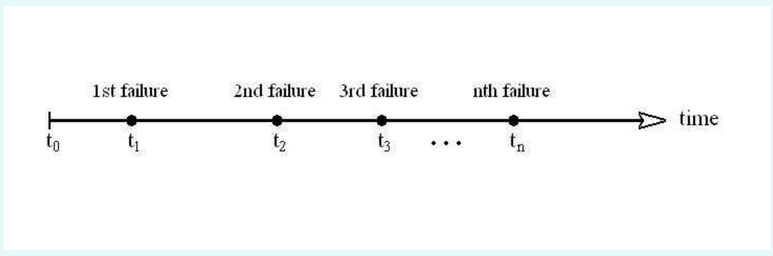
A Hidden Semi-Markov Model for the Occurrences of Water Pipe Bursts

Theodoros Economou
Trevor Bailey, Zoran Kapelan

Introduction

Water companies (especially in the UK) have a need for proper and accurate predictions of water pipe bursts or blockages. In addition to the fact that processes driving the occurrences of pipe failures are complex and often unmeasurable, the available historical data is often scarce and unreliable.

Engineering experience suggests that water pipes may be regarded as repairable components that deteriorate in time. Often, pipe failure data are expressed failure times within an observation period, therefore a common modelling approach is to assume that failures occur as a stochastic process.



In particular, the *non-homogenous Poisson process* (NHPP) is an appropriate model mainly because its intensity function $\lambda(t;\boldsymbol{x})$ varies with time (and possibly other covariates) so the process has a *time-dependent failure rate*.
Water pipes are buried underground and as such, the mechanisms that give rise to the failures are influenced by all kinds of external processes that are usually unobserved. A sudden change in climate for instance may cause a pipe to enter a state where it is failing more than usual for a short period of time. In this work we therefore propose use of hidden Markov models (HMM) in conjunction with the NHPP in order to capture situations where an underlying latent process is affecting the state of a pipe.

Holding Times

In an HMM, the parameters of a model vary according to the (discrete) states of an unobserved Markov chain with transition matrix $\boldsymbol{P}=\{p_{ij}\}$ and initial distribution $\boldsymbol{\pi_i}$ where $i,j\in\{1,2,\ldots,S\}$ the state space of the chain. Combining this HMM with the NHPP results in an NHPP-HMM failure rate which depends on the state that the chain is in, i.e. $\lambda(t,s;\boldsymbol{x})$. By doing this, the resulting process that generates the failures allows for both serial dependence and overdispersion (MacDonald and Zucchini, 1997).

It can be shown that for a discrete Markov chain, the length of time τ that a state i occupies is implicitly distributed as:

$$h_i(\tau)=(p_{ii})^{\tau-1}(1-p_{ii})$$

In other words, the *holding times* are *geometrically distributed*. As a result, the holding times inherit the *'memoryless'* property of the geometric distribution i.e.:

$$Pr(\tau>S+T|\tau>T)=Pr(\tau>S)$$

This effectively means that while in a given state, the behaviour of the Markov chain remains unaffected by the length of time that has passed. In the context of the NHPP however, this consequence can be unattractive since the failure rate which is driven by the chain does depend on length of time that has passed potentially limiting the flexibility of the NHPP-HMM.
Here we consider a hidden semi-Markov formulation in which the distribution of the holding times can be specified explicitly. We assume a *negative binomial* distribution for $h_i(\tau)$.

Hidden semi-Markov NHPP

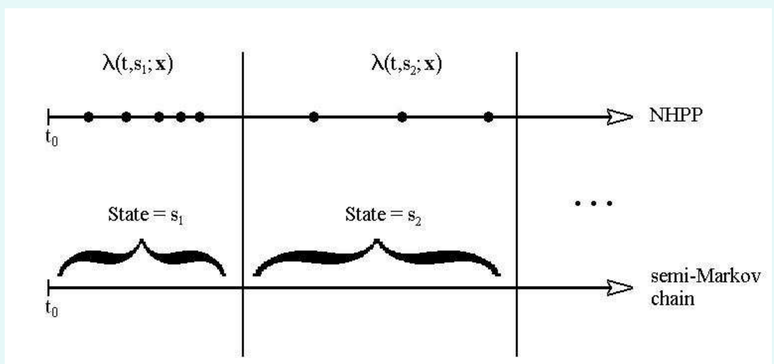
For the failure rate of the NHPP we assume a power law formulation:

$$\lambda(t,s;\boldsymbol{x})=\theta_s t^{\theta_s-1}\exp(\boldsymbol{\beta x})$$

where the shape parameter θ_s is state dependent and $\boldsymbol{\beta}=(\beta_0,\beta_1,\ldots,\beta_q)$ are the parameters of possible covariates $\boldsymbol{x}=(1,x_1,\ldots,x_q)$
Assuming that m failures occurred in $(t_0,t_{end}]$ at times $t_0<t_1\leq\ldots\leq t_m\leq t_{end}$, the likelihood of the NHPP is

$$\ell(t_0,t_1,\ldots,t_{end};s)=\prod_{k=1}^{m}\lambda(t_k,s;\boldsymbol{x})e^{-\Lambda_s(t_0,t_{end})}$$

where $\Lambda_s(t_0,t_{end})=\int_{t_0}^{t_{end}}\lambda(u,s;\boldsymbol{x})du$.
Also consider a realisation of the semi-Markov chain being observed in $(t_0,t_{end}]$ which starts at state s_1 holding for a time τ_1 then going to state s_2 holding for τ_2 and so on until it reaches s_N, the state held up to t_{end}.



The joint likelihood of these two processes, $L(t_0,t_1,\ldots,t_{end};\tau_1,\ldots,\tau_N,s_1,\ldots,s_N)$, is:

$$\pi_{s_1}h_{s_1}(\tau_1)\ell(t\in\tau_1;s_1)\prod_{j=2}^{N}p_{s_{j-1}s_j}h_{s_j}(\tau_j)\ell(t\in\tau_j;s_j)$$

where $t\in\tau_j$ represents the failure times that occurred in the interval τ_j.
However, in the NHPP-HSMM the chain is not being observed therefore the likelihood of the data (i.e. the failure times), $L(t_0,t_1,\ldots,t_{end})$, is given by summing the above likelihood over all possible states and holding times

$$\sum_{\tau_1=1}^{\infty}\cdots\sum_{\tau_N=1}^{\infty}\sum_{s_1=1}^{S}\cdots\sum_{s_N=1}^{S}L(t_0,t_1,\ldots,t_{end};\tau_1,\ldots,\tau_N,s_1,\ldots,s_N)$$

The complexity of the likelihood as well as the fact that in some cases it may be sensible to consider some of the variables as random effects, favour the employment of *MCMC* as the estimation mechanism.
The evaluation of $L(t_0,t_1,\ldots,t_{end})$ is prohibitively *computationally intensive* for any reasonable length of the observation period $(t_0,t_{end}]$. Fortunately, one can employ the idea of *recursive algorithms* used in the HMM literature (Rabiner, 1989) to efficiently evaluate the likelihood.

Forward Algorithm

The idea behind forward recursion is to consider a variable $\alpha_T(j)$ sequentially at each discrete time step $T\in\{1,\ldots,t_{end}-t_0\}$, where:

$$\alpha_T(j)=\Pr(t_0,t_1,\ldots,T;\text{ chain is in }j)$$

i.e. the probability of the data up to time T and the chain being in state j at T. One can then compute $\alpha_T(j)$ recursively:

$$\alpha_1(j)=\pi(j)h_j(1)\ell(t\in(0,1])$$

$$\alpha_2(j)=\pi(j)h_j(2)\ell(t\in(0,1,2])+$$

$$\sum_{i=1}^{S}\alpha_1(j)p_{ij}h_i(1)\ell(t\in(t_1,t_2])$$

$$\alpha_3(j)=\pi(j)h_j(3)\ell(t\in(0,1,2,3])+\cdots$$

etc.

The likelihood is then given by

$$L(t_0,t_1,\ldots,t_{end})=\sum_{j=1}^{S}\alpha_{t_{end}}(j)$$

Note that for the calculation of $\alpha_{t_{end}}(j)$, we use the survival function instead of $h_j()$ itself, effectively accounting for *right censoring* since we would not realistically expect the chain to change state exactly at t_{end}.
Once the likelihood is efficiently calculated, *Metropolis-Hastings* can then be used (Scott, 2002) to estimate the parameters of $\lambda(t_k,s;\boldsymbol{x})$, the transition matrix, the initial probability distribution and the parameters of the specified holding times.
The proposed model is currently being applied to a Canadian distribution network of water pipes. This network consists of 1349 pipes with 5425 recorded failures in the period 1945-2003.

References

MacDonald, I. and Zucchini, W. (1997). Hidden Markov and other models for discrete-valued time series. *Chapman and Hall.*

Rabiner, L. R. (1989). A tutorial on hidden Markov models and selected application in speech recognition. *Proceedings of the IEEE*, **77**, 257-285.

Scott, S. L. (2002). Bayesian methods for hidden Markov modelsa: Recursive computing in the 21st century *Journal of the American Statistical Association*, **97**, 337-351.