Constructing a correlated sequence of matrix exponentials with invariant first-order properties

Kenneth Mitchell

Computer Science Telecommunications Program, University of Missouri – Kansas City, 5100 Rockhill Road, 57OC STB, Kansas City, MO 64110, USA

Abstract

In this paper, we demonstrate a method for developing analytic Markovian traffic sources in which the correlation structure can be arbitrarily constructed leaving the marginals invariant. We construct a simple model based on empirical data sets and show the effects of changing the autocorrelation on the behavior of network traffic. © 2001 Elsevier Science B.V. All rights reserved.

Keywords: Autocorrelations; Matrix exponential Distributions; Arrival processes; Communications networks; Markov modeling; Queueing theory

1. Introduction

Models of networks using the assumption of renewal arrival traffic frequently result in buffer designs which may significantly under-predict losses when compared with actual traffic conditions [8,20]. Short- and long-range dependencies have been observed in LAN traffic [8] as well as TCP applications over wide-area networks [6]. VBR video traffic in particular has been shown to have strong correlations [1,19,21]. These correlations cause much higher buffer overflow than that predicted in most analytic models. The degree of such impact and how the autocorrelation structure can be incorporated into performance models of these systems is not clear. Attempts to quantify the impact of autocorrelated network traffic have been made difficult due to the fact that most analytic traffic source models have not been able to isolate the effects of the autocorrelations from the marginals.

To shed some light on this important problem, we examined one of the Bellcore Ethernet traces produced by Leland and Wilson et al. In 1989 and 1990, Leland and Wilson [11] collected traces of several hundred million Ethernet packets from Bellcore’s Morristown Research and Training Center. These packet traces formed the basis of the conclusion in several papers [9,10,25–27], that renewal models of network traffic may poorly reflect the actual traffic in communications networks.

Eramilli [2] performed simulation experiments in which the inter-arrival times of the Bellcore data were shuffled, allowing him to isolate the effects of the near-range and long-range autocorrelations on buffer occupancy. Shuffling the data in different ways
destroys either the near-range or the long-range auto-
correlations. The actual inter-arrival times, and there-
fore the marginal distribution of the empirical trace,
is not affected. These experiments, along with our
own [14], demonstrate that it is the ordering of the
inter-arrival times which affects the autocorrela-
tion structure. With this in mind, we set out to develop
analytical source models in which we can control
the autocorrelations analytically while leaving the
marginals invariant. The resulting process is similar
to the discrete autoregressive (DAR) model discussed
in Heyman et al. [3] and the transform expand
sample (TES) process discussed in Jagelman [4]. This
model, however, is a continuous time analytic model
and lends itself readily to Markov chain analytic
techniques [15].

In Section 2, we briefly describe our techniques for
creating correlated sequences of matrix exponential
processes which leave the marginals invariant. In Sec-
tion 3, we apply these techniques to construct a simple
two state Markovian model of Ethernet traffic. Section
4 shows some results of our studies and we conclude
the paper in Section 5.

2. Constructing correlated processes

2.1. Matrix exponential distribution

A matrix exponential (ME) distribution is defined
[12] as a probability distribution with representation
\( (p, B, e) \), i.e.,
\[
F(t) = 1 - p \exp(-Bt)e', \quad t \geq 0, \tag{1}
\]
where \( p \) is the starting operator for the arrival process,
\( B \) is the process rate operator, and \( e' \) is a summing op-
erator. The minus sign in \( \exp(-Bt) \) represents a na-
tural generalization from a scalar exponential process
to a vector process. The order of the representation is
indicated by the dimension of the matrix \( B \), and the
degree of the distribution \( F(t) \) is the minimal order of
all its representations. The \( n \)th moment of the matrix
exponential distribution is given by
\[
E[X^n] = n!pV^n e', \tag{2}
\]
where \( V \) is the inverse of \( B \).

The class of matrix exponential distributions is iden-
tical to the class of distributions that possess a rational
Laplace–Stieltjes transform. Matrix exponential distri-
butions are dense in the set of all distributions, so any
distribution can be approximated arbitrarily close with
a matrix exponential distribution. The class of ma-
trix exponential distributions is closed under mixtures,
convolutions, and order statistics of such distributions.
A matrix exponential distribution may have an under-
lying probabilistic interpretation (i.e. a phase-type
distribution), but is not necessarily limited to such an
interpretation. Phase-type distributions form a proper
subset of matrix exponential distributions.

A representation is not unique. If \( F(t) \) has represen-
tation \( (p, B, e) \) then \( (pX, X^{-1}BX, X^{-1}e) \) is also
a representation for any non-singular matrix \( X \). The
only limitations on \( (p, B, e) \) stem from the require-
ment that \( F(t) \) must form a distribution function. Con-
sequently, \( (pe') \) is a real number between zero and one
and the eigenvalues of \( B \) must either be positive reals
or must come in complex conjugate pairs with a pos-
tive real part. As there are no prescribed structural
or domain restrictions on the components \( (p, B, e) \),
one can choose a physically based representation (i.e.
phase-type), or an algorithmic representation.

2.2. Matrix exponential sequence

A sequence of ME random variables \( T_1, T_2, \ldots \) such
that the joint probability density over any finite se-
quence of consecutive inter-event times is given by
\[
f_{T_1, T_2, \ldots, T_n}(t_1, t_2, \ldots, t_n) = \pi(0)\exp(-Bt_1)\exp(-Bt_2)L\cdots\exp(-Bt_n)Le',
\]
where \( \pi(t) \) is a vector representing the internal state of
the process at time \( t \). The matrix \( L \) is the (non-zero)
event transition rate operator which generates an event
and starts the next interval in the appropriate starting
state. All matrices and vectors are assumed to be of
finite dimension \( m \). This process is interpreted as a
stream of events (in this case, successive departures
from a matrix exponential process) occurring at times
\( t_1, t_1 + t_2, t_1 + t_2 + t_3, \ldots \), and inter-event times \( t_1, t_2, \ldots \).
Note that if the rank of \( L \) is one, then the sequence
\( (T_1, T_2, \ldots, T_n) \) is a renewal process. An infinitesimal
characterization of the vector process \( \pi(t) \) analogous
to the Poisson process is given by Lee et al. [7].

- If at time \( t \) the system is in state vector \( \pi(t) \) and
  no events occur between \( t \) and \( t + h \), then at time

$t + h$ the system will be in state vector $\pi(t + h) = \pi(t)(I - Bh + o(h))$

- If at time $t$ the system is in state vector $\pi(t)$ and a single event occurs between $t$ and $t + h$, then at time $t + h$ the system will be in state vector $\pi(t + h) = \pi(t)(Lh + o(h))$

This allows the closed-form representation of the state of the system to be $\pi(t) = \pi(0) \exp(Q^t)$, for $t > 0$, where $Q^* = L - B$.

There are several ways in which such a sequence can be constructed. For illustrative purposes we use a “phase-type” viewpoint and review the construction of a Markov arrival process (MAP). We start with a continuous time Markov chain with rate matrix $Q^*$ and invariant steady-state vector, $\pi$, with $\pi Q^* = 0$, $\pi e' = 1$. Depending on the application, certain transitions represent events of interest. By putting these in a matrix $L$, we can write $Q^*$ as $Q^* = Q^* - L + L$, and define $B = L - Q^*$. Internal transition rates are represented by the progress rate operator $B$ and represent transitions that are not of special interest. The vector describing the state of the process immediately after a special event depends on the state immediately before the event and the transition which created the special event, thus $\pi(r^+) = \pi(t)L$. The process just described is a MAP with $D_0 = B$ and $D_1 = L$, see Neuts [17,18].

A correlated sequence of matrix exponentials is more general than a MAP in that we allow imaginary phases to exist (i.e. Coxian distributions). The elements of the vectors $\pi(t)$, and matrices $B$ and $L$ may not have a probabilistic interpretation, but vector valued functions on the state $\pi(t)$ do. The matrices $B$ and $L$ which we construct are of the form which can be used directly in vector balance equations for Markovian queueing models. For details see [12,16,24]. The joint density function of the first $n$-successive inter-arrival intervals is a sequence of matrix exponentials with representation given in Eq. (3).

We assume that the process is in equilibrium, the steady state being represented by the vector $\pi$. In equilibrium, the starting vector $\pi(0)$ is the steady state at embedded arrival points and is denoted by the vector $p$.

$$p = \frac{\pi L}{\pi Le'}.$$  \hspace{1cm} (7)

The vector $\pi$ is also the residual vector for the process till the next event and is related to $p$ by the expression

$$\pi = \lambda p V,$$  \hspace{1cm} (8)

where $V$ is the inverse of $B$, and $\lambda$ is the mean arrival rate for the process.

The covariance of the sequence of matrix exponentials (3) is given as

$$\text{cov}[X_0, X_k] = pVVL_k Ve' - (pVe')^2.$$  \hspace{1cm} (9)

See Lipsky [13] for a complete derivation or [16,18] for a short review. If the process is assumed to be covariance stationary, the autocorrelation is obtained by dividing $\text{cov}[X_0, X_k]$ by the variance

$$\text{var}[X_0] = 2pV^2e' - (pVe')^2.$$  \hspace{1cm} (10)

Define the matrix $Y$ as

$$Y = VL$$  \hspace{1cm} (11)

and note that the autocorrelations decay matrix geometrically as $pVLY^k Ve'$. Note in particular that $L$ only appears in the autocorrelation equation and not in the moment equation (2). A renewal process can be expressed as an uncorrelated sequence of matrix exponentials (3) by making the following assignments:

$$L_r := Be',$$  \hspace{1cm} (12)

$$\pi(0) := p.$$  \hspace{1cm} (13)

The rank of the $L_r$ operator is 1 as a result of the product of the column vector $e'$ and the row vector $p$.

2.3. Constructing a correlated process with invariant marginals

Generally, we want to construct a correlated sequence with varying autocorrelations from known point processes. Let a given matrix exponential distribution induce a renewal sequence $(p, B, e)$ which can be represented as a sequence of uncorrelated matrix exponentials using (12) and (13). An autocorrelated process can then be constructed by introducing an $L_r$ (or $Y_r$) such that $p$ and $B$ remain invariant as follows.
First, find the residual vector \((\pi)\) for renewal version of the process using Eq. (8). Using Eqs. (6) and (12) construct \(Q^* = Be'p - B\). Now consider the equation \(\pi Q^* = 0\). For a given \(Q^*\) its solution is unique, but for a given \(\pi\), the solution is not, and each \(Q^*\) which satisfies the equation results in a process with a different autocorrelation structure as it leaves \(\pi, p\), and \(B\) invariant.

We are interested in developing parsimonious models, so we start with a single parameter \(\gamma\), leading to a geometrically decaying covariance. Define the operator \(L_{\gamma}\) by

\[
L_{\gamma} = (1 - \gamma)Q^* + B, \quad -1 \leq \gamma < 1. \tag{14}
\]

By use of Eqs. (6) and (12), \(L_{\gamma}\) can be expressed in the following form:

\[
L_{\gamma} = (1 - \gamma)(Be'p - B) + B. \tag{15}
\]

From Eq. (11), this \(L_{\gamma}\) induces a transition matrix \(Y_{\gamma}\),

\[
Y_{\gamma} = (1 - \gamma)e'p + \gamma I, \quad -1 \leq \gamma < 1. \tag{16}
\]

We now show that the marginal density of the \(k\)th random variable \(f_{T_k}(t_k)\) at equilibrium remains invariant for any \(\gamma, -1 \leq \gamma < 1\). The expression for the marginal density of the \(k\)th random variable in a sequence of matrix exponentials from Eq. (3) is

\[
f_{T_k}(t_k) = \int_0^\infty \int_0^\infty \cdots \int_0^\infty f_{T_{k-1},t_2,...,t_1}(t_1,t_2,...,t_n) \times (dt_1,...,dt_{k-1},dt_{k+1},...,dt_n)
\]

\[= p(V^k)e^{(1 - \gamma)e'p + \gamma I}) \exp(-Bt_k) \exp(V(t_{k+1} - t_k)e') \exp(V(t_{k+2} - t_{k+1})e') \cdots \exp(V(t_n - t_{n-1})e') \]

\[= p(Y_{\gamma}e^{(1 - \gamma)e'p + \gamma I}) \exp(-Bt_k), \tag{17}
\]

Using the definition of \(Y_{\gamma}\), the following expressions are obtained:

\[
pY_{\gamma} = p((1 - \gamma)e'p + \gamma I) \exp(-Bt_k) L_{\gamma} \exp(V(t_{k+1} - t_k)e') \exp(V(t_{k+2} - t_{k+1})e') \cdots \exp(V(t_n - t_{n-1})e') \]

\[= (1 - \gamma)e'p + \gamma I)e', \quad -1 \leq \gamma < 1. \tag{18}
\]

\[Y_{\gamma}e' = ((1 - \gamma)e'p + \gamma I)e' \]

\[= (1 - \gamma)e' + \gamma e' = e', \quad -1 \leq \gamma < 1. \tag{19}
\]

Substituting expressions (18) and (19) into (17) yields

\[
f_{T_k}(t_k) = p \exp(-Bt_k) L_{\gamma}e' = p \exp(-Bt_k) BY_{\gamma}e'. \tag{20}
\]

Therefore the marginal distribution is invariant with respect to \(\gamma\).

For any value of \(-1 \leq \gamma < 1\), the vectors \(\pi, p, B\) remain invariant. We call \(\gamma\) the measure of persistence of the process, and as \(\gamma \to 1\), the slope of the decay of the autocorrelations decreases. Also, as \(\gamma \to 0\), the time spent in each state increases although the proportion of time spent in each state remains the same. When \(\gamma = 0\), the process is a renewal process and \(Y_{\gamma} = e'p\). For \(\gamma < 0\), the lag autocorrelations alternate in sign.

The renewal operator \(Y_{\gamma} = e'p\) has a single eigenvalue of \(pe' = 1\) and an eigenvalue of 0 with multiplicity \(m - 1\), where \(m\) is the order of representation. The correlated operator \(Y_{\gamma}\) has a single eigenvalue of 1 and eigenvalue \(\gamma\) of multiplicity \(m - 1\). Thus

\[
Y_{\gamma} = (1 - \gamma)e'p + \gamma I. \tag{21}
\]

The expression for the lag-\(k\) autocorrelation becomes

\[
\text{corr}[X_0,X_k] = c\gamma^k, \tag{22}
\]

where the constant \(c\) is

\[
c = (pV^2e' - (pVe')^2)(2pV^2e' - (pVe')^2). \tag{22}
\]

The allowable values for \(\gamma\) are bound such that \(-1 \leq \gamma < 1\) and \(|c\gamma| \leq 1\). In this paper we will assume \(0 \leq \gamma < 1\).

3. Constructing a two-state correlated process from empirical data

We want to determine the relative impact of autocorrelation on a simple \(G/M/1/k\) performance model. To do this, we analyzed such a system as constructed above and performed our own shuffling experiments on the available Bellcore data. By randomizing blocks of size \(n\), we can study the relative impact of short-range versus long-range autocorrelation, e.g., if \(n = 100\), we eliminate short-range autocorrelation up to lag-100 but long-range autocorrelation is not affected. As \(n\) increases, we eliminate longer-range autocorrelation until \(n \to \infty\) and the entire trace is in random order. For this latter case, the trace should have no autocorrelation structure and should behave as a pure renewal interarrival stream — which is exactly what happens. Our results corroborate those of Erramilli [2] who performed similar shuffling experiments on Bellcore Ethernet traces.
Fig. 1 shows the packet inter-arrival time autocorrelation in one of the Bellcore Ethernet traces. Packet inter-arrival time autocorrelation is plotted to lag-1000 together with that of a renewal stream constructed from the marginals. Clearly, an analytical model with a renewal arrival stream will not be an effective surrogate for a system with autocorrelated arrivals such as those found in the Bellcore Ethernet packet trace, even if the renewal stream matches the first-order statistics of the Bellcore trace.

The slow rate of decay in autocorrelations is significant. Erramilli [2] found the rate of decay in long-range-dependent traffic to be a power of the lag time. This characteristic cannot be duplicated in any Markovian source model, but only approximated to some extent.

Recent papers however have shown that long-range dependency in traffic does not always dominate performance. Jelenkovic [5] has observed what are called weakly stable systems in which arrival processes have states which generate traffic with mean inter-arrival times that are shorter than the mean system service time. In this scenario the short-range dependencies tend to dominate. Models which use heavy tailed distributions to modulate sources exhibiting long-range dependency have also found these tendencies in what have been termed as blow up points [22,23]. Based on these observations, it may be possible to limit the density of the tail (and thus the order of representation) in the modulating process. In this example we show that reasonable approximations can be constructed which perform well over a wide range of conditions.

Using standard techniques, we computed a phase-type hyper-exponential ($H_2$) distribution taken from samples of the Bellcore trace. For our example, we use a matrix exponential representation $(p, B, e)$. Thus,

\[ p = [0.2854710695, 0.7145289305], \]
\[ B = \begin{bmatrix} 0.3747871858, & 0 \\ 0, & 2.998297486 \end{bmatrix}, \]
\[ e' = \begin{bmatrix} 1 \\ 1 \end{bmatrix}. \]

The renewal process represented by Eqs. (23)–(25) above has $\lambda = 1.0$ and $C^2_x = 3.223610$. Now we want to make a matrix exponential point process with a specific autocorrelation structure. Using Eq. (16), the expression for $Y_\gamma$ is

\[ Y_\gamma = (1 - \gamma)e'p + \gamma I = \begin{bmatrix} a + b\gamma, & b - b\gamma \\ a - b\gamma, & b + a\gamma \end{bmatrix}, \]

where $a = 0.2854710695$ and $b = 0.7145289305$.

By adjusting the values for $\gamma$, we obtain matrix exponential point processes with varying autocorrelation structures. Since the Bellcore trace has been shown to exhibit long-range dependencies, the lag autocorrelation of our model will always drop below that of the empirical data at some point. Table 1 shows the $\gamma$ values for these varying autocorrelation lags at which our model intersects and falls below that of the empirical trace.

### Table 1

<table>
<thead>
<tr>
<th>Model lag</th>
<th>$\gamma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.58003</td>
</tr>
<tr>
<td>10</td>
<td>0.88248</td>
</tr>
<tr>
<td>100</td>
<td>0.96949</td>
</tr>
<tr>
<td>1000</td>
<td>0.99675</td>
</tr>
</tbody>
</table>

4. Numerical results

Fig. 2 compares packet losses from our analytic models for varying $\gamma$ values with losses resulting from a simulation model using the Bellcore Ethernet
Fig. 2. Log. plot of packet loss probabilities for a two-state approximation vs. Bellcore data into a buffer of size 100 with an exponential server.

Fig. 3. Log. plot of packet loss probabilities for a two-state approximation vs. Bellcore data into a buffer with an exponential server, $\rho = 0.7$. 
trace. This figure shows that as the autocorrelation lag increases, losses from the $G/M/1/k$ model increase. Losses from the renewal inter-arrival stream remain very low. It is important to note that when the analytical model matches the lag-1 autocorrelation ($\gamma=0.58$), losses are still very low. For this particular traffic it appears that matching the slope of the decay of the autocorrelations farther out in the tail are more important than matching the lag-1 autocorrelation. Also, the range of losses experienced in the analytic model is the result of altering the autocorrelation structure only. The marginal distribution remains exactly the same.

Fig. 3 shows packet loss probabilities from a simulation model of the $G/M/1/k$ system and utilization $\rho = 0.7$ with our analytic model for buffer sizes in the range of 50–150. This figure shows that a value of $\gamma$ can be chosen such that the performance behavior can be closely modeled in the range of buffer sizes under study. In fact, the decay of the autocorrelations in the analytical model can be made as arbitrarily close to 0 as possible, making the modeling of the behavior of large buffers possible. For a value of $\gamma = 0.99$, packet loss probabilities are overestimated, but the slope of the decay in packet loss probabilities most closely matches that of the Bellcore Ethernet trace. Of course, as with any Markovian model which tries to approximate long-range-dependent behavior, the slope will eventually diverge from that of actual empirical data.

5. Conclusion

To be effective, models must incorporate the second-order statistics (autocorrelations) of the systems they purport to emulate. We have developed techniques to allow correlated source models to be constructed which allow the autocorrelation structure to be modified while leaving the marginals invariant. The traffic we have studied has been shown to have long-range dependencies and it is evident that matching the lag-1 autocorrelation is not enough. It appears that matching near-term autocorrelation is not as important as matching the slope of the decay of the autocorrelation structure at some point farther out in the tail.

References