Proceedings of the 2018 Winter Simulation Conference
M. Rabe, A. A. Juan, N. Mustafee, A. Skoogh, S. Jain, and B. Johansson, eds.

MODELLING AND GENERATING NONHOMOGENEOUS POISSON PROCESSES USING A SPLINE FUNCTION

Lucy E. Morgan

STOR-i Centre for Doctoral Training in Partnership with Industry
Lancaster University
Lancaster, LA1 4YR, UK

ABSTRACT

Approaches to modelling nonhomogeneous Poisson processes (NHPPs) commonly use piecewise representations of the rate function. In reality, real-world rate functions are unlikely to take a piecewise form and therefore bias is introduced. We propose a spline function representation using a large number of knots. The resulting function, being both smooth and highly flexible, is able to take on a wide variety of functional shapes reducing the bias between it and the true process. Due to the added flexibility we control overfitting, and thus variability, by adding a penalty to the NHPP log-likelihood. Our approach optimizes the spline coefficient and penalty parameter combination by minimizing a modified AIC score. Our approach also leads to a simple method for arrival generation from the resulting spline function.

1 SPLINE FUNCTIONS

Suppose we observe a NHPP, $N(t)$, with true rate function $\lambda^c(t)$ on the interval $[0,T]$ over m days. We represent the rate function using a spline function; a linear combination of n d-degree B-spline basis functions,

$$\lambda(t) = \sum_{k=1}^{n} c_k B_{k,d}(t), \tag{1}$$

see (De Boor 1978). Here $B_{k,d}(t)$ is the k^{th} d-degree B-spline and c_k is its coefficient. B-splines have local support over $d+2$ knots and at any time t, only $d+1$ B-splines are non-zero. We allow the number of knots, and thus the number of B-splines, to be large allowing the resulting spline function to be highly flexible. Once the knots have been placed the B-spline functions are fixed and it is the spline coefficient vector, $\boldsymbol{c} = \{c_1, c_2, \ldots, c_n\}$, that completely determines the shape of the resulting spline function. It is therefore $\boldsymbol{c}$ that we aim to optimize using the penalized log-likelihood.

2 FITTING THE SPLINE FUNCTION

When fitting the spline function we consider the penalized log-likelihood, $l_p(\cdot)$. The penalty is a measure of the curvature of the rate function; it controls overfitting of the NHPP and thus the variability. Conditional on m days of observations, $\boldsymbol{t} = \{t_{11}, t_{12}, \ldots, t_{ma_m}\}$ where a_i denotes the number of arrivals on day i, the penalized log-likelihood is

$$\begin{aligned} l_p(\lambda(\boldsymbol{t};\boldsymbol{c})) &= l(\lambda(\boldsymbol{t};\boldsymbol{c})) - \theta \int_0^T \{\lambda''(u;\boldsymbol{c})\}^2 du \\ &= \sum_{i=1}^{m} \sum_{j=1}^{a_i} \log\left(\lambda(t_{ij};\boldsymbol{c})\right) - m \int_0^T \lambda(y;\boldsymbol{c}) dy - \theta \int_0^T \{\lambda''(u;\boldsymbol{c})\}^2 du \end{aligned}$$

where θ is the penalty parameter that controls the amount of penalization and $l(\cdot)$ is the unpenalized log-likelihood.

To find the optimal rate function, $\lambda(t)$, we search for the optimal combination $\{\theta,\boldsymbol{c}\}$ by minimizing a modified AIC score

$$\mathrm{AIC}_{mod} = -2\, l(\lambda(\boldsymbol{t};\boldsymbol{c})) + 2\, e,$$

where e is the effective degrees of freedom, see (Gray 1992).

3 OPTIMISING THE SPLINE COEFFICIENTS

For a fixed penalty value, θ, we find the optimal spline coefficients $\boldsymbol{c}$ by maximizing the penalized log-likelihood using a trust region approach, see (Conn, Gould, and Toint 2000). By convention the trust region approach is a minimization algorithm, we therefore solve the equivalent problem of minimizing the negative penalized log-likelihood. The trust region approach takes a local approximation of the penalized log-likelihood in the form of a second-order Taylor series expansion,

$$m(\boldsymbol{p}_h) = l_p(\boldsymbol{c}_h) + g(\boldsymbol{c}_h)^T \boldsymbol{p}_h + \frac{1}{2}\boldsymbol{p}_h^T H(\boldsymbol{c}_h)\boldsymbol{p}_h,$$

where $\boldsymbol{p}_h$ is the vector of proposed steps for each spline coefficient in iteration h, $\boldsymbol{c}_h$ is the current value of the spline coefficients and $g(\boldsymbol{c}_h)$ and $H(\boldsymbol{c}_h)$ are the gradient and Hessian of the penalized log-likelihood respectively. The key is in choosing each step $\boldsymbol{p}_h$ within some region in which we expect the Taylor series approximation to hold well. Known as the trust region sub-problem, the objective of the optimization in step h is

$$\min\, m(\boldsymbol{p}_h), \quad \text{subject to: } ||\boldsymbol{p}_h|| \leq \Delta$$

where Δ is the radius of the trust region and $||\cdot||$ denotes Euclidean distance. This is a convex quadratic optimization problem. The algorithm stops when the proposed step falls below some threshold ε. Once we have the optimal spline coefficients, $\boldsymbol{c}$, we can assess the modified AIC score of the spline fitted with penalty θ. The spline function with the minimum score is chosen.

4 ARRIVAL GENERATION

Once the optimal combination $\{\theta,\boldsymbol{c}\}$ has been found we generate arrivals by taking advantage of the composition of the spline function. Note that the maximum of (1), the spline function, is not clear but, as soon as the knots have been placed, the maximum of each individual B-spline is known. Further, when uniformly spaced B-splines, otherwise known as cardinal B-splines, are used this maximum is the same. Using the fact that the sum of n Poisson processes is again a Poisson process, we treat each of the n B-splines as an independent Poisson process, scaled by the spline coefficients, and generate arrivals from them using an efficient thinning algorithm, see (Klein and Roberts 1984). We can then treat the superposition of arrivals from all n scaled B-splines as arrivals from the spline function, $\lambda(t)$. When cardinal B-splines are used, this procedure simplifies as all n B-splines are transformations of the first B-spline and thus all arrival times can be generated using the first B-spline scaled by the n spline coefficients in turn.

REFERENCES

Conn, A. R., N. I. Gould, and P. L. Toint. 2000. *Trust Region Methods*, Volume 1. Philadelphia: Society for Industrial and Applied Mathematics.

De Boor, C. 1978. *A Practical Guide to Splines*, Volume 27. New York: Springer-Verlag.

Gray, R. J. 1992. “Flexible Methods for Analyzing Survival Data Using Splines, with Applications to Breast Cancer Prognosis”. *Journal of the American Statistical Association* 87 (420): 942–951.

Klein, R. W., and S. D. Roberts. 1984. “A Time-Varying Poisson Arrival Process Generator”. *Simulation* 43 (4): 193–195.