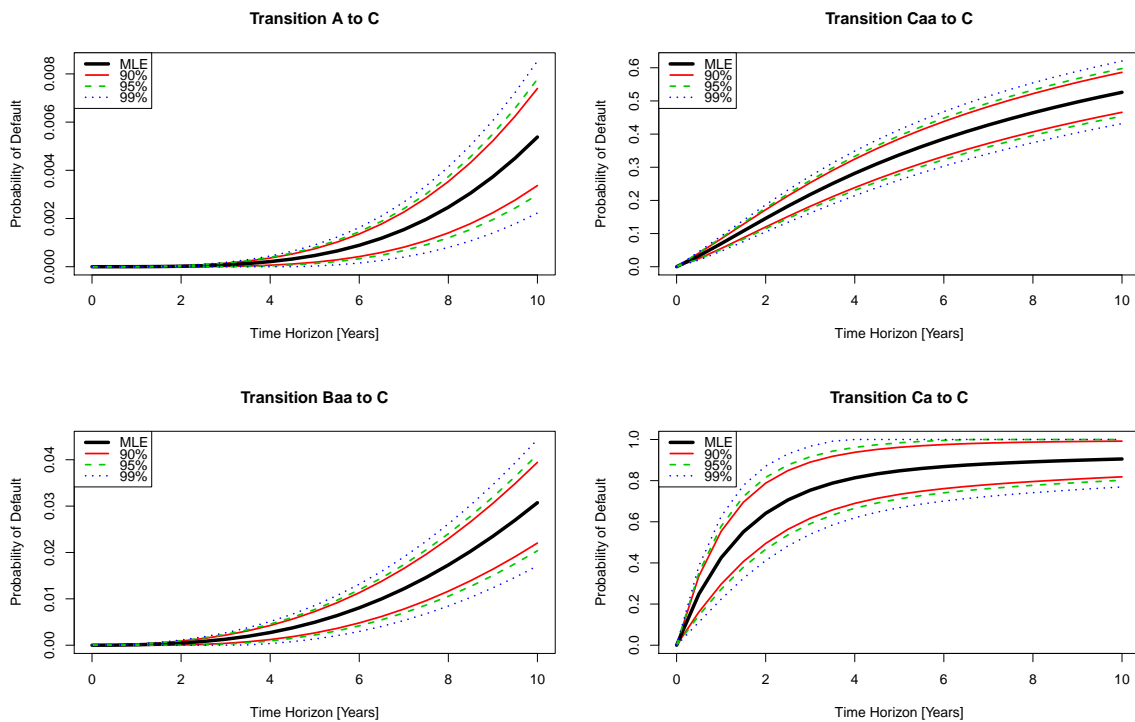
Figure 2.   Confidence Intervals as maps of time for Discrete-Time Transitions into the Default Category C over 10 years-Moody's Corporate Rating Discrete-Time Transitions 2016

it with the rating they were pre-default. This implies that the amount of "information" obtained from each year is similar. We plot the results in Figure 3.
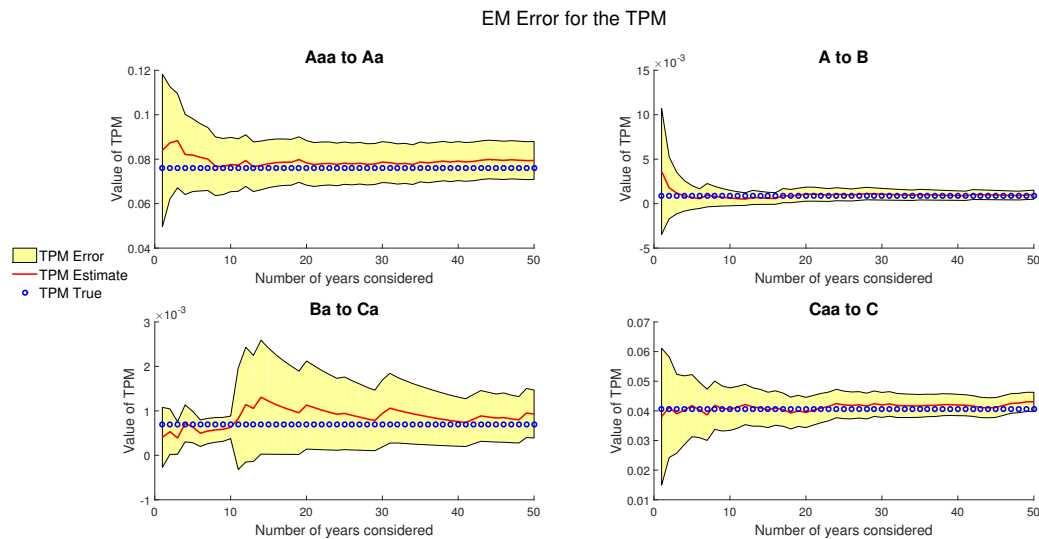
Figure 3. Estimated value in some TPM entries and 95% confidence interval as the amount of data increases.

One observes that in most cases the errors in the TPM behave as expected. The surprising result is the Ba to Ca entry whose error increases. As alluded above, one can only understand the error in the TPM by understanding the underpinning error of the generator estimation. Although, in theory, the Ba to Ca transition depends on all entries in the generator we know that certain entries have a greater impact. We, therefore, look at the error in some important generator entries, Figure 4. From Figure 4 it is clear that the main contributor to the error is (unsurprisingly) the Ba to Ca
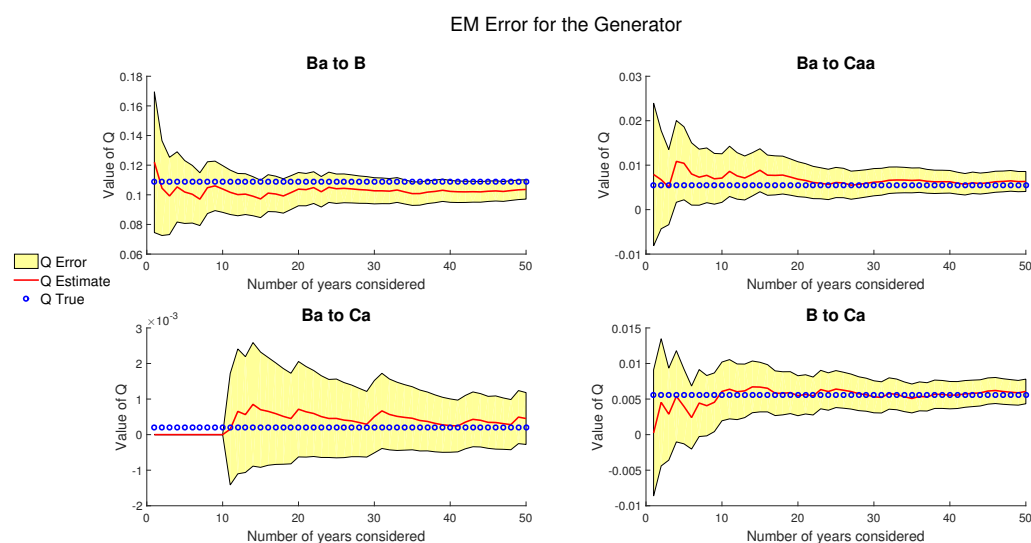
Figure 4. Estimated value in the generator and 95% confidence interval as the amount of data increases.

entry. Initially, we need to wait for a transition from Ba to Ca to happen which increases the likelihood and hence the uncertainty surrounding the estimate. Moreover, it then takes several more years of data before the estimate becomes more stable. This uncertainty in the generator then propagates to uncertainty in the TPM entries, and one observes the extremely strong correlation between the TPM entry and the corresponding generator entry. Due to this, the error in the Ba to Ca transition probability is much larger than the other estimates, even after 50 years of observation. This behaviour in the CTMC modelling is not ideal (and the IFRS 9 regulation exacerbates the effect), but it shows some of the challenges in obtaining good estimates and errors for small probabilities (rare events), namely that the model is still sensitive to individual observations. One can use this to assess the sensitivity in the model, for example, adding one observation of a company defaulting and then recomputing the probabilities and their associated errors will provide an idea of the sensitivity.

4. Extending Markov Processes to Capture Rating Momentum

In this section, we work with the continuously observed data case and hence can broaden our scope of models (we are no longer restricted to Markov models). In the previous section, we highlighted many good features of the EM algorithm, in particular, that one could derive closed-form expressions for the errors. However, the EM algorithm does not generalize well as one quickly runs into difficulties when using models that have more complex likelihoods. This is the case when we generalize to point processes. Before detailing the model we are proposing let us start by showing that the data (see Section 3) contains non-Markov features.

4.1. *Testing for non-Markovian phenomena*

In Lando and Skodeberg (2002)'s analysis of Standard and Poor's rating data set, the authors tested the presence of rating momentum. For consistency and completeness, we show that *rating momentum* behaviour is also present in Moody's data set.

The test follows a standard semi-parametric hazard model approach developed in Andersen *et al.* (1991) (see also Andersen *et al.* (2012)). The basic idea is to test whether the intensity (from leaving the state) is influenced by previous transitions, that is, we model the intensity for any given

firm, n in state i as,

$$\lambda_{in}(t) = q_i(t)\exp(cZ_n(t)),$$

where q is an unspecified "baseline" intensity[1], Z contains information relating to the firm and c is the coefficient we estimate. One important point here is that we are often dealing with censored observations (many firms stop being rated after a while), hence using hazard models is useful since we have access to the theory of partial likelihoods which can handle censored observations, see Cox and Oakes (1984). One can then for example set the covariate Z as,

$$Z_n(t) = \begin{cases} 1, & \text{if firm } n \text{ was downgraded to its current state,} \\ 0, & \text{otherwise.} \end{cases}$$

Hence in this setting the Markov assumption is equivalent to the null hypothesis $c = 0$. The general statistical framework including fitting c by maximising the partial likelihood is covered in Andersen *et al.* (1991) and (Lando and Skodeberg 2002, Appendix A), but we do not discuss these further here.

The result from this analysis can be seen in Table 2 – we can see a statistically significant downward momentum effect (i.e. the null hypothesis is rejected on standard significance levels α of 10%, 5% or 1%) but no significant upward momentum behaviour in the Moody's data. These findings are consistent with those of Lando and Skodeberg (2002).

	coefficient	p-value
downward momentum	0.33010	<0.0001
upward momentum	-0.01487	0.68153

Table 2.   Likelihood ratio test for downward and upward momentum.

4.2.   *Our new Model to capture Rating Momentum*

As one can see from Table 2 there is very strong evidence that downward momentum exists in the data. Let us now describe a tractable methodology, using *marked point processes* that can capture this effect. Readers unfamiliar with point processes can consult Appendix C for further details.

The likelihood of a single realisation of a marked point process is given in (Daley and Vere-Jones 2003, p.251), namely,

$$L = \prod_{i=1}^{N_g(T)} \lambda_g(t_i) f(k_i|t_i) e^{-\int_0^T \lambda_g(u)\mathrm{d}u}, \tag{4.1}$$

where we use the following notation, N_g is the set of times at which events occur, λ_g is the intensity, k is the mark and f is the so-called *mark's distribution*. The subscript g is a common notation used to imply that this is the intensity of the ground process, i.e. we are only considering the events of interest. Setting $\lambda = q_i$ and $f = q_{ij}/q_i$ we recover the likelihood of a CTMC and hence one can see that these processes are generalizations of Markov processes.

To incorporate rating momentum into such models we draw inspiration from Hawkes processes and change the intensity of the model for appropriate rating changes. The basic idea is to start with a CTMC (with generator matrix $\mathbf{Q}$), which acts as a baseline intensity, then add a non-Markov

---

[1]Observe that we are not assuming that the baseline is time homogeneous in the test.

component which is a self-excitation intensity decaying exponentially[2]. That is, any downgrade observed increases the intensity of then future downgrades for a certain while. We also introduce two types of momentum, one if the company downgrades from investment-grade (Baa and better) and another if the company downgrades from a speculative-grade (this modelling choice is further discussed in Section 4.4 and 4.5). Using the same notation as before, given the state space $\{1, \ldots, h\}$ such that state h (default) is absorbing, we model the intensity of the stochastic process X at time t as follows,

$$\lambda_g(t) = \sum_{j=1}^{h-1} q_j \mathbb{1}_{\{X(t)=j\}} + \sum_{m=1}^{2} \sum_{\tau \in \tau_m(t)} \beta_m \alpha_m e^{-\beta_m(t-\tau)},$$

where m denotes investment or speculative downgrade, $\tau_m(t)$ is the set of downgrade times (of type m) prior to time t and α_m and β_m correspond to the intensity and memory of the "momentum" in each case. One can note that the intensity of the stochastic process drops (returning to the baseline intensity) as more time elapses since the previous downgrade and this rate is controlled by β. In particular, this allows one to include empirically observed effects such as the momentum's influence reducing over time (see Couderc (2008)).

In this set up we add only four parameters to the $\approx (h-1)^2$ parameters of the CTMC case; the effectiveness of this parsimony is substantiated below (see Section 4.4). To the best of our knowledge, no other model we are aware of captures the momentum effect so simply. Further parameters and extensions can be introduced, nonetheless, we focus only on this model. Its analysis is found in Section 4.3.1 and 4.5.

We work under the following modelling assumptions which, we believe to be sufficiently reasonable and keep the model parsimonious (most of these can be easily lifted and the model extended).

(i) We only consider downward momentum. Since upward momentum is not as statistically significant (Table 2) we do not consider it.

(ii) There are two types of momentum, *investment* and *speculative*.
Companies being downgraded from investment grades (numerically these are the ratings from 1 to $(h-1)/2$) feel the investment momentum and remaining downgrades are affected by speculative momentum.

(iii) Finally (not easy to remove) no points occurred prior to time 0, the so-called edge effects. This essentially means that companies do not have momentum when they are initially rated.

REMARK 4.1 (Prudent Estimation) *Since we only consider momentum as a purely negative effect, if we assume a company has no momentum when it initially does then we will obtain more conservative numbers for the downgrades. Therefore in calibration, if one does not use a full history of a company's rating change the model will be more prudent.*

With these assumptions let us define the mark's distribution. We take the following marked distribution (for $X(t_i) \in \{1, \ldots, h-1\}$, t_i is the time of the ith jump),

$$f(X(t_i^-)|t_i) = \frac{\sum_{j,k=1}^{h} q_{jk} \mathbb{1}_{\{X(t_i^-)=j,\ X(t_i)=k\}}}{\lambda_g(t_i)} \Big(\mathbb{1}_{\{X(t_i)<X(t_i^-)\}}$$
$$+ \mathbb{1}_{\{X(t_i)>X(t_i^-)\}} \frac{1}{N_j} \sum_{m=1}^{2} \sum_{\tau \in \tau_m(t_i)} \beta_m \alpha_m e^{-\beta_m(t_i-\tau)} \Big),$$

---

[2]This is a common and well-understood form to use in Hawkes processes, see Bacry *et al.* (2015).

where we denote by t_i^- the time immediately prior to the ith jump and N_j is the number of states one can downgrade to i.e. $N_j = \sum_{k>j} \mathbb{1}_{\{q_{jk}>0\}}$. Substituting the intensity and mark distribution into (4.1), yields the following expression for the likelihood,

$$L = \prod_{i=1}^{N_g(T)} \left\{ \left(\sum_{j,k=1}^{h} q_{jk} \mathbb{1}_{\{X(t_i^-)=j,\ X(t_i)=k\}} + \frac{1}{N_j} \sum_{m=1}^{2} \sum_{\tau \in \tau_m(t_i)} \beta_m \alpha_m e^{-\beta_m(t_i-\tau)} \right) \mathbb{1}_{\{X(t_i)>X(t_i^-)\}} \right.$$

$$\left. + \sum_{j,k=1}^{h} q_{jk} \mathbb{1}_{\{X(t_i^-)=j,\ X(t_i)=k\}} \mathbb{1}_{\{X(t_i)<X(t_i^-)\}} \right\}$$

$$\times \exp\left(- \int_0^T \sum_{j=1}^{h-1} q_j \mathbb{1}_{\{X(u)=j\}} + \sum_{m=1}^{2} \sum_{\tau \in \tau_m(u)} \beta_m \alpha_m e^{-\beta_m(u-\tau)} \mathrm{d}u \right). \qquad (4.2)$$

Note that the likelihood is for the information regarding one company. We can construct the likelihood of multiple companies by taking the product, but it is worthwhile noting that this assumes independence among companies. This is unlikely to be true due to business cycles etc, however, these correlated systemic effects can be introduced into risk modelling using the methods from McNeil and Wendin (2007). Hence, we concentrate purely on the idiosyncratic effect of rating momentum.

The integral involving the momentum (last integral in (4.2)) can be simplified, to

$$\int_0^T \sum_{\tau \in \tau_m(u)} \beta_m \alpha_m e^{-\beta_m(u-\tau)} \mathrm{d}u = \sum_{\tau \in \tau_m(T)} \alpha_m \left(1 - e^{-\beta_m(T-\tau)} \right).$$

Unlike the CTMC case this likelihood is complex and there appears to be no real simplification, the main reason for this is the time and history dependence amongst jumps for which simplifications of the form $q_{ij}^{\mathbf{K}_{ij}}$ are no longer possible. We proceed forward by relying on Markov Chain Monte Carlo (MCMC) techniques to estimate the parameters.

4.3. *An MCMC calibration algorithm for the model*

In the CTMC setting as considered in Bladt and Sørensen (2005), Bladt and Sørensen (2009) and dos Reis and Smith (2018) the data augmentation step for the CTMC was costly making the algorithm extremely slow compared to other algorithms. In our setting, we have access to a complete data set and this expensive step is avoided. Moreover, the likelihood we deal with is complex and thus MCMC (see Gilks *et al.* (1996)) is one of the few methods that can deliver reasonable estimations.

The basic set up of MCMC is to estimate the parameter(s) θ through its posterior distribution given some data D, typically denoted $\pi(\theta|D)$. In general, one cannot access this posterior distribution and direct Monte Carlo simulation is not possible as one does not know the normalizing constant. MCMC gets around this by observing through Bayes' formula that,

$$\pi(\theta|D) \propto L(D;\theta)\pi(\theta),$$

where L is the likelihood and $\pi(\theta)$ is the prior distribution of θ. It is then possible to sample from this distribution using the Metropolis-Hastings algorithm with some proposal distribution.

Let $\mathbf{X}$ denote the set of all company transitions. We are interested in obtaining the joint distribution $\pi(\mathbf{Q}, \boldsymbol{\alpha}, \boldsymbol{\beta}|\mathbf{X})$ where $\mathbf{Q}$ is the matrix with the baseline intensities and jump probabilities (has the same form as a generator matrix of a CTMC) and $\boldsymbol{\alpha} := (\alpha_1, \alpha_2)$, $\boldsymbol{\beta} := (\beta_1, \beta_2)$ are the

momentum parameters. Since we assume the prior distribution of $\mathbf{Q}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ to be independent, Bayes' theorem implies that,

$$\pi(\mathbf{Q}, \boldsymbol{\alpha}, \boldsymbol{\beta}|\mathbf{X}) \propto \pi(\mathbf{X}|\mathbf{Q}, \boldsymbol{\alpha}, \boldsymbol{\beta})\pi(\mathbf{Q})\pi(\boldsymbol{\alpha})\pi(\boldsymbol{\beta}) = L\pi(\mathbf{Q})\pi(\boldsymbol{\alpha})\pi(\boldsymbol{\beta}),$$

where L is the likelihood defined in (4.2). The full conditional distribution of each parameter is obtained by conditioning on knowledge of all other parameters.

For the priors, firstly for $\mathbf{Q}$, we assume that the initial transitions carry no momentum hence we can set the prior as the CTMC maximum likelihood estimate (MLE) based on the initial transitions. We therefore set the prior as exponential with the mean being the MLE. For $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, we use a Gamma random variable with a reasonable variance as the prior. This is to reflect that we have far less knowledge for these parameters but do not expect them to be either zero or too large.

The next issue we tackle is how to simulate from the full conditional distribution. Dealing with the parameters of the model first, their full conditional distributions are clearly not standard distributions so we use the single-component Metropolis-Hastings algorithm. As always with Metropolis-Hastings we need to define a good proposal function. In order to avoid a high number of rejections, we take our proposal as a Gamma random variable with mean as the current step and a small variance. In effect, this creates a random walk type sampling scheme that is always nonnegative. Therefore, if we denote the set of parameters by γ and the proposal distribution by ψ (which can depend on the current parameters), the nth step acceptance probability of a proposed point γ_s given the current γ'_s is given by,

$$\frac{\pi(X|\gamma_s, \gamma_{n,-s})\pi(\gamma_s)\psi(\gamma'_s|\gamma_s)}{\pi(X|\gamma'_s, \gamma_{n,-s})\pi(\gamma'_s)\psi(\gamma_s|\gamma'_s)},$$

where $\gamma_{n,-s}$ denotes the set of parameters at the nth update not including the s parameter.

4.3.1. Model Calibration. Now that we have the necessary tools, we can calibrate our model using Moody's data set. Running 11000 MCMC iterations (taking 1000 burn in) we obtain the following results[1]. For the Markov style "base" component,

$$\mathbf{Q} = \begin{pmatrix} \text{Aaa} & \text{Aa} & \text{A} & \text{Baa} & \text{Ba} & \text{B} & \text{Caa} & \text{Ca} & \text{C} \\ -0.0869 & 0.0836 & 0.0031 & 0 & 0.0002 & 0 & 0 & 0 & 0 \\ 0.0117 & -0.1088 & 0.0942 & 0.0025 & 0.0003 & 0.0001 & 0 & 0 & 0 \\ 0.0006 & 0.0240 & -0.0938 & 0.0666 & 0.0017 & 0.0007 & 0.0002 & 0 & 0 \\ 0.0002 & 0.0016 & 0.0387 & -0.0947 & 0.0496 & 0.0040 & 0.0006 & 0.0000 & 0 \\ 0.0001 & 0.0006 & 0.0033 & 0.0636 & -0.1774 & 0.1060 & 0.0037 & 0.0001 & 0 \\ 0.0000 & 0.0003 & 0.0012 & 0.0035 & 0.0503 & -0.1610 & 0.1012 & 0.0040 & 0.0004 \\ 0 & 0.0002 & 0.0001 & 0.0013 & 0.0048 & 0.1028 & -0.1976 & 0.0622 & 0.0261 \\ 0 & 0 & 0.0018 & 0.0029 & 0.0050 & 0.0447 & 0.1346 & -0.2838 & 0.0948 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{pmatrix},$$

and for the momentum parameters,

$$\boldsymbol{\alpha} = (0.031, 0.1291) \qquad \text{and} \qquad \boldsymbol{\beta} = (3.5234, 1.7095).$$

One interesting observation arising from calibration is the difference of momentum parameters across the investment and the speculative downgrades. There is apparently more momentum in the speculative downgrades than in the investment downgrades, namely, the momentum intensity is larger and lasts longer in speculative grades[2].

---

[1]The MCMC algorithm, written in MATLAB, took ≈ 8.5 hours to run on a Intel Xeon E7-4660 v4 2.2GHz processor.
[2]Note that both $(0.1 \approx) \alpha_1\beta_1 < \alpha_2\beta_2 (\approx 0.2)$ and $\beta_1 > \beta_2$.

15

This may seem counter-intuitive, however, setting a credit rating ultimately involves combining information from various sources and making a judgement on the exposure of that company (sovereign) to different risks. As discussed in (Couderc 2008, Chapter 5 and 6), there appears to be a noticeable difference on which information influences downgrades/defaults for investment-grade and speculative-grade obligors. This points towards an intrinsic difference between these classes of ratings and thus it is not too surprising that our momentum model also shows a difference. From a practical point of view, the model suggests that a downgrade in a speculative-grade company is more damming for future performance, the information that influences speculative-grade rating changes implies deeper issues within the company and hence higher chances of further downgrades/default.

4.4.   *Bayesian Information Criterion*

Let us give some justification for the use of this model. We have argued that a point process style model is a strong choice and to keep the model as robust and simple as possible we added four extra "momentum parameters" (with relation to the CTMC model). We believe four to be the optimal choice due to the fact that only adding two parameters does not yield as good a fit to the observed data and adding parameters to every rating does not seem appropriate, since we do not have enough transitions across all ratings to obtain a reliable fit. We therefore did not consider more momentum groups than investment and non-investment grade.

As we have access to a full data set, one can directly calculate the MLE for the $\mathbf{Q}$ generator matrix of the Markov model setting. Therefore we can test our momentum model against the purely Markov model.

The Markov model is a particular case of our momentum model, set $\alpha_i = 0$ and β_i a constant for $i \in 1, 2$. Hence, a priori the non-Markov model stands to fit the data better (in the sense of achieving a likelihood at least as large). The question we look to answer is, are we actually capturing the data better or just overfitting? To do this we calculate the Bayesian Information Criterion (BIC), it is a common test used in statistics for model selection and is known to penalize model complexity more than other statistical tests, such as the Akaike information criterion (see (Claeskens and Hjort 2008, Chapter 3)). We believe this feature makes the BIC a good test to justify our more complex model. The BIC for a model M can be written as (some authors use the negative of this)

$$\mathrm{BIC}(M) = 2 \log\left(L(M|D)\right) - \log(n)\mathrm{dim}(M),$$

where n refers to the number of data points and $\mathrm{dim}(M)$ is the number of parameters in the model. From a given set of models, the model with the largest BIC is taken as the better one. Naturally, the indicator of how much "better" one model is over another is the difference in the BIC, where a BIC difference strictly greater than 10 is taken as very strong evidence of the model superiority.

	BIC
Difference	$138.5 \gg 10$

Table 3.   The BIC difference between the non-Markov and Markov model on the Moody's dataset.

The result in Table 3 gives us confidence that our non-Markov model captures reality better without overfitting and with sufficient parsimony with relation to the Markov (CTMC) one.

4.5.   *Examples and testing*

Probabilities of default as maps of time: Markov Vs. non-Markov. One important aspect of the non-Markov theory is how it impacts the estimates for the TPM and the transition probabilities.

REMARK 4.2 (Obtaining transition probabilities and model simulation)   *In the standard Markov set up, the TPM is calculated using (3.1). In the non-Markov set up we do not have such a simple relation, hence we are forced to use Monte Carlo techniques. In this case, we prescribe multiple companies in each rating at the start (we used a total of 10^7) and simulate individual transitions according to the point process model. By recording the rating of each company at various points in time (see below) we can then build transition matrices over several time horizons in the same way one builds an empirical TPM.*

The simulation of our momentum model is similar to that of a standard CTMC, i.e. based on the current state one simulates a "jump time" then simulates the new state to jump into. The main difference here is the added complexity of the time and history dependence that exists in our momentum model. To simulate the jump time of a fixed company we use the standard accept/reject method introduced in (Ogata 1981, Algorithm 2) for varying intensities. For each accepted jump time, we then calculate the transition probabilities based on this time and simulate the jump to the next state. We then repeat this process for each company up until the time horizon required.

It is of particular interest to understand how the evolution in time of the probabilities of default change when using the CTMC Markovian and our non-Markovian model. Using the calibrated model, Figure 5 details the probabilities of defaults for the various ratings as maps in time.
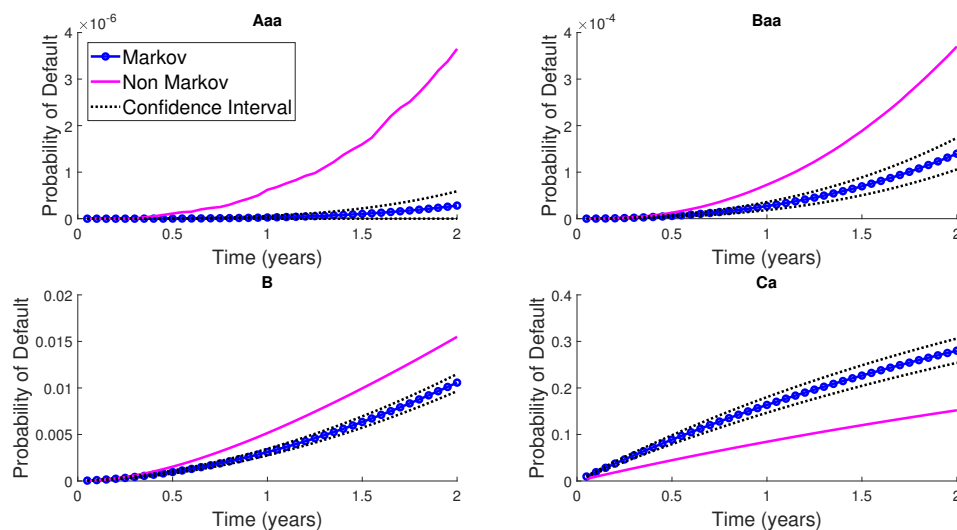

Figure 5. The probability of default given by each model for various ratings as a function of time.

The first observation one can make from Figure 5 is, the non-Markov model produces higher probabilities of default, except for the Ca rating (the non-Markov default probability is also lower for rating Caa). The reason for this is precisely the non-Markovianity in the data. In a Markov framework, all companies in the same rating are treated the same, consequently, it is unlikely that an investment-grade company will continue to downgrade quickly while the non-Markov model allows for this.

On the other hand, companies may enter rating Ca before defaulting, hence in the momentum model, some companies in this rating are carrying an extra term making default more likely. This implies we can account for a larger number of defaults while keeping the **Q** matrix Ca to C entry smaller. This is not the case in the Markov model and thus to produce enough defaults from Ca one makes the **Q** matrix entry larger. Consequently, the Markov model overestimates the default probability for obligors initially rated Ca.

Probabilities of default: Empirical Vs. Markov Vs. non-Markov. To test how reliable these results are, we can compare one-year probabilities of default as estimated from each calibrated model compared to that we observe from the data. To do so, we fix some time horizon T (one-year here) and consider all companies that have either defaulted or not withdrawn by this period. We

then build an empirical TPM over this horizon based on the company's rating at time zero and T. Concentrating solely on probabilities of default we obtain the results in Table 4.

Ratings \ Model	Investment-Grade				Speculative-Grade			
	Aaa	Aa	A	Baa	Ba	B	Caa	Ca
Empirical	0	0	0	0.0004000	0.0005	0.0012	0.0064	0.0563
non-Markov	1×10^{-6}	4×10^{-6}	0.0000125	0.0000734	0.0011	0.0052	0.0298	0.0845
Markov	3×10^{-8}	2.5×10^{-7}	4.86×10^{-7}	0.0000271	0.0002	0.0031	0.0407	0.1635
# Companies per rating at $t = 0$	413	1313	2232	2318	2021	4504	1333	59

Table 4.   Comparing one-year probability of defaults of each model against the empirical observations. For reference we add explicitly the number of companies per rating at starting time $t = 0$ as in the Moody's data set.

The results in Table 4 are interesting because they highlight stark differences in the models. Starting with the investment-grade, unfortunately, we do not have enough data to fully assess default probabilities at this level. The only grade for which a default within a year is observed is Baa and is higher than what both models predict. One reason the momentum model may not capture this probability as well is the way we have set up the momentum parameters, i.e. an investment and speculative set, and Baa is at the turning point. On the other hand, this number is estimated from a smaller number of defaults so is subject to a larger error. Comparing the Markov and non-Markov, it is unsurprising that our model makes investment-grade defaults more likely.

For the speculative-grades, one observes that Ca and Caa firms have lower one-year default probabilities in the non-Markov model and these estimates are closer to the empirical observations. This is exactly due to the reason mentioned previously, companies downgrading into Ca and Caa "poison" the data in the Markov setting. Implying that in a Markov world a company initially rated Caa or Ca is viewed to be riskier than it actually is.

The difference between the models may have a large impact on a bank's capital requirements for regulation. Although the non-Markov model makes most ratings riskier than the Markov model, we feel it provides a more accurate reflection of default risk.

REMARK 4.3 (Limitations from censored data)   *Unfortunately, in our study, we are limited to small-time horizons due to censored data. Namely, since the default is absorbing, as soon as a company defaults, we keep that information up to the terminal time. However, many companies are only rated over a few years before withdrawing and therefore if we look at empirical TPMs over longer horizons they are built with less (non-default) data. Since we do not want to use the Markov assumption, there does not appear to be a way to incorporate this lost data. Therefore we can only obtain "accurate" numbers on short time scales.*

5.   Summary

In the first part of this paper we have shown how one can evaluate errors in the transition matrices of continuous-time Markov chains at the level of discretely observed data using new closed-form expressions. These results reduced the computation of confidence intervals to less than one half of the time needed by current approaches. Moreover, and of practical importance, by employing the Delta method, our results provide an intuitively interpretable understanding of uncertainty in the model output, the probabilities of default.

In the second part, we have shown the significance of being able to capture non-Markov effects in rating transitions. Comparing against empirical probabilities of default and the classical Markov chain model, one finds a tendency for the Markov chain model to overestimate on some speculative-grades and underestimate on investment-grades. We address this issue by providing a parsimonious model that better captures default probabilities (where empirically observed). Moreover, the non-Markov model points towards significantly higher probabilities of default for investment-grades, where such values are not empirically observed, thus making it more prudent.

We believe that the model we present provides a more accurate view of reality and hence should be considered in credit risk modelling. These observations further highlight the importance of understanding so-called *model risk* and its potential impact in quantitative risk analysis in general.

References

Altman, E. and Kao, D.L., The implications of corporate bond ratings drift. *Financial Analysts Journal*, 1992, **48**, 64–75.

Andersen, P.K., Borgan, O., Gill, R.D. and Keiding, N., *Statistical models based on counting processes*, 2012, Springer Science & Business Media.

Andersen, P.K., Hansen, L.S. and Keiding, N., Non-and semi-parametric estimation of transition probabilities from censored observation of a non-homogeneous Markov process. *Scandinavian Journal of Statistics*, 1991, pp. 153–167.

Bacry, E., Mastromatteo, I. and Muzy, J.F., Hawkes processes in finance. *Market Microstructure and Liquidity*, 2015, **1**, 1550005.

Bielecki, T.R., Crépey, S. and Herbertsson, A., Markov chain models of portfolio credit risk. In *The Oxford Handbook of Credit Derivatives*, chap. 10, pp. 327–382, 2011, Oxford University Press.

Bladt, M. and Sørensen, M., Statistical inference for discretely observed Markov jump processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2005, **67**, 395–410.

Bladt, M. and Sørensen, M., Efficient estimation of transition rates between credit ratings from observations at discrete time points. *Quantitative Finance*, 2009, **9**, 147–160.

Cappé, O., Moulines, E. and Rydén, T., *Inference in hidden Markov models*, Springer Series in Statistics, 2005, Springer, New York, With Randal Douc's contributions to Chapter 9 and Christian P. Robert's to Chapters 6, 7 and 13, With Chapter 14 by Gersende Fort, Philippe Soulier and Moulines, and Chapter 15 by Stéphane Boucheron and Elisabeth Gassiat.

Carey, M. and Hrycay, M., Parameterizing credit risk models with rating data. *Journal of Banking & Finance*, 2001, **25**, 197 – 270.

Christensen, J.H., Hansen, E. and Lando, D., Confidence sets for continuous-time rating transition probabilities. *Journal of Banking & Finance*, 2004, **28**, 2575–2602.

Claeskens, G. and Hjort, N.L., Model selection and model averaging. *Cambridge Books*, 2008.

Couderc, F., Credit Risk and Ratings: Understanding Dynamics and Relationships with Macroeconomics. PhD thesis, Ecole Polytechnique Fédérale de Lausanne, 2008.

Cox, D.R. and Oakes, D., *Analysis of survival data*, 1984, Routledge.

Daley, D.J. and Vere-Jones, D., *An introduction to the theory of point processes. Vol. I*, Second , Probability and its Applications (New York), 2003, Springer-Verlag, New York, Elementary theory and methods.

Daley, D.J. and Vere-Jones, D., *An introduction to the theory of point processes. Vol. II*, Second , Probability and its Applications (New York), 2008, Springer, New York, General theory and structure.

D'Amico, G., Janssen, J. and Manca, R., Downward migration credit risk problem: a non-homogeneous backward semi-Markov reliability approach. *Journal of the Operational Research Society*, 2016, **67**, 393–401.

Dassios, A. and Zhao, H., Exact simulation of Hawkes process with exponentially decaying intensity. *Electronic Communications in Probability*, 2013, **18**, 1–13.

dos Reis, G. and Smith, G., Robust and consistent estimation of generators in credit risk. *Quantitative Finance*, 2018, **18**, 983–1001.

Frydman, H. and Schuermann, T., Credit rating dynamics and Markov mixture models. *Journal of Banking & Finance*, 2008, **32**, 1062–1075.

Gilks, W.R., Richardson, S. and Spiegelhalter, D.J., Introducing Markov Chain Monte Carlo. *Markov chain Monte Carlo in practice*, 1996, **1**, 19.

Inamura, Y., Estimating continuous time transition matrices from discretely observed data. Technical report, Citeseer, 2006.

Kalbfleisch, J. and Lawless, J.F., The analysis of panel data under a Markov assumption. *Journal of the American Statistical Association*, 1985, **80**, 863–871.

Knight, K., *Mathematical statistics*, Chapman & Hall/CRC Texts in Statistical Science Series, 2000, Chapman & Hall/CRC, Boca Raton, FL.

Koopman, S.J., Lucas, A. and Monteiro, A., The multi-state latent factor intensity model for credit rating transitions. *Journal of Econometrics*, 2008, **142**, 399–424.

Korolkiewicz, M.g.W., A dependent hidden Markov model of credit quality. *Int. J. Stoch. Anal.*, 2012, pp. Art. ID 719237, 13.

Kreinin, A. and Sidelnikova, M., Regularization algorithms for transition matrices. *Algo Research Quarterly*, 2001, **4**, 23–40.

Lando, D. and Skodeberg, T.M., Analyzing rating transitions and rating drift with continuous observations. *Journal of Banking and Finance*, 2002, **26**, 423–444.

Lehmann, E.L. and Casella, G., *Theory of point estimation*, Second , Springer Texts in Statistics, 1998, Springer-Verlag, New York.

Løffler, G., Avoiding the rating bounce: why rating agencies are slow to react to new information. *Journal of Economic Behavior & Organization*, 2005, **56**, 365–381.

McNeil, A.J., Frey, R. and Embrechts, P., *Quantitative Risk Management: Concepts, Techniques and Tools*, 2005 (Princeton: Oxford).

McNeil, A.J. and Wendin, J.P., Bayesian inference for generalized linear mixed models of portfolio credit risk. *Journal of Empirical Finance*, 2007, **14**, 131–149.

Nickell, P., Perraudin, W. and Varotto, S., Stability of rating transitions. *Journal of Banking & Finance*, 2000, **24**, 203–227.

Oakes, D., Direct calculation of the information matrix via the EM. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 1999, **61**, 479–482.

Ogata, Y., On Lewis' simulation method for point processes. *Information Theory, IEEE Transactions on*, 1981, **27**, 23–31.

Ogata, Y., Statistical models for earthquake occurrences and residual analysis for point processes. *Journal of the American Statistical association*, 1988, **83**, 9–27.

Pfeuffer, M., ctmcd: An R Package for Estimating the Parameters of a Continuous-Time Markov Chain from Discrete-Time Data. *The R Journal*, 2017, **9**, 127–141.

Van Loan, C., Computing integrals involving the matrix exponential. *Automatic Control, IEEE Transactions on*, 1978, **23**, 395–404.

Wilcox, R., Exponential operators and parameter differentiation in quantum physics. *Journal of Mathematical Physics*, 1967, **8**, 962–982.

Appendix A: Fundamentals of Discretely Observed Markov Processes

The EM algorithm for this problem (See Section 3) is discussed in detail in Bladt and Sørensen (2005) and dos Reis and Smith (2018) and we encourage the reader to consult these texts for further information. For

completeness, we present a brief review of the EM algorithm for the setting of continuous-time Markov chains.

For convergence of the EM algorithm, one works under the following assumption.

ASSUMPTION A.1 (Element constraint) *Similar to Bladt and Sørensen (2005), we will use a manual space constraint to obtain the convergence. Take $1 > \epsilon > 0$, such that for $i \neq j$, $q_{ij} < 1/\epsilon$. Moreover, we assume adjacent mixing, namely, for $i \in \{2, \ldots, h-1\}$, $q_{i,i\pm 1} > \epsilon$ and $q_{1,2} > \epsilon$.*
We denote the space of generator matrices which satisfy this condition as Λ_ϵ.

This assumption is a trivial constraint when one works in credit risk as it requires that: (a) firms can be upgraded or downgraded by one rating which is clearly the case; and (b) that changes in ratings do not happen too fast which is also the practical case.

Let $(X(t))_{t \geq 0}$ be a stochastic process over the finite state space $\{1, \ldots, h\}$. Associated to $X(t)$ is, for i, j in the state space, $\mathbf{K}_{ij}(t)$ the number of jumps from i to j in the interval $[0, t]$ and by $\mathbf{S}_i(t)$ the holding time of state i in the interval $[0, t]$. The EM algorithm is then given by,

  (i) Take an initial intensity matrix $\mathbf{Q}$ and a small positive value ϵ, so $\mathbf{Q} \in \Lambda_\epsilon$.

 (ii) While the convergence criterion is not met and $\mathbf{Q} \in \Lambda_\epsilon$,
  (1) E-step: calculate $\mathbb{E}_{\mathbf{Q}}[\mathbf{K}_{ij}(T)|\mathbf{P}]$ and $\mathbb{E}_{\mathbf{Q}}[\mathbf{S}_i(T)|\mathbf{P}]$.
  (2) M-step: set $q'_{ij} = \mathbb{E}_{\mathbf{Q}}[\mathbf{K}_{ij}(T)|\mathbf{P}]/\mathbb{E}_{\mathbf{Q}}[\mathbf{S}_i(T)|\mathbf{P}]$, for all $i \neq j$ and set q_{ii} appropriately.
  (3) Set $\mathbf{Q} = \mathbf{Q}'$ (where $\mathbf{Q}'$ is the matrix of q's) and return to E-step.

(iii) End while and return $\mathbf{Q}$.

By (dos Reis and Smith 2018, Theorem 2.10), provided the algorithm does not hit the boundary of Λ_ϵ, we obtain convergence (in distribution and parametric) to a stationary point. Typically the E-step in the EM algorithm needs to be calculated numerically, however dos Reis and Smith (2018) following Van Loan (1978) and Inamura (2006) obtained the following result.

PROPOSITION A.2 *Let e_i be the column vector of length h which is one at entry i and zero elsewhere, further let us define the $2h$-by-$2h$ matrices $\mathbf{C}_\gamma^{(\alpha\beta)}$ and $\mathbf{C}_\phi^{(\alpha)}$ as,*

$$\mathbf{C}_\gamma^{(\alpha\beta)} := \begin{bmatrix} \mathbf{Q} & q_{\alpha\beta}\mathbf{e}_\alpha\mathbf{e}_\beta^\intercal \\ 0 & \mathbf{Q} \end{bmatrix} \quad and \quad \mathbf{C}_\phi^{(\alpha)} := \begin{bmatrix} \mathbf{Q} & \mathbf{e}_\alpha\mathbf{e}_\alpha^\intercal \\ 0 & \mathbf{Q} \end{bmatrix} \quad \alpha, \beta \in \{1, \cdots, h\}.$$

Consider a CTMC X observed at n time points $0 \leq t_1 < t_2 < \cdots < t_n$; denote by y_s the state of the chain at time t_s, i.e. $y_s := X(t_s)$. Then, the expected jumps and holding times across observations are,

$$\mathbb{E}_{\mathbf{Q}}[\mathbf{K}_{ij}(t)|\mathbf{y}] = \sum_{s=1}^{n-1} \frac{\left(\exp(\mathbf{C}_\gamma^{(ij)}(t_{s+1} - t_s))\right)_{y_s, h+y_{s+1}}}{(\exp(\mathbf{Q}(t_{s+1} - t_s)))_{y_s, y_{s+1}}},$$

$$\mathbb{E}_{\mathbf{Q}}[\mathbf{S}_i(t)|\mathbf{y}] = \sum_{s=1}^{n-1} \frac{\left(\exp(\mathbf{C}_\phi^{(i)}(t_{s+1} - t_s))\right)_{y_s, h+y_{s+1}}}{(\exp(\mathbf{Q}(t_{s+1} - t_s)))_{y_s, y_{s+1}}}.$$

When one only has access to an observed sequence of TPMs $\mathbf{P}$ with equal observation length we obtain,

$$\mathbb{E}_{\mathbf{Q}}[\mathbf{K}_{ij}(T)|\mathbf{P}] = \sum_{u=1}^{M} \sum_{s=1}^{h} \sum_{r=1}^{h} \mathbf{P}_{sr}^u(t) \frac{\left(\exp(\mathbf{C}_\gamma^{(ij)}t)\right)_{s, h+r}}{(\exp(\mathbf{Q}t))_{s, r}},$$

$$\mathbb{E}_{\mathbf{Q}}[\mathbf{S}_i(T)|\mathbf{P}] = \sum_{u=1}^{M} \sum_{s=1}^{h} \sum_{r=1}^{h} \mathbf{P}_{sr}^u(t) \frac{\left(\exp(\mathbf{C}_\phi^{(i)}t)\right)_{s, h+r}}{(\exp(\mathbf{Q}t))_{s, r}},$$

where $M = T/t$ (the number of observations) and $\mathbf{P}^u$ is the TPM of the u-th observation.

Roughly speaking, the above formula is taking each row in the TPM to contain equal amounts of information (observations). When one knows the number of transitions between the states $\mathbf{N}$, then $\mathbf{P}_{sr}^u(t)$ is replaced by $\mathbf{N}_{sr}(u)$, where $\mathbf{N}_{sr}(u)$ is the number of observed transitions in observation u.

The M-step is just the ratio of these two quantities and thus the results yield closed-form expressions for the EM algorithm's steps making the algorithms much faster (see results in dos Reis and Smith (2018)).

Appendix B: Proof of Theorem 3.3

The proof relies on the multivariate delta method, see (Lehmann and Casella 1998, Theorem 8.16).

PROPOSITION B.1 (Delta Method) *Let* $(X_{1\nu}, \ldots, X_{s\nu})$, $\nu = 1, \ldots, n$, *be* n *independent* s-*tuples of random variables with* $\mathbb{E}[X_{i\nu}] = \xi_i$ *and* $cov(X_{i\nu}, X_{j\nu}) = \sigma_{ij}$. *Let* $\bar{X}_i$ *denote the empirical mean,* $\bar{X}_i := \sum_\nu X_{i\nu}/n$, *and suppose that* h *is a real-valued function of* s *arguments with continuous first partial derivatives. Then,*

$$\sqrt{n}\Big(h(\bar{X}_1, \ldots, \bar{X}_s) - h(\xi_1, \ldots, \xi_s)\Big) \xrightarrow{Dist} \mathcal{N}(0, v^2), \quad v^2 = \sum_i \sum_j \sigma_{ij} \frac{\partial h}{\partial \xi_i} \frac{\partial h}{\partial \xi_j}, \quad provided\ v^2 > 0.$$

We now have the necessary material to prove our result.

Proof of Theorem 3.3. The assumption of asymptotic normality implies the expectation and covariance assumption of Proposition B.1. Moreover, it follows from standard results in likelihood based inference that $\boldsymbol{\sigma} \approx -\mathbf{H}(\hat{\mathbf{Q}})^{-1}$ (see (Knight 2000, Chapter 5.4)).

For the partial derivatives of the probability matrix, it follows immediately by arguments in Section 3.1. Also note that this representation implies that the first partial derivatives of p_{ij} exist and are continuous.

To complete the proof, we need only to show that the RHS of (3.4) is strictly positive. Firstly, at a maximum $\mathbf{H}$ is negative definite (hence $\mathbf{H}^{-1}$ is also negative definite), therefore it is enough to have that $\partial p_{ij}/\partial \mathbf{V}_{\hat{\mathbf{Q}}} \neq 0$ around the MLE. Observing the latter is one of the theorem's assumptions concludes the proof. $\qquad\square$

Appendix C: Overview of Point Processes

Let us discuss how we look to embed history dependence into the model. We are interested in Hawkes processes (a specific type of *self-exciting* point process) which have intensities of the form

$$\lambda_t = \mu + \int_0^t \phi(t - s)\mathrm{d}N_s\,.$$

Hawkes processes are used to model many different phenomena, from earthquake occurrence to high frequency trading, see Ogata (1988) and Bacry *et al.* (2015). Setting $\phi = 0$ yields a constant intensity and this is equivalent to the Markov setting. However, ϕ allows us to vary the intensity with past events which is key for *momentum* since past downgrades influence future transitions. As described in the introduction, a Hawkes process is just a counting process (generalising a Poisson process), hence it would imply that a rating transition had occurred and not which rating we have moved into. The latter being key, we consider processes which take values on some state space: such processes are known as *marked point process (MPPs)*, see (Daley and Vere-Jones 2003, Section 6.4). MPPs are point processes on a product space $\mathcal{T} \times \mathcal{K}$, that is, we return a set of values, $\{t_k, \kappa_k\}$ for $k = 1, 2, \ldots$, where one thinks of t_k as the event time of the point process (with intensity λ) and κ_k of the "mark" associated to the event. These notions are what we shall use, but in general t_k can be multidimensional e.g. to include spatial dependence. In our case we have $\kappa_k \in \{1, \ldots, h\}$ namely, it denotes the ratings which ensure our marked point process to be well defined.

The likelihood of a single realisation of a MPP, is given in (Daley and Vere-Jones 2003, p.251),

$$L = \prod_{i=1}^{N_g(T)} \lambda_g^*(t_i) f^*(\kappa_i|t_i) e^{-\int_0^T \lambda_g^*(u)\mathrm{d}u}\,,$$

where we have the following notation, N_g is the set of events occur, λ_g is the intensity and f is the so-called *mark's distribution*. The * symbolises that the intensity and mark distribution depend on previous events. Namely, the intensity at time t_i, $\lambda_g^*(t_i)$ depends on the previous events, $\{(t_1, \kappa_1), \ldots, (t_{i-1}, \kappa_{i-1})\}$. Also note the distinction that $\lambda_g^*(t_i)$ does not depend on the mark κ_i, but the mark κ_i is allowed to depend on time t_i. The subscript g is a common notation used to imply this is the ground process, which in our case is simply the timing of the upgrades/downgrades. By allowing the intensity and hence the number of jumps and the mark distribution to depend on previous events we can easily change the probability of upgrade/downgrade and thus embed rating momentum into the process. Further details on likelihoods of MPP can be found in (Daley and Vere-Jones 2003, Section 7.3).

One reason that we believe MPPs are a good choice for this particular problem is that one can view them as a natural generalisation of CTMCs. This is apparent from the likelihood since, letting $\lambda = q_i$ and $f = q_{ij}/q_i$ we recover the likelihood of a CTMC.