

ARRIVAL RATE APPROXIMATION BY
NONNEGATIVE CUBIC SPLINES

Farid Alizadeh^a
Nilay Noyan^c

Jonathan Eckstein^b
Gábor Rudolf^d

RRR 46-2004, DECEMBER, 2004

RUTCOR

Rutgers Center for
Operations Research
Rutgers University
640 Bartholomew Road
Piscataway, New Jersey
08854-8003
Telephone: 732-445-3804
Telefax: 732-445-5472
Email: rrr@rutcor.rutgers.edu
<http://rutcor.rutgers.edu/~rrr>

^aRUTCOR Rutgers, the State University of New Jersey 640 Bartholomew Road Piscataway, NJ 08854-8003, USA. E-mail: alizadeh@rutcor.rutgers.edu

^bBusiness School and RUTCOR Rutgers, the State University of New Jersey 640 Bartholomew Road Piscataway, NJ 08854-8003 USA. E-mail: jeckstei@rutcor.rutgers.edu

^cRUTCOR Rutgers, the State University of New Jersey 640 Bartholomew Road Piscataway, NJ 08854-8003 USA. E-mail: noyan@rutcor.rutgers.edu

^dRUTCOR Rutgers, the State University of New Jersey 640 Bartholomew Road Piscataway, NJ 08854-8003 USA. E-mail: grudolf@rutcor.rutgers.edu

ARRIVAL RATE APPROXIMATION BY NONNEGATIVE CUBIC SPLINES

Farid Alizadeh, Jonathan Eckstein, Nilay Noyan, Gábor Rudolf

Abstract. This paper describes an optimization method to approximate the arrival rate function of a non-homogeneous Poisson process based on observed arrival data. We estimate the arrival rate function by cubic splines, a standard class of versatile, twice-differentiable functions. We develop an optimization model that estimates cubic splines via the maximum likelihood principle. A critical feature of the model is that the splines are constrained to be everywhere nonnegative. We guarantee this constraint by using a result of Nesterov characterizing nonnegative polynomials via positive semidefinite matrices. We also describe versions of our model allowing for periodic arrival rate functions and input data of limited precision. We present numerical results using both real-life and artificially generated data sets.

Acknowledgements: Farid Alizadeh, Nilay Noyan, and Gábor Rudolf gratefully acknowledge the support of NSF through Grant # NSF-CCR-0306558 and of ONR through Contract # N00014-03-1-0042.

1 Introduction

Our goal in this paper is to provide a method to approximate the arrival rate of a non-homogeneous Poisson process based on observed arrival data. Such a method has applications in diverse areas including the management of periodically updated databases (see [4] and [5]). We estimate the arrival rate function by nonnegative cubic spline functions and the maximum likelihood principle, and ensure the nonnegativity of the approximating splines by applying a characterization of nonnegative polynomials due to Nesterov [8] based on positive semidefinite matrices. The next two sections present the necessary background information on non-homogeneous Poisson processes, cubic splines and nonnegative polynomials. The optimization model and its variants are described in Section 4, while Section 5 is concerned validating the model and determining the most appropriate number of spline knots. Numerical results — based on both real-life and artificially generated data sets — are presented in Section 6.

2 Likelihood Functions for a Non-homogenous Poisson Process

2.1 The basic case

Assume that we are given the set of observed arrival times $\mathbf{t} = (t_0, \dots, t_n)$, where $t_0 < t_1 < \dots < t_n$. For an arrival rate function $\lambda : [t_0, t_n] \rightarrow \mathbb{R}_+$ the likelihood of an arrival at time t_j given that there was an arrival at time t_{j-1} is $\lambda(t_j) \exp(-\int_{t_{j-1}}^{t_j} \lambda(t) dt)$. Using the convention $\ln(0) = -\infty$ the log-likelihood function is given by

$$L(\mathbf{t}, \lambda) = \sum_{j=1}^n \left[\ln \lambda(t_j) - \int_{t_{j-1}}^{t_j} \lambda(t) dt \right] = \sum_{j=1}^n \ln \lambda(t_j) - \int_{t_0}^{t_n} \lambda(t) dt. \quad (1)$$

2.2 Discrete data

In many practical situations, one does not have *exact* arrival time information, but instead data of the following form: the time interval of interest I is divided into subintervals $I = \bigcup_{j=1}^k I_j$, and we are given the number of arrivals n_j in each $I_j = [q_{j-1}, q_j]$; let $\mathbf{n} = (n_1, \dots, n_k)$ and $\mathbf{n} = \sum_{j=1}^k n_j$. For an arrival rate function $\lambda : I \rightarrow \mathbb{R}_+$ the probability of having n_j arrivals in I_j is $\frac{1}{n_j!} \left(\int_{q_{j-1}}^{q_j} \lambda(t) dt \right)^{n_j} \exp \left(- \int_{q_{j-1}}^{q_j} \lambda(t) dt \right)$. The log-likelihood function in this case is

$$L_d(\mathbf{n}, \lambda) = \sum_{j=1}^k \left[n_j \ln \left(\int_{q_{j-1}}^{q_j} \lambda(t) dt \right) - \ln n_j! \right] - \int_{q_0}^{q_k} \lambda(t) dt. \quad (2)$$

Note that the terms $\ln \mathbf{n}_j!$ are independent of λ , and therefore can be ignored for purposes of optimization. Our experimental data in Section 6 are in this form, although with very short, one-second periods I_j .

Another possibility is to artificially create a sequence of arrival times by introducing \mathbf{n}_j arrivals in every interval I_j ($j = 1, \dots, k$) and evaluate $L(\mathbf{t}, \lambda)$ as given in (1). In our numerical experiments, these arrivals were uniformly distributed random points in the respective intervals.

2.3 Periodic arrival rate functions

In this section, we consider the case in which the arrival rate of the Poisson process is periodic, that is $\lambda(t) = \lambda_0(t \pmod{T})$ for some period $T > 0$ and $\lambda_0 : [0, T) \rightarrow \mathbb{R}_+$. For simplicity, let us assume that we are given the arrival times $0 < t_1 < \dots < t_n < cT$ for the time period $[0, cT]$, $c \in \mathbb{N}$. Now, the log-likelihood function can be written in the form

$$L(\mathbf{t}, \lambda_0) = \sum_{j=1}^n \ln \lambda_0(t_j \pmod{T}) - c \int_0^T \lambda_0(t) dt. \quad (3)$$

For discrete data in the interval $I = [0, cT]$, the log-likelihood function is given by

$$L_d(\mathbf{n}, \lambda_0) = \sum_{j=1}^k \left[\mathbf{n}_j \ln \left(\int_{q_{j-1}}^{q_j} \lambda_0(t \pmod{T}) dt \right) - \ln \mathbf{n}_j! \right] - c \int_0^T \lambda_0(t) dt. \quad (4)$$

3 Nonnegative Cubic Splines

3.1 The spline property

A *cubic spline* is a continuous piecewise polynomial function constructed from third-order polynomials, constrained to have continuous first and second derivatives. The breakpoints separating the pieces of a spline are called *knots*. Let us consider a cubic spline function P on the interval $(0, T)$, with knots $0 = \mathbf{a}_0 < \mathbf{a}_1 < \dots < \mathbf{a}_m = T$ and coefficients $P_k^{(i)}$ ($i = 1, \dots, m$, $k = 0, 1, 2, 3$). For a point $t \in [\mathbf{a}_{i-1}, \mathbf{a}_i]$ the value of P is

$$P(t) = P^{(i)}(t) = \sum_{k=0}^3 P_k^{(i)}(t - \mathbf{a}_{i-1})^k. \quad (5)$$

The spline property that the P should have respectively continuous values, continuous first derivatives, and continuous second derivatives is expressed by the following equalities for $i = 1, \dots, m - 1$:

$$P_0^{(i+1)} - P_0^{(i)} - P_1^{(i)}(\mathbf{a}_i - \mathbf{a}_{i-1}) - P_2^{(i)}(\mathbf{a}_i - \mathbf{a}_{i-1})^2 - P_3^{(i)}(\mathbf{a}_i - \mathbf{a}_{i-1})^3 = 0 \quad (6)$$

$$P_1^{(i+1)} - P_1^{(i)} - 2P_2^{(i)}(a_i - a_{i-1}) - 3P_3^{(i)}(a_i - a_{i-1})^2 = 0 \quad (7)$$

$$2P_2^{(i+1)} - 2P_2^{(i)} - 6P_3^{(i)}(a_i - a_{i-1}) = 0. \quad (8)$$

If one also wants P to be a periodic with period T , then the values and derivatives of P at 0 and T should match, and thus one has the equalities

$$P_0^{(1)} - P_0^{(m)} - P_1^{(m)}(a_m - a_{m-1}) - P_2^{(m)}(a_m - a_{m-1})^2 - P_3^{(m)}(a_m - a_{m-1})^3 = 0 \quad (9)$$

$$P_1^{(1)} - P_1^{(m)} - 2P_2^{(m)}(a_m - a_{m-1}) - 3P_3^{(m)}(a_m - a_{m-1})^2 = 0 \quad (10)$$

$$2P_2^{(1)} - 2P_2^{(m)} - 6P_3^{(m)}(a_m - a_{m-1}) = 0. \quad (11)$$

3.2 Nonnegativity

In order for us to be able to interpret a function P as the arrival rate of a non-homogeneous Poisson process, it has to be everywhere nonnegative. The following theorem is a special case of a result of Nesterov given in [8].

Theorem 1 $P^{(i)}(t) = P_0^{(i)} + P_1^{(i)}(t - a_{i-1}) + P_2^{(i)}(t - a_{i-1})^2 + P_3^{(i)}(t - a_{i-1})^3 \geq 0$ for all $t \in (a_{i-1}, a_i)$ if and only if there exist parameters $x_i, y_i, z_i, s_i, v_i, w_i \in \mathbb{R}$ such that

$$\begin{aligned} P_0^{(i)} &= (a_i - a_{i-1})s_i \\ P_1^{(i)} &= x_i - s_i + 2(a_i - a_{i-1})v_i \\ P_2^{(i)} &= 2y_i - 2v_i + (a_i - a_{i-1})w_i \\ P_3^{(i)} &= z_i - w_i \\ x_i z_i &\geq y_i^2, \quad s_i w_i \geq v_i^2 \\ x_i, z_i, s_i, w_i &\geq 0. \end{aligned} \quad (12)$$

4 Optimization Models

4.1 The basic case

Our basic approach is to maximize the log-likelihood function subject to the constraint that λ is a nonnegative spline with knots $a_0 < a_1 < \dots < a_m$. Consider for example the case of a periodic arrival rate with period T , where we are given the exact arrival times $0 < t_0 < \dots < t_n < cT$ in the interval $[0, cT]$. Let us introduce the notation $\bar{t} = t \pmod{T}$ and define i_j such that $\bar{t}_j \in [a_{i_j-1}, a_{i_j}]$. Then we have the following model:

$$\max_{\mathbf{t}, \mathbf{P}} L(\mathbf{t}, \mathbf{P}) = \sum_{j=1}^n \ln \left[\sum_{k=0}^3 P_k^{(i_j)} (\bar{t}_j - \mathbf{a}_{i_j-1})^k \right] - c \sum_{i=1}^m \sum_{k=0}^3 P_k^{(i)} \frac{(\mathbf{a}_i - \mathbf{a}_{i-1})^{k+1}}{k+1}$$

subject to

$$\begin{aligned} P_0^{(i+1)} - P_0^{(i)} - P_1^{(i)}(\mathbf{a}_i - \mathbf{a}_{i-1}) - P_2^{(i)}(\mathbf{a}_i - \mathbf{a}_{i-1})^2 - P_3^{(i)}(\mathbf{a}_i - \mathbf{a}_{i-1})^3 &= 0, \quad i = 1 \dots m-1, \\ P_0^{(1)} - P_0^{(m)} - P_1^{(m)}(\mathbf{a}_m - \mathbf{a}_{m-1}) - P_2^{(m)}(\mathbf{a}_m - \mathbf{a}_{m-1})^2 - P_3^{(m)}(\mathbf{a}_m - \mathbf{a}_{m-1})^3 &= 0, \\ P_1^{(i+1)} - P_1^{(i)} - 2P_2^{(i)}(\mathbf{a}_i - \mathbf{a}_{i-1}) - 3P_3^{(i)}(\mathbf{a}_i - \mathbf{a}_{i-1})^2 &= 0, \quad i = 1 \dots m-1, \\ P_1^{(m)} - 2P_2^{(m)}(\mathbf{a}_m - \mathbf{a}_{m-1}) - 3P_3^{(m)}(\mathbf{a}_m - \mathbf{a}_{m-1})^2 &= 0, \\ 2P_2^{(i+1)} - 2P_2^{(i)} - 6P_3^{(i)}(\mathbf{a}_i - \mathbf{a}_{i-1}) &= 0, \quad i = 1 \dots m-1, \\ 2P_2^{(1)} - 2P_2^{(m)} - 6P_3^{(m)}(\mathbf{a}_m - \mathbf{a}_{m-1}) &= 0, \end{aligned} \tag{13}$$

$$\begin{aligned} P_0^{(i)} - (\mathbf{a}_i - \mathbf{a}_{i-1})s_i &= 0, \quad i = 1 \dots m, \\ P_1^{(i)} - x_i - s_i + 2(\mathbf{a}_i - \mathbf{a}_{i-1})v_i &= 0, \quad i = 1 \dots m, \\ P_2^{(i)} - 2y_i - 2v_i + (\mathbf{a}_i - \mathbf{a}_{i-1})w_i &= 0, \quad i = 1 \dots m, \\ P_3^{(i)} - z_i - w_i &= 0, \quad i = 1 \dots m, \\ x_i z_i - y_i^2 &\geq 0, \quad i = 1 \dots m, \\ s_i w_i - v_i^2 &\geq 0, \quad i = 1 \dots m, \\ x_i, z_i, s_i, w_i &\geq 0, \quad i = 1 \dots m. \end{aligned}$$

4.2 Properties of the optimization problem

Notice that the value of $P(\mathbf{t})$ for any given \mathbf{t} , along with the equalities (5)-(11) describing the spline property, is linear in terms of the coefficients defining P . Also, since for $(\tau_1, \tau_2) \subset (\mathbf{a}_{i-1}, \mathbf{a}_i)$

$$\int_{\tau_1}^{\tau_2} P(\mathbf{t}) d\mathbf{t} = \int_{\tau_1}^{\tau_2} P^{(i)}(\mathbf{t}) d\mathbf{t} = \sum_{k=0}^3 P_k^{(i)} \frac{(\tau_2 - \tau_1)^{k+1}}{k+1} \tag{14}$$

is a linear expression of the coefficients, so is $\int_{\tau_1}^{\tau_2} P(\mathbf{t}) d\mathbf{t}$ for any interval $(\tau_1, \tau_2) \subset (0, T)$. The nonnegativity condition (12) features two quadratic inequalities (along with linear equalities and inequalities) in terms of the coefficients and the parameters. The quadratic inequalities are equivalent to the following convex constraints:

$$x_i - \frac{y_i^2}{z_i} \geq 0, \quad s_i - \frac{v_i^2}{w_i} \geq 0, \quad i = 1 \dots m.$$

Therefore, the set of feasible points is convex. Since the objective function to be maximized is concave, (13) describes a convex nonlinear programming problem, so locally optimal solutions are also globally optimal.

4.3 Variations of the basic model

For the different scenarios described in Section 2 we obtain variations of the model (13). If the arrival rate is not periodic, the conditions corresponding to (9)-(11) are omitted and the objective $L(\mathbf{t}, \mathbf{P})$ is given by substituting (5) into (1). For the case of discrete data, the objective $L_d(\mathbf{t}, \mathbf{P})$ is obtained by maximizing the objective function (14) subject to the same constraints.

5 Verifying the model

5.1 Determining the optimal number of knots by k-folding

The results obtained from our models depend strongly on the number of knots used for the spline function. Selecting too small a number of knots overly restricts the family of functions available to estimate the arrival rate, while selecting too large a number might overfit the data. To determine the number of knots appropriate for a given data set, we use the technique of k-folding, also known as the leave-one-out cross validation method [10]. We now describe the procedure for the case of periodic arrivals. We first divide the set of observed complete time periods into k subsets D_1, \dots, D_k of equal size. Then, for each $i = 1, \dots, k$, estimate the arrival rate, leaving out D_i from our data set. Next, we examine how well this estimated arrival rate P_i describes the behavior of the process in the interval D_i by evaluating the appropriate log-likelihood function $L_{(d)}(D_i, P_i)$ (*i.e.* we consider only the data not used in obtaining the estimate). We then choose the number of knots for which, over a suitably large number of repetitions of the k-folding process, the average value of $L_{(d)}(D_i, P_i)$ is the highest.

5.2 Sensitivity to Interval Lengths for Discrete Data

In the case of discrete data, the results we obtain are also influenced by the length of the intervals I_j . Generally speaking, smaller interval lengths correspond to more precise information. We evaluate the influence of the interval length on the estimation process as follows: we start with a data set with small interval lengths. We then modify these data by merging some adjacent intervals (implying some loss of information), and apply our estimation method to the modified data set. In order to compare the approximate arrival rate functions obtained from different applications of the merging procedure, we evaluate their log-likelihood functions (2) on the original, non-merged data.

5.3 Testing on a known arrival rate function

Another way of testing the model’s performance is to consider (possibly generated) data coming from a Poisson process with a known arrival rate function, and compare the estimated arrival rate function to the original. The results can be used to determine the amount of data necessary to obtain a reasonably good fit.

6 Numerical Results

6.1 Results on a real data set

In this section we present results obtained for a data set of approximately 10,000 e-mail arrivals over a period of 446 days. The arrival times are given as integer seconds, so we have discrete information with one-second intervals. For the purpose of creating our figures, we calculated a simple step function approximation of the arrival rate by dividing the time period into 64 equal intervals, and estimating λ via (number of arrivals/length of interval) in each interval. The optimization models were written in AMPL, and solved by KNITRO (see [1] and [9]) on the NEOS servers (see [2], [3] and [6]).

6.1.1 Selecting the Appropriate Version of the Model

We used two versions of the model, respectively assuming daily and weekly periodicity. We also considered both a discrete model with one-second intervals, and a model with “simulated” exact arrival times spaced uniformly in the one-second interval of the data set. We produced splines with $m = 3, 6, 12, 24, 48, 96$ equidistant knots for the daily and $m = 7, 14, 21, 42, 84, 168$ equidistant knots for the weekly version. To compare the quality of solutions obtained from two versions, we evaluated the log-likelihood function as given in (4) for all splines.

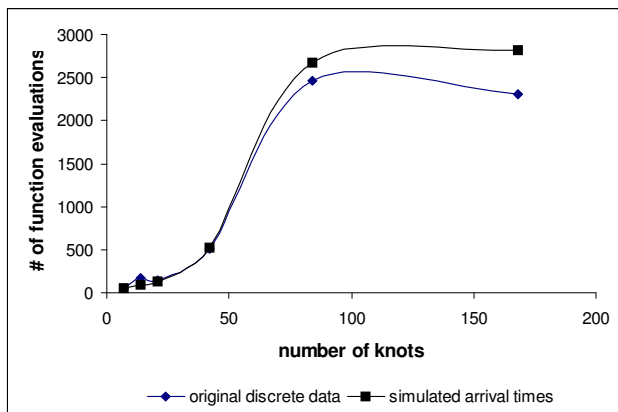


Figure 1

Using exact generated arrival times did not significantly change the results while resulting in a slight increase in computational time (see Figure 1).

Based on these results we select the version of our model using the original discrete data.

6.1.2 Periodicity

For e-mail arrivals, it is natural to assume either daily or weekly periodicity. Figure 2 shows our method's 48-knot spline estimate for daily periodicity, compared to the step-function approximation. In the remainder of this section, we consider the case of weekly periodicity, as the results provide a more detailed description of the behavior of the arrival process.

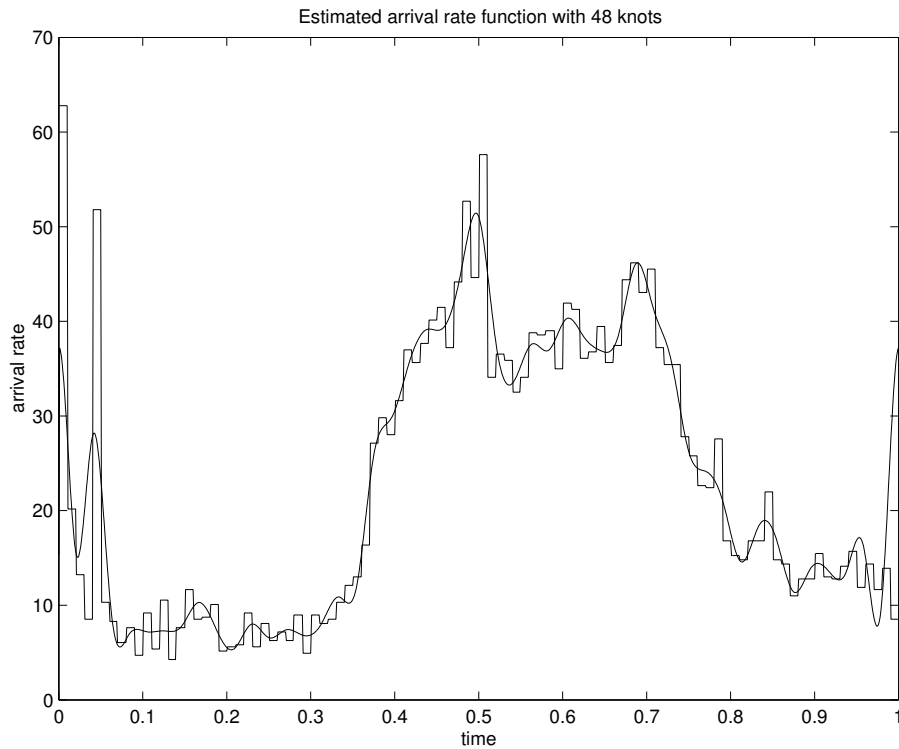


Figure 2

6.1.3 Determining the number of knots

In order to determine the appropriate number of knots to use for the approximating splines we performed the k-folding procedure described in Section 5.1. Figure 3 shows the average of results from 10 different 5-foldings.

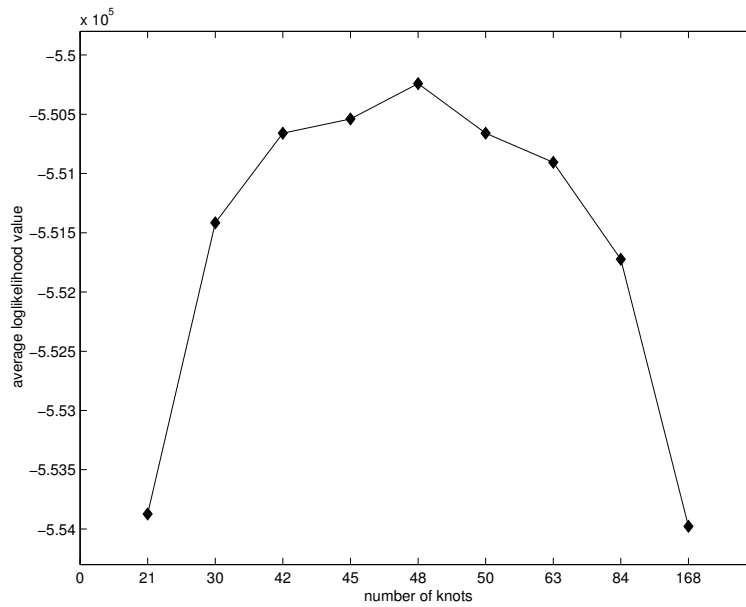


Figure 3

Based on these results, we decided to use 48 knots. Figure 4 shows an estimate using 14 knots (which does not provide sufficient detail to describe the arrival process), while Figure 5 shows an estimate using 336 knots, resulting in overfitting the data. Our best estimate (using 48 knots) is presented in Figure 6.

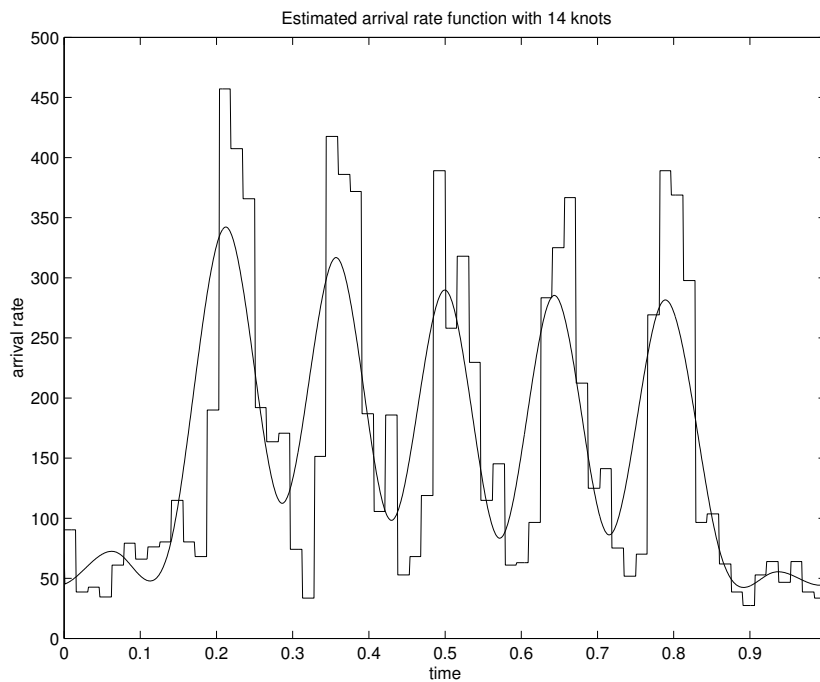


Figure 4

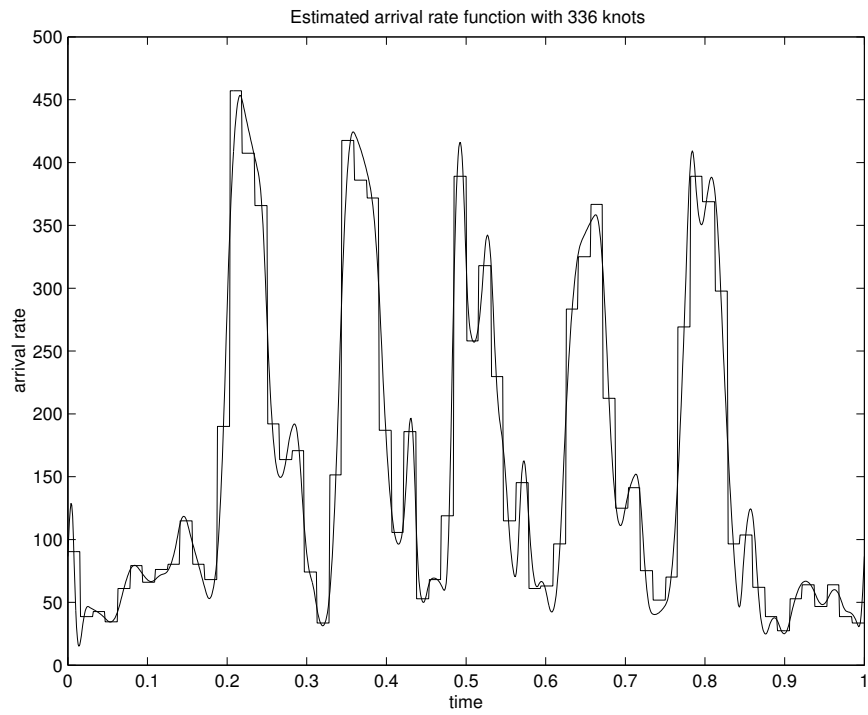


Figure 5

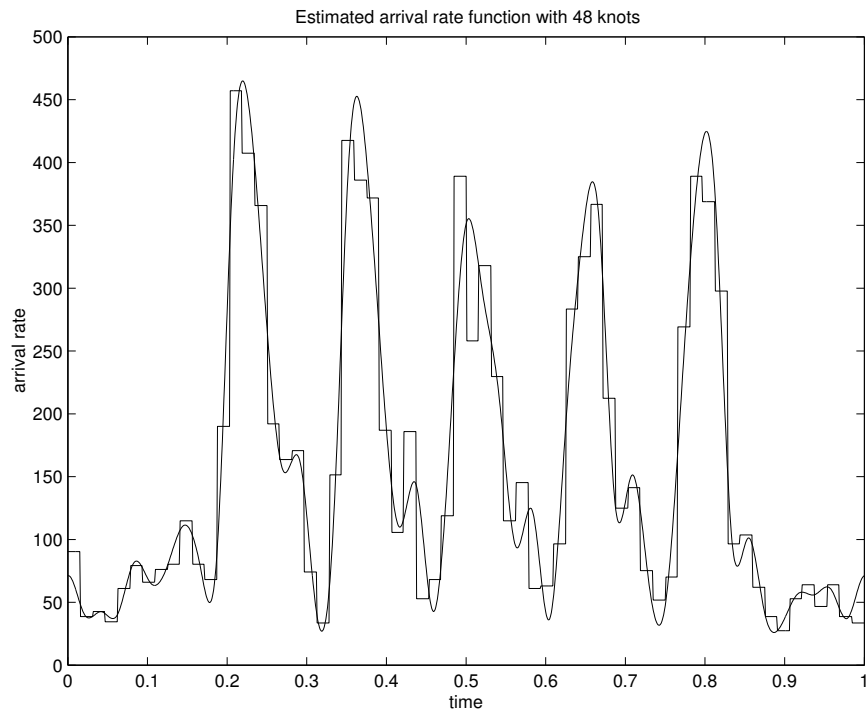


Figure 6

6.1.4 Robustness

We applied the merging procedure described in Section 5.2 to test the sensitivity of our model to the length of the intervals in the data set. Figures 7 and 8 show the results obtained by merging adjacent time intervals of overall length s .

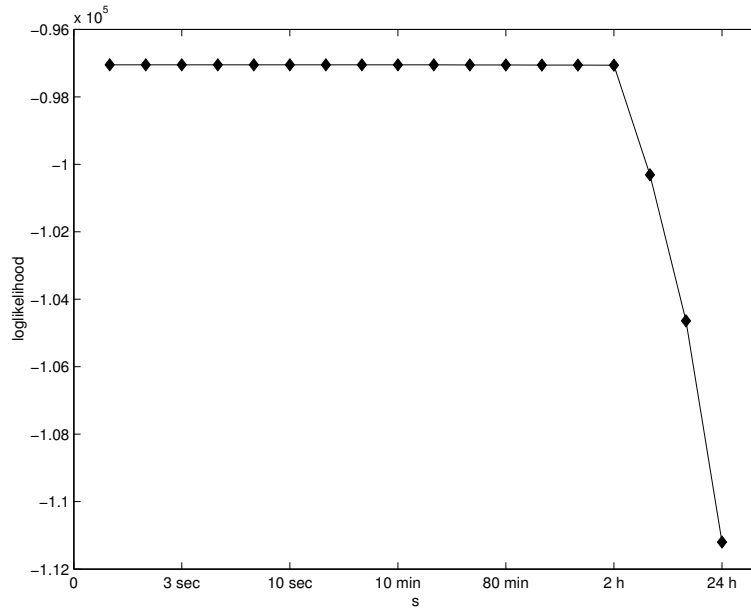


Figure 7

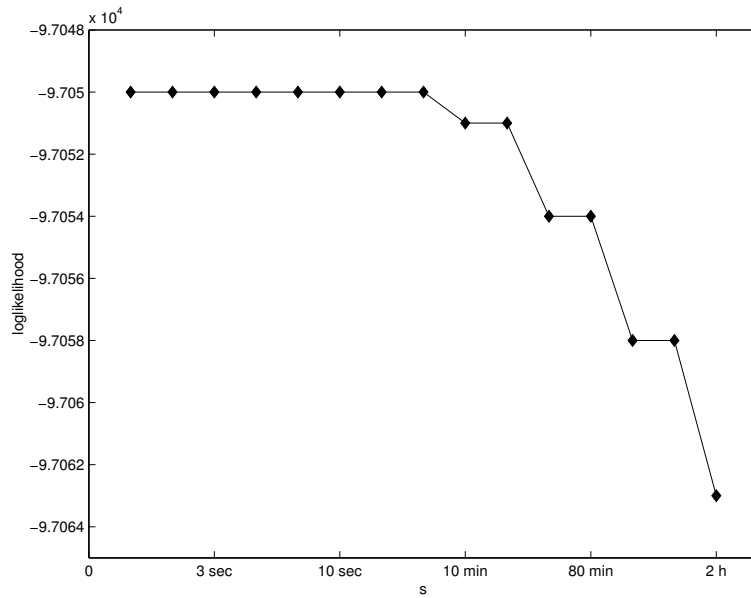


Figure 8

It is interesting to note that, as can be seen in Figure 7, a sharp drop in the objective value occurs when the estimating splines reach zero and the inequalities in the non-negativity constraints become binding. Figure 8 shows the section of the graph before this drop in more detail, while Figure 9 shows an estimating spline after the drop. It is apparent that a discrete model with arrival periods of up to 5 minutes does not materially affect the fit.

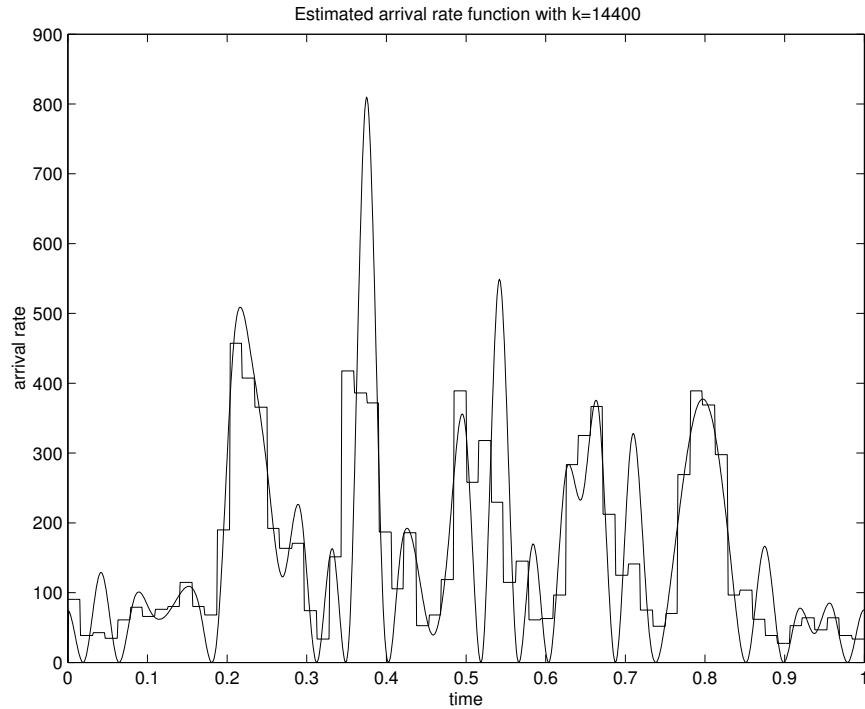


Figure 9

6.2 Results on data sets generated from a known arrival rate function

As suggested in section 5.3, we tested our method by randomly generating data sets for time periods of different lengths, using known periodic arrival rate functions with a period of 1. Figure 10 shows the original arrival rate functions; λ_1 is a cubic spline with 6 knots, while $\lambda_2(t) = 100(\sin(2\pi t) + 1)$.

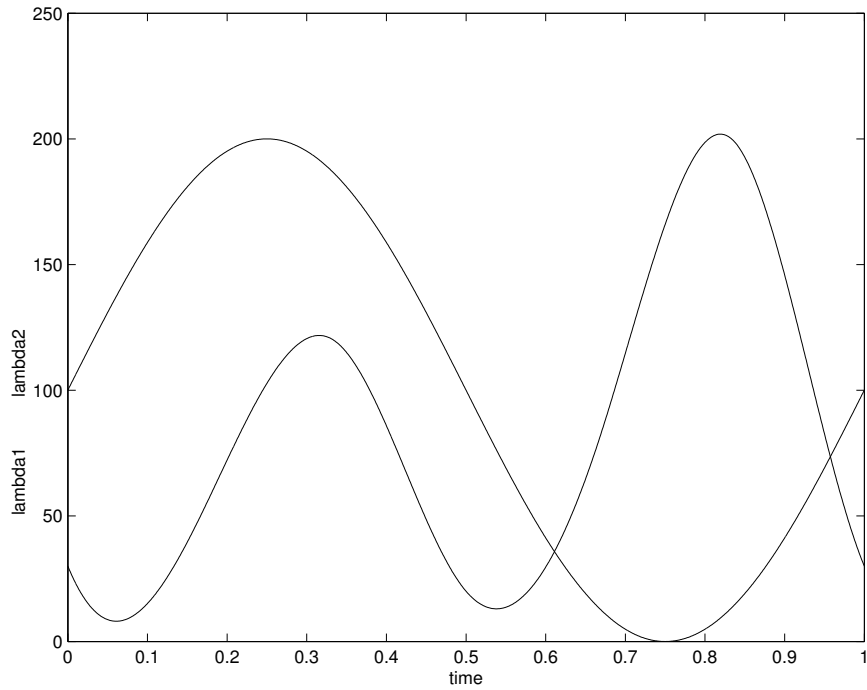


Figure 10

We estimate both of these functions with our method by using 6-knot splines. To measure the accuracy of an estimate λ^* , we compute the L1-norm, L2-norm and the maximum of the absolute value of the difference $\lambda_i - \lambda^*$ ($i = 1$ or 2). Figure 11 show the results of these experiments for λ_1 , while Figure 12 gives the results for λ_2 ; the x axis shows the logarithm of the length of time for which data was generated.

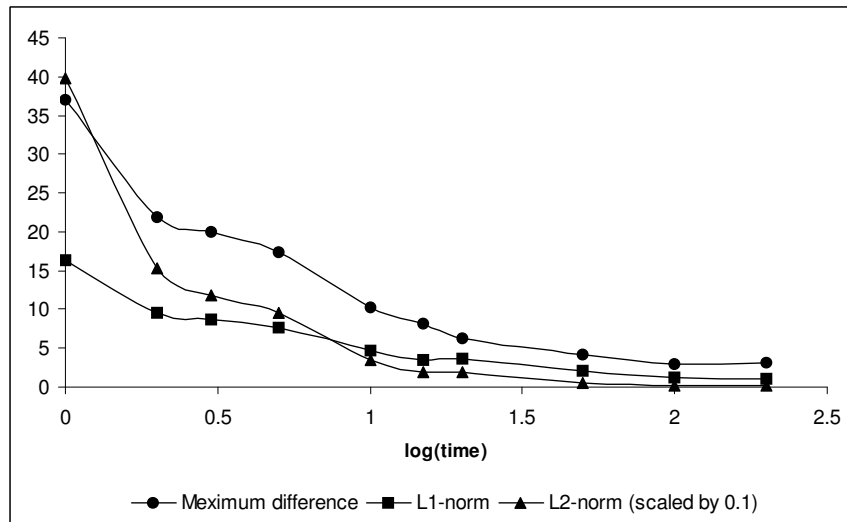


Figure 11

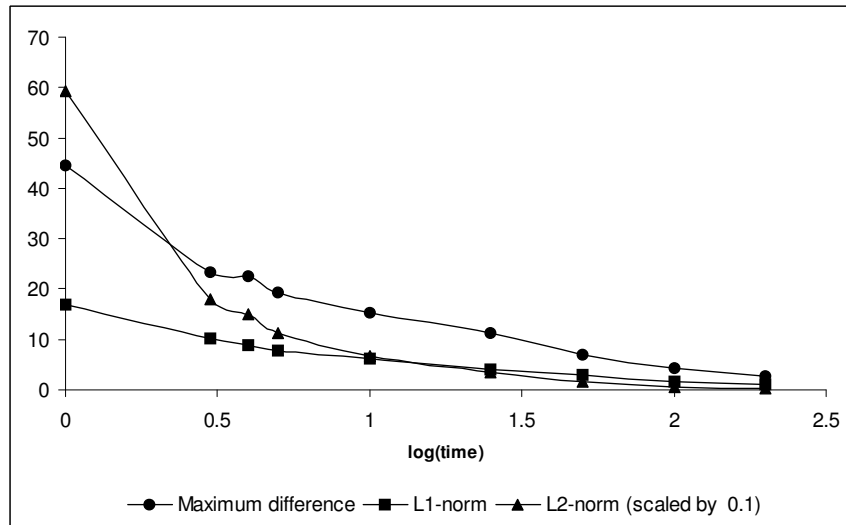


Figure 12

6.3 Illustrating the importance of the nonnegativity constraints

The distinguishing feature of our model is the way we ensure the nonnegativity of the function estimating the arrival rate. While in some cases the maximum likelihood criterion might by itself ensure nonnegativity, this is not the case for small data sets or arrival rate functions which can get close to zero.

To illustrate this phenomenon, we first consider a small data set of 550 e-mail arrival times. Figure 13 shows the 42-knot spline estimation of the arrival function provided by the basic version of our model while Figure 14 shows the results when the nonnegativity constraints are omitted.

Finally, we consider a large data set artificially generated from the arrival rate function $\lambda(t) = 200[\sin(6\pi t) + 0.8]_+$ (see Figure 15). Figure 16 shows estimating splines obtained from arrival data for 40 time periods, with and without requiring nonnegativity.

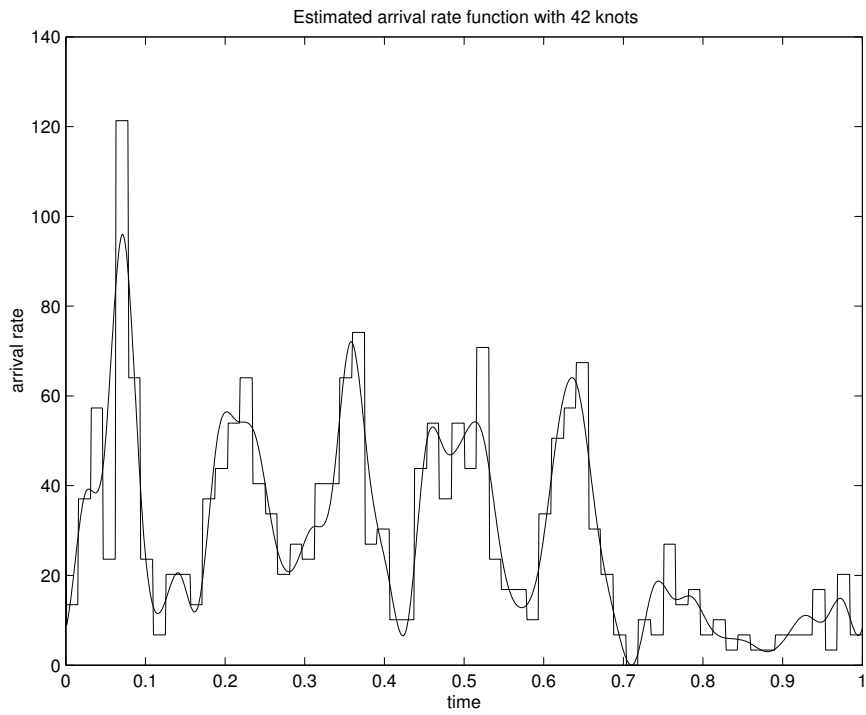


Figure 13

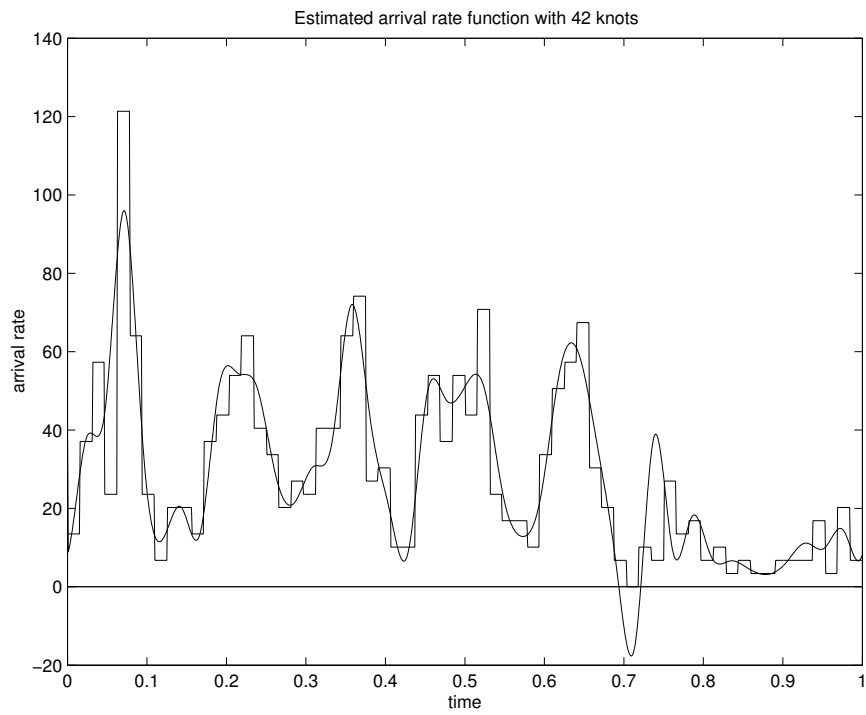


Figure 14

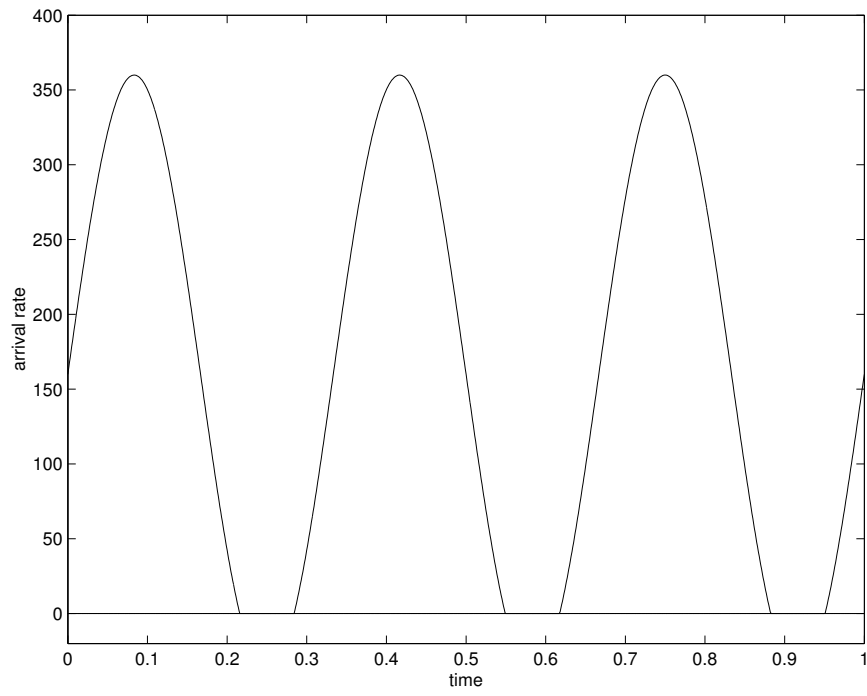


Figure 15

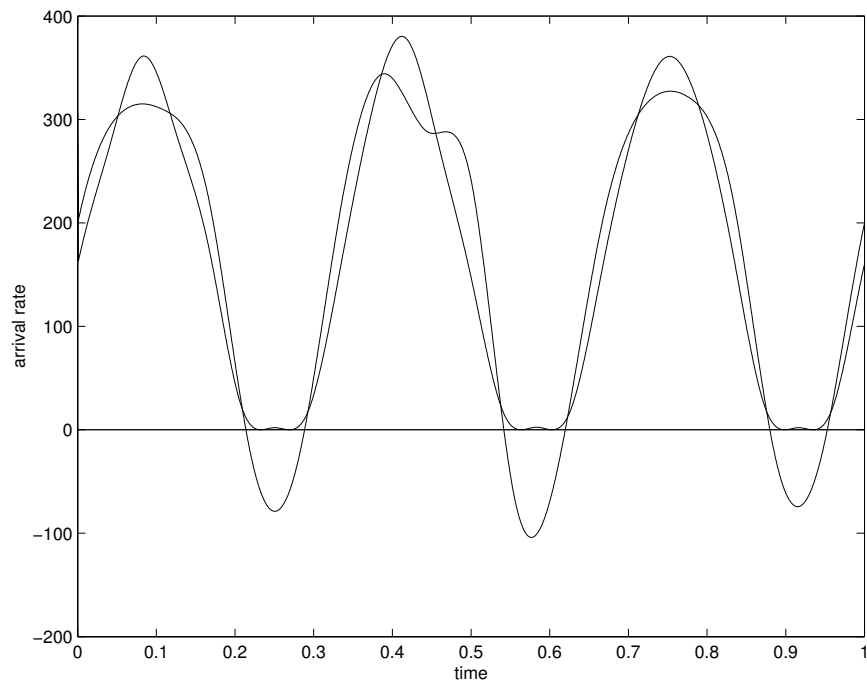


Figure 16

References

- [1] Byrd, R., M. E. Hribar, and J. Nocedal. 1999. An Interior Point Method for Large Scale Nonlinear Programming. *SIAM J. Optimization*, 9 (4), 877–900.
- [2] Czyzyk, J., M. Mesnier, and J. Mor. 1998. The NEOS Server. *IEEE Journal on Computational Science and Engineering* 5 , 68–75.
- [3] Dolan, E. 2001. The NEOS Server 4.0 Administrative Guide. Technical Memorandum ANL/MCS-TM-250, Mathematics and Computer Science Division, Argonne National Laboratory.
- [4] Gal, A. and J. Eckstein. 2001. Managing Periodically Updated Data in Relational Databases: A Stochastic Modeling Approach. *Journal of the ACM* 46(6), 1141–1183.
- [5] Gal, A., J. Eckstein and Zachary Stoumbos. 2003. Scheduling of Data Transcription in Periodically Connected Databases. *Stochastic Analysis and Applications* 21(5), 1021–1058.
- [6] Gropp, W. and J. Mor. 1997. Optimization Environments and the NEOS Server. *Approximation Theory and Optimization*, M. D. Buhmann and A. Iserles, eds., Cambridge University Press, 167–182.
- [7] Nesterov, Y. 1997. Structure of non-negative polynomials and optimization problems. CORE Discussion Papers 9749.
- [8] Nesterov, Y. 2000 . Squared functional systems and optimization problems. J. B. G. Frenk, C. Roos, T. Terlaky, S. Zhang, eds. *High Performance Optimization*. Kluwer Academic Publishers, Utrecht, The Netherlands, 405-440.
- [9] Nocedal, J. and R. A. Waltz. 2003. KNITRO User’s Manual. Technical Report OTC 2003/05, Optimization Technology Center, Northwestern University, Evanston, IL, USA. Shao, J. (1993).
- [10] Shao, J. (1993). Linear model selection by cross-validation. *Journal of the American Statistical Association* 88, 486–494.